Institute for
Effective Education
Empowering educators with evidence

Research Schools Network

Pupil use of comparative judgement in descriptive writing

Notre Dame High School, Sheffield

Problem: What challenges does your school have that need to be addressed?

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

"Despite its centrality to the work of schools and teachers, there is in fact little high-quality research related to marking."

"There is an urgent need for more studies so that teachers have better information about the most effective marking approaches."

- ***A marked improvement?*** (Elliott, et al., 2016)

Making judgements about the work of peers, comparing example worked solutions and using exemplars in lessons have all been demonstrated to have a positive impact on pupil learning. Previous research has shown 13-, 14- and 15-year-old pupils are able to make reliable comparative judgements in science and maths. The current evaluation will explore whether using comparative judgement software can reduce teacher workload and will investigate the impact of this approach on pupil outcomes.

Innovation: How will the innovation help improve the problem you have identified and benefit teachers and learners?

No More Marking is an online comparative judgement platform that allows work to be assessed holistically, producing a score for each pupil. Research (Jones & Wheadon, 2015; McMahon & Jones, 2014) indicates that lower secondary school pupils are surprisingly adept at assessing

their peers' work using comparative judgement. Our innovation is substituting teacher marking and feedback for a structured lesson where the pupils use No More Marking to make multiple comparative judgements on the work concerned in order to better understand "what a good one looks like". Teachers often use exemplars, but they tend not to do so comparatively (Evans & Swan, 2014) – this innovation addresses that.

Existing evidence: What evidence is there that this innovation will improve outcomes?

Pupils' competence when assessing peers' work using comparative judgement has been demonstrated for the case of mathematics (Jones & Wheadon, 2015) and science (McMahon & Jones, 2014). We are confident this finding will be replicated for the case of descriptive writing.

Qualitative evidence has been published as to the value of pupils comparatively judging peers' work for learning (Jones & Alcock, 2014; Seery et al., 2012). Moreover, educational and psychological research has demonstrated the promise of comparing example worked solutions for learning (eg. Evans & Swan, 2014; Pachur & 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.

Research question or hypothesis: What effect will the intervention, implemented for how long, with which pupils, have on what outcomes?

Our primary hypothesis is that the use of No More Marking by pupils for two cycles of work over a period of one to two months, specifically in KS3 descriptive writing, will lead to pupil outcomes in descriptive writing that are comparable to the use of conventional teacher marking.

Our secondary hypothesis is that undertaking peer assessment with comparative judgement for two cycles of work over a period of one to two months, result in greater gains in KS3 descriptive writing from pre-test to post-test than the use of conventional written marking.

Additionally, we hypothesise that the use of No More Marking by pupils will reduce teacher workload (measured by teacher perception of their own workload).

Method: Include sample, design, measures, intervention, process evaluation and analysis

Sample / participants

Our evaluation will involve approximately 600 KS3 pupils.

The KS3 pupils will comprise 24 complete classes of Years 7, 8 and 9 pupils from four secondary schools. Two classes from each year group will take part from each participating school. The prior ability of each class will be measured using KS2 scores. We will also collect pupil level data to allow us to interrogate our results by gender, prior ability, etc.

A letter will be sent to the parent of each participating child, enabling them to withdraw their child's data from the analysis.

Participating schools are listed in the appendices. We have ensured that pupils in these schools have not been previously exposed to comparative judgement software.

Design and assignment to condition

Four classes from each year group will be assigned to control and four classes to treatment. We will use the prior ability data to ensure that the ability across the four control and treated groups in each year group is broadly similar.

Measures

Pupils will produce an initial piece of work, an immediate post-test piece of work and a final delayed post-test piece of work. This will be the same piece of work for all year groups involved so that an overall effect size can be calculated. The first piece of work will serve as a pre-test. The pre-test and both post-tests will be ranked using No More Marking by participating teachers. in order to remove bias, No More Marking allows 'blind marking' to ensure that no teacher makes any comparison between one of their pupils and a pupil from another class. The volume of tests requiring comparison means that it is not practically possible for participating teachers to undertake all comparisons (in terms of cover costs and implications). We will therefore employ appropriately qualified PhD students to undertake those comparisons that are still to be made at the end of the moderation event. The comparative judgement will include grade descriptors that act as anchors to provide an absolute rank as well as a comparative ranking. Additionally, we will employ a subject expert to moderate a sample of 20% of the pre-test and delayed post-tests to

ensure that there is no inbuilt bias as a result of using comparative judgement to measure the effect of comparative judgement.

We will investigate reliability using internal consistency and split-halves techniques (Jones & Alcock, 2014). We will investigate validity by calculating the correlation of the outcomes of the lesson 2 peer assessment activity with (i) teacher judgements and (ii) existing achievement data (SATs scores in English).

Intervention

The trial will be a sequence of seven lessons delivered by pupils' usual teachers in normal class time, with a further delayed post-test carried out later. The treated group will receive four 'treatments' in lessons two, three, five and six.

Participating teachers will be contractually obliged to attend pre-trial training and we will produce a research manual for each participant.

- **Lesson one**: 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 will be used as a pre-test.
- **Lesson two:** pupils evaluate examples of work produced by older pupils. The treated group will do this via a structured No More Marking Lesson. The control group will be free to do this in any way that they wish, provided that they avoid the pupils making multiple comparisons. We will provide some 'traditional' suggestions as to how they could do this.
- **Lesson three:** pupils reflect on their own first drafts. The treated group will do this via a structured No More Marking Lesson. We will include the examples of work that were used in lesson two alongside each treated group's own work in the set of work that the pupils make comparative judgements on. The control group will receive feedback on their draft and the teacher will be free to lead this reflective session in any way that they wish, provided that they avoid the pupils making multiple comparisons. We will provide some 'traditional' suggestions as to how they could do this.
- GAP of at most one month.
- **Lesson four:** 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 five**: pupils evaluate examples of work produced by older pupils. The treated group will do this via a structured No More Marking Lesson. The control group will be free

to do this in any way that they wish, provided that they avoid the pupils making multiple comparisons. We will again provide some 'traditional' suggestions as to how they could do this.

- **Lesson 6:** pupils reflect on their own first drafts. The treated group will do this via a structured No More Marking Lesson. We will include the examples of work that were used in lesson five alongside each treated group's own work in the set of work that the students make comparative judgements on. The control group will receive feedback on their draft and the teacher will be free to lead this reflective session in any way that they wish, provided that they avoid the students making multiple comparisons. We will provide some 'traditional' suggestions as to how they could do this.
- **Lesson seven:** pupils produce a written answer to a descriptive writing exam question (again in the style of GCSE English Paper 1 Section B). This will serve as the immediate post-test.
- GAP of as close to four weeks as is manageable in the context of the school timetable.
- **Lesson eight:** pupils produce a further written answer to a descriptive writing exam question (again in the style of GCSE English Paper 1 Section B). This will serve as the delayed post-test.

We will commission a subject specialist to develop lesson resources and to translate the GCSE band descriptors into appropriate prose statements for KS3 pupils. These prose statements will be used as an assessment framework throughout (for both students and teachers).

Process evaluation

We will collect qualitative and quantitative data from all participants following lesson three and six via questionnaires containing Likert-type and open-text questions.

The questionnaires are included in the appendices.

Additionally we will observe a sample of both control and treated feedback lessons to provide further qualitative data.

Data analysis

All data will be anonymised by assigning an ID number to each pupil, teacher and school. Any identifying marks on written work will be removed prior to uploading to the No More Marking website.

We will compare the pre-test and post-test rankings of control and treated groups to calculate an overall effect size for the intervention. We will also calculate effect sizes within each school and

year group.

The No More Marking website generates scores and further control statistics (internal consistency and other measures) which will be downloaded as csv files. The main analysis will be conducted using standardised methods from the literature (Jones & Alcock, 2014) using the software package R. Correlations will be calculated to establish the validity and reliability of assessment outcomes, and Bayesian analysis will be conducted to establish the learning gains of the intervention and control groups.

Conclusion: What will happen if your innovation improves outcomes, or not? What are the limitations of your evaluation?

If the innovation is successful we would look to secure funding to extend the trial over a wider range of age groups and subjects. If the innovation is not successful, we will interrogate the qualitative data for any clues that suggest there is promise in the innovation, but that the methodology chosen was a barrier for its effectiveness.

Limitations are:
We do not anticipate demoralisation of the control group teachers, but nevertheless will take steps to minimise this. We will run separate training events for teachers of control and treated groups in order to do so.

Deviation of participating teachers from the methodology. We will run a training session for all teachers and produce a research manual that describes in detail each aspect of the study. A sample of four lessons (two intervention, two control) will be observed by an independent education expert to evaluate conformity to the intervention and control design.

Diffusion of treatment. Research evidence suggests that comparing examples of pupil work is not common practice (Evans & Swan, 2014). To minimise this limitation we will explicitly request control teachers not to compare pupils' work.

We will communicate our findings via Huntington Research School, our TSA website and our project report to the IEE.

References

Carroll W M (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.