Evaluation of The OTTO Club: a two-arm randomised controlled trial
Statistical Analysis Plan

Evaluator (institution): National Centre for Social Research (NatCen)
Principal investigator(s): Helena Takala

Template last updated: August 2019

PROJECT TITLE	Evaluation of The OTTO Club: a two-arm randomised controlled trial
DEVELOPER (INSTITUTION)	The OTTO Club
EVALUATOR (INSTITUTION)	National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR(S)	Helena Takala
PROTOCOL AUTHOR(S)	Ekaterina Stoilova, Enes Duysak, Alina Haque
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Pupils aged 5-6 in Year 1, Key Stage 1
NUMBER OF SCHOOLS	140
NUMBER OF PUPILS	3,724[1]
PRIMARY OUTCOME MEASURE AND SOURCE	Accuracy of letter formation – Alphabet Writing task in The Detailed Assessment of Speed of Handwriting (DASH-2)
SECONDARY OUTCOME MEASURE AND SOURCE	1. Ability to form letters through the correct process – Alphabet Writing task in the DASH-2 2. Fine motor control – Drawing Circles task in the Movement Assessment Battery for Children (MABC-3) 3. Postural control – supine flexion and prone extension postures from the Clinical Observations of Motor and Postural Skills (COMPS-2) 4. Handwriting confidence and motivation to practise – bespoke survey measure 5. End of year writing attainment – Interim teacher assessments for Key Stage 1

[1] Estimated number at randomisation (see Sample size calculations overview)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]	DATE	12 May 2026

Table of contents

Introduction

The OTTO Club is a programme designed by a team of Occupational Therapists (OTs) to improve Year 1 children's handwriting, as well as the underlying skills of postural and fine motor control, and increase confidence and motivation to practise handwriting. The OTTO Club trains teachers and teaching assistants (TAs) with the aim of:

a. building their understanding of how postural stability, fine motor control, and writing techniques affect handwriting quality;
b. increasing their knowledge, skills and confidence to implement related activities within handwriting lessons, and
c. ultimately improving pupils' motor and handwriting skills.

Over the 10 weeks of the programme, a trained teacher and/or TA delivers a 45-60-minute The OTTO Club literacy lesson to their Year 1 class in place of their usual weekly literacy or handwriting lesson, alongside daily 10 – 15-minute follow-up activities to reinforce the lesson. The programme uses fun and engaging activities in the lesson and follow-up activities to build the skills that children need for effective handwriting such as posture, pencil grasp, and manual dexterity. It can be run as a whole-class intervention (main programme) and optionally as an additional targeted intervention, which follows the same format with a selected group of pupils who require further support or practice once the whole-class intervention is completed.

NatCen is conducting an efficacy trial of The OTTO Club, consisting of an impact evaluation (IE) and an implementation and process evaluation (IPE). The programme will be delivered in a sample of 140 primary schools in England between November 2025 and June 2026. More details on the intervention and the evaluation are available in the study protocol (Takala et al., 2025).

This Statistical Analysis Plan (SAP) details the design and respective analysis planned for the IE.

Design overview

The impact evaluation of The OTTO Club will be conducted as a two-arm cluster randomised controlled efficacy trial, with randomisation at the school level and pupils as the unit of analysis. Table 1 summarises the trial design.

Table 1. Trial design

Trial design, including number of arms		Two-arm cluster randomised controlled trial (The OTTO Club and teaching as usual)
Unit of randomisation		School
Stratification variables (if applicable)		School-level handwriting time[2] (low, medium, high, NA) and a binary indicator for whether the school is in an Education Investment Area
	variable	Accuracy of letter formation

[2] This is the amount of time a class spent handwriting in a typical week, collected from teachers before randomisation.

Primary outcome	measure (instrument, scale, source)	Detailed Assessment of Speed of Handwriting, 2nd Edition (DASH-2; Barnett et al., 2024) Alphabet Writing task, number of correctly formed letters in one minute
Secondary outcome(s)	variable(s)	1. Ability to form letters through the correct process 2. Fine motor control 3. Postural control 4. Confidence in handwriting and motivation to practise handwriting 5. End of year writing attainment
	measure(s) (instrument, scale, source)	1. DASH-2 Alphabet Writing task, number of letters formed via the correct process (using OT professional judgement) 2. Movement Assessment Battery for Children, 3rd Edition (MABC-3; Henderson & Barnett, 2023) Drawing Circles task, number of correctly formed circles 3. Clinical Observations of Motor and Postural Skills, 2nd Edition (COMPS-2; Wilson et al., 2000), supine flexion posture and prone extension posture, combined duration (in sec) for which the two positions are held 4. Bespoke attitudes questionnaire, confidence subscale sum score (1-5) and motivation subscale sum score (1-5) 5. Interim teacher assessments for Key Stage 1 (pupil working towards the expected standard (0) or working at the expected standard or at greater depth (1))
Baseline for primary outcome	**variable**	Accuracy of letter formation
	measure (instrument, scale, source)	DASH-2 Alphabet Writing task, number of correctly formed letters in one minute
Baseline for secondary outcome	**variable**	1. Ability to form letters through the correct process 2. Fine motor control 3. Postural control 4. Confidence in handwriting and motivation to practise handwriting 5. Accuracy of letter formation
	measure (instrument, scale, source)	1. DASH-2 Alphabet Writing task, number of letters formed via the correct process (using OT professional judgement) 2. MABC-3 Drawing Circles task, number of correctly formed circles 3. COMPS-2 supine flexion posture and prone extension posture, combined duration for which the two positions are held (in sec)

		4. Bespoke attitudes questionnaire, confidence subscale sum score (1-5) and motivation subscale sum score (1-5) 5. DASH-2 Alphabet Writing task, number of correctly formed letters in one minute

Research questions

The IE will answer the following research questions:

Primary research question:

- RQ1. What is the impact of The OTTO Club on 5-6-year-olds' accuracy of letter formation?

Secondary research questions:

- RQ2. What is the impact of The OTTO Club on 5-6-year-olds' ability to form letters through the correct process?
- RQ3. What is the impact of The OTTO Club on 5-6-year-olds' fine motor control?
- RQ4. What is the impact of The OTTO Club on 5-6-year-olds' postural control?
- RQ5. What is the impact of The OTTO Club on 5-6-year-olds' confidence and motivation to practise handwriting?
- RQ6. Does the impact of The OTTO Club on pupils' accuracy of letter formation differ for pupils from disadvantaged backgrounds, as measured by FSM eligibility status?
- RQ7. What is the impact of The OTTO Club on pupils' end-of-year writing attainment at the end of Year 1?

Recruitment and randomisation

A sample of 140 primary schools in England was recruited for the trial by the delivery team and randomised by NatCen into the two trial arms, with half randomly assigned to receive The OTTO Club (n = 70) and half randomly assigned to continue teaching as usual (TAU) (n = 70). Randomisation was stratified by school-level handwriting time[3] and whether the school is located in an Education Investment Area (EIA), to ensure balance of those characteristics between the OTTO and TAU groups. Incentive payments will be offered to schools in both trial arms for the completion of evaluation activities. More details on recruitment are available in the study protocol (Takala et al., 2025).

Randomisation of schools was carried out by the Impact Evaluation team at NatCen using the randtreat command in Stata version 17 in July 2025. Stata .do and .log files were used to record

[3] This information was collected from Year 1 teachers in June 2025 using a template developed by NatCen, where teachers logged the time in minutes their class spent handwriting each day and lesson of a specified week. Schools were split into tertiles based on the total time spent handwriting across all lessons in that week. The four groups will be high, medium and low handwriting time, alongside missing handwriting information.

the process. Researchers were blind to school identity at the time of randomisation and school identifiers were linked back into the data once randomisation was complete.

Outcome measures

The primary outcome will be pupils' accuracy of letter formation, measured using the Alphabet Writing subtest of the Detailed Assessment of Speed of Handwriting (DASH-2; Barnett et al., 2024). This outcome assesses the legibility of letter output (the "product" of handwriting) and will be scored as the number of correctly formed letters of the alphabet in one minute[4].

Secondary outcomes will be pupils' ability to form letters through the correct process, pupils' fine motor and postural control, as well as their confidence in handwriting and motivation to practise handwriting.

- Pupils' ability to form letters through the correct process (the "process" of handwriting) will be measured as the number of letters formed via the correct process in one minute on the DASH-2 Alphabet Writing task, using the professional judgement of OT testers[5] (e.g., forming the lowercase letter 'α' with a counterclockwise circular stroke starting on the right-hand side).

- Fine motor control will be assessed using the Drawing Circles task in the Movement Assessment Battery for Children, 3rd Edition (MABC-3; Henderson & Barnett, 2023) and scored as the number of correctly formed circles (out of 24 in the task). To be counted as correctly formed, circles need to be formed within the bounds of an inner and an outer outline, form a single, continuous line and not cross either boundary.

- Pupils' postural control will be assessed using observations on the supine flexion and prone extension postures subtests in the Clinical Observations of Motor and Postural Skills 2nd Edition (COMPS-2; Wilson et al., 2000), and scored as the combined duration in seconds for which both positions are held (capped at a maximum of 2 minutes per posture)[6].

- Pupils' confidence in handwriting and their motivation to practise handwriting will be assessed via a bespoke attitudes survey, developed by NatCen in collaboration with the delivery team. The survey contains seven items, with three items falling into the confidence sub-scale, and four into the motivation sub-scale (see Appendix A). Responses are given on a 5-point Likert scale, ranging from 1 (No, never) to 5 (Yes, always). Each of the confidence and motivation scores will be a sum across the items belonging to the respective subscale.

In addition, as a follow up for the whole-class intervention, we will explore the impact of The OTTO Club on later writing attainment using teacher assessment of pupil writing attainment at

[4] Repeated letters (e.g., "a a") will be counted only once, and testers will prompt pupils to try and write a different letter down.

[5] OTs will be blind to schools' random allocation, as far as possible.

[6] In our pilot of the measure, we capped the duration for which pupils hold each pose at one minute. While we increase the cap to two minutes for the main testing in the trial to allow for higher-than-expected performance, based on our pilot of the measure, we do not expect most students to hold each pose for longer than a minute. This cap is introduced for logistical reasons associated with testing.

the end of Year 1. This is a non-statutory, in-school summative teacher assessment[7] of pupil performance in writing at the end of Year 1, and for the purpose of this evaluation it will be treated as a binary measure of whether the pupil is working at the expected standard / at greater depth, or whether they are working towards the expected standard. This analysis will constitute an addendum to the main report.

More detail on the outcome measures is provided in the study protocol (Takala et al., 2025).

Data collection

The primary and secondary outcomes will be collected at baseline (October/November 2025) and endline (February 2026).

It was intended that all outcome measure assessments are carried out by a team of occupational therapists (OTs), trained by the research team to carry out testing for the study (see study protocol). However, due to logistical constraints, in some schools, endline testing will be carried out by trained researchers instead. To test for any potential differences in scores across the measures this might have resulted in, we will conduct a sensitivity analysis for the primary outcome, including an indicator of whether each school was tested by an OT or a researcher as a covariate (see Analysis).

Sample size calculations overview

Tables 2 and 3 present power calculations conducted at three stages: at the protocol stage, at randomisation, and at the time of writing (pre-baseline-testing). All power calculations were conducted using PowerUp! (Dong & Maynard, 2013). We assume a Type I error rate of 0.05 and a Type II error rate of 0.20 (i.e., 80% power).

Planned sample size at protocol stage

At the protocol stage, we anticipated a sample of 138 schools to be recruited. Based on figures from the Department for Education for school year 2023/24[8], we anticipated an average of 26.6 pupils per class and an average of 6.5 pupils (24.6%) being eligible for Free School Meals (FSM)[9].

Given the lack of evidence that focuses specifically on handwriting as the evaluation subject, our estimates of pre-post and intracluster correlations (ICCs) were informed by EEF's review of writing practice[10] and represent an average across estimates provided for wider education measures. Specifically, we assumed a pre-post correlation of 0.52 at the pupil level and 0.45 at the school level, and an ICC of 0.12.

[7] More information and guidance on non-statutory KS1 assessments is available at https://www.gov.uk/government/publications/key-stage-1-teacher-assessment-guidance/key-stage-1-teacher-assessment-guidance

[8] Schools, pupils and their characteristics. Academic year 2023/24. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2023-24

[9] Due to targeted recruitment in EIAs, the proportion of FSM eligible pupils in the sample might be higher than this estimate, which would increase statistical power for the FSM subgroup analysis compared to estimates provided below.

[10] Education Endowment Foundation (2024). Writing practice review: Understanding current practice and research priorities in teaching writing. Accessed from https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/writing-practice-review.

Based on our experience with trials involving primary data collection in nurseries and primary schools (e.g., Bury et al., 2022; Basharat et al., 2023), we expected moderate to high attrition from recruitment to endline at both the pupil and the school level. Our power calculations therefore account for 20% school attrition and 20% pupil attrition between randomisation and endline.

Under these assumptions, at protocol we estimated that the trial would be powered to detect a Minimum Detectable Effect Size (MDES) of 0.19 for the whole sample and a MDES of 0.25 for the FSM subsample (see Table 2).

Achieved sample size at randomisation

At randomisation, our recruited sample consisted of 140 schools[11], with 70 randomised into The OTTO Club group and 70 into the TAU group. Retaining the same assumptions of the average number of pupils and the number of pupils eligible for FSM as before, the same anticipated pre-post correlations and ICCs, and the same 20% school and pupil attrition at endline, the trial would be powered to detect a MDES of 0.19 for the whole sample and 0.25 for the FSM subsample (see Table 2).

Table 2. Power calculations – at protocol and randomisation

		Protocol (assuming 20% school and 20% pupil attrition at endline)		Randomisation (assuming 20% school and 20% pupil attrition at endline)	
		Overall	FSM	Overall	FSM
Minimum Detectable Effect Size (MDES)		**0.19**	**0.25**	**0.19**	**0.25**
Pre-test/ post-test correlations	level 1 (pupil)	0.52	0.52	0.52	0.52
	level 2 (school)	0.45	0.45	0.45	0.45
Intracluster correlations (ICCs)	level 2 (class)	0.12	0.12	0.12	0.12
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		21.3	5.2	21.3	5.2
Number of schools	intervention	55	55	56	56

[11] A larger number of schools than planned was recruited to better prepare the trial for potential attrition post-randomisation.

	control	55	55	56	56
	total	110	110	112	112
Number of pupils	intervention	1,175	289	1,192	293
	control	1,175	289	1,192	293
	total	2,350	578	2,384	586

Updated power calculations

At the time of writing this SAP in October 2025, the school sample stands at 125 schools – due to 15 schools withdrawing post-randomisation at the beginning of the school year (n = 8 in The OTTO Club group and n = 7 in the TAU group).

Using the latest figures available from the Department of Education for school year 2024/25[12], we anticipate an average of 26.2 pupils per class and an average of 6.8 pupils (25.7%) being eligible for FSM[13].

Retaining the same assumptions of pre-post correlations and ICCs, and a 20% school and pupil attrition between randomisation and endline, at the time of writing we estimate that the trial would be powered to detect a MDES of 0.19 for the whole sample and 0.25 for the FSM subsample (see Table 3). Given the attrition numbers observed post-randomisation, we also estimated a scenario in which school-level attrition is 30% between randomisation and endline. In this case, the trial would be powered to detect a MDES of 0.20 for the whole sample and 0.27 for the FSM subsample (see Table 3).

Table 3. Updated power calculations

		Pre-baseline (assuming 20% school and 20% pupil attrition at endline)		**Pre-baseline (assuming 30% school and 20% pupil attrition at endline)**	
		Overall	**FSM**	**Overall**	**FSM**
Minimum Detectable Effect Size (MDES)		**0.19**	**0.25**	**0.20**	**0.27**
Pre-test/ post-test correlations	level 1 (pupil)	0.52	0.52	0.52	0.52
	level 2 (school)	0.45	0.45	0.45	0.45
Intracluster correlations (ICCs)	level 2 (class)	0.12	0.12	0.12	0.12

[12] Schools, pupils and their characteristics. Academic year 2024/25. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics./2024-25

[13] Due to targeted recruitment in EIAs, the proportion of FSM eligible pupils in the sample might be higher than this estimate, which would increase statistical power for the FSM subgroup analysis compared to estimates provided below.

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		21	5.4	21	5.4
Number of schools	intervention	56	56	49	49
	control	56	56	49	49
	total	112	112	98	98
Number of pupils	intervention	1,174	302	1,027	264
	control	1,174	302	1,027	264
	total	2,348	604	2,054	528

Analysis

The primary and secondary outcome analyses will use an Intention-to-Treat (ITT) approach to estimate the impact of The OTTO Club on each outcome of interest, and follow EEF statistical analysis guidance (EEF, 2022). The analysis will be implemented in Stata 17.

For each outcome, we will report descriptive statistics, including means and standard deviations of baseline and endline scores, and present histograms of their distributions in each trial arm.

Primary outcome analysis

The primary outcome analysis will estimate the effect of The OTTO Club on pupils' accuracy of letter formation as the primary outcome (**RQ1**), measured as the number of correctly formed letters of the alphabet in one minute on the DASH-2 Alphabet Writing task. We will estimate a **two-level linear mixed effects regression model**[14], to account for pupils (level 1) being clustered within schools (level 2).

The model will include the total number of correctly formed letters at endline as the dependent variable, a binary indicator of treatment allocation (OTTO or TAU) as predictor, and pupils' number of correctly formed letters at baseline and stratification variables (school-level handwriting time and EIA status) as covariates. The basic form of the model is:

$$Number\ of\ correctly\ formed\ letters_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on pupils' accuracy of letter formation. u_j is a school-level random effect and e_{ij} is the error term, both assumed to be normally distributed and uncorrelated with covariates in the model.

Following EEF statistical analysis guidance (EEF, 2022), will report Hedges' g effect sizes along with their 95% confidence intervals (CIs) for all primary and secondary analyses (see Effect size

[14] We will check the distribution of the outcome and conduct sensitivity analyses if required.

calculation). We will report intracluster correlations (ICCs) for the primary outcome analysis (see Intracluster correlations (ICCs)).

Secondary outcome analysis

An ITT approach will be adopted to estimate the impact of The OTTO Club on secondary outcome measures (RQ2-5), with **two-level linear mixed-effects regression models** estimated to account for pupils (level 1) being clustered within schools (level 2).

Ability to form letters through the correct process

To estimate the effect of The OTTO Club on pupils' ability to form letters through the correct process (**RQ2**), measured as the number of letters of the alphabet formed through the correct process in the DASH-2 Alphabet Writing task, we will estimate a **two-level linear mixed effects regression model**, with pupils at level 1 and schools at level 2. The model will include the number of letters formed through the correct process at endline as the dependent variable, a binary indicator for treatment allocation as predictor, as well as the number of letters formed through the correct process at baseline and stratification variables (school-level handwriting time and EIA status) as covariates. The model will also include a random intercept by school. The basic form of the model is:

$$Number\ of\ letters\ formed\ via\ correct\ process_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on pupils' ability to form letters through the correct process. u_j is a school-level random effect and e_{ij} is the error term.

Fine motor control

To estimate the effect of The OTTO Club on pupils' fine motor control (RQ3), measured as the number of correctly formed circles on the MABC-3 Drawing Circles task, we will estimate a **two-level linear mixed effects regression model**, with pupils at level 1 and schools at level 2. The model will include the number of correctly formed circles in the task at endline as the dependent variable, a binary indicator for treatment allocation as predictor, and the number of correctly formed circles at baseline and stratification variables (school-level handwriting time and EIA status) as covariates. The model will also include a random intercept by school. The basic form of the model is:

$$Number\ of\ correctly\ formed\ circles_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on pupils' fine motor control. u_j is a school-level random effect and e_{ij} is the error term.

Postural control

To estimate the effect of The OTTO Club on pupils' postural control (RQ4), measured as the combined duration (in seconds) of maintaining the COMPS-2 supine flexion and prone extension postures, we will estimate a **two-level linear mixed effects regression model**, with pupils at

level 1 and schools at level 2. The model will include the combined duration in seconds of holding both postures at endline as the dependent variable, a binary indicator for treatment allocation as predictor, as well as combined duration in seconds of holding the postures at baseline and stratification variables (school-level handwriting time and EIA status) as covariates. The model will also include a random intercept by school. The basic form of the model is:

$$Combined\ duration\ of\ holding\ the\ postures_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on pupils' postural control. u_j is a school-level random effect and e_{ij} is the error term.

Based on pilot data, we do not expect many pupils to be able to hold each posture for over two minutes. If we find that a considerable proportion of pupils (i.e., > 5%) do, and as such have their combined duration values recorded as the combined 4-minute cap, we will conduct a sensitivity analysis using a censored model and compare the results to the linear mixed effects regression model described above.

Confidence and motivation to practise handwriting

To estimate the effect of The OTTO Club on pupils' confidence in handwriting and motivation to practice handwriting, we will estimate **two-level linear mixed effects regression models**, with pupils at level 1 and school at level 2. Each model will include the respective sum scores for the confidence sub-scale (three items) and the motivation sub-scale (four items), as the dependent variable, a binary indicator for treatment allocation as a predictor, and stratification variables (school-level handwriting time and EIA status) as covariates. The models will include a random intercept by school. The basic form of the models is:

$$Confidence\ /\ Motivation\ score_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Stratum_j + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on confidence and motivation scores. u_j is a school-level random effect and e_{ij} is the error term.

Subgroup analyses

To estimate whether the impact of The OTTO Club on the primary outcome differs for pupils from disadvantaged backgrounds, as measured by FSM eligibility status (RQ6), we will conduct a sub-group analysis. FSM eligibility status will be collected from schools alongside pupil enumeration data at baseline. Following EEF guidance (2022), the sub-group analysis will be carried out both using a sub-sample consisting of only pupils eligible for FSM, and via an interaction term added to the primary outcome model for the whole sample. The outcome for both approaches will be pupils' accuracy of letter formation.

Specifically, we will re-estimate the primary outcome model for the whole sample, with the addition of the FSM eligibility indicator and an interaction term combining FSM eligibility and treatment allocation. The basic form of the model is:

$$Number\ of\ correctly\ formed\ letters_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + \beta_4 FSM\ status_{ij} + \beta_5 Intervention * FSM\ status_{ij} + u_j + e_{ij}$$

Where pupils i are clustered within schools j. β_1 is the estimated treatment effect of the programme on the number of correctly formed letters for the whole sample. β_4 is the difference in the number of correctly formed letters between pupils eligible for FSM and those who are not. β_5 is the difference in the effect of the programme on the number of correctly formed letters between pupils eligible for FSM and those who are not. u_j is a school-level random effect and e_{ij} is the error term.

From this model, we will report the interaction term coefficient β_5 and its CI, alongside their interpretation for RQ6.

Additionally, we will estimate the primary outcome model for a sub-sample consisting only of pupils who are eligible for FSM:

$$Number\ of\ correctly\ formed\ letters_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

Where pupils eligible for FSM i are clustered within schools j. β_0 is the intercept and β_1 is the estimated treatment effect of the programme on the number of correctly formed letters among pupils eligible for FSM. u_j is a school-level random effect and e_{ij} is the error term.

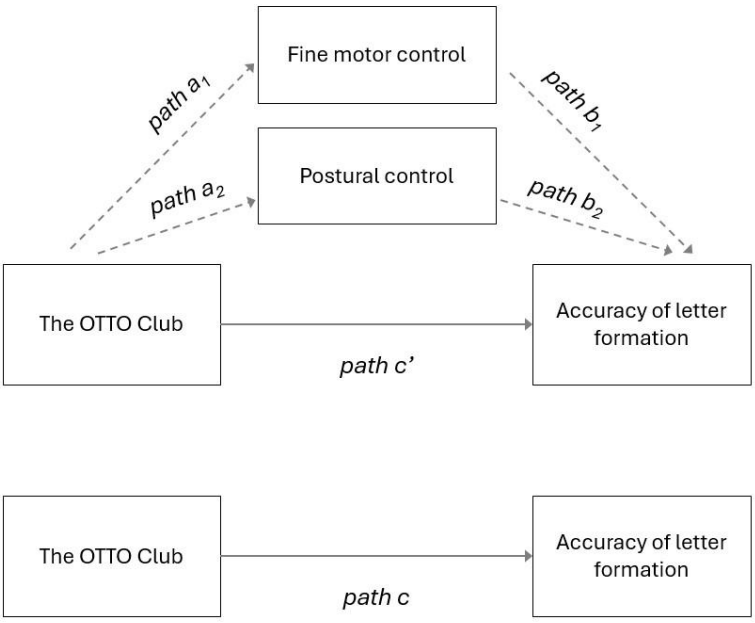
Additional analyses

Mediation analysis

We will conduct an exploratory mediation analysis to understand whether any effect of The OTTO Club on accuracy of letter formation at endline as the primary outcome is mediated by changes in **fine motor skill** and **postural control**[15]. For the purpose of this analysis, fine motor skill and postural control will be treated as independent, parallel mediators (i.e., mediators that do not mutually affect each other causally). Figure 1 displays the parallel causal mediation model.

[15] as two proposed mechanisms by which the intervention could affect the primary outcome of accuracy of letter formation (see study protocol, Takala et al., 2025).

Figure 1. Parallel mediation model



The programme may affect the outcome both directly (path c') and indirectly through two independent mediators (paths a_1*b_1 and a_2*b_2).

The analysis will decompose the ITT estimate for the effect of The OTTO Club on the primary outcome into a) two indirect effects - via changes in i) fine motor skill and ii) postural control, and b) a direct effect of the programme on the primary outcome that cannot be explained by changes in fine motor skill or postural control. We anticipate that any effect of the programme on pupils' letter formation accuracy may be at least partly mediated through improvements in these two skills.

The mediation analysis will adopt a parallel mediation framework, estimating the indirect effect of each mediator simultaneously to account for any correlation between the mediators. The analysis will involve the following steps:

1. Regressing each mediator on The OTTO Club programme. The effect of the programme on each mediator is conventionally referred to as *path a*.
2. Regressing pupils' accuracy of letter formation at endline on The OTTO Club programme and on both mediators simultaneously. The effect of each mediator on the outcome is conventionally referred to as *path b*, while the unique direct effect of the programme (with both mediators accounted for) is referred to as *path c'*. The total effect of the programme on the outcome (direct and via mediators) is referred to as *path c*.
3. Calculating the average causal mediated effect (ACME, *path a*b*) for each mediator, the total ACME across the two mediators, and the proportion mediated effect (that is, the magnitude of the mediated effect relative to the total effect).

To estimate paths *a1* and *a2* as the effect of the programme on each mediator, we will estimate two-level mixed effects linear regression models predicting each mediator from treatment allocation. The model predicting fine motor control will include pupils' number of correctly formed circles in the MABC-3 Drawing Circles task at endline as the dependent variable. The model predicting postural control will include the combined duration in seconds for which pupils

maintain both COMPS-2 postures at endline as the dependent variable. The models will include respective baseline scores and stratification variables (school-level handwriting time and EIA status) as covariates, and a random intercept by school.

$$Fine\ motor\ control_{ij} = \beta_0 + \beta_{1a} Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

$$Postural\ control_{ij} = \beta_0 + \beta_{1b} Intervention_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

β_{1a} and β_{1b} represent the effect of The OTTO Club on each mediator, respectively (i.e., paths *a1* and *a2*).

To estimate paths *b1* and *b2* as the effects of each mediator on the primary outcome of accuracy of letter formation, we will estimate a two-level linear mixed effects regression model predicting the number of correctly formed letters on the DASH-2 Alphabet writing task at endline as the dependent variable. The model will include treatment allocation and both mediators as predictors, as well as pupils' number of correctly formed letters at baseline and stratification variables (school-level handwriting time and EIA status) as covariates, and a random intercept by school.

$$\begin{aligned} Number\ of\ correctly\ formed\ letters_{ij} \\ = \beta_0 + \beta_1 Intervention_j + \beta_2 Fine\ motor\ control_{ij} + \beta_3 Postural\ control_{ij} \\ + \beta_4 Baseline_{ij} + \beta_5 Stratum_j + u_j + e_{ij} \end{aligned}$$

β_2 and β_3 represent the effects of each mediator on the primary outcome (i.e., paths *b1* and *b2*). β_1 represents the direct effect of The OTTO Club on the primary outcome, with both predictors accounted for (i.e., *path c'*).

Drawing on estimates from the two models above, we will calculate the ACME for each mediator as:

$$ACME_1 = path\ a_1 * path\ b_1$$

$$ACME_2 = path\ a_2 * path\ b_2$$

and the total ACME as:

$$ACME_{total} = ACME_1 + ACME_2$$

The total effect of the programme on the primary outcome (direct and via mediators) will be calculated as:

$$Total\ effect\ (path\ c) = Direct\ effect\ (path\ c') + ACME_{total}$$

Finally, the proportion mediated effect will be calculated as:

$$Proportion\ mediated = \frac{ACME_{total}}{Total\ effect}$$

For all steps, we will present unstandardised model coefficients, p values and 95% confidence intervals. The primary effect size interpreted in the mediation analysis will be the proportion mediated effect and its confidence interval[16].

Sensitivity analysis

To test for any systematic differences in outcomes that might have resulted from researchers carrying out endline testing in some schools instead of OTs (see Data collection), a subset of Alphabet Writing scripts will be independently rescored by qualified OTs. Scripts completed by pupils assessed by trained researchers will be scanned and pseudo-anonymised, with all identifying information removed before scoring. The pseudo-anonymised scripts will then be randomly allocated to OTs, who will score them independently using the same standardised guidance applied in the main testing period. OTs will be blinded to the original researcher scores, school allocation and pupil identity. Researcher scores used in the primary analysis will remain unchanged regardless of agreement with OT scores.

Inter-rater reliability between researcher and OT scores will be assessed using a two-way random-effects intraclass correlation coefficient (ICC) for absolute agreement, based on single measures. This approach estimates extent to which scores assigned by different rates agree in absolute terms. To explore potential systematic differences in scoring, mean differences between research and OT scores will also be examined. Agreement between scoring methods will also be explored visually using Bland–Altman plots.

Furthermore, we will conduct a sensitivity analysis for the primary outcome, in which we include an indicator of whether the school was tested by an OT or a researcher as a covariate in the primary outcome model. By comparing findings from this sensitivity analysis to findings from the main analysis, we will be able to discuss any measurement bias arising from these logistical constraints in endline testing.

Furthermore, for logistical reasons, the baseline testing period was extended for one week past some settings' intervention start. This means that for a small number of schools, baseline measures might have been taken in the first week of intervention delivery, where children would have not yet experienced substantial handwriting practice and the intervention would therefore not be expected to have produced measurable improvements in children's handwriting outcomes. To ensure that this does not introduce bias into our analysis, we will conduct a sensitivity analysis, in which we re-estimate the primary outcome model for a sample consisting only of schools whose baseline measures were taken before the start of the intervention, and compare and discuss results.

Follow-up analyses

To estimate the impact of The OTTO Club on pupils' end-of-year writing attainment as a follow-up analysis at the end of Year 1 (RQ7), we will estimate **a two-level logistic regression model** to account for pupils (level 1) being clustered in schools (level 2). The dependent variable will be the binary outcome representing end-of-year teacher assessments of whether each pupil is working towards the expected standard (0) in writing or working at the expected standard / at greater depth (1). The model will include a binary indicator for treatment allocation, baseline accuracy

[16] For the mediation analysis, this will be obtained using quasi-Bayesian estimation in the R package mediation.

of letter formation and stratification variables (school-level handwriting time and EIA status) as covariates, and a random intercept by school. The basic form of the models is:

$$logit\left[P\left(Y_{ij}=1\right)\right]=\beta_0+\beta_1 Intervention_j+\ \beta_2 Baseline_{ij}+\ \beta_3 Stratum_j+u_j$$

Where pupils i are clustered within schools j. $Y_{ij}=1$ is the outcome of achieving/exceeding the expected standard in writing. β_0 is the intercept and β_1 is the log-odds coefficient for the treatment effect, representing the change in log-odds of achieving the expected standard in writing for pupils receiving The OTTO Club compared to TAU. u_j is a school-level random effect.

To help interpretability, the log-odds coefficient from the model will be translated into a marginal effect, representing the change in probability of achieving the expected standard in writing at the end of Year 1 that is associated with receiving The OTTO Club.

Imbalance at baseline

To assess imbalance between The OTTO Club and TAU groups at baseline, we will conduct cross-tabulations at the school and pupil level. Balance will be assessed i) for the sample as randomised (to assess whether randomisation was successful at balancing characteristics) and ii) for the sample analysed (to assess whether attrition has introduced imbalance), following EEF guidance (2022).

At the school level, characteristics assessed for imbalance will be those used as stratifiers at randomisation – school-level handwriting time and EIA status. At the pupil level, we will assess balance for the covariates included in the primary and secondary outcome models – baseline scores for pupils' letter of accuracy formation, ability to form letters through the correct process, fine motor control, postural control, and confidence and motivation to practise handwriting – as well as FSM eligibility as a covariate in the subgroup analysis.

For continuous variables (baseline scores), we will report means and standard deviations. For categorical variables (school-level handwriting time, EIA status, and EYPP eligibility), we will report counts and percentages in each category.

We will report standardised mean differences (SMDs) for all covariates and translate differences as Hedges' g effect sizes for continuous covariates. An SMD and/or effect size greater than 0.05 will be considered as an indication of possible imbalance. Where imbalances are detected, a sensitivity analysis will be estimated for the primary outcome with unbalanced variables included as predictors in the main model.

Missing data

We will follow EEF analysis guidance to address missing data (EEF, 2022). Multiple imputation will be considered, depending on the extent (e.g., if above 5%) and patterns of missingness.

We will report the extent of missing data for all outcomes and covariates across models, and the number of complete cases in the sample. To explore patterns of missing data, we will estimate a two-level logistic regression model, regressing the presence of missing data on variables in the data that might be predictive of missingness. The model will include school-level characteristics (treatment allocation, school-level handwriting time and EIA status) and child-level characteristics (FSM eligibility) as predictors, and a random intercept by school:

$$logit\left[P\left(Missing\ data_{ij} = 1\right)\right] = \beta_0 + \beta_1 Intervention_j + \beta_2 School\ handwriting\ time_j + \beta_3 EIA\ status_j + \beta_4 FSM\ status_{ij} + u_j + e_{ij}$$

If data is missing in a way that is not correlated with either observables and/or unobservable variables, missing observations are considered missing completely at random (MCAR) and complete case analysis should yield unbiased (though less precise) estimates. If data is missing in a way that is correlated with observable variables, then the missing observations are missing at random (MAR). If only the outcome variable in a substantive model is MAR conditional on covariates, these covariates will be included in the model for analysis. If a covariate in a substantive model is MAR conditional on other covariates, analysis will be done after multiple imputation (MI). If the reason for missing data depends on an unobserved variable, even after considering all the information in the observed variables, then the missing observations are missing not at random (MNAR). If data is suspected to be MNAR we will follow suggestions for employing sensitivity analyses as per EEF analysis guidance (2022).

If MI is performed, we will describe the variables used for imputation, the number of imputations performed, and the results of any sensitivity analyses to test assumptions about missing data.

Compliance

Compliance with the intervention will be measured at the school-level[17] as a binary variable (1, 0). For a school to count as compliant with the intervention, the teacher and/or TA of the class taking part in the evaluation must meet all of the following criteria:

i) attended the online training provided by the delivery team,
ii) attended at least one of the two monthly support sessions,
iii) completed at least the first nine out of the ten weekly lessons with their class, and
iv) complete at least three out of the five daily practice sessions with their class, over at least three different days of the week, including the Wacky Words practice session as mandatory each week.

Training and support session attendance data will be collected from the delivery team, and data on the delivery of weekly lessons and daily practice sessions will be collected from teachers, using a delivery log designed by NatCen.

To account for noncompliance (i.e., schools not complying with their intended treatment allocation) and isolate the impact of The OTTO Club for schools that do adhere to their random assignment, we will conduct a Complier Average Causal Effect (CACE) analysis. While compliance is measured at the school level, the unit of the following analysis will be pupils.

Following EEF guidance (2022), we will adopt an instrumental variable (IV) approach (Angrist & Imbens, 1995), using Two Stage Least Squares (2SLS) regression with random treatment allocation as the IV. As the first step of the 2SLS approach, we will regress observed school-level compliance on random treatment allocation as the IV, alongside stratification variables (school-level handwriting time and EIA status) as school-level covariates:

[17] As only one Year 1 teacher/class per school will take part in evaluation activities, this is the same as the teacher/class level for this trial.

$$Compliance_j = \beta_0 + \beta_1 Intervention_j + \beta_2 Stratum_j + e_j$$

Where $Compliance_j$ is the outcome of a school complying with the intervention, and β_1 is the change in log-odds of compliance for schools that is associated with being allocated to the treatment group[18]. e_j is residual variation in school compliance that is not explained by treatment allocation. From this regression model, we derive a predicted probability of compliance for each school that isolates treatment allocation from, and accounts for, unobserved characteristics influencing compliance.

As the second step of the 2SLS approach, this predicted school-level compliance will be included as a predictor in a linear mixed-effects regression model, regressing the primary outcome at endline on predicted compliance instead of treatment allocation. As in the primary outcome analysis, the model will include all covariates included in the main model – pupils' number of correctly formed letters at baseline[19] and stratification variables (school-level handwriting time and EIA status) as school-level covariates and a random intercept by school:

$$Number\ of\ correctly\ formed\ letters_{ij} = \beta_0 + \beta_1 \widehat{Compliance}_j + \beta_2 Baseline_{ij} + \beta_3 Stratum_j + u_j + e_{ij}$$

In this model, the coefficient for predicted compliance, β_1 , represents the CACE - the average causal effect of treatment among complier schools. This effect accounts for noncompliance and isolates the impact of The OTTO Club for schools that would adhere to their random assignment. If there are no confounding factors affecting both compliance and pupils' number of correctly formed letters as the primary outcome, the CACE would equal the ITT estimate of the treatment effect.

Intra-cluster correlations (ICCs)

The ICCs will be estimated for the primary outcome at endline, using variance estimates from i) the primary outcome model and ii) an empty model with no covariates. The ICC for schools ρ_S will be estimated using the following formula (Hedges, 2007):

$$\rho_S = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2}$$

Where σ_B^2 is the variance in outcome between schools (the variance of u_j; σ_u^2), σ_W^2 is the variance within schools (the variance of e_{ij}; σ_e^2) and σ_T^2 is the total variance.

Effect size calculation

Following EEF analysis guidance (2022), we will calculate effect sizes (ES) using the following formula for cluster-randomised trials, as adapted from Hedges (2007):

$$ES = \frac{(\overline{Y_T} - \overline{Y_C})_{adjusted}}{\sqrt{\sigma_u^2 + \sigma_e^2}}$$

[18] In line with EEF guidance, we will report the results from the first stage of the 2SLS alongside with i) the correlation between the IV and the endogenous variable; and ii) their associated F test.

[19] As in the first stage of the 2SLS analysis, compliance is defined and estimated at the cluster level, pupil-level baseline scores are not included in the model, and are only included in the second stage pupil-level model.

Where $(\overline{Y_T} - \overline{Y_C})_{adjusted}$ is the difference in means between The OTTO Club and TAU groups, adjusting for stratification variables and any baseline characteristics as specified in each respective outcome model. This adjusted mean difference from each model takes into account clustering, as estimated two-level models include random intercepts for school.

$\sqrt{\sigma_u^2 + \sigma_e^2}$ is the unconditional (i.e., unadjusted) total standard deviation (i.e., a weighted average of estimates for the two groups), derived from a two-level "null" or "empty" model with no predictors, where σ_u^2 is the variance of the school-level random intercept and σ_e^2 is the variance of residuals. As per EEF guidance[20], this SD accounts for clustering by explicitly separating variance components for each level.

We will also report 95% CIs for the ES.

For the binary follow-up outcome of end-of-year writing attainment, we will translate log-odds coefficients from the logistic regression model into marginal effects (ME), representing the change in probability of achieving the outcome that is associated with receiving the The OTTO Club programme. ME will be calculated as:

$$ME = p(1-p) * \beta_1$$

Where p is the predicted probability of achieving the outcome for The OTTO Club group, and β_1 is the log-odds coefficient from the logistic regression model estimated for the follow-up analysis.

ES will be estimated in Stata 17.

[20] See 8) Account for multi-site trial (MST) considerations. (pp. 9-11), in the EEF Statistical Analysis Guidance (2022). Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1699621596

References

Angrist, J. D., & Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association, 90*(430), 431–442. https://doi.org/10.1080/01621459.1995.10476535.

Barnett, A., Henderson, S. E., & Scheib, B. (2024*). The Detailed Assessment of the Speed of Handwriting, Second Edition (DASH-2)*. Pearson.

Basharat, M., Taylor, I., Duysak, E., Kuo, T. (2023). Stop and Think: Learning Counterintuitive Concepts Statistical Analysis Plan. Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/EEF-SAP-Stop-andThink_v1.pdf?v=1740396636

Bury, J., Marshall, L., Read, H., Roberts, E., Fletcher, A., & Scandone, B. (2022). Hanen Learning Language and Loving It (LLLI) Evaluation Report. Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/HanenLLLI_report_finalised.pdf?v=1740408226

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67. https://doi.org/10.1080/19345747.2012.673143.

Education Endowment Foundation (2022). Statistical Analysis Guidance for EEF Evaluations. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1679395501.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4). 341 – 370. https://doi.org/10.3102/1076998606298043.

Henderson, S. E., & Barnett, A. (2023). *Movement Assessment Battery for Children | Third Edition.* Pearson.

Takala, H., Leonard, H., Duysak, E., Rennick, A., Stoilova, E., Wadsworth, K. (2025). *Evaluation of The OTTO Club: a two-arm randomised controlled trial. Evaluation Protocol.* Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/the_otto_club_-_evaluation_-_protocol_-_v.1.0.0.pdf?v=1760635911.

Wilson, B., Kaplan, B., Pollock, N., & Law, M. (2000). *Clinical Observation of Motor and Postural Skills: Administration and scoring manual - Second Edition.* Therapro, Inc.

Appendix A

Handwriting scale – revised item selection and subscales

Motivation

1. Do you feel good when you do handwriting?
2. Do you like doing handwriting in class?
3. Do you like doing handwriting at home?
4. Do you look forward to handwriting?

Confidence

5. Do the other children in your class have better handwriting than you?
6. Do you make lots of mistakes in handwriting?
7. Are you good at writing letters and words?