Research Schools Network

Institute for **Effective Education**
Empowering educators with evidence

No More Marking

An evaluation of an online comparative judgement platform on teacher workload and pupils' English outcomes in Years 7, 8 and 9

John Coats

Notre Dame High School

February 2019

About IEE Innovation Evaluation Grants

The first four IEE Innovation Evaluation Grants were awarded in February 2017. Funded by the Institute for Effective Education (IEE), these grants supported pilot evaluations of innovations of teaching and learning approaches based on the Research Schools Network's goal of improving the attainment of pupils by increasing the use of evidence-based practices.

Since then a further 26 projects have been successful in their application for an IEE Innovation Evaluation Grant, bringing the total number to 30. The applications we received included a wide range of interesting, school-led innovations – from after-school film clubs to improve the creative writing of Year 5 pupils, to the use of audio feedback with Year 12 pupils – and we were really impressed with the thought that applicants had put into how these innovations could be evaluated.

The evaluations are small-scale, and test the kinds of innovations that schools are interested in. This is very much a "bottom-up" exercise, allowing schools to get some indicative evidence behind real-world initiatives. Many evaluations are now coming to an end, and we are starting to publish reports on the findings. It is important remember that these are small-scale projects, often carried out in one school, so it is not possible to generalise their findings. In fact, the main benefit of the Innovation Evaluation projects may be in the process, rather than the findings.

Contents

Executive summary

Description of the innovation

During intervention lessons, pupils were supported to make multiple comparative judgements of both older pupils' work and the work of their peers. This was done using *No More Marking* software in four lessons over two cycles of work.

Summary of the evaluation

The study involved a total of 24 classes of Year 7–9 pupils from four urban secondary schools with lower than national average proportion of disadvantaged pupils. Two classes from each year group in each school took part. Assignment to treatment was carried out at a whole class level using key stage 2 (KS2) writing scores to minimise the difference in prior attainment between control and intervention cohorts.

The pre-test, immediate post-test and delayed post-test (one month after completing the second cycle of work) were all questions in the style of GCSE English Paper 1 Section B – a descriptive piece of writing based on a visual stimulus.

Summary of results

The study found that the use of *No More Marking* by pupils for two cycles of key stage 3 (KS3) descriptive writing lessons over a period of one to two months, led to pupil outcomes in descriptive writing that are comparable to the use of conventional teacher marking (delayed testing effect size = -0.06, n = 466). However, there was variation in effect sizes at the delayed post-test between boys (-0.23), girls (+0.14) and disadvantaged pupils (-0.20). In almost all cases the effect size was larger for the delayed post-test than the immediate post-test.

The study also found that the use of this intervention reduces teacher perception of their workload (t-test p-value < 0.001) compared to the work involved in conventional marking and feedback. Qualitative pupil responses indicated a greater enjoyment of lessons than normal for the intervention cohort. Pupil responses also indicated that, despite the withdrawal of (often labour-intensive) conventional feedback provided by the teacher, the intervention cohort felt equally able to both describe what a good piece of work looked like, and to produce a better quality piece of work in the future. These results are important because they could have positive implications for teacher retention.

Introduction

Description of the problem

Teachers currently spend a large amount of time marking, with limited evidence on the impact of both the type of feedback given and the time spent on generating the feedback.

Review of existing research

We know that feedback can have a high impact when carried out well. Teachers traditionally spend large amounts of time marking individual pupil work as part of the process of providing feedback to pupils, but there isn't a large volume of quality research to help guide teachers as to how to use the time they are spending marking work most effectively.

Elliott et al. (2016) state that, "despite its centrality to the work of schools and teachers, there is in fact little high-quality research related to marking." (p. 5) and conclude, "there is an urgent need for more studies so that teachers have better information about the most effective marking approaches" (p. 7).

Comparative judgement is a process whereby (usually) two pieces of work are compared side by side and a judgement made as to which is the best. By carrying out lots of judgements of work, with each piece of work being involved in multiple comparisons, it is possible to create a ranking of the work. Judgements are carried out on the computer with work having been scanned in. Software is able to produce normalised scores for pieces of work alongside a numerical rank order.

Pupils' competence when assessing peers' work using comparative judgement has been demonstrated in the case of mathematics (Jones and Wheadon, 2015) and science (McMahon and Jones, 2014). We believe that it is worth asking whether this can be replicated in the case of descriptive writing.

Qualitative evidence has been published as to the value of pupils comparatively judging peers' work for learning (Jones and Alcock, 2014; Seery et al., 2012). Moreover, educational and psychological research has demonstrated the promise of comparing example worked solutions for learning (eg. Evans and Swan, 2014; Pachur and Olsson, 2012). There is experimental evidence that using exemplars in the classroom may be particularly beneficial for lower-achieving pupils (Carroll, 1994) due to reflecting on stronger performing peers' answers.

Description of the innovation

The innovation saw pupils in Years 7–9 using *No More Marking* comparative judgement software over two cycles of work in place of (often labour-intensive) normal assessment and feedback. Pupils used the comparative judgement software to compare older pupils' work and the work of their peers.

Research questions

We had two research questions that we were attempting to answer through this evaluation.

1. Primary research question: What impact does peer comparative judgement, carried out for two cycles of work over a period of one to two months, have on descriptive writing progress of pupils in Years 7, 8 and 9, compared to similar pupils who did not experience comparative judgement?
2. Secondary research question: What impact does peer comparative judgement, carried out for two cycles of work over a period of one to two months, have on teachers' perceived workload, compared to teachers who did not use peer comparative judgement?

Method

Sample

Schools involved in the evaluation were as follows:

TABLE 1

School identifier	Local Authority	School type	% pupil premium
A	Sheffield	Urban	11.3
B	Sheffield	Urban	20.2
C	Rotherham	Urban	24
D	Doncaster	Urban	22.7

Participating schools were asked to provide two English classes from each of Years 7, 8 and 9 to participate in the trial. We asked schools to provide classes in similar pairs if they set by ability; otherwise we provided no criteria for selection.

Table 2 shows the characteristics of pupils participating in the trial.

NB the table below contains a breakdown of those pupils who were used in the final analysis. There was a drop-out of a number of pupils due to either key stage 2 (KS2) data being absent or of pupils being absent for one or more of the three formal assessments. We have only included those pupils in our analysis for whom we have a complete set of data.

TABLE 2

	Control cohort	Intervention cohort
Number of pupils	217	249
Male	49.3%	49.0%
Female	50.7%	51.0%
Pupil premium	10.6%	7.2%

Those pupils (n=466) for whom we have complete data are not necessarily typical of the pupil intake for the schools concerned. In particular the proportion of pupil premium pupils in the dataset used for analysis of results is not representative of the schools involved.

Assignment to condition

Prior attainment data (KS2 writing level) was obtained for pupils from all participating groups prior to assignment to condition.

Entire teaching groups were randomly allocated to the intervention and control cohorts in such a way as to minimise the difference in KS2 writing outcomes between intervention and control groups

Following the initial randomisation the average KS2 writing level of the pupils in the intervention and control groups was checked to ensure that the difference was close to zero.

Innovation

The innovation saw pupils using comparative judgement software in place of traditional teacher marking and feedback. The processes of comparative judgement aim to support pupils in developing an understanding of 'what a good one looks like' that they can then apply when producing their own similar work.

We commissioned a subject specialist to develop lesson resources and to translate the GCSE band descriptors into appropriate prose statements for key stage 3 (KS3) pupils. These prose statements were then used as an assessment framework throughout (for both pupils and teachers).

The trial consisted of a sequence of seven lessons delivered by pupils' usual teachers in normal class time with a further delayed post-test carried out later. The sequence of lessons was as below, with the intervention group receiving four 'treatments' in lessons 2, 3, 5 and 6.

- **Lesson 1:** introduction to the specific objectives followed by pupils producing a first written answer to a descriptive writing exam question (in the style of GCSE English Paper 1 Section B – a descriptive piece of writing based on a visual stimulus and therefore directly relevant to GCSE outcomes). This first draft was used as the pre-test.
- **Lesson 2**: pupils evaluated examples of work produced by older pupils in response to the same question and visual stimulus. The intervention group did this via a structured No More Marking Lesson. The control group were free to do this in any way that they wished, provided that they avoided the pupils making multiple comparisons. We provided some 'traditional' suggestions as to how they could do this.
- **Lesson 3**: pupils reflected on their own first drafts. The intervention group did this via a structured No More Marking Lesson. We included the examples of work that were used in Lesson 2 alongside each intervention group's own work in the set of work that the pupils made comparative judgements on. The control group received feedback on their draft, with the teacher being free to lead this reflective session in any way that they wished, provided that they avoided the pupils making multiple comparisons. We provided some 'traditional' suggestions as to how they could do this.
- Gap of at most one month.
- **Lesson 4**: reminder of the specific objectives followed by pupils producing a written answer to a descriptive writing exam question (again in the style of GCSE English Paper 1 Section B).
- **Lesson 5**: pupils evaluated examples of work produced by older pupils. The intervention group did this via a structured No More Marking Lesson. The control group were free to do this in any way that they wished, provided that they avoided the pupils making multiple comparisons. We again provided some 'traditional' suggestions as to how they could do this.

- **Lesson 6**: pupils reflected on their own first drafts. The intervention group did this via a structured No More Marking Lesson. We included the examples of work that were used in Lesson 5 alongside each intervention group's own work in the set of work that the pupils make comparative judgements on. The control group received feedback on their draft and the teacher was free to lead this reflective session in any way that they wished, provided that they avoided the pupils making multiple comparisons. We provided some 'traditional' suggestions as to how they could do this.
- **Lesson 7**: pupils produced a written answer to a descriptive writing exam question (again in the style of GCSE English Paper 1 Section B). This served as the immediate post-test.
- Gap of as close to four weeks as was manageable in the context of the school timetable.
- **Lesson 8**: pupils produced a further written answer to a descriptive writing exam question (again in the style of GCSE English Paper 1 Section B). This served as the delayed post-test.

Training was provided to all participating teachers at the outset. The intention was to train all participating schools at the same time, but this was not possible due to constraints on individual schools. We therefore visited each participating school to deliver the training separately.

The items covered in training were:

- parameters for the trial
- administration of questionnaires during the trial
- running the final assessment.
- dealing with consent and confidentiality
- what to do if there any problems during the trial
- an opportunity for clarification on any potential issues
- time to explore the risks associated with running the trial – we felt that by better understanding the risks, participants were more likely to adhere strictly to the trial protocols. These risks included demoralisation of control group, diffusion of treatment and managing the novelty effect for the intervention group.

The training was accompanied by the production of a comprehensive training manual. This included all lesson plans and resources produced by our subject specialists.

Outcome measures

The pre-test was carried out in Lesson 1. This was a piece of writing in the style of GCSE English Paper 1 Section B.

The immediate and delayed post-tests were carried out in Lessons 7 and 8 respectively, and were also pieces of writing in the style of GCSE English Paper 1 Section B.

A comparative judgement process was carried out by all participating teachers from the trial. This process was carried out 'blind' in that all pupil work was anonymised prior to the comparative judgements being made. Additionally, the comparative judgement software was set up so that each teacher was prevented from seeing any pieces of work from their own school in any of the judgements that they made. When combined with the fact that each piece of work was seen by multiple teachers in order to obtain the final ranking, the process was believed to be robust in terms of eliminating bias.

All 1,398 tests (pre-tests, immediate post-tests and delayed post-tests) were judged together so that differences could be measured between pre-tests and post-tests. The binary decision data taken from the comparative judgements was statistically modelled by the No More Marking software to produce a score for each piece of work. The scores ranged from -6.99 to +6.05 with a mean of -0.03 and a standard deviation of 2.14.

Each pupil therefore ended up with three scores: one for the pre-test; one for the immediate post-test; and one for the delayed post-test.

Effect sizes could then be calculated by considering the changes in average scores for the cohort concerned, alongside the standard deviation of the data for the cohort concerned. We calculated an effect size for the immediate post-test (looking at changes in the mean scores from pre-test to immediate post-test) and an effect size for the delayed post-test (looking at changes in the mean score from pre-test to delayed post-test).

It was important for us to understand the accuracy of the ranking produced by the comparative judgement process. We therefore commissioned a subject specialist to provide an absolute score for one in five pieces of work from each class, based on the prose descriptors used throughout the trial.

The Spearman's rank correlation coefficients for the sample of absolutely-marked work versus the comparative judgement rankings were as below.

TABLE 3A

n=80	Absolute marking
Pre-test ranking	0.70

TABLE 3B

n=87	Absolute marking
Immediate post-test ranking	0.73

These correlations suggest the ranking produced by the comparative judgement process has in fact produced a correct ranking for the pupil work involved in this study. There is a limitation to this conclusion in that we have used a single subject specialist to mark the sample of work and we have not considered these correlations in relation to typical inter-rater reliability when marking GCSE-style writing.

We considered several outcome measures to evaluate the impact of the intervention on teacher workload. Different school assessment policies have different expectations of teachers. We have made an assumption that workload related to teacher retention is as much about how teachers feel about the amount of work they have as it is a clock-watching exercise. Therefore we asked all participating teachers for their perception of the workload involved via the question:

> Thinking about the amount of time you have spent outside of the classroom for this sequence of lessons, what is your perception of the workload associated with providing

feedback to the pupils compared with the normal amount of time outside of lessons you would expect to spend providing feedback?

1=significantly less time than normal

2=less time than normal

3=same time as normal

4=more time than normal

5=significantly more time than normal

We anticipated that the control teachers would report no change, and the difference between these responses and those of the intervention teachers would help us answer the question about the impact on workload. The teacher survey was conducted following each of Lessons 3 and 6 using the above five-point Likert scale. A free text box was also included in the teacher questionnaires to allow the submission of a general comment.

Process evaluation

We also collected quantitative data from all pupils via Likert scale surveys. This was to give us further insight into the active ingredients of the intervention. The questions specifically addressed the effectiveness of the feedback to the pupils and referenced known characteristics of good feedback as outlined in *What Makes Great Teaching* (Coe, et al., 2014):

- This lesson was enjoyable compared to my normal lessons.
- Because of the last lesson I understand what a good example of work looks like.
- Because of this lesson I understand what a good example of work looks like.
- I could describe to you what a good example of work looks like.
- I know what specific things I need to do to improve my work.
- It will be easy for me to produce a better piece of work now.

Our original intention had been to carry out a number of lesson observations to check for implementation fidelity. Due to an increased spend on training (school constraints meaning multiple visits to schools) we had to limit our observations to one control lesson and one intervention lesson. Both observations adhered to evaluation protocols.

Analyses

Primary research question.

Having obtained pre-test, immediate post-test and delayed post-test scores we were then able to compare the differences of these scores for control and intervention cohorts in order to establish the effect size of the intervention.

This process was carried out for:

- all pupils
- all male pupils
- all female pupils

- all disadvantaged pupils

We calculated an effect size for the immediate post-test (looking at changes in the mean scores from pre-test to immediate post-test) and an effect size for the delayed post-test (looking at changes in the mean score from pre-test to delayed post-test).

Secondary research question.

The quantitative Likert responses for perceived time spent providing feedback were plotted as average scores with 95% confidence intervals, with control and intervention cohorts plotted side by side. We carried out t-tests to see whether there was a significant variation in the responses between control and intervention teachers in any of the cohorts. We plotted the responses for Lesson 3 and Lesson 6 separately so that we could see whether teacher views changed over time as they became more familiar with the process. We also aggregated the teacher responses across Lessons 3 and 6 to give us a larger dataset of responses.

Analysis of the process

The Likert responses from pupils were plotted as average scores with 95% confidence intervals with control and intervention cohorts plotted side by side. We carried out t-tests to explore any significant variation in the responses between control and intervention pupils in any of the cohorts. We plotted the responses for Lesson 3 and Lesson 6 separately so that we could see whether pupil views changed over time as they became more familiar with the process. We also aggregated the pupil responses across Lessons 3 and 6 to give us a larger dataset of responses.

Although our project paperwork asked pupils to identify whether they were responding to questions in Lesson 3 or Lesson 6 on the proformas, we had a small number of returns which did not include the lesson number. These Likert responses were included within the aggregated pupils' responses, but were obviously not used in the analysis for Lesson 3 or Lesson 6 responses.

The qualitative data from the free text box responses was grouped into themes.

Cost

Cost of the project is broken down below.

The total project cost was £19,920.

This can be broken down as £12,700 for costs associated with:

- project management
- employment of subject specialists to create resources, assessment materials and mark schemes
- creation of all training materials
- delivery of training
- subject specialist time absolute marking 20% of the sample
- project evaluation including data manager time
- administration.

And as £7,220 paid to participating schools to cover:

- staff time to attend training

- staff time to attend final comparative judgement moderation
- travel
- administration.

The cost of delivering the intervention (ie, project costs minus costs related to the evaluation) is calculated as £4,105.

This can be broken down as £2,100 cover costs for training teachers, £700 for production of lesson plans and resources, £700 for trainer time (assuming half day training at four schools as per this project) and other administrative, materials and travel costs associated with the training.

Therefore the cost ‘per pupil’ can be calculated as <£10.

Once a teacher has been trained in the intervention, there is no need for additional intervention/ resources/training and therefore the costs per pupil fall significantly over time.

Results

At the end of the trial we had complete data for 466 pupils.

Primary research question

The scores across the pre-test, immediate post-test and delayed post-test ranged from -6.99 to +6.05. The mean score was -0.03 and the standard deviation 2.14.

As would be expected there was an increase in scores from the pre-test to the immediate post-test across the entire study, followed by a drop in the delayed post-test scores.

TABLE 4

N=466	Mean pre-test score	Mean immediate post-test score	Mean delayed post-test score
Whole cohort	-0.13	0.08	-0.04

Table 5 shows the scores for the control and intervention groups and shows the effect size of the intervention for both the immediate and delayed post-tests.

The standard deviation of the immediate post-test scores was 2.12.

The standard deviation of the delayed post-test scores was 2.14.

TABLE 5

N=466	Mean pre-test score	Mean immediate post-test score	Immediate effect size	Mean delayed post-test score	Delayed effect size
Control cohort N=217	-0.54	-0.14	-0.16	-0.38	-0.06
Intervention cohort N=249	0.22	0.27		0.25	

The intervention group appears to have made less progress during the intervention than the control group, but they made up some of that ground between the immediate and delayed post-tests.

The difference in pre-test normalised scores is somewhat surprising as the original allocation to groups was carried out in such a way as to ensure a similar prior attainment of control and

intervention cohorts (based on KS2 writing data). This difference can be attributed to a number of possible factors:

- The actual distribution of the pre-test scores is different from the actual distribution of KS2 writing data – the pre-test scores for a small number of the control cohort were very low.
- The prior attainment data was less 'up to date' for Year 8 and Year 9 pupils than for Year 7 pupils.
- The writing task used in this study was evaluating different skills than those used to inform the KS2 writing scores used for prior attainment.

Similar analysis by gender and disadvantage has been carried out with results presented below. Because the original allocation was based on matching prior attainment for the cohort as a whole, it is possible that prior attainment for pupil groups (male, female, disadvantaged) may not be strictly comparable.

TABLE 6: MALE PUPILS

N=229	Mean pre-test score	Mean immediate post test score	Immediate effect size	Mean delayed post-test score	Delayed effect size
Control N=107	-1.46	-0.73	-0.35	-0.87	-0.23
Intervention N=122	-0.07	-0.11		-0.02	

For the males (N=229) the standard deviation for the immediate post-test scores was 2.17 and for the delayed post-test scores was 2.28.

TABLE 7: FEMALE PUPILS

N=237	Mean pre-test scores	Mean immediate post-test score	Immediate effect size	Mean delayed post-test score	Delayed effect size
Control N=110	0.36	0.44	0.04	0.10	0.14
Intervention N=127	0.49	0.64		0.50	

For the females (N=237) the standard deviation for immediate post-test scores was 1.96 and for the delayed post-test normalised scores was 1.93. Again, the delayed effect size increases from the immediate effect size.

TABLE 8: DISADVANTAGED PUPILS

N=41	Mean pre-test score	Mean immediate post test score	Immediate effect size	Mean delayed post test score	Delayed effect size
Control N=23	-0.41	-0.52	-0.08	-0.89	-0.20
Intervention N=18	0.29	-0.05		-0.60	

The standard deviations for the disadvantaged pupils (N=87) were 1.89 for the immediate post-test scores and 2.03 for the delayed post-test scores. This is the only pupil group we have looked at where the delayed effect size is less positive than the immediate effect size and where there is not an uplift from the pre-test score to the immediate post-test score.

Secondary research question:

Question

Thinking about the amount of time you have spent outside of the classroom for this sequence of lessons, what is your perception of the workload associated with providing feedback to the pupils compared with the normal amount of time outside of lessons you would expect to spend providing feedback?

1=significantly less time than normal

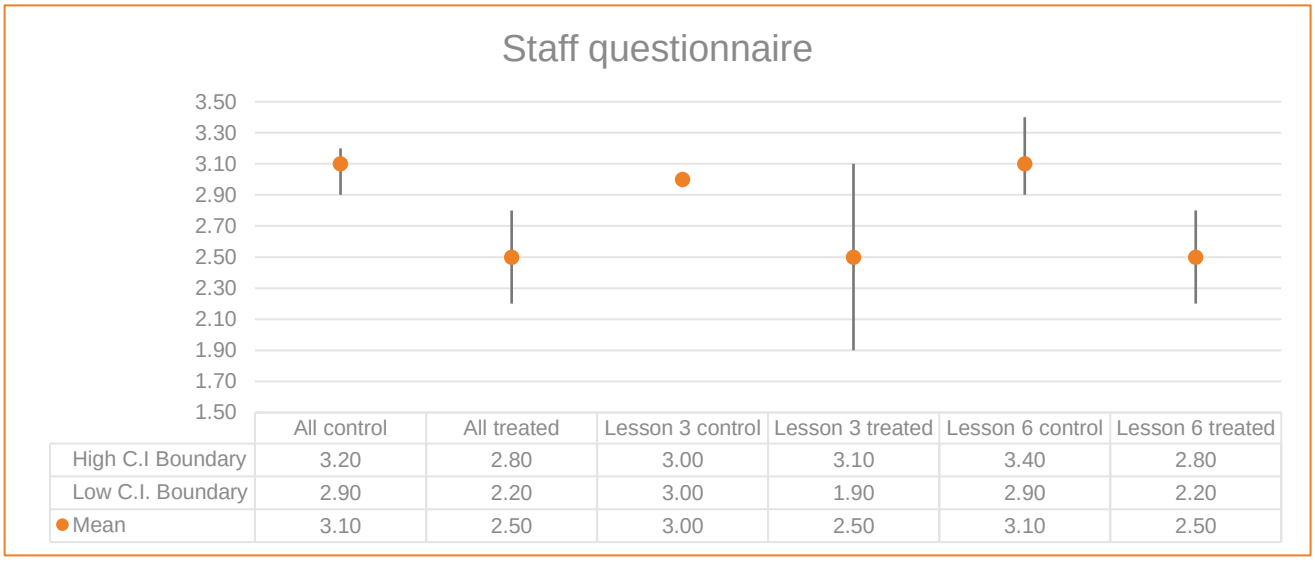
2=less time than normal

3=same time as normal

4=more time than normal

5=significantly more time than normal

CHART 1



	All control	All treated	Lesson 3 control	Lesson 3 treated	Lesson 6 control	Lesson 6 treated
High C.I Boundary	3.20	2.80	3.00	3.10	3.40	2.80
Low C.I. Boundary	2.90	2.20	3.00	1.90	2.90	2.20
Mean	3.10	2.50	3.00	2.50	3.10	2.50

The data suggests that teachers perceive a lower workload associated with peer comparative judgement.

A one-tailed t-test performed on the aggregate responses gives a p-value of < 0.001 and strongly supports this finding.

The difference in staff responses between Lessons 3 and 6 is to be expected as teachers became more efficient as they got used to a new way of working. We would therefore expect a more positive response relating to workload after Lesson 6 than Lesson 3 which was in fact observed. The fact that not all control teachers responded with a '3' suggests that they may have conflated work associated with adherence to the trial protocols with work associated with giving feedback, or that the planned teaching sequence had a heavier marking load than is typical for them.

There were a small number of free text comments from teachers, with only one recorded by the intervention teachers which is included below in its entirety.

> "I think that there were definite benefits to using the software and that pupils enjoyed the experience. On the whole, I heard some effective discussions between pupils when they were comparing work in the computer room and it did encourage them to look at their work through the eyes of an examiner. However, I think they would have benefitted from further discussion on the ranking of the pieces and why they had been ranked in this way. I also felt that using the software did not adequately replace the need for individual feedback on students' work."

Process evaluation findings

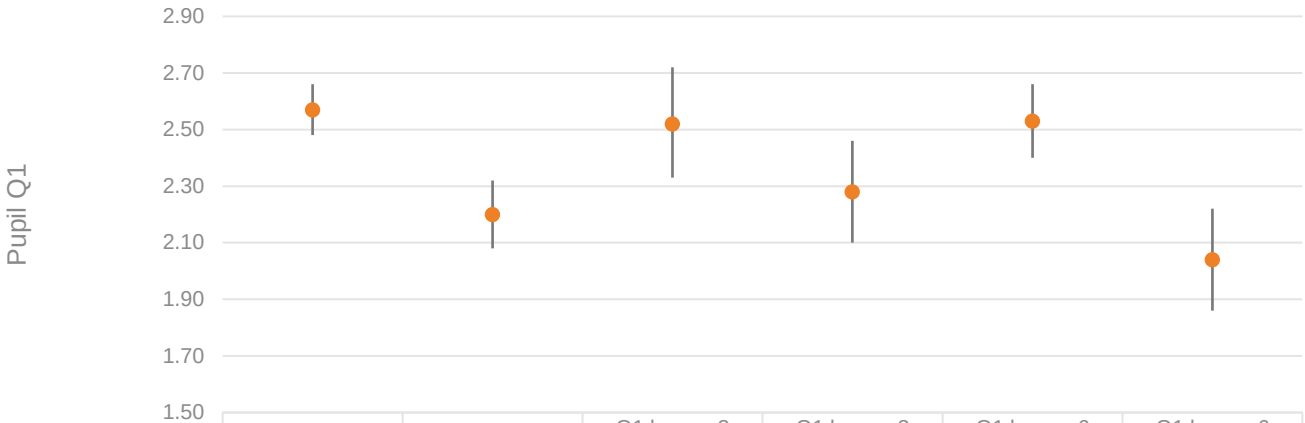
Analysis of pupil responses

The results are presented in the same way as for the teacher responses. For the control cohort N=217 and intervention cohort N=249.

Question 1

> *This lesson was enjoyable compared to your normal lessons*
>
> 1=strongly agree
>
> 2=agree
>
> 3=neither agree nor disagree
>
> 4=disagree
>
> 5=strongly disagree

CHART 2



	Q1 all control	Q1 all treated	Q1 lesson 3 control	Q1 lesson 3 treated	Q1 lesson 6 control	Q1 lesson 6 treated
High C.I. Boundary	2.66	2.32	2.72	2.46	2.66	2.22
Low C.I. Boundary	2.48	2.08	2.33	2.10	2.40	1.86
Mean	2.57	2.20	2.52	2.28	2.53	2.04

Question 2

This question relates to the lesson(s) where pupils were looking at examples of older pupils' work.

Because of <u>the last</u> lesson I understand what a good example of work looks like

1=strongly agree

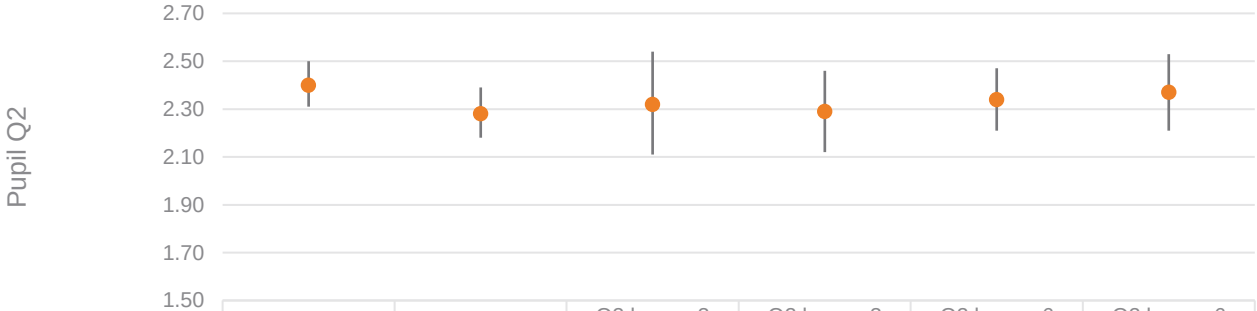
2=agree

3=neither agree nor disagree

4=disagree

5=strongly disagree

CHART 3



	Q2 all control	A2 all treated	Q2 lesson 3 control	Q2 lesson 3 treated	Q2 lesson 6 control	Q2 lesson 6 treated
High C.I. Boundary	2.50	2.39	2.54	2.46	2.47	2.53
Low C.I Boundary	2.31	2.18	2.11	2.12	2.21	2.21
Mean	2.40	2.28	2.32	2.29	2.34	2.37

The combined responses for intervention pupils suggests that intervention groups felt that the

process of comparatively judging older pupils' work was at least as beneficial as 'normal' ways of looking at exemplar work.

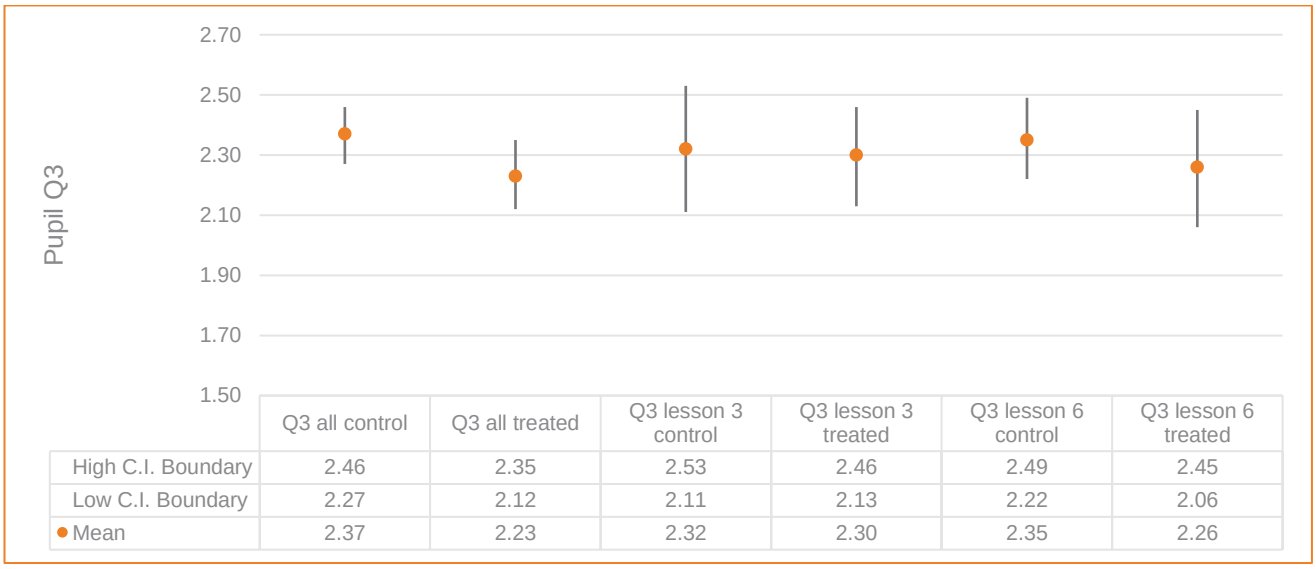
A one-tailed t-test on the aggregate responses gives a p-value <0.05 supporting this finding.

Question 3

This question relates to the lesson(s) where pupils were looking at examples of work from their own class.

> *Because of <u>this</u> lesson I understand what a good example of work looks like.*
>
> 1=strongly agree
>
> 2=agree
>
> 3=neither agree nor disagree
>
> 4=disagree
>
> 5=strongly disagree

CHART 4



	Q3 all control	Q3 all treated	Q3 lesson 3 control	Q3 lesson 3 treated	Q3 lesson 6 control	Q3 lesson 6 treated
High C.I. Boundary	2.46	2.35	2.53	2.46	2.49	2.45
Low C.I. Boundary	2.27	2.12	2.11	2.13	2.22	2.06
Mean	2.37	2.23	2.32	2.30	2.35	2.26

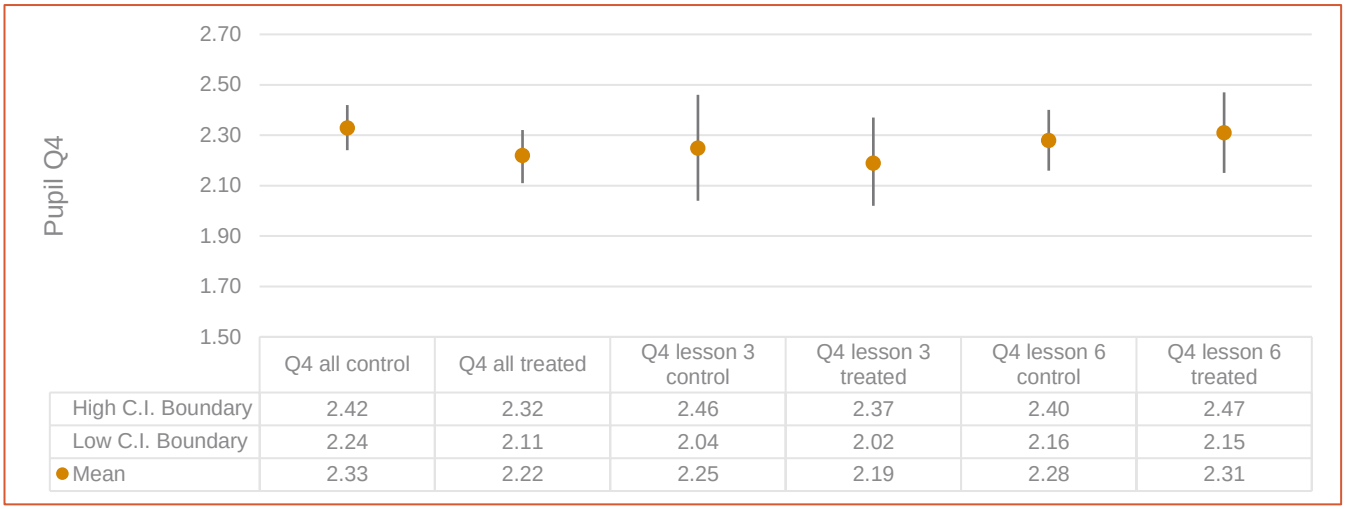
The data plot suggests that pupils may prefer comparative judgement to learn from their peers' work.

A one-tailed t-test on the aggregate responses gives a p-value <0.05 to support this finding.

Question 4

> *I could describe to you what a good example of work looks like*
>
> 1=strongly agree
>
> 2=agree
>
> 3=neither agree nor disagree
>
> 4=disagree
>
> 5=strongly disagree

CHART 5



	Q4 all control	Q4 all treated	Q4 lesson 3 control	Q4 lesson 3 treated	Q4 lesson 6 control	Q4 lesson 6 treated
High C.I. Boundary	2.42	2.32	2.46	2.37	2.40	2.47
Low C.I. Boundary	2.24	2.11	2.04	2.02	2.16	2.15
Mean	2.33	2.22	2.25	2.19	2.28	2.31

These results suggest that pupils receiving the treatment may feel better able to describe what a good piece of work looks like. A one-tailed t-test gives $p > 0.05$ suggesting that this result is not significant.

Question 5

I know what specific things I need to do to improve my work

1=strongly agree

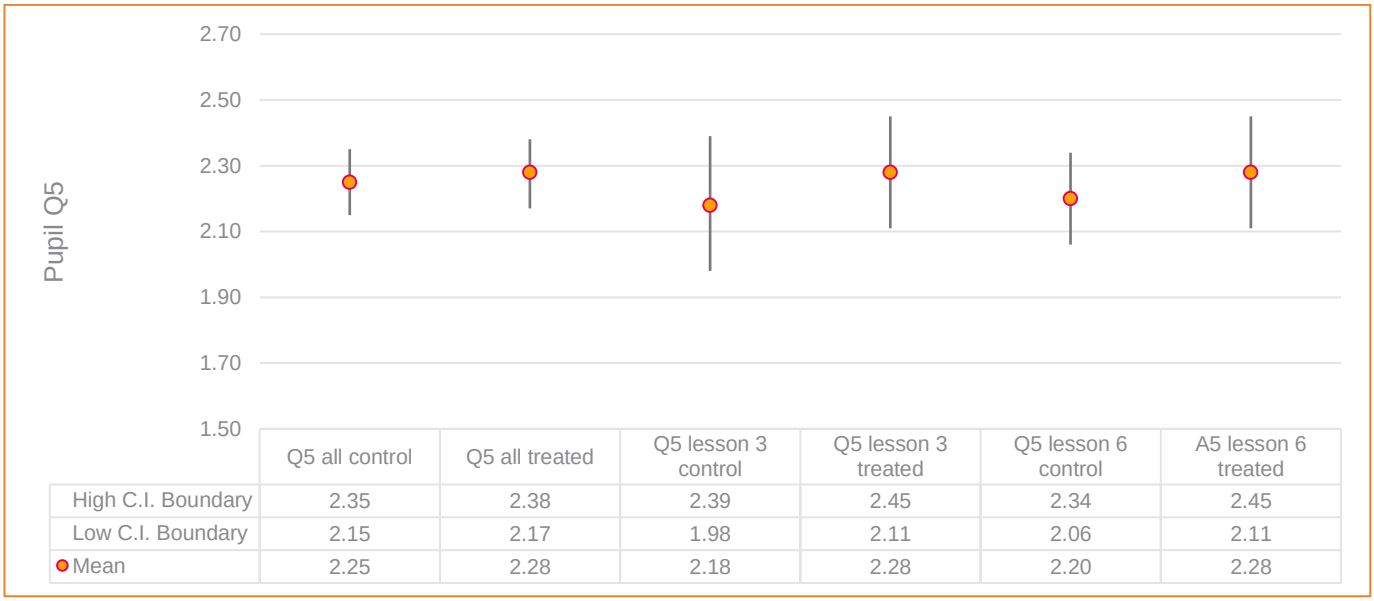
2=agree

3=neither agree nor disagree

4=disagree

5=strongly disagree

CHART 6



	Q5 all control	Q5 all treated	Q5 lesson 3 control	Q5 lesson 3 treated	Q5 lesson 6 control	A5 lesson 6 treated
High C.I. Boundary	2.35	2.38	2.39	2.45	2.34	2.45
Low C.I. Boundary	2.15	2.17	1.98	2.11	2.06	2.11
Mean	2.25	2.28	2.18	2.28	2.20	2.28

The results of this question are extremely powerful. Most marking and feedback involves either direct or indirect advice on how to improve your work. There is only a slight difference in pupil responses between the control cohort (who have received this 'normal' feedback) and the intervention group who haven't, with a one-tailed t-test giving a p-value of 0.74.

The eagle-eyed will note that the aggregated average for the intervention cohort does not look plausible given the separate Lesson 3 and Lesson 6 averages for the intervention cohort. This is due to some of the returns from pupils not identifying whether their response related to Lesson 3 or Lesson 6. We have not been able to include these responses in the separate calculations for Lesson 3 or Lesson 6, but it is obviously appropriate to include these within the aggregated results.

Question 6

It will be easy for me to produce a better piece of work now

1=strongly agree

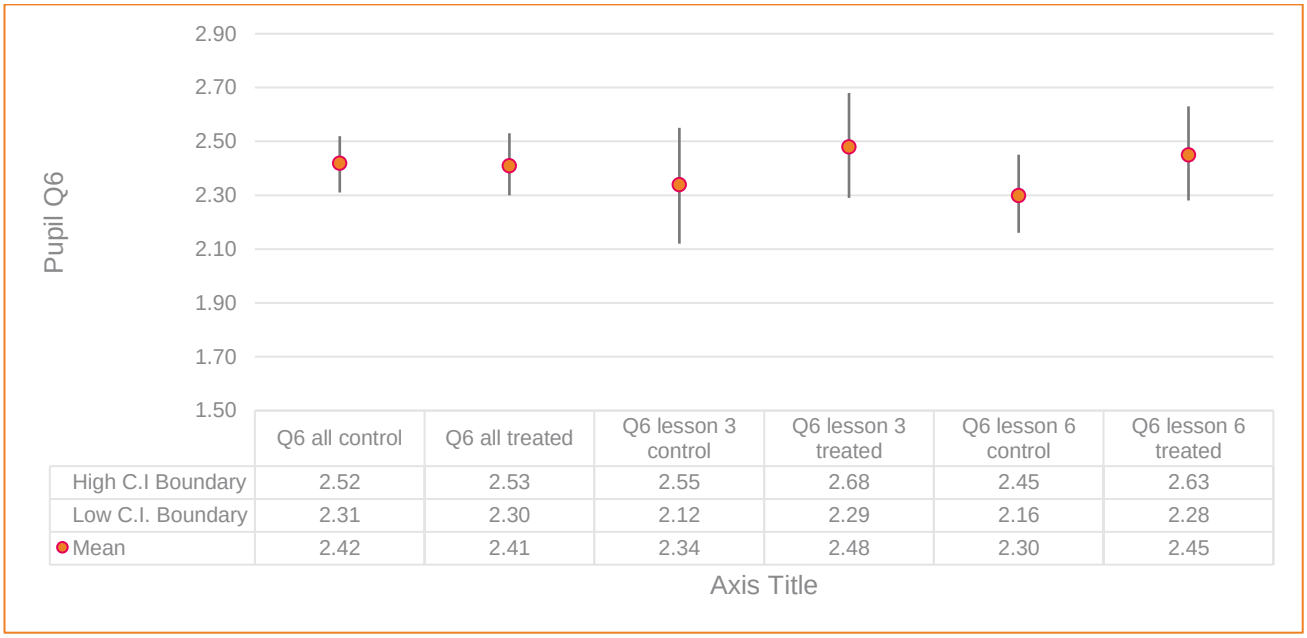
2=agree

3=neither agree nor disagree

4=disagree

5=strongly disagree

CHART 7



	Q6 all control	Q6 all treated	Q6 lesson 3 control	Q6 lesson 3 treated	Q6 lesson 6 control	Q6 lesson 6 treated
High C.I Boundary	2.52	2.53	2.55	2.68	2.45	2.63
Low C.I. Boundary	2.31	2.30	2.12	2.29	2.16	2.28
Mean	2.42	2.41	2.34	2.48	2.30	2.45

Again there is no discernible difference in the aggregated results for pupils who received the treatment and therefore did not receive conventional teacher marking and feedback.

The apparent anomaly between the aggregated control average and the separate averages for the control Lesson 3 and Lesson 6 responses is explained in a similar way to the apparent anomaly for pupil responses to Question 5.

Question 7

Box for free text comments.

Pupil comments for groups receiving treatment can be grouped into the following broad themes:

- Enjoyable tasks. ”I enjoyed doing this lesson and it has inspired me to use lots of different devices.”
- Helpful to see examples of other pupils’ work. “A lot more interesting because you could compare your work with other peoples and recognise weaknesses that could be applied.”
- Legibility of some of the work. “Some of the creative writing was illegible. therefore hard to compare with others.”
- Helpful in making progress. ”I normally put my own words down and am not normally pleased with it so this helped.”

It is interesting to note that there were comments in the control group about the ‘lack of clarity’ or ‘confusing nature’ of the mark scheme that was provided, but no similar comments amongst the intervention cohort. This could suggest that the control teachers relied more heavily on a mark scheme in lessons whereas the teachers of intervention groups did not expose pupils to these in the same way.

There were also comments from the control group about the repetitive nature of the tasks that they were being asked to do. This is interesting as the intervention group were repeating the same sorts of activity but did not comment on this. This may be as a directly related to pupils’ enjoyment of the comparative judgement process and may partly explain why the control cohort’s responses to Question 1 were less positive than those of the intervention cohort.

Discussion/Conclusion

The answer to our primary research question is that, in this study, the use of No More Marking by pupils for two cycles of work over a period of one to two months, specifically in key stage 3 (KS3) descriptive writing, led to pupil outcomes in descriptive writing that are comparable to the use of conventional teacher marking when considering the delayed post-test effect sizes.

This statement masks the fact that when considering the delayed effect sizes, this intervention had a positive effect size for girls vs a negative effect size for boys. The intervention had a larger negative effect size for disadvantaged pupils than for non-disadvantaged pupils. In addition the effect sizes of the intervention were more negative on the immediate post-test than on the delayed post-test.

The answer to our secondary research question is that, in this trial, the use of No More Marking by pupils in the manner described above does reduce teacher workload (measured by teacher perception of their own workload).

The study revealed interesting data from pupils relating to a higher level of engagement with the intervention lesson activities, and pupils' own perception of the increased utility of the intervention lesson activities. Alongside this it is insightful to note that there was no difference in pupil responses between the intervention and control cohorts relating to knowing what to do to improve their work. Labour-intensive traditional marking and feedback is all about trying to make sure pupils know what to do to improve their work. Pupil perception was that they were equally as informed about how to improve their work after carrying out comparative judgement activities that required less teacher input outside of the classroom.

Our study shows the effect sizes becoming less negative over time (delayed post-tests showing intervention pupils making up ground). Good quality teaching is about embedding knowledge, understanding and skills for the long term, rather than training pupils to 'perform' for an exam at a particular point in time and this observation is therefore of interest.

Further evaluation is necessary to establish how easily these results translate to other subjects and key stages. Reduction in workload (without eroding pupil outcomes) is important to retain high quality teachers and the results of this study – namely similar pupil outcomes with a statistically significant reduction in perceived workload for teachers therefore has import to the profession. However, a better understanding is also needed as to why there is such a difference in effect sizes between boys and girls.

Within our own school these results will be shared with heads of department who will be encouraged to trial similar approaches. We will also share the findings with those schools involved in the trial who we anticipate will disseminate these results similarly. We also look forward to sharing our findings with those networks that we are actively engaged with.

In conclusion, this study provides evidence that teachers can reduce their workload without impacting negatively on overall pupil outcomes in the medium term. This study should promote further questions in schools around why teachers are spending large amounts of time providing conventional feedback when other, more efficient methods are available that may have similar effects on pupil outcomes. Given the urgent need to address workload in order to retain teachers in the majority of subject areas, this study should also act as a catalyst for future research in this area, particularly with respect to the differing performance of boys, girls, and disadvantaged pupils.

References

Carroll WM, (1994). Using worked examples as an instructional support in the algebra classroom, *Journal of Educational Psychology*, 86(3), 360-367.

Elliott V et al., 2016. A marked improvement, England: EEF, Oxford University.

Evans S & Swan M, (2014). Developing students' strategies for problem solving in mathematics: the role of pre-designed "sample student work", *Educational Designer*, 2(7). http://www.educationaldesigner.org/ed/volume2/issue7/article25/index.htm

Jones I & Alcock L, (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones I & Wheadon C, (2015). Peer assessment using comparative and absolute judgement, *Studies in Educational Evaluation*, 47, 93–101. https://doi.org/10.1016/j.stueduc.2015.09.004

McMahon S & Jones I, (2015). A comparative judgement approach to teacher assessment, *Assessment in Education: Principles, Policy & Practice*, 22(3), 368–389. https://doi.org/10.1080/0969594X.2014.978839

Pachur T & Olsson H, (2012). Type of learning task impacts performance and strategy selection in decision making, *Cognitive Psychology*, 65(2), 207–240. https://doi.org/http://doi.org/10.1016/j.cogpsych.2012.03.003

Seery N, Canty D & Phelan P, (2012). The validity and value of peer assessment using Adaptive Comparative Judgement in design driven practical education, *International Journal of Technology and Design Education*, 22(2), 205–226.

Contact us

+44 (0)1904 328166 info@the-iee.org.uk
Berrick Saul Building, University of York, York YO10 5DD
Twitter: @IEE_York the-iee.org.uk/

The Institute for Effective Education (IEE) is an independent charity working to improve education for all children by promoting the use of evidence in education policy and practice.

In collaboration with the Education Endowment Foundation (EEF) we support a national Research Schools Network and have developed resources aimed at people on the front line of education.

The Institute for Effective Education is a charity registered in England, charity number 1168744

Institute for
Effective Education
Empowering educators with evidence

Research
Schools Network