PROJECT TITLE	Helping Handwriting Shine
DEVELOPER (INSTITUTION)	University of Leeds
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Dr Ben Styles
TRIAL (CHIEF) STATISTICIAN	Dr Joana Andrade
SAP AUTHOR(S)	Dr Joana Andrade, Dr Ben Styles, Gemma Stone
TRIAL REGISTRATION NUMBER	ISRCTN13315075
EVALUATION PROTOCOL URL OR HYPERLINK	*Protocol*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	March 2019	N/A

SAP changes

- FSM eligibility status provided by the schools used as a Year 5 Experiment stratifier instead of everFSM (as stated in the protocol).
- The Year 2 randomisation was stratified by 4 delivery regions (Darlington, Leeds, Newcastle and Sheffield) not the three stated in the protocol (Leeds, Bradford and the North East).
- FSM subgroup analysis groups specified in terms of FSM, not ever FSM or FSM6.
- FSM6 and ever FSM data will be requested from the National Pupil Database (NPD) in order to evaluate the overlap of the different deprivation measures.
- Exploratory subgroup analysis specified in terms of pupils’ writing speed, not described in the protocol, included in the analysis.

Table of contents

Contents

Introduction

The Helping Handwriting Shine (HHS) trial comprises two experiments within the same randomised controlled efficacy trial. Both experiments have two main arms (intervention and control).

HHS is an intervention developed at the School of Psychology of the University of Leeds with the aim of improving the capability of children who struggle to write fluently to undertake cognitively effortful behaviour, particularly writing. The theory of change (ToC) of HHS hypothesises that if children are taught to automate handwriting they will be able to free cognitive capacity that is being directed to the function of handwriting and, in turn, become more fluent and proficient writers.

Under the HHS set-up children are taught to focus on writing tasks they find difficult, and to repeat and refine them in order to overcome their motor deficits and be able to write effortlessly, both in terms of legibility and writing speed. The intervention is thus structured as the delivery of a set of materials to pupils over 24 thirty-minute sessions (three per week over a period of eight weeks) by one or more trained staff member/s within a school.
HHS is a two stage programme that has two modes of delivery: class based and in small group settings. In stage 1, which took place in October-November 2018, school staff (teachers, teaching assistants or Special Educational Needs coordinators) attended a one-day workshop in order to learn how to correctly deliver the intervention in class-based or small group settings. In stage 2, the school staff trained during stage 1 will deliver the HHS intervention to pupils over two four-week periods.

There is also evidence that the age of the target children and the mode of delivery influence the overall impact of the intervention. The literature suggests that a stronger correlation between automatic handwriting and quality of writing composition is to be found in younger children (six to seven years old) than in older children (ten to 11 years old), possibly because "as writers develop, and write more sophisticated texts, there are other issues which account for more of the variance" (Medwell et al, 2009, pp. 329-344). As such, it is to be expected that generic class-based interventions, although adequate for the younger age group, have a smaller overall impact than for older children whose specific needs can be better met in the context of small group settings.

Taking into account the suitability of the different modes of delivery to different child age brackets, the HHS trial was designed to include a Year 2 Experiment (targeted at six-seven year old children) and a Year 5 Experiment (targeted at nine-ten year old children).

The Year 2 Experiment is a school randomised trial that will evaluate the impact of class-based intervention on the writing ability of younger children. To reflect the nature of the class-based mode of delivery this trial will involve the whole of Year 2 in each school.

The primary research question for the Year 2 Experiment is: *What is the impact of the Helping Handwriting Shine intervention on the comparative judgement measurement scale for writing of children aged six to seven?*

The secondary research questions for the same experiment are:

1. What is the impact of the Helping Handwriting Shine intervention on the writing composition of children aged six to seven?

2. What is the impact of the Helping Handwriting Shine intervention on the handwriting speed of children aged six to seven?

(See Design Overview section below for the specific measures to be used when evaluating the outcomes associated with the Experiment.)

The Year 5 Experiment is a multi-site trial that will evaluate the impact of the HHS intervention through small group settings, on the writing ability of older children who struggle to handwrite. Children who meet the eligibility criteria specified in the Study Design Section will be selected in each of the intervention group schools of the Year 2 Experiment to take part in this trial.

The primary research question for the Year 5 Experiment is: *What is the impact of the Helping Handwriting Shine intervention on the comparative judgement measurement scale for writing of targeted children aged nine to ten?*

And for the same experiment the secondary research questions are:

3. What is the impact of the Helping Handwriting Shine intervention on the writing composition of targeted children aged nine to ten?
4. What is the impact of the Helping Handwriting Shine intervention on the handwriting speed of targeted children aged nine to ten?

(See Design Overview section below for the specific measures to be used when evaluating the outcomes associated with the Experiment.)

The eligibility criteria for the Year 5 trial include two distinct (but not mutually exclusive) groups: pupils who have slow and effortful handwriting, and pupils with illegible handwriting. We will also explore if there is a differential impact of the HHS intervention on each of these two groups.

One of the objectives of the HHS trial is also to evaluate if a higher or lower degree of fidelity to the protocol in the implementation of the programme during Stage 1 and Stage 2 will have an impact on the overall results of the HHS intervention. For this purpose we will also consider the following additional secondary research questions:

1. Is there an association between fidelity and the comparative judgement measurement scale (the primary outcome) in the Year 2 Experiment?
2. Is there an association between fidelity and the target pupils' comparative judgement measurement scale (the primary outcome) in the Year 5 Experiment?

Finally, we will also consider the differential effect, if any, of the intervention on FSM eligible pupils by addressing the research question: *Are effects on writing ability (as indexed by the primary RQs above) different for pupils eligible for FSM when compared to non-eligible children? If so, how?*

Design overview

Trial type and number of arms		Two randomised controlled trials, each with two arms
Unit of randomisation		Experiment 1:School Experiment 2: Pupil
Stratification variables (if applicable)		Experiment 1: Region (Training Hub) Experiment 2: School and FSM eligibility
Primary outcome	variable	Writing ability
	measure (instrument, scale)	Writing Assessment Measure* (comparative judgement true scores[1])
Secondary outcome(s)	variable(s)	Writing composition Handwriting speed
	measure(s) (instrument, scale)	Writing Assessment Measure** (criterion referencing scores) Handwriting Speed Test*

Dunsmuir, Kyriacou, Batuwitage, Hinson, Ingram and O'Sullivan, 2013.
** Wallen, Bonney and Lennox, 2006

Year 2 Experiment school level randomisation: To facilitate the delivery of training to school staff, four different training hubs were set up in different locations: Darlington, Leeds, Newcastle and Sheffield[2].

The assignment of schools to hubs took into account distance to the nearest or second nearest training hub to ensure that the training venues were reachable by the school staff undertaking the HHS training within a reasonable commuting time. Table 1, below, presents the assignment of the 103 schools taking part in the trial to training hubs:

Table 1: Distribution of participating schools in training hubs

Training hub	Number of schools	Percentage of schools	Maximum school-hub distance (miles)
Darlington	24	23%	24
Leeds	30	29%	13
Newcastle	28	27%	15
Sheffield	21	20%	17

The school level randomisation for the Year 2 Experiment took place in early August 2018[3], after baseline testing[4] being stratified by region (assigned training hub). The stratification was introduced to prevent a clumping of intervention schools that could impede workshop delivery. The school-level randomisation is described below, in Table 2:

Table 2: Year 2 Experiment school level randomisation regional strata

[1] See Primary Outcome Measure section below for the definition of comparative judgement true scores (Page 7)

[2] Only three regions (Leeds, Bradford and the North East) were considered in the protocol. During the school recruitment process concerns were raised by the NFER team that not enough schools would be recruited if the trial was to be restricted to the geographical area originally defined in the protocol. With the agreement of the University of Leeds team, the geographical area of recruitment was thus extended to include an additional region.

[3] Some of the participant schools only provided the results of the baseline assessment in the last weeks of July 2018, which lead to the Year 2 randomisation taking place in early August 2018 instead of July 2018, as stated in the protocol.

[4] Baseline testing took place between the 20/6/2018 and 17/7/2018.

Region (training hub)	No. of schools in each arm		Total no. of schools
	HHS intervention	CONTROL	
Darlington	12	12	24
Leeds	15	15	30
Newcastle	14	14	28
Sheffield	10	11	21
Total:	52	51	103

Sampling for the Year 2 Experiment secondary outcomes analysis: The comparative judgement score will be obtained for all scripts at baseline and follow-up. The secondary outcomes for the Year 2 Experiment will be marked not on the full Year 2 cohort but on a Year 2 pupil sample. In each of the participating schools, we will randomly select five[5] of the pupils taking part in the trial to be included in the secondary outcomes sample. The trial is fully powered for the primary outcome only but we acknowledge a discussion of power for secondary outcomes is sometimes useful, particularly as we are using a reduced sample here. Five was used as it is sufficient to allow follow-up of at least one pupil per school and is consistent with the idea that when estimating regression coefficients in a multi-level model, small cluster sizes are adequate[6]. This analysis will have a higher MDES than for the primary outcome (approximately 0.25) but its inclusion is for verification of the comparative judgement method rather than intended to be fully powered.

We have chosen to select the secondary outcomes sample from the pupils present at pre-randomisation baseline assessment instead of selecting from the more restricted group of pupils with pre-test and post-test outcomes data. Although there is a risk that this procedure will lead to attrition, and consequent loss of power, we have decided to adopt it in order to prevent biasing the analyses results[7].

Year 5 Experiment pupil level randomisation: Six schools withdrew from the intervention group after the early August Year 2 school level randomisation. 372 Year 5 pupils from the remaining 46 intervention schools were selected to take part in the trial and were randomised post-baseline testing in the second week of October 2018[8]. The randomisation was stratified by school and eligibility for FSM[9].

The pupil-level randomisation breakdown by FSM eligibility status is described below in Table 3.

Table 3: Year 5 Experiment pupil level randomisation (FSM eligible strata)

[5] In each school the pupils will be randomly selected from its Year 2 cohort taking part in the trial, with no stratification in terms of class or FSM eligibility being taken into account. If a school has less than five pupils taking part in the Year 2 trial, all the pupils will be included in the secondary Year 2 sample.

[6] See Snidjers *et al*, 2005, pp. 1570-1573.

[7] The rationale being that this will ensure not only that bias is not introduced in the analysis but also that the primary and secondary analyses will be performed under similar settings.

[8] Baseline testing took part between the 19/11/2018 and the 1/10/2018 and the randomisation on the 9/10/2018. The baseline assessment scripts of one of the schools taking part on the trial were lost and the school was invited to retake the baseline test. The second assessment took part on the 11/10/2018 , four days before the random allocation result was disclosed to the school

[9] The FSM eligibility information was obtained directly from the schools during Year 5 data collection

FSM status	No. of pupils in each arm		Total no. of pupils
	HHS intervention	CONTROL	
Not eligible	131	130	261
Eligible	55	56	111
Total:	186	186	372

Outcome measures

Primary Outcome Measure

The primary outcome measure for both Experiments will be the Writing Assessment Measure (WAM) marked using Comparative Judgement. The assessments are to be administered by NFER test administrators and marked using Comparative Judgement by external, blind judges, a pool or current/former teachers, via the No More Marking (NMM) platform[10].

Comparative Judgment produces a rank score against a set of scripts without reference to any pre-established criteria or norms. The NMM programme randomly selects pairs of scripts from within each 'task' (either Year 2 baseline; Year 2 follow-up; Year 5 baseline or Year 5 follow-up). Judges are presented with a pair and then asked to choose which one is better, one or the other.

The NMM platform uses the Bradley-Terry model (Hunter, 2014, pp. 384-406) to produce true scores. True scores measure a latent ability[11], in this case writing ability, and are computed from the wins and losses of a script against other scripts[12]. To produce results with a high level of reliability, ten judgements are made per script (known from previous work by No More Marking to produce a very reliable measure; see Pollitt 2012). True scores are measured in a scale that is linear, robust to missing data, has estimates of precision, detects misfit, and the parameters of the objects being measured can be separated from the measurement instrument being used. Although there is no underlying assumption of normality, since true scores measure a latent writing ability they are generally normally distributed and are standardized to have a mean of zero and a standard deviation of two.

The NMM platform also provides the following information throughout and after the judging process, which enables monitoring of judges and the iterative completion of scores:

- Judge Infit: A measure of consistency between judges. Judges will be excluded from judging if their 'infit' parameter is greater than a pre-defined threshold (1.2).
- Inter-rater reliability: The correlation between the scale produced by half the judges and the scale produced by the other half. The platform takes four random halves, and reports the mean and standard deviation of the four replications.

[10] https://www.nomoremarking.com/
[11] See Hunter, 2014, pp. 384-406.
[12] The computation of true scores also takes into account the scores of the scripts that the script was judged against (Hunter, 2014, pp. 384-406).

Secondary Outcome Measure (1)

The measure for the first secondary outcome, writing composition, for both Experiments will be the Writing Assessment Measure (WAM) marked using criterion referenced scores (in Appendix A) but excluding the handwriting element. (Dunsmuir *et al*, 2015, pp. 1-18).

Second Secondary Outcome Measure (2)

The measure for the second secondary outcome, handwriting speed, for both Experiments will be Handwriting Speed Test raw scores. These correspond to the number of letters pupils are able to write per minute when writing at their usual writing speed (Wallen *et al*, 1996, pp. 141-144).

Follow-up

As of December 2018, four schools have withdrawn from the intervention group and a fifth school has pulled out from the Year 5 intervention but not from the Year 2's. This brings the total number of schools in the intervention group to 45 at Year 2 and 44 at Year 5. No schools have withdrawn from the control group, leaving the total number of schools as 55.

Sample size calculations overview

Experiment 1 (Year 2)		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.18	0.21	0.18	0.20*
Pre-test/ post-test correlations	level 1 (pupil)	0.65	0.65	0.65	0.65
	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-	-
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		two	two	two	two
Average cluster size		37	11	37	11
Number of schools	intervention	50	50	52	52
	control	50	50	51	51
	total	100	100	103	103
Number of pupils	intervention	1850	574	1874	559*
	control	1850	574	1979	591*
	total	3700	1148	3853	1150*

*Assuming the same proportion of FSM eligible pupils reported by Schools for Year 5 displayed in Table 3 above, 111 out of 372 pupils (approximately 30 per cent).

Experiment 2 (Year 5)		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.23	0.23 to 0.41*	0.23	0.40**
Pre-test/ post-test correlations	level 1 (pupil)	0.65	0.65	0.65	0.65
	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		two	two	two	two
Average cluster size		-	-	-	-
Number of schools	intervention	-	-	-	-
	control	-	-	-	-
	total	50	50	46	46
Number of pupils	intervention	185	54 to 185*	186	56**
	control	185	54 to 185*	186	56**
	total	370	108 to 370*	372	111**

* Based on the assumption that the rates of ever FSM and FSM eligibility are similar and that the probability of an everFSM pupil being eligible for the trial is identical to the overall probability of being eligible, the expected number eligible per school is 2.16. Under the assumption that the rates of everFSM and FSM eligibility are similar and that all eligible pupils are FSM eligible, the expected number per school is 7.4 as per the main sample size calculation. The true value will lie somewhere in between.
** Assuming the proportion of FSM eligible pupils reported by Schools for Year 5 displayed in Table 3 above, 111 out of 372 pupils (approximately 30 per cent).

The power calculations were performed with the calculations for a simple randomised design being adjusted for pre-post correlation and design effect using the Kish formula (Kish, 1965). All the calculations were performed assuming 80% power and alpha=0.05.

In the absence of a writing trial pilot, parameters for sample size were estimated using comparable EEF studies and materials. The 2015 EEF table of intra-cluster correlations (Education Endowment Foundation, 2015) suggests a value of 0.109 for Key Stage 1 English in the North East and the 2013 Pre-testing in EEF Evaluations paper (Education Endowment Foundation, 2013) suggests a correlation of 0.73 between Key Stage 1 and Key Stage 2 English. The Grammar for Writing evaluation (Torgerson et al., 2014a) had a school-level ICC of 0.26 and the class-level ICC was 0.32. It used a predicted KS2 writing level as the baseline measure but the correlation was low at 0.54. The Calderdale Improving Writing Quality evaluation (Torgerson et al., 2014b) had a school-level ICC of only 0.04 for the extended writing task but this was based on only a sub-group of primary school children who went on to secondary schools within the trial. This trial also used a predicted KS2 writing level as the baseline and the correlation was also low at 0.35. Based on the research mentioned above, and to remain realistically conservative, we have adopted the values of 0.15 and 0.65 for the ICC and pre-post correlation, respectively.

According to a 2016 meta-analysis of handwriting interventions[13] (Santangelo and Graham, 2016) handwriting instruction was associated with a rather large effect size of 0.84 on the quality of student writing. On the other hand, recent studies provide evidence that EEF trials are underpowered (Sanders and Ni Chonaire, 2015 and Lortie-Forgues, 2017) and considering an effect size of 0.15, double the median effect size of EEF trials to date, is a reasonable assumption.

Considering an ICC of 0.15, a pre-post correlation of 0.65, and an effect size of 0.15 would result in trials requiring 140 (70 versus 70) schools. However, an efficacy trial on 140 schools would be too costly, and entail a risk of diluting the intervention through limited delivery capacity. Taking into account these practical considerations we have settled for an effect size of 0.18 for the Year 2 trial and an effect size of 0.23 for the Year 5 trial[14], both within what is expected from previous meta-analysis of handwriting interventions.

These adjusted effect sizes require 100 schools (50 versus 50) for the Year 2 experiment and 370 Year 5 pupils (185 versus 185) for the Year 5 pupil randomised experiment.

Given that the Year 2 intervention is delivered at class level, some form of teacher effect is to be expected. It is possible that including teacher-level variance improves the model, but this is conceptually equivalent to including another baseline measure as a covariate, which is discouraged in the analysis guidance. As the unit of randomisation is the school and we are measuring every pupil in the school for the primary outcome (or randomly sampling within a school for secondary outcomes), we get an unbiased estimate of the school means in the model.

Analysis

The primary and secondary analyses will follow EEF guidelines for both the Year 2 and the Year 5 Experiments.

Primary outcome analyses

The primary analyses for both experiments will be intention-to-treat.

For Year 2 a multilevel random intercepts model with two levels (school and pupil) will be used to account for cluster randomisation. The main analysis will investigate if the attendance of a class that received the HHS intervention had an effect on pupils' writing ability. This will be determined by fitting a model with the dependent variable as writing ability post-intervention as measured by the Writing Assessment Measure (WAM) comparative judgement true scores described above.

To control for prior writing ability, pupil-level WAM comparative judgement true scores assessed at baseline will be included in the model as a covariate. The model will also

[13] It should be taken into account that the (Santangelo and Graham, 2016) meta-analysis included non-randomised designs.

[14] We reasonably expect the Year 5 intervention, a small groups intervention targeted at selected pupils, to be more effective than the class level Year 2 intervention.

contain a dummy variable for region (training hub), to reflect the Year 2 stratified randomisation.

The two level random intercepts model is given by:

$$Y_{ij} = \beta_{0j} + \beta_1 \text{intervention}_j + \ \beta_2 \text{baseline WAM}_{\ ij} + \boldsymbol{\beta}\text{region}_j + \epsilon_{ij}$$

Where Y_{ij} is the post-intervention WAM comparative judgement true score of pupil *i* in school *j*, $\boldsymbol{\beta_{0j}}$ is the intercept in school *j,* intervention_j is the school-level intervention/control dummy variable*,* $\text{baseline WAM}_{\ ij}$ is the baseline WAM comparative judgement true score of pupil *i* in school *j*, and region_j is a dummy variable for the training hub assigned to school *j*.

The model will be run in R (version 3.4.1) using the package 'nlme'.

Since the Year 5 experiment is a multi-site efficacy trial, as per EEF 2018 guidelines we will be using a fixed effects single level model. The primary analysis will determine if receiving the HHS intervention in a small group setting had an effect on the writing ability of pupils who struggle to produce fluent handwriting due to a deficit in fine motor skills. For this purpose we will fit a single-level regression model with the dependent variable as writing ability post-test as measured by WAM comparative judgement true scores.

Similarly to Experiment 1, WAM comparative judgement true scores assessed at baseline will be included in this model. Dummy variables for school and FSM status will also be included in the model to reflect the stratified Year 5 randomisation.

The regression model is given by:

$$Y = \beta_0 + \beta_1 \text{intervention} + \ \beta_2 \text{baseline WAM} + \beta_3 \text{FSM} + \boldsymbol{\beta}\text{school} + \epsilon$$

Where Y is the post-intervention WAM comparative judgement true score of the pupil, intervention is the intervention/control dummy variable, baseline WAM is the baseline-intervention WAM comparative judgement true score of the pupil, and FSM and school are dummy variables for the pupil's FSM eligibility status and school, respectively.

The model will be run in R (version 3.4.1).

Secondary outcomes analyses

For the secondary analyses we will use an identical ITT approach to the analyses of the primary outcomes described above: fit two-level models (pupil and school) with random intercepts to account for cluster randomisation for the Year 2 Experiment, and single level fixed effects models for the multi-site Year 5 Experiment.

Analyses of secondary outcome (1): writing composition

The first secondary outcome analysis will assess if the attendance of a class that received the HHS intervention had an effect on pupils' writing composition. For this purpose we will fit a model whose dependent variable is writing composition post-intervention as measured by the Writing Assessment Measure (WAM) criterion referencing scores previously described. The baseline covariate for the model will consist of pre-test pupil-level WAM comparative judgement true scores, and the randomisation stratifier-covariate will be the same regional indicator included in the Year 2 primary ITT model.

The two level random intercepts model is given by:

$$Y_{ij} = \beta_{0j} + \beta_1 \text{intervention}_j + \ \beta_2 \text{baseline WAM}_{\ ij} + \boldsymbol{\beta}\text{region}_j + \epsilon_{ij}$$

Where Y_{ij} is the post-intervention WAM criterion reference score of pupil *i* in school *j*, $\boldsymbol{\beta}_{\mathbf{0}j}$ is the intercept in school *j*, intervention_j, is the school-level intervention/control dummy variable, $\text{baseline WAM}_{\;ij}$ is the baseline WAM comparative judgement true score of pupil *i* in school *j*, and region_j is a dummy variable for the training hub assigned to school *j*.

The model will be run in R (version 3.4.1) using the package 'nlme'.

The secondary analysis will determine if receiving the HHS intervention in a small group setting had an effect on the writing composition of the Y5 pupils selected to take part on the trial. To investigate this we will fit a single-level regression model with the dependent variable as writing composition post-intervention as measured by WAM criterion referencing scores.

Pre-test WAM comparative judgement true scores will be included in this model to control for baseline composition. The stratifier indicator dummy variables included in the primary ITT model, school and FSM eligibility status, will also be included in the first secondary outcome model (for experiment 1) to reflect the stratified Year 5 randomisation.

The regression model is given by:

$$Y = \beta_0 + \beta_1\text{intervention} + \ \beta_2\text{baseline WAM} + \beta_3\text{FSM} + \boldsymbol{\beta}\text{school} + \epsilon$$

Where Y is the post-intervention WAM criterion reference score of the pupil, intervention is the intervention/control dummy variable, baseline WAM is the baseline WAM comparative judgement true score of the pupil, and FSM and school are dummy variables for the pupil's FSM eligibility status and school, respectively.

The model will be run in R (version 3.4.1).

Analyses of secondary outcome (2): handwriting speed

The second secondary outcome analyses will assess if a Year 2 pupil being in a class that received the HHS intervention or a Year 5 pupil receiving the intervention under a small group setting had an effect on their handwriting speed.

As was the case for the primary and first secondary outcome analyses, in the context of the Year 2 Experiment we will account for cluster randomisation by running multi-level (pupil and school) models with random intercepts, while in the context of the multi-site Year 5 Experiment we will run single level fixed effects models.

The analysis of the second outcome has to take into account that this outcome is evaluated in terms of a count ("letters per minute"). As such we will fit multilevel and regression models that are appropriate to model count data: Poisson regression models or negative-binomial regression models. The choice of which models to apply will be dictated by the characteristics and distribution of handwriting speed raw scores data collected in the Experiments[15].

The analyses will be run in R (version 3.4.1) using the 'lme4' or the 'R2MLwiN' packages. The choice of package will be determined by which model is suited to the characteristics of the underlying distribution of the writing speed variable.

For both Experiments we will run the analyses by fitting models with post-intervention handwriting speed raw scores as the dependent variable, controlling for handwriting speed

[15] Poisson regression models assume that the mean and variance of the underlying data are identical, a condition that can be relaxed for negative-binomial models. The validity of the assumption that the variance equals the mean of the distribution will be tested before models are implemented.

at baseline by including pupil-level handwriting speed raw scores measured pre-intervention as a covariate. Covariates that reflect the stratified randomisations in the trial, a regional indicator for the Year 2 Experiment, and school and FSM status indicators for the Year 5 Experiment, will also be included in the models.

Subgroup analyses

The primary outcome models for both the Year 2 and Year 5 experiments will be modified for the FSM pupils analyses specified in the protocol. Power analyses will also be performed to determine if subgroup analyses are underpowered. In accordance to the EEF 2018 guidelines, underpowered subgroup analyses will be reported as exploratory.

We have deviated from the standard EEF procedure of using ever FSM/ FSM6 as a deprivation indicator considering FSM eligibility instead. This decision was motivated by the necessity to avoid collinearity in our regression models, since both ever FSM and FSM6 are known to be highly correlated with the FSM eligibility indicator already included as a stratifier in the Year 5 Experiment models. As a remedial measure we will include in the trial report tables describing the level of overlap of the different measures.

We will approach the analyses in two distinct ways: we will run models with interaction terms (i.e. models that include both the FSM indicator and the product of the FSM indicator and randomised group), and we will run separate primary outcome models on just the FSM eligible pupils. Both approaches conform to the EEF 2018 guidelines.

The Year 2 multilevel level random intercepts model with interaction terms is given by:

$$Y_{ij} = \beta_{0j} + \beta_1 \text{intervention}_j + \ \beta_2 \text{baseline WAM}_{\ ij} + \beta_3 \text{FSM}_{\ ij} +$$

$$+\beta_4 \text{FSM}_{\ ij} * \text{intervention}_j + \boldsymbol{\beta}\text{region}_j + \epsilon_{ij}$$

With $\text{FSM}_{\ ij}$ being a dummy variable for pupil *i* in school *j*’s FSM eligibility status and the remaining variables as described in the Primary Analysis section.

The Year 5 regression model with an extra FSM*group interaction term is given by:

$$Y = \beta_0 + \beta_1 \text{intervention} + \ \beta_2 \text{baseline WAM} + \beta_3 \text{FSM} + \beta_4 \text{FSM} * \text{intervention}_j + \boldsymbol{\beta}\text{school} + \epsilon$$

With the model’s variables being the ones introduced in the Primary Analysis section.

The criteria for eligibility for the Year 5 trial include both pupils who are slow and effortful writers and pupils whose handwriting is illegible. To evaluate the differential impact of the small groups setting HHS intervention on slow and effortful writers versus children whose handwriting is illegible we will also run modified primary outcome models with interaction terms (handwriting speed raw scores at baseline and the product of the handwriting speed raw scores at baseline and randomised group will be included in the model as covariates)[16].

And the model with handwriting speed interaction terms is given by:

$$Y = \beta_0 + \beta_1 \text{intervention} + \ \beta_2 \text{baseline WAM} + \beta_3 \text{FSM} + \beta_4 \text{baseline handwriting} +$$

$$+\beta_5 \text{baseline handwriting} * \text{intervention} + \boldsymbol{\beta}\text{school} + \epsilon$$

[16] This subgroup analysis, not being pre-specified in the original Protocol, will be reported as a post-hoc exploratory analysis.

Where $\text{baseline handwriting}$ is the pupil’s handwriting speed raw score measured at baseline and the remaining variables as defined in the Primary Analysis section.

No subgroup analyses will be performed for the secondary outcomes.

Imbalance at baseline

For both the Year 2 and the Year 5 Experiments, we will explore imbalance at baseline in terms of the primary outcome model covariates (pre-test writing ability and randomisation stratification indicators) for analysed groups.

To evaluate the imbalance in terms of randomisation stratifiers. We will produce contingency tables with number and proportion of cases for the control and intervention groups and each stratifier. For the Year 2 Experiment we will consider the breakdown in terms of the number and proportion of schools in the different region groups, while for the Year 5 Experiment we will consider the breakdown in terms of the number and proportion of pupils in the different schools and FSM status. Chi-squared tests will also be performed to enquire if there are significant differences between the intervention and the control groups.

To evaluate the imbalance in terms of pre-test writing ability we will be comparing differences in means between the intervention and control groups and reporting them as effect sizes as specified in the *Effect size calculation* section below. For the Year 2 Experiment we will be fitting a two-level (school and pupil) model, while for the Year 5 we will be performing a t-test.

Missing data

The Year 2 and the Year 5 HHS interventions are demanding for the participating schools in terms of mobilising staff and resources, and also call for a high level of engagement from pupils, particularly Year 5s.

It is to be expected that the measurement attrition rates for both Experiments will be in excess of five per cent and strategies to tackle the problem of missing data will be considered.

After evaluating to what extent data are missing and counting the number of complete cases, we will proceed to identify patterns of missingness in outcome variables. Note that given the design of the Experiments only pupils who were assessed at baseline for the different outcomes were included in the trial, and so we are not expecting to find missing cases in the data corresponding to any of the covariates of the different models to be run (baseline outcomes and stratification variables whose values are already known). As such, we will not investigate missingness in terms of any variables other than outcome variables.

To test if outcome data are not missing completely at random (MCAR) we will carry out Little’s MCAR test (McKnight at al 2007 pp.93-94), whose null hypothesis is that data is MCAR. Little’s MCAR test can be performed in R (version 3.4.1) using the package ‘BaylorEdPsych’.

If we reject the hypothesis of MCAR data, we will then investigate missingness patterns by means of substantive models for the different outcome variables. For the Year 2 Experiment the outcome substantive model will be a two-level (pupil and school) logistic model with baseline outcome, region and randomisation group indicators as covariates; while for Year 5 we will fit a logistic regression model with baseline outcome, school, FSM and randomisation group indicators as covariates.

After this stage the analyses will follow the road-map from EEF 2018 analysis guidance[17].

If necessary, sensitivity analysis would build on a multi-level multiple imputation and can be implemented in R (version 3.4.1) using the packages MICE and smcfcs.

Compliance

The compliance model was designed jointly by the evaluator and the developer, and will consist of three measures of compliance.

1. Number/length of handwriting sessions delivered

This measure is included because it gives an indication of dosage (see Humphrey et al, 2016). It rates the number and length of all handwriting sessions delivered by the trained teacher. NFER provide teachers with a log in which they must note information on length, date, etc. about all sessions delivered. For the Year 2 experiment, teachers record this information at class level. For the Year 5 experiment, teachers record this information at pupil level.

Each session is rated as not delivered/less than 30 mins/30 mins/more than 30 mins. Total intervention delivery time per school will be calculated from this rating by summing each session length according to the following rule:

Rating given by teacher per session	Time estimate to be used in measure
Not delivered	0
Less than 30 mins	20
30 mins	30
More than 30 mins	40

As this data reflects actual contact time between the children and the intervention, this will form the main Complier Average Causal Effect analysis.

Taking into account the nature of the HHS intervention, school level for Year 2 and pupil level for Year 5, the Year 2 Experiment's measure will be a school level dosage measure and the Year 5 Experiment's a pupil level dosage measure.

Year 2 Experiment school-level in the intervention group schools: Average total length of HHS sessions delivered per class[18] .

Year 5 Experiment pupil-level: Total length of small group setting HHS sessions.

To evaluate if there is an association between the dosage of pupils with the HHS interventions and writing ability we will treat the total intervention delivery time measure defined above as a pseudo-continuous dosage measure and adopt the instrumental variables approach (IV) prescribed by the EEF 2018 guidelines (Angrist & Imbens 1995, pp.

[17] We are working under the expectation that there will be will no missing values among the models' covariate under MAR, and that it will be possible to obtain valid estimates by including covariates predictive of non-response in the substantive models. The models' interpretation is conditional on these covariates being included.

[18] The number of sessions delivered to Year 2 classes in each of the intervention schools will be recorded by each classroom teacher delivering the HHS intervention in a fidelity/dosage log. For each of the intervention group schools, the Year 2 dosage measure will be computed as:
$\frac{total\ length\ of\ sessions\ across\ teachers}{number\ of\ Y2\ classes}$.

431-442). The instrumental variable regressions by two-stage least squares with group allocation as the instrumental variable models will be fit using the function ‘ivreg’ from the R package ‘AER’ and the estimation of causal effects will be done resorting to the ‘ivpack’ package. The analyses will, as before, be run in R (version 3.4.1).

In a systematic review of handwriting interventions performed in 2011 Hoy *et al* concluded that interventions that included less than 20 practice sessions were ineffective (Hoy *et al*, 2011). Based on the review’s finding we will also define the following dichotomous (Y/N) compliance variables of whether no fewer than 20 sessions were delivered/attended:

Year 2 Experiment school-level: All the Y2 classes in the school had at least 20 out of 24 class based HHS interventions delivered. This variable takes the value “N” for schools in the control group.

Year 5 Experiment pupil level: The pupil attended at least 20 out of 24 small group setting HHS interventions. This variable takes the value “N” for pupils in the control group.

We will once again use instrumental variables (IV) approaches with group allocation as the instrumental variable, as suggested in (Angrist & Imbens 1995, pp. 431-442), to enquire if there is an association between delivering/attending at least 20 sessions and the different trials’ primary outcomes. We will be using R (version 3.4.1) and the packages mentioned earlier in this section to perform the analyses.

2. Attendance at training (high, medium, low)

This measure will form an additional compliance analysis, and is included because it gives an indication of training dosage for teachers. It is measured at school level across both experiments (not separately), collected at training sessions via a register of attendance and then passed to the evaluator by the trainers. We will provide descriptive stats for this measure to indicate compliance.

- High – at least one member of staff attends who will be the deliverer for each year group, and one member of SLT.
- Medium – only one member of staff attends, who will be the deliverer for one year group, intending to cascade to the other. A member of SLT also attends the training.
- Low – only one member of staff attends, who may/may not be a deliverer, intending to cascade to all others. No member of SLT attends.

3. Extent to which schools use the programme after the eight week delivery

This measure will also form an additional compliance analysis, measured at school level across both experiments (not separately). We will provide descriptive stats for this measure to indicate compliance. It is included because it gives an indication of reach and responsiveness (see Humphrey *et al*, 2016). A relatively unusual aspect of this intervention is the substantial gap for ‘embedding’ between the formal intervention period and the follow-up testing. As approaches towards and integration of the intervention may vary dramatically during this period, this will be tracked at a school level using an NFER-provided log to be completed monthly in the period between the end of the eight week formal intervention and the testing in June 2019. Trained teachers complete it monthly. It requires teachers to indicate to what extent they have used the techniques and materials from the intervention in the month of completion. This data will then be compiled and cut scores applied to reflect:

- Regularly
- Occasionally
- Not at all

Intra-cluster correlations (ICCs)

For the Year 2 Experiment school-level ICCs will be estimated from the variance of the random intercept and residual variance of the multi-level models by means of the formula:

$$ICC = \frac{\sigma^2_{intercepts}}{\sigma^2_{intercepts} + \sigma^2_{residuals}}$$

ICCs at baseline will be computed considering random intercepts two-level (school and pupil) models with no covariates, and post-test ICCs will be derived from the primary ITT model and secondary ITT model for the first secondary outcome described above (writing composition).

Effect size calculation

As advised by the EEF 2018 guidelines, we will be reporting effect Hedges' gs as effect sizes. These are calculated according to the formula:

$$\boldsymbol{g} = \frac{\boldsymbol{\overline{o}_i} - \boldsymbol{\overline{o}_c}}{\boldsymbol{s}^*}$$

With $\boldsymbol{\overline{o}_i} - \boldsymbol{\overline{o}_c}$ corresponding to the difference between the intervention and control group in terms of the mean value of the outcome being assessed, and $\boldsymbol{s}^*$ corresponding to the pooled standard deviation[19] of the outcome.

For both Experiments, the numerator for the effect size calculation will be the coefficients of the intervention group from the regression models (single level for Year 5, multi-level for Year 2). As prior ability is one of the covariates included in the models, we will be using unconditional variance from the corresponding models without covariates as denominators. The effect size thus computed is equivalent to Hedges'g.

Confidence intervals for each effect size will be computed by multiplying the standard errors of the intervention group by the left-tailed inverse of the Student's t-distribution with a probability of 2.5% and the number of degrees of freedom associated to the intervention group. The confidence intervals for the standard errors will be converted to effect size confidence intervals using the same formula as the effect sizes themselves.

[19] The pooled standard deviation is computed as $s^* = \sqrt{\frac{(N_i-1)s_i^2 + (N_c-1)s_c^2}{N_i+N_c-2}}$ with N_i and N_C being the number of elements in the intervention and control groups, and s_i and s_c the standard deviations of the outcome measured in the intervention and control groups.

References:

Angrist, J.D., and Imbens, G.W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. American Statistical Association 90 (430).

Dunsmuir, S., Kyriacou,M., Batuwitage, S., Hinson, E., Ingram, V. and O'Sullivan, S. (2015). An evaluation of the writing Assessment Measure for children's narrative writing. Assessing Writing 23.

Education Endowment Foundation (2013). *Pre-testing in EEF Evaluations*. London: EEF [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/Pre-testing_paper.pdf [23 March 2018].

Education Endowment Foundation (2015). *Intra-cluster Correlation Coefficients*. London: EEF [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/ICC_2015.pdf [23 March 2018]

Hoy, M.M.P., Egan, M.Y., and Feder, K.P. (2011). A Systematic Review of Interventions to Improve Handwriting. Canadian Journal of occupational therapy 78 (1).

Hunter, D.R. (2004). MM algorithms for generalized Bradley-Terry models. Annals of Statistics 32(1).

Kish, L. (1965). Survey Sampling. New York: Wiley.

Lortie-Forgues, H. (2017) What can we learn from RCTs in Education? A meta-analysis of RCTs Commissioned by the EEF and IES Paper presented at the RCTs in the Social Sciences Conference, University of York, September 2017.

McKnight, P.E., McKnight, K.M., Sidani, S., and Figueredo, A. J. (2007). Missing Data: A gentle introduction. The Guildford Press.

Medwell, J., Strand, S., and Wray, D. (2009). The links between handwriting and composing for Y6 children. Cambridge Journal of education 39 (3).

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy and Practice 19 (3).

Sanders, M. and Ni Chonaire, A. (2015). *"Powered to Detect Small Effect Sizes": You Keep Saying That. I Do Not Think It Means What You Think It Means.* (Working Paper No. 15/337). Bristol: CMPO [online]. Available: http://www.bris.ac.uk/media-library/sites/cmpo/documents/WP15337_Web_Version.pdf [23 March 2018]

Santangelo, T. & Graham, S. (2016) A Comprehensive Meta-analysis of Handwriting Instruction. Educational Psychology Review 28: 225-265.

Snijders, Tom A.B. (2005) 'Power and Sample Size in Multilevel Linear Models'.
In: B.S. Everitt and D.C. Howell (eds.), Encyclopedia of Statistics in Behavioral Science (3).Chicester: Wiley.

Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C. and Jefferson, L. (2014a). *Grammar for Writing. Evaluation Report and Executive Summary*. London: EEF [online]. Available: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/FINAL_EEF_Evaluation_Report_-_Grammar_for_Writing_-_February_2014.pdf [23 March 2018]

Torgerson, D., Torgerson, C., Mitchell, N., Buckley, H., Ainsworth, H., Heaps, C. and Jefferson, L. (2014b). *Improving Writing Quality. Evaluation Report and Executive Summary.*

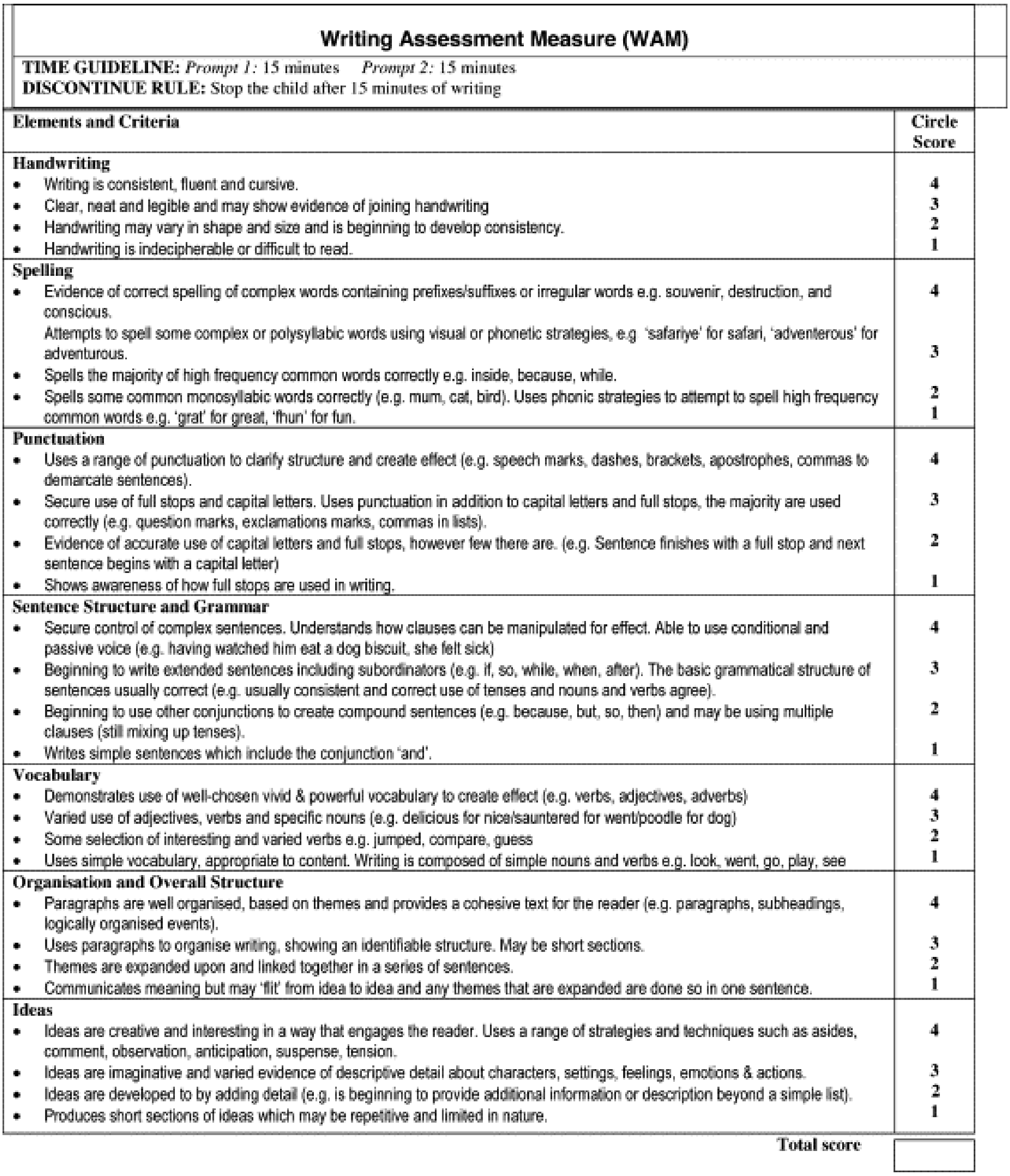
London: EEF [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Evaluation_Report_-_Improving_Writing_Quality.pdf [23 March 2018]

Wallen, M., Bonney, M.A., Lennox, L. (2006). The Handwriting Speed Test. Australian Occupational Therapy Journal 60 (5).

Appendices

APPENDIX A

Writing Assessment Measure marking scheme (Dunsmuir, Kyriacou, Batuwitage, Hinson, Ingram and O'Sullivan, 2013).



Writing Assessment Measure (WAM)

TIME GUIDELINE: *Prompt 1:* 15 minutes *Prompt 2:* 15 minutes
DISCONTINUE RULE: Stop the child after 15 minutes of writing

Elements and Criteria	Circle Score
Handwriting	
• Writing is consistent, fluent and cursive.	4
• Clear, neat and legible and may show evidence of joining handwriting	3
• Handwriting may vary in shape and size and is beginning to develop consistency.	2
• Handwriting is indecipherable or difficult to read.	1
Spelling	
• Evidence of correct spelling of complex words containing prefixes/suffixes or irregular words e.g. souvenir, destruction, and conscious.	4
Attempts to spell some complex or polysyllabic words using visual or phonetic strategies, e.g 'safariye' for safari, 'adventerous' for adventurous.	3
• Spells the majority of high frequency common words correctly e.g. inside, because, while.	2
• Spells some common monosyllabic words correctly (e.g. mum, cat, bird). Uses phonic strategies to attempt to spell high frequency common words e.g. 'grat' for great, 'fhun' for fun.	1
Punctuation	
• Uses a range of punctuation to clarify structure and create effect (e.g. speech marks, dashes, brackets, apostrophes, commas to demarcate sentences).	4
• Secure use of full stops and capital letters. Uses punctuation in addition to capital letters and full stops, the majority are used correctly (e.g. question marks, exclamations marks, commas in lists).	3
• Evidence of accurate use of capital letters and full stops, however few there are. (e.g. Sentence finishes with a full stop and next sentence begins with a capital letter)	2
• Shows awareness of how full stops are used in writing.	1
Sentence Structure and Grammar	
• Secure control of complex sentences. Understands how clauses can be manipulated for effect. Able to use conditional and passive voice (e.g. having watched him eat a dog biscuit, she felt sick)	4
• Beginning to write extended sentences including subordinators (e.g. if, so, while, when, after). The basic grammatical structure of sentences usually correct (e.g. usually consistent and correct use of tenses and nouns and verbs agree).	3
• Beginning to use other conjunctions to create compound sentences (e.g. because, but, so, then) and may be using multiple clauses (still mixing up tenses).	2
• Writes simple sentences which include the conjunction 'and'.	1
Vocabulary	
• Demonstrates use of well-chosen vivid & powerful vocabulary to create effect (e.g. verbs, adjectives, adverbs)	4
• Varied use of adjectives, verbs and specific nouns (e.g. delicious for nice/sauntered for went/poodle for dog)	3
• Some selection of interesting and varied verbs e.g. jumped, compare, guess	2
• Uses simple vocabulary, appropriate to content. Writing is composed of simple nouns and verbs e.g. look, went, go, play, see	1
Organisation and Overall Structure	
• Paragraphs are well organised, based on themes and provides a cohesive text for the reader (e.g. paragraphs, subheadings, logically organised events).	4
• Uses paragraphs to organise writing, showing an identifiable structure. May be short sections.	3
• Themes are expanded upon and linked together in a series of sentences.	2
• Communicates meaning but may 'flit' from idea to idea and any themes that are expanded are done so in one sentence.	1
Ideas	
• Ideas are creative and interesting in a way that engages the reader. Uses a range of strategies and techniques such as asides, comment, observation, anticipation, suspense, tension.	4
• Ideas are imaginative and varied evidence of descriptive detail about characters, settings, feelings, emotions & actions.	3
• Ideas are developed to by adding detail (e.g. is beginning to provide additional information or description beyond a simple list).	2
• Produces short sections of ideas which may be repetitive and limited in nature.	1
Total score	