Pragmatic cluster randomised controlled trial of contextualised grammar teaching and small group teaching to improve the writing skills of 11 year old children

Natasha Mitchell, Hannah Ainsworth, Hannah Buckley, Catherine Hewitt, Laura Jefferson, David J Torgerson, Carole J Torgerson.

York Trials Unit
Department of Health Sciences
University of York
YORK YO10 5DD

Dr Natasha Mitchell
Ms Hannah Ainsworth
Ms Hannah Buckley
Dr Laura Jefferson
Professor David Torgerson

School of Education
Leazes Road
Durham University
DURHAM DH1 1TA

Professor Carole Torgerson

Correspondence to: Carole Torgerson

carole.torgerson@durham.ac.uk

Keywords: grammar teaching; writing skills; randomised controlled trial; small group teaching

Abstract

Introduction: We evaluated two interventions: a contextualised grammar teaching intervention – *Grammar for Writing* - to assess whether it improved 11 year old children's writing skills; and a small group literacy intervention to assess whether or not this was effective.

Design and method: We used a pragmatic cluster randomised trial with partial split plot design. Independent concealed randomisation was undertaken at the class level, and, within the intervention group, children were also individually randomised to receive the whole class intervention plus a small group intervention or to receive the intervention in a whole class setting only. The main outcomes were writing and reading assessed by the Progress in English 11 (Long Form) test (GL Assessment).

Results: In 2013, 55 schools in England, each with two classes, were recruited and randomised. Within each school, the two classes were randomly allocated to receive either the intervention or the control condition. After randomisation, 2 schools withdrew, leaving 53 schools, 106 classes and 2510 pupils. We observed an effect size (ES) of 0.10 favouring the *Grammar for Writing* classes; however, this was not statistically significant (95% confidence interval (CI) -0.10 to 0.31). Pupils randomised to the small groups had an increased literacy score when compared with the control classes (ES = 0.24, 95% CI 0.00 to 0.49) and when compared with the intervention children taught in the whole class (ES = 0.21, 95% CI 0.04 to 0.38).

Conclusion: There is little evidence that this form of contextualised grammar teaching had an effect on 11 year old children's writing skills. There was some evidence of an effect for small group teaching.

Background

In the United Kingdom (UK), the move to the first year of secondary school at age 11-12 (year 7) from the final year at primary school at age 10-11 (year 6) is commonly known as the 'transition' from primary to secondary school. The equivalent in the United Sates is the transition from elementary school to middle or junior high school between the 5th and 6th grades. In the UK transition starts early in year 6 and does not end until sometime after pupils have settled into their new school (Evangelou et al, 2008). Recent Ofsted figures suggest attainment in English in the UK at age 11 at the end of key stage 2 (KS2) has remained static: 79% of pupils achieved the expected level 4 or above in 2005, with a slight rise to 82% in 2011 (Ofsted, 2012). The figures for 2012 and 2013 were 75% and for 2014 79% (DfE, 2014). In terms of writing standards, there is some evidence of a slight improvement overall: in 2014 attainment increased by 2 percentage points at level 4 or above and 3 percentage points at level 5 or above (DfE, 2014).

The *Grammar for Writing* intervention is a curriculum intervention aimed at improving writing skills by providing contextualised grammar teaching. The intervention in this trial was a modified version of an existing grammar intervention aimed at improving writing skills in older children, with the modified version targeting lower attaining writers in year 6. An evaluation of the existing intervention using a trial design was undertaken by the developers (Jones et al, 2012; Myhill et al, 2012), which found some evidence that it was effective in enhancing writing performance in year 8 pupils. Myhill et al (2012) also found the intervention benefited sub-groups differentially - improving higher attaining writers' writing

more, compared with struggling writers' writing. However, this trial had a number of limitations. First, the authors did not use intention to treat analysis because they removed a school allocated to the comparison group due to 'poor' implementation. Second, the authors did not adjust for the clustered nature of the data, i.e., the statistical analysis assumed individual randomisation, when cluster randomisation had been used so this would have produced biased standard errors (Bland, 2010).

Our evaluation using a trial design was developed in the light of the evidence from the previous developer-led trial and focused on a younger year group, thus adding to the evidence-base around interventions to improve writing in this age group. In addition, our study also sought to add to the evidence base for small group teaching. The Education Endowment foundation's Toolkit provides summaries of the effectiveness of educational interventions using evidence from an overview of available meta-analyses in the specific topic areas (EEF, 2015). The Toolkit reviewed the evidence on small group teaching and found moderate impact for moderate cost based on limited evidence (EEF, 2015). The Toolkit also reviewed the evidence on one-to-one tuition and concluded that it provides modest impact for high cost based on extensive evidence (EEF, 2015). There is some evidence from the US for the effect of smaller class sizes positively affecting children's performance most notably demonstrated by the Tennessee classroom experiment, which showed that when primary school children were randomly allocated to be taught in smaller classes, pupils in the smaller classes performed significantly better than those in the larger classes (Mosteller, 1995). More recently, evidence from the Toolkit on reducing class size suggested that this intervention provides low impact for very high cost based on moderate evidence (EEF, 2015). We are unaware of any

randomised trial of small group teaching being conducted in a United Kingdom (UK) setting, however. Consequently, we evaluated *Grammar for Writing* both in a full class context and in a small group setting to enable us to ascertain whether small group teaching is an effective intervention in the UK context.

Our trial was funded by the Education Endowment Fund (EEF), and the developers of the intervention based at the University of Exeter were responsible for developing and delivering the *Grammar for Writing* intervention and for recruiting primary schools. The full report to the funders is available on the EEF website: http://educationendowmentfoundation.org.uk/projects/grammar-for-writing/

Research questions

The primary research question was 'What is the effectiveness of the *Grammar for Writing* intervention compared with a 'business as usual' control group on the writing skills of participating children?'

A secondary question was 'Does teaching a subgroup of children in small groups in addition to whole class teaching lead to better outcomes compared with teaching them using a whole class only approach.

Methods

Trial Design

A pragmatic cluster randomised trial with split plot design was used. Recruitment targeted schools with two year 6 classes. The classes were randomised into two

groups: a *Grammar for Writing* group and a 'business as usual' control group; one class in each school was randomised to receive the intervention and one class was randomised to receive the control condition. Within the intervention classes, individual children who met the inclusion criteria were randomised to receive either the whole class form of the intervention alone or to receive the whole class intervention plus a small group intervention. This design is known as a partial split plot; it is a variant of a factorial design, due to its combination of cluster and individual randomisation. A cluster randomised design was required in this case as the intervention was class-based, which precluded the use of individual randomisation. However, the addition of the individual level randomisation allowed for further investigation into the effect of *Grammar for Writing* when delivered as a small group intervention. The design, therefore, allowed us to examine the class level effects of *Grammar for Writing* by comparing the intervention classes with the control classes. Additionally, it allowed us two further comparisons of interest. First, it meant that we could disentangle any 'small group' treatment effects by comparing the outcomes for the small group pupils in the intervention group with the outcomes for their peers, who were not in small groups, in the control class. Second, we could also ascertain whether there was any additional advantage of delivering *Grammar for Writing* in a small group *in addition* to whole class teaching compared with whole class teaching alone: this was assessed by comparing those in the intervention classes who were randomised to 'small group' with those randomised to 'whole class'.

The trial was designed, conducted and reported to CONSORT standards (Altman et al, 2011) in order to minimise all potential threats to internal validity, such as

selection bias and a range of post randomisation biases (Cook and Campbell, 1969; Shadish, Cook and Campbell, 2002; Torgerson and Torgerson, 2008). In this way, unbiased estimates of the impact of the intervention are provided.

Recruitment

The evaluation team (University of York and Durham University) and the implementation team (University of Exeter), in collaboration with the National Association of Teachers of English (NATE), jointly provided information documentation about the trial to schools. Schools which wanted to take part were asked to sign an 'Agreement to participate' form to ensure they agreed to all the trial related procedures. Schools with high proportions of pupils eligible for free school meals and pupils achieving level 3 or borderline level 4 in English and, with, ideally, two year 6 classes, were targeted for recruitment to the trial.

Participating primary schools informed parents of all pupils in year 6 about the study using material provided by the evaluation team and the University of Exeter. Parents had the opportunity to withdraw their child's data from being used in the evaluation (so-called 'opt out consent') prior to randomisation. Participating primary schools then shared pupil data with the evaluation team (including pupil name, unique pupil number (UPN), gender, date of birth (DoB), free school meals (FSM) status, English as an additional language (EAL) status, key stage 2 (KS2) English teacher assessment from Dec 2012).

Eligibility

School inclusion criteria: Primary schools were eligible to take part in the trial if they agreed to all trial procedures, including: informing parents; provision of pupil data; randomisation; and implementation of the intervention as allocated.

Pupil inclusion criteria: Within the intervention class, pupils were eligible for individual randomisation if they were expected to achieve level 3c, level 3b, level 3a, level 4c or level 4b in English by the end of key stage 2 (based on teacher assessment).

School exclusion criteria: Primary schools were excluded from participating in the trial if they did not agree to all points listed in the 'Agreement to participate' form or if they were not able to carry out testing at the end of the intervention period.

Pupil exclusion criteria: Pupils were excluded from individual randomisation if they were expected to achieve below level 3 or above level 4b in English by the end of key stage 2. Exclusion also occurred if parents/guardians returned an opt-out form to the school, and in these instances no data were provided to the evaluators. Those predicted to achieve below level 3 were excluded from testing as it was thought the post-testing could have caused undue anxiety.

Intervention

The *Grammar for Writing* intervention was designed by the implementation team from the University of Exeter. It involved a continuing professional development (CPD) day for all teachers (in the intervention condition) which was developed and delivered by Exeter University, and the use of teaching materials with embedded grammar teaching, with the aim of improving writing. The implementation team developed 15 sequential guided writing sessions; and the embedded grammar aspects encouraged pupils to make connections between a linguistic feature and the effect it

has in writing (Jones et al, 2012). The intervention focused on encouraging pupils to actively make grammatical choices which would affect how their writing would communicate to the reader; it did not focus on pupil's grammatical errors or any inaccuracies (Jones et al, 2012). Year 6 classes randomised to the intervention used their literacy class time to deliver the intervention. As above, eligible pupils within the intervention class were individually randomised to 'whole class' or to 'small group'. The intention was that individuals randomised to 'small group' would receive the intervention in the whole class setting and additional intervention delivery in a small group. [However, we cannot be certain that this small group teaching occurred in all schools as there was no fidelity assessment in terms of adherence to the trial design.] Pupils randomised to the 'business as usual' group received their usual literacy lessons as planned by their teachers.

Outcomes

Writing and reading achievement, assessed through the Progress in English (PiE) 11: Second Edition Long Form (LF) test (GL Assessment), were the literacy outcomes. The test includes both narrative and non-narrative exercises and assesses both reading and writing skills including areas such as spelling, grammar and comprehension. The Progress in English test was the only test available to the evaluation team (in order to comply with EEF testing policy) which included a writing component. Tests were marked by GL Assessment blind to allocation (i.e., markers did not know whether test papers were from the intervention or control pupils).

Primary outcome

The primary outcome was extended writing score which refers to the combined raw score on the two extended writing tasks (exercises 5 and 6) from the PiE 11 LF. Exercise 5 had a total possible 20 marks and involves writing a persuasive letter. Exercise 6 had a total possible 12 marks and assesses informative writing. Overall, the extended writing task score could be in the range 0 to 32, with a higher score representing higher attainment.

Secondary outcome

Reading score, the combined raw score on the reading tasks (exercises 3, 4, 3x and 4x), was used as the secondary outcome. Exercise 3 (comprising exercises 3 and 3x) had a total possible 19 marks and assessed reading comprehension of a narrative. Exercise 4 (comprising exercises 4 and 4x) had a total possible 13 marks and assessed non-narrative reading comprehension. Overall, reading score could range between 0 and 32, with a higher score representing better attainment.

Spelling and grammar score, the combined raw score on the spelling and grammar tasks (exercises 1 and 2) was chosen as a further secondary outcome. Exercise 1 had a total possible 10 marks and assessed spelling. Exercise 2 had a total possible 10 marks and assessed grammar. This means the spelling and grammar score combined could range from 0 to 20, with a higher score representing higher attainment.

Fidelity

Fidelity was assessed for every intervention class in the trial using a measure devised by the implementation team. The measure consists of three component scores relating to use of grammar terms, linking grammar effects in writing and using talk

to develop discussion about choices and effects. Each of these components was rated between 1 and 3 with 1 corresponding to 'rarely', 2 corresponding to 'partially as planned' and 3 corresponding to 'as planned.' As such, the fidelity score could range between 0 and 9, with higher scores corresponding to higher fidelity.

Delivery of outcomes

Teachers were asked to deliver the outcome tests. They were not blind to the group allocation of the children. However, they were asked to deliver the test under 'exam' conditions with the pupils in the classes sitting the test at the same time.

Sample size

For the purposes of calculating the sample size it was assumed 60 schools would be recruited with an average of 54 pupils per school; this would result in a total sample size of 3240 pupils. Assuming 27 pupils per class and an intra-cluster correlation coefficient of 0.19, the design effect would be 5.94. When divided into the total sample size, this produces an *effective sample size* of 546 pupils. However, assuming a pre- and post-test correlation of 0.70 the effective sample size increases to 1070. We allowed for an attrition rate of 10%, meaning the final effective sample size was 964 pupils. This allowed a difference of 0.18 standard deviations to be detected, with 80% power (2p = 0.05) in the writing scores of the intervention and control classes, should one exist.

The focus of this trial was on pupils who were performing between level 3c and level 4b; therefore the sample size calculation was based on this subgroup of children. For the individually randomised component of the trial, it was assumed that there would

be approximately 8 children per class in the 60 classes (480 pupils in total) and there would be a pre- and post-test correlation of 0.70 which would increase the effective sample size from 480 to 942. We allowed for an attrition rate of 10% which gave an effective sample size of 848 meaning a difference of 0.20 of a standard deviation (80% power; 2p = 0.05) in writing scores could be detected between the two randomised groups, if such a difference existed. If there were a modest intra-cluster correlation of 0.05 remaining, despite individual randomisation, then the effective sample size might decline to 630 participants as there would be a design effect of 1.35. This effective sample size would allow for detection of an effect size of 0.23 standard deviations (80% power, 2p = 0.05), should one exist.

Randomisation

Randomisation was conducted at two levels: class and individual. At the class level one class was randomised to receive the intervention and one class continued with 'business as usual' within each school. This randomisation was conducted using stratification by school with a fixed block size of 2. Further randomisation within the intervention class was conducted at the individual level for pupils predicted to achieve between level 3c and level 4b in KS2 writing. Eligible pupils were assigned to either receive the whole class form of the intervention only or to receive the whole class intervention plus small group intervention through deterministic minimisation within schools. Minimisation is a technique that ensures balance between the groups by using an arithmetical algorithm. The algorithm calculates the balance on specified variables after each individual has been allocated such that the next allocated individual minimises any chance imbalance between the groups (Torgerson & Torgerson, 2008). Gender and predicted KS2 writing level were used as

minimisation factors with two and three levels respectively. As each small group needed to contain between 4 and 6 pupils and due to the fact that class size varied, different allocation ratios were used depending on the number of eligible pupils in the intervention class at each school. In total, five allocation ratios were employed as below:

Number eligible pupils	Allocation ratio (whole: small)
6 or fewer	1:2
7-11	1:1
12-18	2:1
19-24	3:1
25-30	4:1

Both the class and individual level random assignments were conducted by the trial statistician (HB) based at York Trials Unit. Class randomisation was conducted in Stata® version 12 (Stata Corporation, College Station, Texas, USA); individual level minimisation was conducted using minimPy (http://sourceforge.net/projects/minimpy/). The class level allocation occurred first; individual level allocation occurred after the trial statistician received pupil baseline data. At each stage of randomisation the evaluation team provided this information to the implementation team for them to disseminate this information to the schools.

Analysis

Analysis was conducted in Stata® version 13 (Stata Corporation, College Station, Texas, USA) using the principles of intention to treat, meaning that all classes and pupils were analysed in the group to which they were randomised irrespective of whether or not they actually received the intervention and irrespective of implementation fidelity. Statistical significance was assessed at the 5% level unless

otherwise stated. Effect sizes were calculated and are presented alongside 95% confidence intervals. Effect size was defined as:

$$\Delta = \frac{\beta_{intervention}}{\sigma_{\varepsilon}}$$

where $\beta_{intervention}$ is the difference in mean score between the intervention and control groups and σ_{ε} is the residual standard deviation.

The test and outcomes were examined for ceiling or floor effects using summary statistics and graphical representations. Intra-cluster correlation coefficients (ICCs) were estimated and are presented alongside 95% confidence intervals.

Cluster Level Analysis

Primary analysis

The primary objective of this part of the trial was to investigate the effectiveness of the *Grammar for Writing* intervention on the writing skills of all pupils at level 3 and above. The difference in writing scores between pupils in the intervention classes and those in the 'business as usual' classes was compared using a multilevel regression analysis to allow for the hierarchical nature of the data. The model used extended writing score as the response variable with group allocation, gender, FSM status, English as an additional language (EAL) status, month of birth and predicted KS2 score included as fixed effects. School and class were included as random effects.

Secondary analyses

The primary analysis was repeated a total of four times. The first repetition used reading score as the response and the second used spelling and grammar score in

order to assess the impact of the intervention in terms of the secondary outcomes. The third analysis compared pupils allocated to receive additional small group teaching of the intervention with those in the control group at levels 3c, 3b, 3a, 4c or 4b. The effect of the intervention in terms of extended writing score was also analysed in the sub-group of pupils who were eligible for FSM through the inclusion of an interaction term in a final iteration of the primary analysis; for this analysis statistical significance was assessed at the 10% level.

Individual Level Analysis

Primary analysis

The primary objective of this trial was to investigate the effectiveness of the small group form of the intervention on the writing skills of eligible pupils. The difference in writing scores between pupils allocated to the whole class plus small group intervention and those to whole class intervention only was compared using a multilevel regression analysis with extended writing score as the response variable. Group allocation, gender, FSM status, EAL status, month of birth and predicted KS2 score were used as fixed effects in the model with class as a random effect. Although the trial was randomised at the individual level, because children were taught in classes or small groups there would still have been a clustering of outcomes, hence the need to use multilevel regression methods.

Secondary analyses

The primary analysis was repeated four times, first with reading score as the response and second with spelling and grammar score (to assess the impact of the intervention in terms of the secondary outcomes). A third analysis compared those

in the intervention class allocated to the whole group only with those in the control class. The effect of the intervention in terms of extended writing score was also analysed in the sub-group of pupils who were eligible for FSM through the inclusion of an interaction term in a final iteration of the primary analysis.

Results

Recruitment and follow-up of participants

The implementation team, in collaboration with the National Association of Teachers of English (NATE), recruited schools and pupils. School recruitment took place between January and March 2013. Four geographical areas in the UK were targeted: Sheffield, London, West Midlands and the South West. Originally it was proposed to recruit 60 schools with two classes per school. However, due to time constraints it was only possible to recruit 55 schools each with two classes. A total of 4 schools (8 classes) withdrew from the trial; two withdrawals occurred post cluster level randomisation and two occurred post individual level randomisation (during or after the intervention delivery period). This left 51 schools with 102 classes involved in the trial at the point of testing (Figure 1). At the time of class randomisation 2510 pupils were included in the trial. At the start of intervention delivery, 2500 pupils were involved in the trial (of whom 2394 were eligible for testing). By the testing period 2424 pupils remained and of these 2318 were eligible for testing (i.e., predicted to achieve level 3 or above).

INSERT Figure 1: CONSORT flow diagram ABOUT HERE

Baseline characteristics

There was a large proportion of missing data (a minimum of 47.3% missing for each variable) in relation to school level characteristics. The mean school size was around 439 pupils (SD 158.37). Around a third of pupils in the recruited schools were eligible for FSM and approximately 45% were of minority ethnic origin. Both of these percentages are considerably higher than the national UK averages in January 2013 which were reported at 19.2% and 28.5% respectively (Department for Education, 2013).

The mean class size in both the intervention and control classes was 23.6 pupils. The average number of pupils predicted to achieve between level 3c and level 4b was consistent between the control and intervention classes (between 14 and 15 pupils). Of the 53 schools for which pupil data were provided, classes at 32 schools involved pupils who were predicted to achieve below level 3. Not all schools included classes with pupils predicted below level 3 as some were organised in mixed attainment groups (44 schools) and others were organised in literacy groups (8 schools).

Table 1 shows baseline characteristics by cluster level allocation (i.e., intervention and control). In relation to the demographic characteristics of FSM status, pupil premium (PP) status, EAL status, month of birth and predicted KS2 writing level at baseline, proportions of pupils within each category were similar between the intervention and control arms both as randomised and as analysed in the primary cluster level analysis.

INSERT Table 1: Baseline pupil level characteristics ABOUT HERE

Of the 2394 pupils eligible for inclusion in the primary cluster level analysis, 412 (17.2%) were missing the primary outcome of extended writing score. The most common reason for missing primary outcome data was that none of the extended writing questions had been attempted (42.7% of 412); this included potential absence. Over one quarter of missing data was due to partial completion of relevant questions (27.4%) with school withdrawal and a test paper collection error being the reason for the remaining missing data (18.4% and 11.4% respectively).

Table 2 shows baseline characteristics by individual level allocation (i.e., whole class intervention only and whole class intervention plus small group form of intervention). In relation to the demographic characteristics of FSM status, PP status, EAL status, month of birth and predicted KS2 writing level at baseline, proportions of pupils within each category were similar between those allocated to remain in the whole group and those randomised to receive the additional form of the intervention. This is the case both as randomised and as analysed in the primary individual level analysis.

INSERT Table 2: Baseline characteristics of pupils who were individually randomised ABOUT HERE

Of the 777 pupils eligible for inclusion in the primary individual level analysis, 146 (18.8%) were missing the primary outcome of extended writing score. The most common reason for missing primary outcome data was that none of the extended writing questions had been attempted (45.2% of 146); this included potential absence. Over one quarter of missing data was due to partial completion of relevant questions (26.0%) with school withdrawal and a test paper collection error being the reason for the remaining missing data (21.2% and 7.5% respectively). A similar approach to missing data was taken as described above.

Summary statistics of data relating to the teachers were collected after the continuing professional education (CPD) days. Teacher data were collected from 53 of the 55 schools involved in the cluster level randomisation. At one school, given the re-randomisation of pupils into three classes (further details provided in the fidelity section below), data are recorded for two control teachers. This is also the case for another school where two staff taught one control class. In one further school, two teachers co-taught the intervention class and data were provided for both.

Data were available on 54 intervention teachers and 55 control teachers. The proportion of males in the intervention arm was slightly higher than in the control arm: 20.4% compared with 16.4%. An assessment of grammar knowledge was conducted by the implementation team using a test with a score ranging from 0 to 30 with higher marks relating to a higher level of grammar knowledge. Data relating to

this test are missing for 15 of the 109 teachers (13.8%), all of whom were teaching control classes. The mean grammar knowledge score was similar between arms, although a higher proportion of intervention teachers had fewer than 5 years' teaching experience (40.7% compared with 16.4%). The distribution of teacher age was fairly similar between the intervention and control groups; however, there were more intervention teachers aged between 26 and 30 years than control teachers (35.2% compared with 20.0%) and more control teachers aged between 36 and 40 than intervention teachers (20.3% compared with 3.6%). Missing data were more common in relation to control teachers, with over a quarter of data missing on each variable.

Outcomes and analysis

The test and outcomes were assessed for ceiling or floor effects using histograms and summary statistics; no evidence for either effect was found. Due to a collection error, post-test data were available for 50 of the 51 schools which remained in the trial at the point of testing. Pupils who were predicted to achieve below level 3 in KS2 writing at baseline are excluded from all analyses. Based on the results from these 50 schools, intra-cluster correlation coefficients (ICCs) were estimated (Table 3). These were somewhat larger than the one used for the sample size calculation estimates (i.e., 0.19). The correlation between outcome and the predicted KS2 level was also lower than expected (Spearman's Rho 0.54).

INSERT Table 3: Estimated intra-cluster correlation coefficients (ICCs) ABOUT HERE

Cluster level analysis

Raw, unadjusted mean post-test scores are presented in Table 4 by trial arm. Scores were similar in both allocated groups for all outcomes, as were proportions of pupils completing the relevant sections of the test for each outcome.

INSERT Table 4: Unadjusted average scores for the intervention and control groups ABOUT HERE

Results from the primary and secondary cluster level analyses are presented in Table 5. The number of pupils included in each analysis is shown alongside the adjusted difference in mean score between the allocated groups and associated effect sizes.

INSERT Table 5: Results from primary and secondary cluster level analyses ABOUT HERE

Primary analysis

There was little evidence of a difference in the primary outcome of extended writing score between the allocated groups (p=0.30), with an effect size of 0.10 (95% CI: -0.10 to 0.31), which was not statistically significant.

Secondary analyses

There was little evidence of a difference between the randomised groups in terms of reading or spelling and grammar score (p=0.14 and p=0.88 respectively). A reading effect size of 0.10 (95% CI: -0.03 to 0.24) and a spelling and grammar effect size of 0.01 (95% CI: -0.14 to 0.16) were found.

Control versus small group form of intervention

There was some evidence (p=0.05) of a difference in extended writing score between the allocated groups when control pupils between level 3c and level 4a were compared with intervention pupils receiving the additional small group teaching with an effect size of 0.24 (95% CI: 0.00 to 0.49).

Subgroup analysis

Despite there being no evidence of an overall intervention effect for grammar teaching, there was some evidence of a statistically significant interaction between allocated group and FSM status (p=0.08), suggesting the intervention had a different effect on FSM and non-FSM pupils. Table 6 shows the marginal mean extended writing scores for those receiving FSM and not receiving FSM by trial arm: the scores are higher for pupils not eligible for FSM than for those eligible to receive them. This suggests, therefore, that if grammar teaching is effective, it is more effective among pupils not receiving FSM.

INSERT Table 6: Marginal mean extended writing scores for FSM and non-FSM pupils ABOUT HERE

Individual Level Analysis

Raw, unadjusted mean post-test scores are presented in Table 7 by individual level allocation for those in the intervention class. Scores were similar in both allocated groups for all outcomes

INSERT Table 7: Average writing scores comparing small group versus whole class ABOUT HERE

Results from the primary and secondary individual level analyses are presented in Table 8. The number of pupils included in each analysis is shown alongside the adjusted difference in mean score between the allocated groups and associated effect sizes.

Primary analysis

There was some evidence (p=0.02) of a difference in extended writing score between the allocated groups, with a statistically significant effect size of 0.21 (95% CI: 0.04 to 0.38).

INSERT Table 8: Results from primary and secondary individual level analyses ABOUT HERE

Secondary analyses

There was no evidence of a difference between the randomised groups in terms of reading score (p=0.94) and little evidence of a difference in spelling and grammar

score (p=0.11). A reading effect size of -0.01 (-0.20 to 0.18) and a spelling and grammar effect size of 0.14 (95% CI: -0.03 to 0.31) were found.

Subgroup analysis

There was no evidence of a statistically significant interaction between allocated group and FSM status (p=0.54) suggesting that the small group effect did not have a differential effect dependent on FSM status.

Primary analysis with exclusion of intervention small group

The intention to treat cluster analysis demonstrated an effect size of 0.10. However, because a 'small group effect' was potentially driving this non-significant effect size we repeated the analysis removing pupils who had been randomised to have the small group intervention. The analysis was conducted on 1772 pupils and there was no evidence of a difference in extended writing score between the allocated groups, with a non-significant increase of 0.20 marks for those in the intervention group compared with those in the control group (p=0.57, 95% CI: -0.48 to 0.87). This relates to an effect size of 0.06 (95% CI: -0.15 to 0.28).

Fidelity

Fidelity scores were available for 52 of the 55 randomised schools. Two scores were missing due to school withdrawal before the start of the intervention and the third was missing due to withdrawal of a school during the intervention and before fidelity assessment. A fidelity score was available for the fourth school which withdrew from the trial due to the timing of the fidelity assessment. The minimum fidelity score recorded was 4 out of 9. The maximum and most frequently recorded score was 9,

with 56.4% of schools being judged to have delivered as planned in relation to all three components. The mean fidelity score was 8.2 (SD 1.27) and the median score was 9.

One school requested three teaching groups after both levels of randomisation had occurred. The school allowed the evaluation team to create the new teaching groups randomly but this meant that pupils potentially did not receive the condition to which they were originally assigned at both the individual and cluster level. At one school data were provided on incorrect classes. This was only discovered after cluster randomisation and after the first CPD day. The randomisation of the correct classes to intervention or control resulted in the previous intervention teacher teaching the control class and vice versa, hence the allocations were switched in practice. This meant that eligible pupils in the control class needed to be individually randomised for practical reasons. One school did not have enough pupils eligible for individual randomisation in the intervention class to teach a small group due to pupil extraction, hence all pupils received the intervention at the class level.

Conclusions and implications

We undertook a large pragmatic randomised controlled trial of *Grammar for Writing* in year 6 pupils. Our data suggest only a relatively small effect size (approximately 0.10 standard deviations difference) on the GL Assessment measure between the classes randomised to receive the intervention and those continuing with 'business as usual'. This difference was not statistically significant, with a 95% confidence interval ranging from -0.10 to 0.31, suggesting that this difference may have occurred by chance. Indeed, when the small group children were excluded from the

intervention group, the effect size was reduced to 0.06 of a standard deviation difference.

When children were taught in small groups there was a larger effect size of between 0.21 to 0.25, which did not materially differ in the comparisons between small groups versus large intervention groups or small groups versus large control groups. This suggests, therefore, that the difference found in the small group *Grammar for Writing* intervention was a consequence of children being taught in small groups *per se* rather than due to any intrinsic benefit of teaching *grammar* in small groups.

Although we found little evidence that *Grammar for Writing* was effective, as measured by the GL Assessment outcome in year 6 pupils, we found that teaching children in small groups of about 4-6 children per group improved writing skills by around a quarter of a standard deviation compared with similar children taught in class sizes of approximately 25. This finding supports previous evidence that small group teaching is effective (EEF, 2013), although this benefit needs to be set against the increased cost of teaching children in small groups. However, there remains an alternative explanation for the impact of small groups. Whilst some schools delivered small group teaching within the same time allocation for literacy, other schools may have delivered additional teaching. Consequently, the apparent benefit of small group teaching may be due to additional teaching and not entirely due to being taught in small groups.

Strengths

In the design and conduct of our study we used best practice as defined by the CONSORT guidelines for randomised controlled trials. Importantly, we used

independent concealed allocation to ensure that the schools and children were allocated without the possibility of bias. We used the principles of intention to treat by including all consenting children and schools in the final analysis. We pre-specified our main outcome and wrote a statistical analysis plan before we observed the data. We also used an independent testing company to mark the test papers, blind to the allocated group.

Limitations

Although our trial was relatively large with over 100 classes and more than 2400 pupils, it was not possible to recruit to the target of 120 classes in 60 schools. Furthermore, the actual intra-cluster correlation coefficients (ICCs) were somewhat larger than our predicted ICCs, which would have reduced our statistical power. In terms of attrition, we lost 4 schools after randomisation. Two schools withdrew from the study after cluster level randomisation, two after individual level randomisation and post-test data were not retrieved by the testing company for a fifth school. We also lost a number of pupils, for the main outcome, who did not complete all of relevant questions on the post-test, so these were excluded from the analysis. However, we do not think that these post-randomisation exclusions are likely to have introduced bias, as there is no reason to link their loss to the intervention. Selection bias due to attrition was unlikely (see baseline tables, where there is little difference between the analysed groups).

Although the test papers were marked blindly, they were delivered to the children by teachers who were not blind to group allocation. To reduce the possibility of teacher bias we gave instructions that the children should sit the tests under 'exam'

conditions. However, we cannot exclude the possibility that teachers may have given inappropriate help to some children whilst sitting the test.

The design meant that both the intervention and control classes were nested within the same school. Consequently we cannot completely exclude the possibility that the relatively small effect size difference between the groups may be a consequence of 'contamination' or 'spill over' between the intervention and control teachers. If this did occur, however, it would suggest the transmission of the intervention between intervention and control teachers was as effective as the dedicated CPD training sessions that intervention teachers attended. Furthermore, because the effect size after removing the small group effect was so slight (0.06), a significant proportion of control teachers must have been 'contaminated'.

Generalisability of results

A wide range of schools across England were recruited; consequently, our findings should be applicable to most English primary schools, particularly those in inner-city, urban areas or schools with a high proportion of pupils belonging to minority ethnic groups or pupils eligible for FSM.

Further research

Our study did not find sufficient evidence to support the use of *Grammar for Writing* in year 6 pupils. We did find some evidence, however, showing small group teaching to have modest effects among year 6 pupils who were between levels 3c and 4b in the standard assessment tests (SATs). In our study these children only had approximately one term's exposure to being taught in small groups. It might be useful to look at,

say, a full year's exposure to small group teaching for these children and to estimate the effectiveness and cost-effectiveness of such an approach.

Conclusion

In conclusion, we found a small (effect size 0.10) impact of *Grammar for Writing* in our intention to treat analysis, which was not statistically significant and was, in part, explained by the small group impact of a subsample of children. Small group teaching may have had a modest benefit and would merit further study.

Acknowledgements

We acknowledge and thank the Education Endowment foundation for funding the trial reported here; the Developer of the intervention at the University of Exeter who was responsible for the recruitment of the schools and for training the teachers and delivering the intervention; and all schools, teachers and pupils involved in the trial.

References

Altman DG, Moher D & Schulz KF (2012). Improving the reporting of randomised trials: the CONSORT Statement and beyond. Statist. Med., 31: 2985–2997.

Bland JM. The analysis of cluster-randomised trials in education. *Effective Education* 2010; **2**: 165 – 180

Cook TD & Campbell D. (1969). Quasi-experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin.

Department of Education (2013). Schools, pupils and their characteristics (Reference SFR21/2013). Statistical First Release.

Department of Education (2014) National curriculum statistics at key stage 2 in England (Reference SFR 30/2014). Statistical First Release:

Education Endowment Foundation (EEF), (2013). Small group tuition, The Sutton-Trust-EEF Teaching and Learning Toolkit.

Evangelou, M., et al. (2008). What Makes a Successful Transition from Primary to Secondary School?: Findings from the Effective Pre-school, Primary and Secondary Education 3-14 (EPPSE) Project, Department for Children, Schools and Families.

Jones S, Myhill DA, Bailey T. (2012) Grammar for writing? An investigation of the effects of contextualised grammar teaching on students' writing. Reading and Writing DOI 10.1007/s11145-012-9416-1

Mosteller F. (1995) The Tennessee Study of Class Size in the Early School Grades. The Future of Children, Vol. 5. No. 2, Critical Issues for Children

Myhill DA, Jones SM, Lines H & Watson A. (2012). Rethinking grammar: the impact of embedded grammar teaching on students' writing and students' metalinguistic understanding. Research Papers in Education, 27:2, 139-166.

Ofsted (2012). Moving English forward: action to raise standards in English (110118).

Schagen, I. a. (2004). *But What Does It Mean? The Use of Effect Sizes in Educational Research.* Slough: NFER.

Shadish WR, Cook TD & Campbell DT. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA, US: Houghton, Mifflin and Company.

Torgerson DJ & Torgerson CJ. (2008). Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction. Palgrave Macmillan.

Torgerson CJ, Wiggins A, Torgerson DT, Ainsworth H, Barmby P, Hewitt C, Jones K, Hendry V, Askew M, Bland M, Coe, R, Higgins S, Hodgen J, Hulme C & Tymms P. (2011). The Every Child Counts Independent Evaluation Report. Department of Education.

Figure 1: CONSORT flow diagram

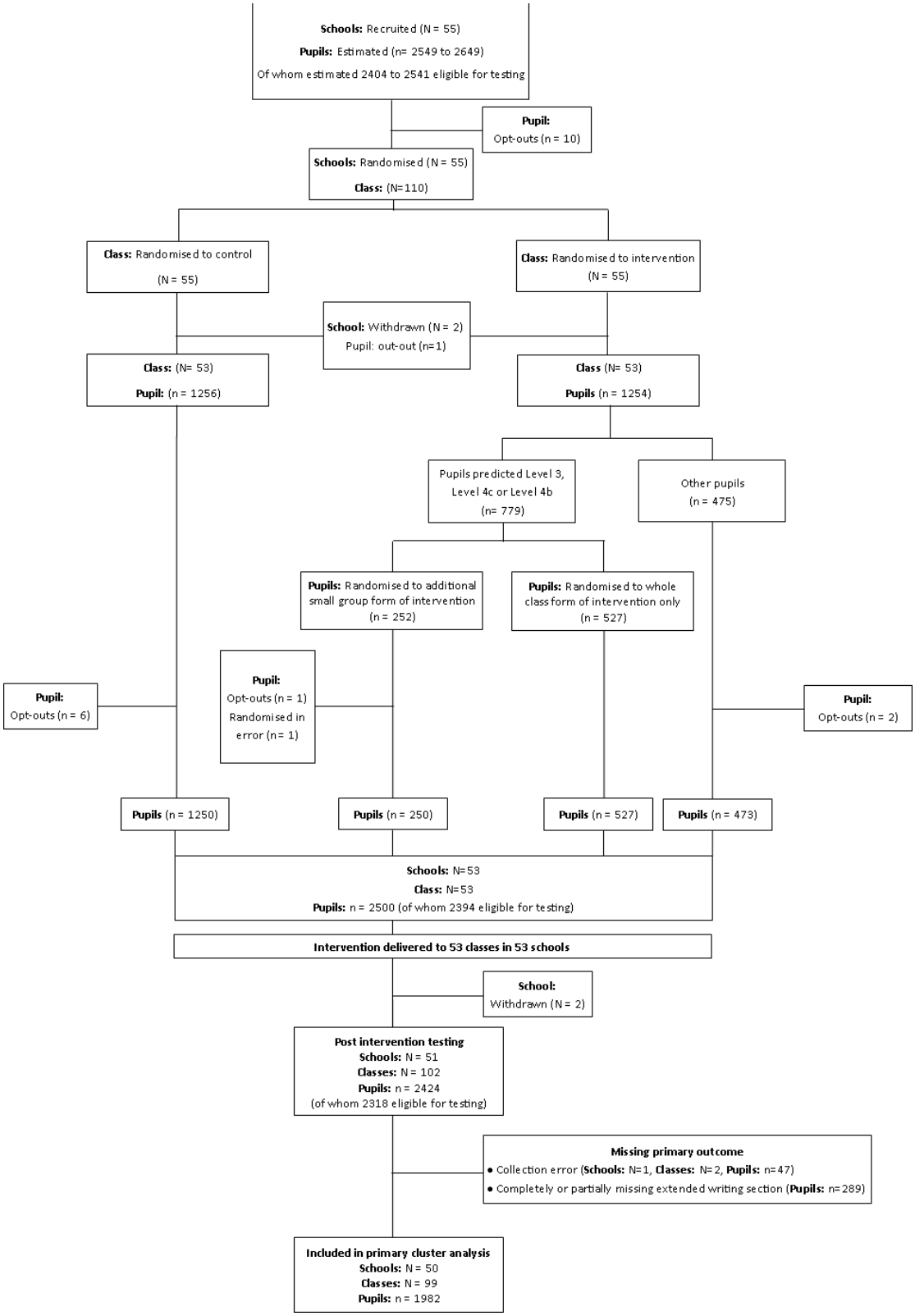

Table 1: Baseline pupil level characteristics

	As randomised (All level 3 or above) Frequency (%)		**As analysed (primary cluster analysis)** Frequency (%)	
	Intervention	**Control**	**Intervention**	**Control**
	n =1194	**n** = 1200	**n** = 1004	**n** = 978
Gender				
Male	609 (51.0)	617 (51.4)	507 (49.5)	500 (51.1)
Female	585 (49.0)	583 (48.6)	497 (50.5)	478 (48.9)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
FSM				
Eligible	420 (35.2)	392 (32.7)	328 (32.7)	295 (30.2)
Not eligible	774 (64.8)	808 (67.3)	676 (67.3)	683 (69.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Pupil premium				
Eligible	446 (37.4)	425 (35.4)	348 (34.7)	324 (33.1)
Not eligible	748 (62.7)	775 (64.6)	656 (65.3)	654 (66.9)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
English as an additional language				
EAL	494 (41.4)	516 (43.0)	408 (40.6)	423 (43.3)
Non-EAL	700 (58.6)	684 (57.0)	596 (59.4)	555 (56.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Predicted KS2 writing level				
Level 3c	48 (4.0)	65 (5.4)	41 (4.1)	45 (4.6)
Level 3b	78 (6.5)	80 (6.7)	62 (6.2)	62 (6.3)
Level 3a	128 (10.7)	113 (9.4)	93 (9.3)	88 (9.0)
Level 4c	278 (23.3)	256 (21.3)	228 (22.7)	203 (20.8)
Level 4b	245 (20.5)	248 (20.7)	207 (20.6)	209 (21.4)
Level 4a	189 (15.8)	185 (15.4)	168 (16.7)	165 (16.9)
Level 5 or above	228 (19.1)	253 (21.1)	205 (20.4)	206 (21.1)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Month of birth				
Sept – Nov	331 (27.7)	315 (26.3)	281 (28.0)	260 (26.6)
Dec – Feb	305 (25.5)	299 (24.9)	260 (25.9)	230 (23.5)
Mar – May	278 (23.3)	293 (24.4)	234 (23.3)	245 (25.1)
Jun – Aug	280 (23.5)	293 (24.4)	229 (22.8)	243 (24.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Table 2: Baseline characteristics of pupils who were individually randomised

	As randomised Frequency (%)		**As analysed (primary individual level analysis)** Frequency (%)	
	Small	**Whole**	**Small**	**Whole**
	n = 250	**n** = 527	**n** = 210	**n** = 421
Gender				
Male	144 (57.6)	295 (56.0)	121 (57.6)	233 (55.3)
Female	106 (42.4)	232 (44.0)	89 (42.4)	188 (44.7)
Missing	0 (0.0)	(0.0)	0 (0.0)	0 (0.0)
FSM				
Eligible	98 (39.2)	225 (42.7)	74 (35.2)	171 (40.6)
Not eligible	152 (60.8)	302 (57.3)	136 (64.8)	250 (59.4)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Pupil premium				
Eligible	105 (42.0)	233 (44.2)	80 (38.1)	175 (41.6)
Not eligible	145 (58.0)	294 (55.8)	130 (61.9)	246 (58.4)
Missing	0 (0.0)	0 (0.0)	(0.0)	0 (0.0)
English as an additional language				
EAL	108 (43.2)	232 (44.0)	89 (42.4)	185 (43.9)
Non-EAL	142 (56.8)	295 (56.0)	121 (57.6)	236 (56.1)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Predicted KS2 writing level				
Level 3c	18 (7.2)	30 (5.7)	17 (8.1)	24 (5.7)
Level 3b	19 (7.6)	59 (11.2)	16 (7.6)	46 (10.9)
Level 3a	44 (17.6)	84 (15.9)	31 (14.8)	62 (14.7)
Level 4c	86 (34.4)	192 (36.4)	72 (34.3)	156 (37.1)
Level 4b	83 (33.2)	162 (30.7)	74 (35.2)	133 (31.6)
Level 4a	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Level 5 or above	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Month of birth				
Sept – Nov	57 (22.8)	126 (23.9)	50 (23.8)	100 (23.8)
Dec – Feb	60 (24.0)	132 (25.0)	48 (22.9)	107 (25.4)
Mar – May	61 (24.4)	120 (22.8)	53 (25.2)	98 (23.3)
Jun – Aug	72 (28.8)	149 (28.3)	59 (28.1)	116 (27.6)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Table 3: Estimated intra-cluster correlation coefficients (ICCs)

	n	School ICC (95% CI)	Class ICC (95% CI)
Total raw score	1977	0.21 (0.13 to 0.28)	0.27 (0.21 to 0.34)
Primary outcome (extended writing score)	2033	0.26 (0.17 to 0.34)	0.32 (0.25 to 0.39)

Table 4: Unadjusted average scores for the intervention and control groups

	n (%)	Unadjusted mean (SD)
Pupils predicted level and above		
Intervention	1194 (100.0)	-
Control	1200 (100.0)	-
Extended writing score		
Intervention	1004 (84.1)	22.8 (4.85)
Control	978 (81.5)	22.6 (4.88)
Reading score		
Intervention	867 (72.6)	18.4 (5.18)
Control	847 (70.5)	18.2 (5.22)
Spelling and grammar score		
Intervention	1025 (85.8)	11.4 (5.08)
Control	1051 (87.6)	11.5 (4.92)

Table 5: Results from primary and secondary cluster level analyses

	n	Difference in means* (95% CI)	Effect size (95% CI)
Extended writing score	1982	0.34 (-0.30 to 0.98)	0.10 (-0.10 to 0.31)
Reading score	1714	0.38 (-0.12 to 0.88)	0.10 (-0.03 to 0.24)
Spelling and grammar score	2076	0.04 (-0.44 to 0.51)	0.01 (-0.14 to 0.16)
Control versus small group form of intervention (Extended writing score)	817	0.78 (-0.01 to 1.56)	0.24 (0.00 to 0.49)

* (Intervention – Control)

Table 6: Marginal mean extended writing scores for FSM and non-FSM pupils

	Intervention	Control
	n = 328	**n** = 295
Eligible for FSM	22.0 (95% CI: 21.3 to 22.7)	21.7 (95% CI: 20.9 to 22.4)
Not eligible for FSM	23.1 (95% CI: 22.4 to 23.8)	22.7 (95% CI: 22.0 to 23.5)

Table 7: Average writing scores comparing small group versus whole class

	n (%)	Unadjusted mean (SD)
Pupils predicted level and above		
Small group	250 (100.0)	-
Whole class	527 (100.0)	-
Extended writing score		
Small group	210 (84.0)	21.7 (4.36)
Whole class	421 (79.9)	20.9 (4.30)
Reading score		
Small group	167 (66.8)	16.2 (5.49)
Whole class	336 (63.8)	16.2 (4.44)
Spelling and grammar score		
Small group	210 (84.0)	9.7 (5.08)
Whole class	449 (85.2)	9.2 (4.78)

Table 8: Results from primary and secondary individual level analyses

	n	Difference in means* (95% CI)	Effect size (95% CI)
Extended writing score	631	0.67 (0.12 to 1.23)	0.21 (0.04 to 0.38)
Reading score	503	-0.03 (-0.75 to 0.69)	-0.01 (-0.20 to 0.18)
Spelling and grammar score	659	0.48 (-0.10 to 1.07)	0.14 (-0.03 to 0.31).