

Project Title: Effects of EEF interventions on protection from COVID-19 learning loss
Statistical Analysis Plan

Evaluator (institution): University of Oxford

Principal investigator(s): Dr Ariel Lindorff



Template last updated: August 2019

PROJECT TITLE	<i>Effects of EEF interventions on protection from COVID-19 learning loss: A secondary analysis of data from the EEF Data Archive</i>
EVALUATOR (INSTITUTION)	University of Oxford
PRINCIPAL INVESTIGATOR(S)	Dr Ariel Lindorff
PROTOCOL AUTHOR(S)	Dr Ariel Lindorff, Professor Lars Malmberg, Youchuan Ma, Nicole Sin Hang Law, Julian Gardiner
TRIAL DESIGN	17 included trials, some variation in trial design (see Appendix A for details of each included trial)
TRIAL TYPE	Multiple trials included; some efficacy, some effectiveness (see Appendix A)
PUPIL AGE RANGE AND KEY STAGE	Various; in original trials, age 3-11, Early Years through KS2; in longitudinal outcomes from NPD data, ages 6-16, KS1, KS2, KS4
NUMBER OF SCHOOLS	Varies; analysis includes up to 17 different trials (see Appendix B); across trials up to 1,576 schools though in practice some will overlap <i>Overlap will be noted in the report; not possible to calculate until data are accessed</i>
NUMBER OF PUPILS	Varies; analysis includes up to 17 different trials (see appendix B); across trials up to 38,318 though in practice some may overlap <i>Overlap will be noted in the report; not possible to calculate until data are accessed</i>
PRIMARY OUTCOME MEASURE AND SOURCE	Short-term primary outcomes as in original trials (varies across Maths, Reading, Writing or a combination of these; early literacy for Early Years trials) Long-term primary outcomes: Various including GCSE subjects corresponding to short-term outcome subject areas in each trial (e.g. Maths, English); KS2 Reading or Maths, KS1 English Reading/Writing.
SECONDARY OUTCOME MEASURE AND SOURCE	Pupil attendance

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]	18/06/2024	N/A

Table of contents

SAP version history	1
Introduction.....	2
Trial inclusion criteria	4
Power calculations	4
Analysis	4
Primary outcome analysis.....	4
Secondary outcome analysis	8
Subgroup analyses	9
Correcting for multiple inferences	10
Missing data	10
Compliance	11
Intra-cluster correlations (ICCs).....	11
Limitations/complications to inference	11
References	13
Appendix A	15
Appendix B	16

Introduction

This secondary analysis project investigates whether and to what extent EEF interventions delivered in the years prior to the COVID-19 pandemic (2015/16, 2016/17, 2017/18, 2018/19) as well as those delivered – even if disrupted – during the pandemic (2019/20, 2020/21) had protective effects against learning loss over the period of COVID-19 disruption to normal school operations, teaching and learning. The analyses conducted as part of this project aim to answer the following questions:

- Research Question (RQ) 1: What are the longitudinal impacts of different Early Years and Primary Schools EEF interventions delivered in pre-pandemic (2015/16-2018/19) and pandemic (2019/20-2020/21) years on the educational attainment and attendance of all pupils and pupils eligible for Free School Meals (FSM)?
 - RQ 1A: To what extent, if any, did these EEF interventions have protective effects against learning loss over the time period in which COVID-19 disrupted teaching and learning?
- RQ 2: What are the longitudinal impacts of different broad categories of Early Years and Primary EEF interventions delivered in pre-pandemic years on the educational attainment and attendance of all pupils and pupils eligible for Free School Meals (FSM)?

- RQ 3: How does the long-term impact of individual EEF interventions relate to their short-term impact as measured in original EEF trials?

It is worth noting that the included trials vary in terms of a wide range of features in their design and delivery. They may not be directly comparable where their primary outcomes or theories of change differ substantially. As such, we look individually at each trial to assess its longitudinal (RQ1) and protective effects (RQ1A), as well as looking at broader categories by phase and primary outcome to understand any patterns according to these groupings (RQ2).

Appendix A provides details about the included trials. The majority were two-arm cluster-randomised controlled trials with random allocation at the school or setting level, but there was some variation (four trials were randomly allocated at the pupil or family level). Full inclusion criteria are detailed in the following section.

The concept of “learning loss” over the period of COVID-19 disruption is a challenging one to operationalise, and doing so necessitates a choice of fairly strong assumptions, particularly because there is no comparison group available from the same time period made up of pupils who did not experience COVID-19 disruption. As part of this project, the research team conducted a comprehensive review of approaches taken to conceptualising and measuring COVID-19-related learning loss as well as fade-out of intervention effects over time, and the below analytical strategy was designed taking the learnings from this review process into account and weighing the strengths and limitations of different approaches. We recognise that the approach taken here is not without limitations in its own right, and so we will endeavour to be transparent here and in later project outputs about the assumptions adopted for analytical purposes, and the limitations and caveats of our approach.

Broadly, that approach begins by investigating longitudinal impact of the included EEF interventions using the first available post-COVID-19-disruption statutory assessment outcomes in the NPD for the pupils who participated in the included trials. In some cases these assessment outcomes are quite distal from the original trials, but in the absence of other longitudinal outcomes they are the best information available to understand long-term impact of these interventions. This provides some initial intuition about any possible differences in the progress of pupils in intervention and control groups. However, it does not allow us to untangle long-term effects in general, which may have manifold and complex drivers, from COVID-19 learning loss and protective effects against this in particular.

To better hone in on COVID-19 learning loss, our strategy involves computing “expected” longitudinal educational outcomes for pupils who participated in the included EEF interventions, using a model to predict these based on data from pre-COVID cohorts of pupils who took the equivalent national assessments (KS1, KS2 or GCSE/KS4) in 2017/18 or 2018/19. We then operationalise “COVID-19 learning loss” as the difference between actual and expected outcomes for each pupil who participated in the relevant EEF interventions, and use this as a key outcome variable in subsequent analyses. This learning loss measures assumes reasonable similarity across cohorts in how they would have progressed between key stage

assessments (e.g. KS2 to KS4), which is a strong assumptions that we cannot test directly and consequently an important limitation of our analysis. Thus, to some extent the results must be interpreted with caution. Nonetheless, our judgment was that this approach provides the best approximation of COVID-19 learning loss that could be operationalised using the available data.

Trial inclusion criteria

The first inclusion criterion concerned relevance to the research questions. That is, included trials must have occurred before or during COVID-19. Further, for the pupil age group participating in a given trial, there had to be an available longitudinal outcome in the NPD from either 2021/22 or 2022/23. Longitudinal outcomes considered included key stage 1 (KS1), key stage 2 (KS2) and key stage 4 (KS4) attainment in the subject area(s) corresponding most closely to each trial's primary outcome measure. Attainment beyond KS4 was not considered, as omission of pupils not taking A-levels would bias results.

Trials were not explicitly excluded based on primary outcome measure, but it is worth noting that all included trials ended up having outcomes of either literacy (including literacy and language, reading, and/or writing), mathematics, or both. Efficacy and effectiveness studies were included, but pilot studies were excluded. Studies with a waitlist design were not automatically excluded, but evaluation reports were examined to extract information regarding the nature of any waitlist or similar design. Based on this examination, no studies included a waitlist design that would have directly affected the pupils that participated in the trial, where typically schools in the control group were given an opportunity to implement a given intervention with the following year's cohort of the same age.

Power calculations

Post-hoc power calculations will be performed using actual sample sizes once the data are available. Metadata available through evaluation reports and the EEF Data Archive catalogue at the time of writing did not include sufficient detail to be able to complete these calculations in full.

Analysis

This project will involve several stages of analysis. Data cleaning and checking will be conducted in IBM SPSS, while the main analyses will be conducted in R largely using the *lme4* (Bates et al., 2015) and *metafor* (Viechtbauer, 2010) packages. As this project will use linked EEF Data Archive and National Pupil Database (NPD) pupil-level data, all analysis will be carried out within the Office for National Statistics (ONS) Secure Research Service (SRS) environment via Assured Organisational Connectivity (AOC), with all members of the research team holding up-to-date full Accredited Researcher status according to ONS requirements.

Primary outcome analysis

Primary outcomes for the original included individual trials can be found in Appendix B. These outcomes are quite diverse, so that as noted above we will first look at

each trial individually before endeavouring to synthesise results across broad groupings of trials through meta-analytic methods.

STAGE 1: LONG-TERM ANALYSIS FOR INDIVIDUAL TRIALS

The initial stage of analysis (answering the first research question above) will replicate the original analysis from each trial's evaluation with longitudinal outcomes as identified for each trial in Appendix B, including preserving the same clustering e.g. of pupils in schools and/or classes as in the original primary outcome analysis and including pupil records based on Intention to Treat (ITT) as per EEF's guidance on *Longitudinal analysis of EEF trials* (2019). Where the identified longitudinal outcomes are multiple (e.g. where a trial had both Maths and Reading as primary outcomes), separate models will be estimated for each longitudinal outcome of interest to allow for the possibility of differential long-term effects by subject area, correcting for multiple inference as appropriate (see below section on "Correcting for multiple inference" for details).

Where a trial's original evaluation deviated from the most recent *Statistical analysis guidance for EEF evaluations* (EEF, 2022a), the primary model estimated at this stage will be defined according to the most recent guidance, and a secondary model will replicate the original analysis as per EEF longitudinal analysis guidance (2019).

Information on pupil background characteristics and assessment outcomes will be obtained from the National Pupil Database (NPD), including prior attainment, gender, Free School Meals (FSM) eligibility (within the last 6 years; EVERFSM_6), Special Educational Needs (SEN) provision, and English as an Additional Language status (EAL; coded from LanguageGroupMajor where "Other than English" is coded as 1 and "English" is coded as 0, with "Unclassified" coded as missing). This information will be obtained both for the pupils participating in the included EEF trials and for the entire population of pupils at relevant key stages in the cohorts taking their KS1, KS2 or KS4 assessments in 2017/18 and 2018/19.

Given the aims of the project, data from pre-COVID-19 cohorts of pupils who sat the KS1, KS2 and GCSE/KS4 assessments in 2017/18 and 2018/19 will be used to calculate estimates of expected outcomes for pupils, taking their prior attainment into account (e.g. year 1 phonics screening check as prior attainment for a KS1 literacy longitudinal outcome, KS1 as prior attainment for a KS2 longitudinal outcome; KS2 as prior attainment for a KS4 longitudinal outcome). Equation 1 below provides the basis for estimating expected attainment:

$$Y_{EXPECTED} = \beta_0 + \beta_1 Pretest + e \text{ (Equation 1)}$$

where $Y_{EXPECTED}$ is the available K2 or K4 outcome, β_0 is the intercept, β_1 is the coefficient corresponding to prior attainment, and e is the error term. A multilevel model is not used here, but cluster-robust standard errors are implemented to account for the clustering of pupils within schools; this is because this model is used for the purpose of prediction in a different cohort. This model will be estimated for both the 2017/18 and 2018/19 data, for each relevant key stage (KS1, KS2, KS4)

and for literacy and mathematics outcomes (where “literacy” here encompasses subject foci including English at KS4, Reading and Grammar, Punctuation, and Spelling at KS1 and KS2). To avoid over-controlling, we do not include additional background characteristics as covariates here for two reasons: First, the potential influence of individual pupil background factors is already picked up implicitly in the pre-test measure at least to some extent, as many of these background factors are time invariant and will already have influenced the pre-test outcome. Second, background characteristics associated with policy-relevant achievement gaps are explicitly investigated at a later stage of analysis, where controlling for them here might unhelpfully complicate subsequent substantive interpretation. The 2017/18 and 2018/19 results will be compared to assess stability across pre-COVID assessment years before determining which or what combination of both to use as a predictive model to compute expected pupil scores. Following on from this, an expected score will be generated for each pupil in each trial included in this project, and an estimate of COVID-19 learning loss will be computed at the individual level by subtracting the actual from the expected scores for each pupil (learning loss = expected score – actual score). This learning loss “score” (Y_{LOSS}) becomes a key outcome for further analysis, to facilitate comparisons across the treatment and control groups for trials included in this study. Prior to the main analysis, we will conduct individual- and school-level descriptive analysis of the variation and range in learning loss scores by group (intervention, control) and subgroup (e.g. FSM, gender, EAL, SEN) as well as overall. Previous studies have reached mixed conclusions about the extent of learning loss in different age groups and in literacy and mathematics due to the COVID-19 pandemic, but there is some reasonably secure evidence that primary school pupils had somewhere between 2 to 3.5 months’ less progress in mathematics by Autumn-Spring 2021 than expected based on previous cohorts’ performance, and somewhere between 2 and 3 months’ less progress in reading, with the maths and reading attainment gaps between pupils with and without FSM widening by 0.5 to 1 month compared to pre-pandemic cohorts (EEF, 2022b). Learning loss was also found in some studies to be greater for pupils with EAL, pupils with SEN, and girls (Renaissance Learning & Education Policy Institute, 2022). Less is known about recovery from learning loss by the time point(s) at which this study looks at longitudinal outcomes of interventions, i.e. 2021/22 or 2022/23.

We recognise that this approach implicitly assumes some theoretical trajectory in common between cohorts of pupils before and during/after the COVID-19 pandemic, absent the pandemic itself. It is not possible to provide empirical evidence on this comparison given the lack of a “control” group not exposed to pandemic-related educational disruption. This assumption is a strong one, but allows us to make some sense of how individual pupils might have been expected to achieve, absent the effects of the pandemic, with the caveat that there may be unaccounted-for differences between the cohorts that we cannot test for in the absence of the effects of COVID-19 on the pupils who experienced the pandemic.

Next, following an approach similar to Ashraf et al. (2021a; 2021b), each trial is analysed using a two-level multilevel model (pupils at level 1 within schools at level 2) to account for the clustering of pupils in schools (or settings, in early years trials),

specified below and yielding a consistent measure of effect size on COVID-19 learning loss across trials,

$$Y_{LOSSij} = \beta_0 + \beta_1 Pretest_{ij} + \beta_2 Group_{ij} + u_j + e_{ij} \text{ (Equation 2)}$$

where Y_{LOSSij} is the learning loss score explained above for school j and pupil i , β_0 is the intercept, β_1 is the slope between pre- and post-tests scores, β_2 is the adjusted difference between intervention and control groups where $Group_{ijk}$ takes on a value of 1 for the intervention group and 0 for the control group, between-school variability is represented by $u_j \sim N(0, \sigma_u^2)$ and residual variance is represented as $e_{ij} \sim N(0, \sigma_e^2)$, with σ_u^2 representing the school-level variance and σ_e^2 the pupil-level variance.

Outcomes will not be z-score standardised in the initial analysis of longitudinal outcomes for each trial. Standardising scores to have a mean of 0 and variance of 1 would be problematic in this instance, as this would make comparisons or computations across different assessment timepoints difficult to interpret, where an individual pupil's standardised score is then conditioned on the mean at a given time point, and changes in the mean between time points complicate interpretations of change at the individual level. Given we will look at groups of trials using the same longitudinal outcome measures (learning loss in terms of KS1, KS2 or KS4 literacy or maths achievement) in the subsequent meta-analysis detailed below, using unstandardised outcomes should be sufficient at this stage.

STAGE 2: META-ANALYSIS ACROSS (GROUPS OF) TRIALS

The effect size for each trial (k), both for the longitudinal outcome and for the COVID-19 learning loss measure, will be calculated as in Equation 3.

$$ES_k = \frac{\beta_{2k}}{\sqrt{\sigma_{uk}^2 + \sigma_{ek}^2}} \text{ (Equation 3)}$$

with upper and lower confidence interval bounds estimated using the same equation as above but with the upper and lower bounds of the confidence interval for β_{2k} as the right-hand side numerators, respectively (Equations 4 and 5, below).

$$CI_{k_lower} = \frac{Lower(\beta_{2k})}{\sqrt{\sigma_{uk}^2 + \sigma_{ek}^2}} \text{ (Equation 4)}$$

$$CI_{k_upper} = \frac{Upper(\beta_{2k})}{\sqrt{\sigma_{uk}^2 + \sigma_{ek}^2}} \text{ (Equation 5)}$$

In order to compute a weighted average of effect sizes, still following Ashraf et al. (2021a), the standard errors of the effect sizes from each included trial are computed as in Equation 6.

$$SE_k = \frac{CI_{k_upper} - CI_{k_lower}}{3.92} \text{ (Equation 6)}$$

where 3.92 is the width of a 95% confidence interval in standard errors (Higgins, Li & Deeks, 2023). As SE is computed based on multilevel model results, this statistic inherently accounts for between-school variability.

We use a random-effects approach to meta-analysis, as this takes into consideration the heterogeneity (τ^2) between EEF trials included in the study (Higgins, Li & Deeks, 2023). We use the REML estimator of τ^2 in the *metafor* package in R following the advice of Veroniki et al. (2016). Using this and the estimated effect size for each trial, we compute a pooled effect size for similar outcomes (within literacy and within mathematics) according to Equation 7, below.

$$ES_{pooled} = \frac{\sum_{h=1}^K W_k ES_k}{\sum_{h=1}^K W_k} \text{ (Equation 7)}$$

within which the weight for each trial based on effect size variability and between-trial heterogeneity is $W_h = (SE_h^2 + \tau^2)^{-1}$ (Equation 8).

Ideally we would have taken an individual participant data (IPD) meta-analysis approach, but the number of trials possible to include in this study was too small, especially when grouping by outcome subject area, to allow for this approach in which *trial* would need to constitute a level of analysis. The number of trials is still small for a two-stage meta-analytic approach, but mitigates some of the trial-level sample-size issue that makes IPD meta-analysis unfeasible. No approach entirely resolves the challenge of comparing or synthesising results across a diverse set of trials, of course, but there is at least some basis for this comparison and synthesis where primary outcomes fall within clear groupings (e.g. literacy, mathematics).

STAGE 3: COMPARISON OF EFFECT SIZES ACROSS SHORT- AND LONG-TERM EFFECTS

To answer the last research question, standardised effect sizes will be directly compared (short-term, long-term) for each included trial and for groups of trials for which pooled estimates are feasible and interpretable (literacy and mathematics). For short-term outcomes, this will involve replication of the original analysis for each included trial (or re-analysis according to the latest statistical guidance; EEF, 2022a). There is some variation across trials in the original analysis, but the approach will be consistent within trial across the short- and long-term outcome analyses.

Secondary outcome analysis

The secondary outcome for all trials will be attendance, where this is considered to be a proxy for engagement and constitutes the only readily available non-cognitive measure relevant to the project that can be obtained from the NPD administrative data. Attendance in the NPD is available for each term (Autumn, Spring and Summer, or Terms 1 through 3) of a given year in terms of the number of absences and total sessions possible to attend, where the attendance outcome used in this project will be percent of possible sessions attended in a given year, computed via Equation 9, below:

$$Attendance_{Year} = \frac{\sum_{Term=1}^3 (Sessions\ possible_{Term} - Sessions\ absent_{Term})}{\sum_{Term=1}^3 Sessions\ possible_{Term}} * 100 \text{ (Equation 9)}$$

While the extant literature is limited on the longer-term effects of the COVID-19 pandemic on attendance, there is evidence that at least some effect on attendance rates persisted after the end of school lockdowns in England. Specifically, between 2017/18 and 2021/22, overall absences for pupils without FSM rose from 4.3 to 6 percent, and for pupils with FSM from 7.6 to 9.7 percent (Eyles, Lilywhite & Major, 2023). We might reasonably expect, therefore, at most some small effect of COVID-19 pandemic disruption on attendance, and an effect of at least some interventions in terms of relevant protective effects, particularly where interventions' theories of change include either some direct consideration for attendance or some emphasis on engagement.

All secondary outcome analysis will follow the same steps specified above for primary outcome analysis, and for the same groupings of trials (e.g. by age group and primary outcome subject area). Quality of attendance data during the period of COVID-19 lockdowns, and to some extent immediately after this period, may be somewhat limited due to constraints on school resources and impact of the pandemic on the usual DfE data collection from schools (particularly affecting 2019/20 and 2020/21), however, this will mostly affect our analysis of Early Years trials while attendance data for other trials should remain fairly reliable.

As an ancillary analysis, and to determine whether analyses need to be in some way adjusted to account for differential effects of the COVID-19 pandemic on schools, we will also take an initially-descriptive look at the “TermlySessionsCoronavirus” data from the NPD for Autumn 2020 through Summer 2022, which record how many sessions each pupil was absent for reasons specifically related to COVID-19, aggregated to school-level proportions by year. These data are only used tentatively in this analytic plan because their reliability and completeness are somewhat in question, where schools may not have reported this information consistently in the relevant terms and years.

Subgroup analyses

The main subgroup analysis of interest in this project concerns the group of pupils eligible for FSM, specifically as identified by the EVERFSM6 indicator (of whether a pupil has been eligible for FSM within the last 6 years).

Subgroup analysis will follow the same approach taken for primary and secondary analysis described above, but looking at effect sizes of trial by adding FSM*Group interaction terms to the models run in the main analysis. Further subgroup analyses may be undertaken separately for other groups (e.g. by gender, SEN, EAL) based on the evidence suggesting that the effects of COVID-19 on these groups were particularly pronounced (Renaissance Learning & Education Policy Institute, 2022). The effect size for FSM-eligible pupils will be calculated using the sum of the standardised coefficients of $Group_{ij}$ and of the interaction term ($EverFSM_{6ij} * Group_{ij}$) for each included trial, based on the model specified in Equation 10:

$$Y_{Lossij} = \beta_0 + \beta_1 Pretest_{ij} + \beta_2 Group_{ij} + \beta_3 EverFSM_{6ij} + \beta_4 (EverFSM_{6ij} * Group_{ij}) + u_j + e_{ij} \quad (\text{Equation 10})$$

Where the sample size allows, subgroup analysis – for each of these subgroups – will also be conducted by following the same analytic procedure as in the main analysis but on the subsample corresponding to the relevant subgroup (e.g. where EverFSM_6=1, SEN provision =1, EAL = 1, or Gender=1). This is likely to be possible for some studies but not others, where in some of the included studies at least some of these analyses would be underpowered.

Where possible to use both approaches, the resulting effect sizes for subgroups (e.g. FSM) should be similar; results from both approaches will be compared and reported where appropriate.

Correcting for multiple inferences

The *crctStepdown* package (Watson, 2024) in R will be used to facilitate the implementation of the Romano-Wolf correction, where appropriate, in order to control the family-wise error rate. This package allows for relatively speedy computation while also accommodating multilevel models run using the *lme4* package in R. Correction for multiple inference will only be undertaken where trials are being analysed independently; such adjustment is not necessary where inference is based on evidence pooled across trials (Brookes et al., 2001).

Missing data

Missing data patterns will be assessed across the including studies, and missing data treatment will be decided based on these patterns.

We will report descriptive information on each variable included in any analysis within the project both for individual trials and overall across any groups used for pooling (e.g. across studies with literacy and mathematics outcomes), and missingness will be reported within intervention and control groups as well as overall both for individual trials and for relevant groups of trials (with literacy and with mathematics outcomes). Reporting will take the form of counts and percentages of complete cases on each variable. We will also report total percentages of complete cases by trial for an analytic sample if listwise deletion were used, i.e. percent of data included in an analysis dropping any case with any missing information on an included variable, for primary, secondary and subgroup analysis.

Missingness patterns can be Missing Completely At Random (MCAR), where the probability of missingness is not associated with observed nor with unobserved variables; Missing At Random (MAR), where the probability of missingness is associated with observed variables; or Missing Not At Random (MNAR), where the probability of missingness is associated with unknown information (e.g. unobserved variables, unknown mechanisms). If data are MCAR, testable using Little's MCAR test, then missingness is random, and listwise deletion does not risk biasing results. If data are MAR, we might expect biased results if the proportion of data missing is substantial, and so must consider some missing data treatment to mitigate this bias. In our case, we will consider multiple imputation if the proportion of missing data on independent variables is greater than 5% within a given study or pooled analysis *and* where our initial assessment of missing data patterns suggests data are MAR. It is

not possible to test for whether data are MNAR or MAR, but an assessment of patterns in the missing data and consideration of plausible missingness mechanisms should suffice to reach a credible conclusion to this end. Multiple imputation will use all of the independent variables included in the main analysis and will use 20 imputations (van Buuren, 2018). No imputation will take place for outcome variables as this is generally not recommended. As is typical for multiple imputation, coefficients and standard errors will be pooled using Rubin's rules (Rubin, 1987). If multiple imputation is used, results will be compared to complete case analysis, and both will be reported with an explanation of the implications of any differences between the two. Multiple imputation will use the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011) in R.

Compliance

Depending on the quality and completeness of data on compliance in the EEF Data Archive for the trials included in this secondary analysis, compliance will be included as a covariate/factor in multi-trial models investigating whether and to what extent various intervention features are associated with any effects that are protective against COVID-19 learning loss. Given compliance may not be measured consistently across trials, this variable will likely have to be simplified to a binary indicator for insufficient compliance across a trial sample, based on a review by the research team of how compliance was defined and measured in each trial. An initial review of this information (noting trials in which this is unclear, missing, etc.) is included in Appendix B.

Intra-cluster correlations (ICCs)

ICCs will be calculated in the usual way for empty models and for each of the primary and secondary analysis models using the variance at a given level as a proportion of total variation in the model.

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \text{ (Equation 11)}$$

where σ_u^2 represents the variance at the school level for a given trial and σ_e^2 represents variance at the pupil level. ICC can be interpreted as the proportion of variation in the outcome that lies at the school level.

Limitations/complications to inference

Evaluating longitudinal outcomes across a number of trials introduces multiple complications to analysis as well as interpretation. One particular issue in educational interventions is that where there is an ethical commitment not to deprive the control group of an intervention likely to have a positive effect, a waitlist design (in which the control group receives the intervention or a version of it following trial completion) or some version of this design, is commonly used. As indicated in Appendix A, nine of the 17 included trials included some version of a waitlist design or at least the option for control schools to implement the intervention after the trial.

In cases where a waitlist or modified waitlist design was used, control schools did not implement the intervention with the same cohort of pupils targeted in the relevant trial, but in trials after which control schools had the option to implement if they wanted after the intervention, no data was available regarding whether and with which pupils they may have done so. We did not exclude trials where this information was unknown, but note this as a limitation of our inferences as it would be for any analysis of longitudinal outcomes of trials without information on subsequent engagement with similar interventions in participating control settings.

The available data do not allow us to have any direct measure of differential COVID-19 disruption across schools, which is also an important limitation. It is likely that there was some considerable variation in disruption to learning between schools, not least because of differences in schools' responses to the pandemic, needs of their specific populations of pupils, and resourcing of alternative learning provision.

The distal nature of some of the longitudinal outcomes relied upon in this research also limits the strength of inferences; much can happen between KS2 and KS4. However, as noted above, we use the most proximal information available in the form of the next available statutory assessment following pandemic-related disruption, and we suggest that this offers some useful insight even if caveated according to the complications introduced by other unmeasured influences on learning between the original trials and longitudinal outcomes.

Other important caveats to the inferences that can be drawn from this project's analyses are linked to the necessity of adopting strong assumptions already noted above in this statistical analysis plan. That is, assuming that learning trajectories are similar across cohorts may be problematic, but difficult to impossible to directly assess for the cohorts of interest. We mitigate this to some extent by looking at trajectories in two pre-COVID cohorts to assess stability, but this does not guarantee that the strong assumption across other cohorts holds true.

Finally, we might expect different types of interventions to differ in their longitudinal effects according to outcomes, implementation features, etc. Although we do not put forward strong hypotheses about these aspects, our initial review of intervention features (Appendix B) affords us the opportunity to consider whether to conduct additional analyses that pool results (within outcome) according to different features if and when this is possible, i.e. where numbers of included trials are sufficient to compare across relevant categories (e.g. of type of implementer or intervention feature). Thus, it may be possible to examine this sort of variation in protective effects empirically, though initially we focus on comparisons across literacy and mathematics outcomes rather than intervention features on the principle that these are readily available within the data without importing our own constructed variables from the review and we can reasonably expect to have sufficient numbers of trials to include in pooled estimates by these outcome groupings.

References

- Ashraf, B., Singh, A., Uwimpuhwe, G., Coolen-Maturi, T., Einbeck, J., Higgins, S. & Kasim, A. (2021a). Individual participant data meta-analysis of the impact of EEF trials on the attainment of pupils eligible on Free School Meals: 2011-2019. Education Endowment Foundation & Durham University Research Methods Centre. Available at <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/meta-analysis-of-the-impact-of-eef-trials-on-fsm-pupils-2011-2019>
- Ashraf, B., Singh, A., Uwimpuhwe, G., Higgins, S. & Kasim, A. (2021b). Individual participant data meta-analysis of the impact of educational interventions on pupils eligible for Free School Meals. *British educational research journal*, 47(6), 1675-1699. <https://bera-journals.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/berj.3749>
- Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. [doi:10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01)
- Brookes, S.T., Whitley, E., Peters, T.J., Mulheran, P.A., Egger, M. & Davey Smith, G. (2001). Subgroup analysis in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. *Health Technology Assessment*, 5(33), 1–56. <https://doi.org/10.3310/hta5330>
- Education Endowment Foundation (2019). *Longitudinal analysis of EEF trials*. Available at https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/longitudinal_guidance.pdf?v=1718693265
- Education Endowment Foundation (2022a). *Statistical analysis guidance for EEF evaluations*. Available at <https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1718693265>
- Education Endowment Foundation (2022b). *The impact of COVID-19 on learning: A review of the evidence*. Available at https://d2tic4wvo1iusb.cloudfront.net/production/documents/guidance-for-teachers/covid-19/Impact_of_Covid_on_Learning.pdf?v=1720527233
- Eyles, A., Lilywhite, E. & Major, L.E. (2023). The rising tide of school absences in the post-pandemic era. *London School of Economics and Political Science Blog*. Accessed 09/07/2024 at https://blogs.lse.ac.uk/politicsandpolicy/the-rising-tide-of-school-absences-in-the-post-pandemic-era/#_ftn1
- Higgins, J. P. T., Li, T. & Deeks, J. J. (2023). Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (Eds). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4 (updated August 2023). Available from www.training.cochrane.org/handbook.
- Renaissance Learning & Education Policy Institute (2022). *Understanding progress in the 2020/21 academic year: Extension report covering the first half of the autumn term 2021/22*. DfE.
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd Ed.). Chapman and Hall.
- van Buuren S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. [doi:10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)
- Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. <https://doi.org/10.18637/jss.v036.i03>
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J, Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D. & Salanti, G. (2016). Methods to estimate the between-study variance and

its uncertainty in meta-analysis. *Research synthesis methods*, 7(1), 55-79.
<https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1164>

Watson, S. (2024). crctStepdown: Univariate analysis of cluster trials with multiple outcomes. R package version 0.5.2, <https://CRAN.R-project.org/package=crctStepdown>

Rubin, D.B. (1987). *Multiple imputation for non-response in surveys*. Wiley & Sons.

Appendix A

EEF Project Name	In archive?	EEF report link	Type of trial	Randomisation?	Start of delivery	Pre-test academic year	Post-test academic year	School phase	Age group(s)	Year groups(s)	Primary outcome	Impact - months progress	COVID-19 affected?	Longitudinal outcome	Long. Outcome year	Assessment feasible? Y/N	Notes	Waitlist?
ScratchMaths	Y	https	Efficacy	School	2015/16	2014/15	2016/17	Primary	9 to 11	5 to 6 (sa	Maths	0 N		GCSE 2021/	Y		Yes, k	
FRIENDS	Y	https	Efficacy	School	2015/16	2015/16	2016/17	Primary	9 to 10	5	Combined Ma	0 N		GCSE 2021/	Y		Optio	
Changing Mindsets (re-grant)	Y	https	Effectiveness	School	2016/17	2012/13	2016/17	Primary	10 to 11	6	Maths, readi	0 N		GCSE 2021/	Y		Yes, k	
Grammar for Writing (re-grant)	Y	https	Effectiveness	School	2016/17	2012/13	2016/17	Primary	10 to 11	6	Writing	0 N		GCSE 2021/	Y		Optio	
Accelerated Reader (re-grant)	Y	https	Effectiveness	School	2016/17	2014/15 a	2017/18 a	Primary	8 to 10	4 and 5	Reading	0 N		GCSE 2022/	Y	Longi	Yes, k	
Catch Up Literacy (re-grant)	Y	https	Effectiveness	School	2016/17	2013/14 a	2016/17	Primary	8 to 10	4 and 5	Reading	0 N		GCSE 2022/	Y	Longi	No	
Integrating English	Y	https	Efficacy	School	2016/17	2016/17	2017/18	Primary	9 to 11 (s	5 and 6 (	Writing	-1 N		GCSE 2022/	Y		No	
Reciprocal Reading	Y	https	Efficacy	School	2017/18	2016/17	2017/18	Primary	8 to 9 for	4 (univer	Reading	0 N		GCSE 2022/	Y	Target	No	
Maths Counts	Y	https	Efficacy	Pupil	2016/17	2014/15, 2	2016/17	Primary	7 to 11	3 to 6	Maths	2 N		GCSE 2022/	Y	Miss	No bu	
Catch Up Numeracy (re-grant)	Y	https	Effectiveness	School	2016/17	2015/16	2017/18	Primary	7 to 10	3 to 5	Maths	0 N		GCSE 2022/	Y	Only	No bu	
Learning Language and Loving It	Y	https	Efficacy	School	2019/20	2019/20	N/A	Early years	3 to 4	N	Literacy	N/A Y		KS1 R 2022/	Y	Only	No	
The REAL Programme	Y	https	Efficacy	Family	2019/20	2019/20	2020/21	Early years-Pr	3 to 5	N to R	Literacy	0 Y		KS1 R 2022/	Y	Impa	No	
Family Skills	Y	https	Efficacy	School	2016/17	2016/17	2016/17	Primary	4 to 5	R	Literacy	0 N		KS2 R 2022/	Y		No	
onebillion: app-based maths le	Y	https	Efficacy	School	2017/18	2017/18	2017/18	Primary	5 to 6	1	Maths	3 N		KS2 R 2022/	Y		Yes, k	
Lexia Reading Core5®	Y	https	Efficacy	Pupil	2018/19	2017/18	2018/19	Primary	6 to 7	2	Reading	1 N		KS2 R 2022/	Y		Optio	
Speech Bubbles	Y	https	Efficacy	Pupil	2018/19	2017/18	2018/19	Primary	5 to 7	1 and 2	Reading	-1 N		KS2 R 2022/	Y	Only	No	
REACH Primary	Y	https	Efficacy	School	2019/20	2018/19	2020/21	Primary	7 to 8	3	Reading	-2 Y		KS2 R 2022/	Y	assu	No	

Please note: We have included the above table as an Excel worksheet object to preserve full details recorded in the process of trial selection; these details will not be fully accessible if this document is printed in hardcopy.

Appendix B

[illegible]

Please note: We have included the above table as an Excel worksheet object to preserve full details recorded in the process of trial selection; these details will not be fully accessible if this document is printed in hardcopy.