

Archive Research Study Plan

WhatWorked Education

Dr Emma Dobson

Updated September 2026

Project title	Assessing the Feasibility of Component Network Meta-Analysis Using the Education Endowment Foundation Trial Data
Researcher (institution)	WhatWorked Education
Principal investigator(s)	Dr Emma Dobson
Study plan author(s)	Dr Emma Dobson, Dr Akansha Singh, Rahil Khowaja, Germaine Uwimpuhwe, Professor Steve Higgins
Study design	Methodological archive research study using secondary analysis of EEF evaluation data, incorporating multilevel meta-analysis and prototype Component Network Meta-Analysis (cNMA).
Pupil age range and Key stage	Primary and Secondary School (approximately KS1–KS4; age range determined by eligible EEF trials included within the archive).
Number of interventions included	Approximately 120+ EEF-funded interventions (final number subject to eligibility screening and feasibility criteria).
Primary outcome measure and source	Primary attainment outcomes (Reading, Mathematics and Science) reported within eligible EEF impact evaluations and archived EEF datasets, harmonised onto a common standardised mean difference (Hedges' g) metric.
Secondary outcome measure and source	Where available: additional attainment outcomes, subgroup estimates (including disadvantage-related outcomes such as FSM eligibility or low prior attainment), intervention component coding using the EEF Core Components Taxonomy, trial metadata and archived evaluation reports.

Executive Summary

Purpose of the Study

Understanding which educational programmes improve attainment is only the first step towards improving educational practice. Increasingly, policymakers, schools and research funders require evidence that explains what works, for whom, under what circumstances, and why. While the Education Endowment Foundation (EEF) has built one of the world's largest repositories of randomised controlled trials in education, most evidence synthesis has necessarily focused on interventions as complete programmes. Consequently, programmes are typically compared as whole packages, providing limited insight into the individual components that contribute to their effectiveness. This study seeks to determine whether the EEF Data Archive can support a new generation of evidence synthesis that moves beyond programme-level evaluation, towards understanding the contribution of individual intervention components. Specifically, the study will assess the feasibility of applying Component Network Meta-Analysis (cNMA) to the archive by combining intervention component coding with harmonised attainment outcomes from EEF-funded evaluations. For this feasibility assessment, the required information consists primarily of project-level data, such as effect sizes, their variances, and relevant study and intervention characteristics. These data can be extracted and organised using EPPI-Reviewer (Thomas et al., 2023), and are used in this study as a proxy for the project-level data available within the EEF Data Archive.

This is a methodological feasibility study. The principal objective is not to estimate definitive causal effects for individual components, but to determine whether the existing archive possesses sufficient data quality, completeness, variation and connectedness to support robust component-level synthesis. Where feasibility criteria are met, the study will develop a prototype cNMA capable of estimating indicative associations between intervention components and pupil attainment outcomes. Where criteria are not met, the study will identify the specific methodological and data limitations preventing implementation and provide evidence-informed recommendations for improving future commissioning, evaluation reporting and archive development. The project therefore has two complementary aims. First, it will establish whether cNMA is currently feasible using the EEF archive. Second, irrespective of analytical feasibility, it will generate practical recommendations that strengthen the future value of the archive as a strategic evidence resource for understanding the active ingredients of effective EEF-funded programmes.

Study Workflow

- Stage 1 establishes access to the project level data extracted using EPPI-Reviewer as a proxy for the information that would be obtained from the EEF Data archive and determines whether sufficient component coding and outcome data exist to support further analysis.
- Stage 2 harmonises intervention components and effect sizes, constructs the analysis dataset, develops the component dictionary and freezes the final analysis sample following formal quality assurance and governance review.

- Stage 3 undertakes a progressive statistical assessment of the evidence, with each stage addressing a distinct question about the feasibility of component-level synthesis. Initially, multilevel meta-analysis will be used to assess the overall structure and heterogeneity to provide a suitable basis for component/dimension-level analysis. Component-level models then assess whether individual components or broader component dimensions are associated with variation in intervention effects. Where sufficient variation and connectivity exist, network diagnostics assess whether components can be meaningfully compared across the evidence base. Bayesian estimation will be then used to quantify uncertainty in component effects
- Stage 4 translates methodological and feasibility findings into practical recommendations for EEF commissioning, evaluation design, reporting guidance and archive development. In the scenario where a full cNMA is not feasible, the recommendations will also identify appropriate alternative analytical pathways based on the specific data and network limitations encountered.
- Stage 5 produces the final technical report, executive briefing documents, technical appendix, reproducible statistical code and recommendations for future methodological development.

Decision Framework

Progression through this workflow is conditional upon achieving predefined methodological standards (Figure 1). At multiple stages the project may legitimately conclude that additional data, alternative analytical strategies or future archive development are required before component-level synthesis can be undertaken reliably. Four decision gates will be used at key points in the study:

1. The first decision gate evaluates archive readiness, assessing whether sufficient component coding completeness, primary outcome availability and precision information exist to support analysis. Where necessary, additional evaluation reports will be sourced or Secure Research Service access considered before determining the final analysis sample.
2. The second decision gate follows component harmonisation and dataset preparation. Here the study evaluates whether rare components should be collapsed, whether conceptual overlap requires reclassification and whether the frozen dataset satisfies all predefined quality assurance criteria before modelling commences.
3. The third decision gate assesses statistical feasibility following preliminary meta-analysis and network diagnostics. Only where adequate network connectedness, sufficient component variation and acceptable model diagnostics are demonstrated will the study proceed to proof-of-concept Component Network Meta-Analysis. The purpose is both to evaluate feasibility and, where the evidence permits, generate substantive component-level findings. Alternative analytical pathways, including restricted network analyses or conventional meta-analysis, remain available where assumptions are not satisfied.

4. The final decision gate governs interpretation. Findings will only be translated into substantive recommendations where estimates demonstrate acceptable robustness, stability and interpretability across sensitivity analyses. Where uncertainty remains substantial, conclusions will focus on methodological feasibility and archive development rather than component effectiveness.

If a “NO” decision is reached at any decision gate, the project will not proceed to the subsequent analytical stage. Instead, the reasons for not progressing will be documented and the project will move directly to Stages 4 and 5: translation of the findings into EEF ways of working, followed by reporting and dissemination.

Expected Outputs

Regardless of whether prototype Component Network Meta-Analysis proves feasible, the project will deliver a comprehensive set of methodological and practical outputs. These include: a comprehensive assessment of the readiness of the EEF archive for component-level evidence synthesis; a harmonised, quality-assured analytical dataset linking intervention components with attainment outcomes; a documented Component Dictionary describing coding decisions, harmonisation procedures and any component aggregation undertaken; a prototype Component Network Meta-Analysis (where feasibility criteria are met); Bayesian summaries and exploratory analyses of component combinations¹, moderators and interaction effects (where supported by the available evidence); recommendations for improving future EEF commissioning, evaluator reporting standards and archive development; guidance on how component-based evidence synthesis could support future iterations of the Teaching and Learning Toolkit and wider evidence mobilisation activities; and a full technical report, executive briefing papers, technical appendix and reproducible statistical analysis scripts.

Collectively, these outputs will establish not only whether Component Network Meta-Analysis is currently feasible to analyse EEF funded trials, but also provide a practical roadmap for how educational evidence synthesis can evolve from comparing interventions as complete programmes towards understanding the core components that underpin effective teaching and learning.

¹ All possible combination of components and interactions will be considered and the most saturated model will contain approximately 10 studies per characteristic (component/interaction), in line with the [Cochrane handbook](#).

OVERVIEW OF THE STUDY WORKFLOW

Five Operational Stages and Four Decision Gates

Movement through the study is conditional on predefined methodological criteria rather than timelines.
This ensures analytical decisions are transparent, reproducible and evidence-led.



Figure 1. Study workflow for ‘Assessing the Feasibility of Component Network Meta-Analysis Using the Education Endowment Foundation trial data’

Background and Study Rationale

Policy Context

The EEF was established to improve educational outcomes for children and young people, particularly those experiencing socio-economic disadvantage, through the generation, synthesis and mobilisation of high-quality evidence. Since its establishment in 2011, EEF has commissioned one of the world's largest portfolios of independent educational evaluations, funding hundreds of efficacy and effectiveness trials across all phases of education (Education Endowment Foundation, 2025). Alongside primary research, EEF has invested in evidence synthesis, implementation guidance, methodological innovation and research infrastructure to ensure that evidence can be translated into meaningful improvements in educational practice. Central to EEF's mission is the principle that evidence should not only be scientifically rigorous but also useful for decision-making. Schools, policymakers and programme developers require evidence that supports decisions about which approaches to adopt, how they should be implemented and how interventions can be adapted responsibly for different educational contexts. Consequently, EEF's work extends beyond evaluating individual programmes to developing evidence resources that help practitioners understand what works, for whom, under what circumstances and why. Resources such as the *Teaching and Learning Toolkit*, thematic guidance reports and evidence reviews exemplify this commitment to transforming research findings into practical guidance for schools and educational settings (EEF, 2026).

As EEF's portfolio of completed evaluations has expanded, so too has the opportunity to generate new knowledge through secondary analysis of existing evidence. Rather than viewing each evaluation as an isolated contribution, EEF increasingly seeks to maximise the long-term value of its research investments by enabling cumulative learning across studies. This reflects a broader movement within education and other applied disciplines towards research infrastructures that support making the most effective use of existing evidence, methodological innovation and increasingly sophisticated approaches to evidence synthesis (Ha, 2023; Veroniki et al., 2026). At the same time, expectations of educational evidence have evolved. Earlier generations of evidence synthesis focused primarily on establishing whether programmes improved pupil outcomes on average (Ha, 2023; Pellegrini et al., 2025). While this remains a fundamental objective, there is growing recognition that educational decision-making also requires evidence about how interventions achieve impact, which features appear most influential and how successful approaches can be adapted without compromising effectiveness (Skivington et al., 2021; Education Endowment Foundation, 2024; Veroniki et al., 2026). Understanding these questions is particularly important in education, where interventions are routinely modified to reflect differences in school context, available resources and pupil need.

This shift is consistent with EEF's increasing emphasis on implementation and evidence mobilisation. EEF's implementation guidance recognises that educational outcomes are shaped not only by the choice of intervention but also by how interventions are introduced, enacted and sustained within schools (Education Endowment Foundation, 2024; Waddell & Sharples, 2020). Consequently, evidence that identifies transferable principles across programmes may be as valuable as evidence demonstrating the effectiveness of individual

interventions. This may also have important implications for equity. Not all schools and settings have equal access to the resources, capacity or funding required to adopt named programmes, particularly where these involve substantial financial or implementation requirements. Identifying the underlying principles or components associated with effectiveness may therefore enable schools to apply evidence-informed approaches in ways that are feasible within their own contexts, rather than relying solely on access to specific programmes (Clark & Saxberg, 2012). Moreover, by identifying components that are associated with improved outcomes across multiple programmes and contexts, this approach may also contribute to more generalisable knowledge about effective practice. Rather than focusing solely on whether individual programmes are effective, it can help identify features of practice that may contribute to effectiveness across different interventions and settings. Such knowledge has the potential to strengthen programme design, improve future commissioning, enhance evidence resources and support schools in making more informed decisions about adapting evidence-informed practice for their own contexts.

These developments have important implications for EEF's future evidence strategy. As its archive of completed evaluations continues to grow, so does the opportunity to investigate questions that cannot be answered by individual trials alone. In particular, the accumulation of comparable intervention data creates the possibility of examining recurring characteristics across programmes to better understand the components that contribute to educational impact. Determining whether the existing evidence base is sufficiently mature to support this next generation of evidence synthesis represents an important strategic question for EEF and provides the policy context for the present study.

From Programme Effectiveness to Understanding Mechanisms

Educational interventions are rarely simple, single-component treatments. Most programmes comprise multiple interacting elements, including professional development, curriculum resources, instructional strategies, assessment practices, implementation support, coaching, parental engagement and digital technologies. Each of these is implemented with varying levels of intensity, fidelity and adaptation to local context, meaning that programmes sharing the same name may differ substantially in their practical delivery. Equally, interventions developed independently may incorporate many of the same instructional features despite being classified as different programmes.

Traditional evidence synthesis has been successful in determining whether educational interventions improve attainment on average. Systematic reviews and conventional meta-analyses have provided increasingly robust estimates of programme effectiveness, enabling schools, policymakers and funders to compare broad intervention approaches and make evidence-informed decisions. However, these methods typically treat each intervention as an indivisible package. While this is appropriate for estimating overall programme effects, it provides limited insight into which aspects of an intervention are responsible for its observed impact (Ha, 2023; Rücker et al., 2020; Veroniki et al., 2026). This distinction has become increasingly important as the educational evidence base has matured. Programme-level estimates can establish whether an intervention works on average, but they cannot readily distinguish which recurring features contribute to effectiveness across different programmes.

The EEF Teaching and Learning Toolkit ('Toolkit')

The EEF Toolkit synthesises evidence on a range of educational approaches, such as feedback, metacognition and self-regulation, and small-group tuition, to support schools in making evidence-informed decisions about teaching and learning. These categories provide an accessible summary of the evidence for broad approaches, but necessarily bring together interventions that may differ considerably in their design and delivery. The proposed component-level analysis complements this approach by examining the specific features that make up interventions and investigating whether particular components, or combinations of components, are associated with improved pupil outcomes. In doing so, the work could provide an additional layer of evidence to inform the Toolkit, helping to explain not only whether broad approaches tend to be effective, but which features of those approaches may contribute to their effectiveness. Different interventions categorised as "feedback", "metacognition" or "small-group tuition" may contain substantially different combinations of instructional activities, implementation strategies and supporting resources, while interventions allocated to different Toolkit categories may nevertheless share common underlying components. As a consequence, programme-level syntheses may obscure important similarities between successful interventions while attributing effectiveness to programme labels rather than to the characteristics that genuinely drive educational improvement.

Understanding these core components also supports evidence mobilisation. Schools often need to adapt interventions to reflect their context, available resources and the needs of their pupils. Identifying which components are fundamental to an intervention's underlying mechanism, and which can be adapted without undermining its intended effects, could therefore provide more actionable guidance for schools. This knowledge could also help programme developers and funders distinguish between core components that should be preserved and more adaptable features that can be tailored to different settings. This is consistent with EEF's implementation guidance, which recognises that effective implementation involves balancing fidelity to an intervention's essential components with appropriate adaptation to context.

A growing body of research supports this shift towards mechanism-oriented evidence synthesis. Studies have shown that interventions sharing similar programme labels may operate through different instructional mechanisms, while programmes developed independently often derive their effectiveness from common underlying features. For example, a meta-analysis of school-based social and emotional learning interventions found that several widely implemented programme components were not consistently associated with improved outcomes, whereas certain less frequently used components demonstrated stronger associations with programme effectiveness (Mertens et al., 2020). These findings suggest that programme-level synthesis alone may overlook important mechanisms that explain why some interventions perform better than others. Mechanism-oriented evidence therefore complements, rather than replaces, programme-level evidence. Programme-level synthesis remains essential for determining whether broad intervention approaches improve educational outcomes. However, examining interventions at the level of their constituent components has the potential to generate more transferable knowledge about effective educational practice,

informing intervention development, implementation guidance and evidence mobilisation across a wider range of educational contexts.

Component Network Meta-Analysis

One methodological approach that has emerged to address this challenge is Component Network Meta-Analysis (cNMA). cNMA extends conventional network meta-analysis by representing interventions as combinations of components rather than indivisible treatment packages (Rücker et al., 2020). Rather than comparing named programmes alone, cNMA estimates indicative marginal associations between recurring intervention components and educational outcomes while allowing interventions to share common features across a connected evidence network. Importantly, cNMA does not assume that components operate independently or universally. Recent methodological developments emphasise that meaningful interpretation depends upon explicit modelling assumptions, sufficient variation in intervention combinations and careful consideration of interactions between components and context (Petropoulou et al., 2021; Tsokani et al., 2023). Consequently, before such approaches can contribute routinely to educational evidence synthesis, it is necessary to establish whether existing research infrastructure contains sufficiently complete and consistent information to support their application. Addressing that question provides the central motivation for the present study.

Educational applications of cNMA remain relatively limited but are growing. Recent network-based syntheses have demonstrated that combinations of instructional strategies can outperform individual approaches and that no single instructional strategy is consistently superior across all educational contexts (Peng et al., 2024). Emerging applications in literacy and mathematics similarly suggest that educational effectiveness depends upon particular configurations of intervention components rather than individual instructional techniques in isolation (Peng et al., 2025). These studies provide encouraging proof of concept but have generally focused on relatively narrow intervention domains or specific subject areas. Whether comparable analyses can be undertaken across a large, heterogeneous dataset of educational interventions remains largely unknown. The principal knowledge gap is therefore no longer whether component-oriented evidence synthesis is conceptually valuable, but whether existing educational research infrastructure is sufficiently mature to support it.

Strategic Alignment

The EEF Data Archive provides a unique opportunity to address this gap. It currently contains data from almost 140 EEF-funded evaluations involving more than 1.8 million pupils, together with primary and secondary outcomes, treatment allocation, compliance indicators, missing data information, archived reports and supporting analytical documentation (Uwimpuhwe et al., 2025). By systematically describing the constituent elements of educational programmes, the EEF's core components taxonomy creates the potential to investigate recurring intervention characteristics across multiple evaluations rather than considering programmes solely as individual packages. For the present feasibility study, we use project-level information extracted using EPPI-Reviewer (Thomas et al, 2023) as a proxy for the corresponding project-level data that could be extracted from the EEF Data Archive. This provides the summary-level

information required for the proposed analyses, including effect sizes and their associated uncertainty, while avoiding the additional data-access and governance requirements associated with working directly with the EEF Archive data. However, the availability of project-level data and a component taxonomy does not in itself guarantee that component-level synthesis using component network meta-analysis (cNMA) is feasible. Previous work has demonstrated that interventions within the EEF archive can be coded at component level and that relevant outcome data can be extracted and harmonised. The key remaining uncertainty is therefore whether the resulting evidence base has sufficient structure and variation to support robust cNMA. This project will assess, for example, whether there are sufficient studies contributing non-zero weight to relevant comparisons, whether there is adequate variation in the presence and combination of identified components, and whether components occur with sufficient frequency and connectivity across the evidence network to support reliable estimation of component effects. These questions will be examined empirically as part of the feasibility assessment before proceeding to full cNMA.

Research Questions and Design

This study is a secondary analysis of trial-level data extracted using EPPI-Reviewer, used as a proxy for the corresponding information available from the EEF Data Archive. It will examine whether the archive can support component-focused synthesis of educational interventions and, where the necessary conditions are met, produce a proof-of-concept Component Network Meta-analysis (cNMA). The project is designed primarily as a methodological feasibility study. Its purpose is not to assume that component-level effects can be estimated reliably, but to establish whether the available intervention descriptions, component coding, outcomes and evidence networks are sufficiently complete and coherent to support such analyses. Progression through the study will therefore be governed by formal decision gates. These will determine whether the project proceeds to more advanced analyses, requires additional data recovery or coding, restricts the scope of interpretation, or reports feasibility findings without substantive component estimates.

Study Aim

The aim of the study is to assess the feasibility and practical value of applying component-focused evidence synthesis, including cNMA, to EEF trial data. This will include assessing the coverage and consistency of the data available for component-level synthesis, alongside whether there is sufficient variation in identified components and adequate structure within the evidence network to support robust analysis. The study will explore associations between intervention components and attainment outcomes, particularly for pupils experiencing socio-economic disadvantage, and determine how component-based evidence could inform EEF's future commissioning, evaluation, reporting, archiving and evidence mobilisation.

Study Objectives

The study has five objectives:

1. *Assess data readiness.*
Determine whether eligible evaluations contain sufficient information to derive comparable attainment effect estimates.
2. *Evaluate component-network viability.*
Establish whether there is sufficient variation in component configurations, component frequency and network connectedness to support component-level estimation.
3. *Produce a proof-of-concept synthesis.*
Where prespecified feasibility criteria are met, undertake multilevel meta-analysis, component- and dimension-level analyses and a prototype additive cNMA, supported by appropriate diagnostics and sensitivity analyses.
4. *Explore heterogeneity and disadvantage.*
Examine whether subject area, trial characteristics, phase, setting and disadvantage-related variables can be incorporated as moderators where sufficiently complete and consistently defined.
5. *Generate actionable recommendations for EEF.*
Identify implications for commissioning briefs, intervention specification, evaluator reporting, archive development, future synthesis and evidence mobilisation, regardless of whether full cNMA is feasible.

Research Questions

The study will address the following research questions:

- *RQ1: Data readiness*

What proportion of potentially eligible EEF evaluations contain sufficient information to distinguish intervention arms in multi-arm evaluations; identify the primary attainment outcome; derive a harmonised effect estimate; and recover an appropriate measure of statistical uncertainty?
- *RQ2: Component-network viability*

Do evaluations within one or more candidate attainment domains (e.g. maths, literacy or science) provide sufficient numbers of eligible studies and intervention arms; adequate variation in component composition and prominence (variation in whether components are coded as absent, minor or major); sufficiently frequent occurrence of individual components; adequate separation between components that commonly co-occur; and a connected or otherwise defensible evidence network?
- *RQ3: Proof of concept*

Where feasibility criteria are met, can component-focused analyses produce interpretable indicative associations for individual components, broader component dimensions, and selected recurring component combinations? Can these estimates be

obtained with acceptable uncertainty, model performance and sensitivity to alternative analytical decisions?

- *RQ4: Heterogeneity and disadvantage*

Where suitable information is available, can differences in component associations be explored according to subject domain, pupil age or phase, trial and intervention characteristics, and Free School Meal eligibility or comparable disadvantage indicators? What do the availability and quality of these data imply for equity-focused evidence mobilisation?

- *RQ5: EEF ways of working*

How could the findings inform EEF's future commissioning briefs, intervention theory-of-change requirements, evaluator reporting standards, core-component coding procedures, archive structures and metadata, and evidence reviews and Toolkit development? Can component-level findings be translated into clear and interpretable summary statistics suitable for EEF decision-making and evidence communication, such as the estimated likelihood of a positive impact associated with a particular component, or an equivalent decision-relevant metric?

Research Design

The study is a secondary data analysis and methodological feasibility study drawing on completed EEF-funded evaluations. It will combine information from archived impact datasets, evaluation reports, trial metadata and EEF's core components taxonomy. The research will proceed sequentially. Initial work will establish which studies, outcomes and component codes are available and sufficiently complete. A harmonised effect-size dataset and an arm-by-component matrix will then be constructed. Only after these data have been reviewed and formally frozen will the project progress to prototype component-focused analyses.

The staged design is necessary because the central assumptions underpinning cNMA cannot be established fully in advance. Feasibility will therefore be assessed using a combination of pre-specified criteria and empirical diagnostics applied once the structure of the evidence network is known. Where meaningful thresholds can be specified in advance, these will be defined a priori; other criteria, particularly those relating to variation and network structure, will require diagnostic assessment of the observed data.

In particular, the feasibility assessment will examine whether:

- sufficient eligible evaluations are available within common attainment domains. Rather than applying a single minimum number of studies to an entire domain, adequacy will be considered in relation to the number of studies contributing information to the parameters to be estimated. Studies-per-parameter and component-frequency diagnostics will therefore be used to determine whether particular component effects are sufficiently supported by the available evidence;

- reported outcomes permit comparable effect estimates to be derived. This will be assessed through the outcome harmonisation process, including consideration of outcome domain, measurement characteristics and the availability of information required to calculate comparable standardised effect estimates;
- individual components occur frequently enough to be estimated. As a pre-specified screening criterion, components will ordinarily be required to occur in at least three studies contributing non-zero weight to the relevant analysis. Components falling below this threshold will not normally be considered independently estimable, although their treatment will also take account of the wider network structure and conceptual relevance;
- component configurations vary sufficiently across interventions to allow individual components to be distinguished from commonly co-occurring components. Because sufficient variation is a multivariate property of the evidence network, no single numerical threshold will be specified in advance. It will instead be assessed empirically using component-frequency distributions, co-occurrence patterns and distribution analyses to establish whether the observed combinations provide sufficient information to distinguish component effects;
- highly correlated components can be distinguished empirically. Correlation matrices and component co-occurrence tables will be examined, with a correlation of $r > .80$ used as a screening indicator of potentially problematic collinearity. This threshold will trigger further review rather than automatic collapsing or exclusion of components; final decisions will consider both the statistical evidence and whether the components are conceptually distinct; and
- the resulting evidence network is adequately connected to support the proposed analyses. Network connectivity will be assessed after construction of the analytic dataset by examining the network structure, including the presence of disconnected components or sub-networks and whether sufficient direct or indirect evidence links the intervention/component configurations required for the planned comparisons. Where the full network is not connected, analyses will consider whether sufficiently connected sub-networks exist to support meaningful component-level estimation.

The study draws on pupil-level and, where applicable, school-level data held within individual evaluations. However, the principal unit of analysis for the cross-study synthesis will be the intervention arm–outcome effect estimate, nested within its originating evaluation. Each analytic record will represent an estimated intervention effect for a defined evaluation, intervention arm, comparator, attainment outcome and analysis population. To maximise comparability across evaluations, the primary synthesis will use attainment outcomes measured at the end of the intervention, or at the post-intervention assessment point closest to intervention completion where timing differs between studies. Longer-term follow-up estimates will not be included in the primary component-level synthesis. Intervention duration will be extracted and described as a study characteristic, and its distribution across the included evidence base will be examined as part of the assessment of heterogeneity and comparability. Where sufficient data and variation are available, intervention duration may also be explored as a potential source of heterogeneity in secondary analyses.

Where a trial reports more than one eligible primary outcome at the selected post-intervention time point, the outcomes may contribute separate linked records. Their statistical dependence will be represented in the multilevel model so that one evaluation is not treated as several independent studies. For multi-arm trials, each distinguishable intervention arm will be coded separately. Shared comparator groups and correlations between contrasts will be taken into account where the available information permits. The intervention arm will also form the row-level unit of the arm-by-component matrix. The final number of interventions cannot be determined until screening, outcome recovery and component-coding checks have been completed.

The feasibility assessment will determine whether the evidence should be:

- pooled across reading, mathematics and science with subject adjustment;
- analysed within separate subject-specific networks;
- restricted to one or more well-connected subdomains; or
- reported as a feasibility map without substantive network estimation.

The primary outcomes will be post-intervention attainment outcomes designated as primary in the original evaluation or protocol. The initial feasibility assessment will consider all eligible attainment outcomes available across EEF interventions, rather than restricting inclusion to particular subject domains. This may include, for example, general attainment indices, writing outcomes and other measures of educational attainment, alongside reading and literacy, mathematics and science. The study will first assess the feasibility of component-focused synthesis across the available attainment evidence base as a whole. Where sufficient numbers of studies, component variation and network connectivity are available, analyses will also be conducted separately for specific attainment domains, including reading and literacy, mathematics and science.

Primary outcomes may be measured using standardised tests, national assessments, validated curriculum-based instruments or other attainment measures accepted within the original EEF evaluation. Because studies are expected to use different instruments and scales, eligible results will be transformed to a common standardised effect-size metric where appropriate. Where an evaluator has reported an adjusted effect size and corresponding confidence interval or standard error, this estimate will normally be preferred, provided that it is consistent with the intended estimand and based on an appropriate analysis. Where no suitable effect size or confidence interval is reported, or where the associated measure of variance or standard error is missing, the study team will derive one from archived estimates, summary statistics or pupil-level data where access and data quality permit.

Baseline measures will be trial-specific. They may include prior attainment on the same or a related scale, national assessment scores, baseline standardised test scores or evaluator-defined attainment covariates. Baseline measures will be used to understand the derivation and

comparability of adjusted impact estimates and, where individual-trial estimates must be recreated, to reproduce an appropriately adjusted analysis where feasible.

Secondary outcomes may include additional attainment outcomes identified as secondary in the original evaluation, alternative eligible attainment measures, disadvantage-specific attainment estimates, subgroup estimates for pupils eligible for Free School Meals or equivalent indicators, estimates for pupils with low prior attainment, and implementation, compliance or mediation variables where these are sufficiently complete and relevant to interpretation. The dataset supplied for this study contains primary outcome data only; consequently, any secondary outcome or additional subgroup estimates required for these analyses will need to be identified and extracted separately from evaluation reports, archived datasets or other available project documentation. The feasibility and resource implications of this additional extraction will therefore be assessed before undertaking these analyses.

Secondary attainment outcomes will not automatically be combined with primary outcomes. Their inclusion will depend on conceptual comparability, measurement timing, statistical dependence and their role in the original evaluation. If the number or complexity of potential analyses requires the scope to be reduced, analyses relating to socio-economic disadvantage will be prioritised over analyses of secondary outcomes. This reflects the study's primary interest in understanding whether component-level evidence can provide useful insights into attainment outcomes for pupils experiencing socio-economic disadvantage. Non-attainment outcomes, including social-emotional, behavioural and implementation outcomes, will not form part of the main component-effect analysis unless the study scope is formally expanded following the feasibility assessment. They may nevertheless be recorded as contextual information or used to inform recommendations about future archive development.

Data sources

The study will draw on three principal sources of information:

1. EEF impact datasets;
2. EEF evaluation reports and associated materials; and
3. EEF's core components taxonomy and completed or emerging report coding.

EEF impact datasets

EEF impact datasets will be created extracting the relevant primary outcome parameters (including sample size, effect size and associated uncertainties information) from EEF evaluation reports. Impact datasets will support RQ1 by establishing whether comparable effects and uncertainty estimates can be obtained. They will support RQ3 and RQ4 by providing outcome, trial and moderator information for the synthesis.

EEF evaluation reports

Published EEF evaluation reports will provide information that may not be fully represented and will be used to recover missing outcome information, determine the preferred effect estimate

and resolve ambiguities in intervention coding. Where the published report is insufficient, associated protocols, statistical analysis plans, appendices, technical reports and supplementary materials will be sought.

EEF core components taxonomy and coding

The current version of the EEF taxonomy has 78 components distributed across 12 dimensions. *Dimensions* each contain a number of related components. Each *component* represents groups of related practices, called *examples of practice*. In the context of the taxonomy, a component is defined as a small, meaningful unit of an intervention that is observable, replicable and represents a distinct element of the intervention's theory of change. For the purposes of keeping the present project manageable and focused, only dimensions 1-9 will be used (58 components). Dimensions 1-9 relate to pedagogy, whereas dimensions 10, 11 and 12 comprise components relating to setting-level structure, professional development, and implementation.

All EEF evaluation reports have been double coded using trained coders. These 20 EEF colleagues received training consisting of three online sessions with gap tasks as well as a comprehensive package of support (handbook, an expanded version of the taxonomy with detailed coding guidance, and optional drop-in sessions). All reports were coded by two people, with a third coder used to quality assure a random sample.

EEF's core components taxonomy will provide the organising framework for representing interventions. The primary eligibility scope will be pedagogy-related components within Dimensions 1-9. Components are currently coded using a three-point prominence scale:

- 0: the component is absent;
- 0.5: the component is present but minor; and
- 1: the component is present and major within the programme's theory of change.

These values will be treated as structured judgements about programme prominence rather than as exact measures of classroom dosage. The primary analysis will retain the original coding where feasible, while sensitivity analyses will test alternative representations such as present versus absent, major versus not major and alternative treatment of minor components.

Where required variables or sufficient outcome information cannot be obtained from public reports or existing derived datasets, the team will consider whether the affected study is important to network connectedness; whether the missing information is likely to be recoverable; whether its recovery could materially alter feasibility; the proportionality of additional access requirements; and whether the required information could be obtained through a less burdensome source.

Trial inclusion criteria

Eligibility Criteria

An evaluation will be eligible for inclusion where it:

1. is an EEF-funded impact evaluation, including efficacy and effectiveness evaluations;
2. evaluates at least one intervention arm coded with one or more pedagogy-related components within Dimensions 1-9 of the EEF taxonomy and reports, or permits derivation of, at least one eligible post-intervention attainment effect estimate and associated measure of uncertainty;
3. includes a comparator condition that can be identified sufficiently to support estimation and interpretation of the intervention effect.

These criteria represent the minimum information required for an evaluation to contribute to the component-level synthesis. Additional information on the sample, intervention arm, outcome measure, intervention duration and timing of outcome measurement will be extracted where available and used to support interpretation, assessment of heterogeneity and secondary analyses, but incomplete reporting of these characteristics will not automatically result in exclusion where the component coding and required effect estimate are available.

Longer-term follow-up data are not required for eligibility. The primary synthesis will focus on attainment measured at the end of the intervention, or at the post-intervention assessment point closest to intervention completion where assessment timing differs between evaluations.

Randomised controlled trials will form the principal evidence base. Other controlled impact designs will only be considered where they are compatible with the intended estimand and can be incorporated without undermining comparability. Any such inclusion will be agreed with EEF and examined in sensitivity analysis.

Exclusion Criteria

Evaluations will be excluded where:

- no impact data are available;
- no eligible attainment outcome is reported or recoverable;
- the intervention cannot be linked to an eligible taxonomy component;
- the reported result cannot be converted to a compatible effect metric;
- the relevant outcome is measured only before intervention exposure;
- the design does not provide a defensible counterfactual estimate;
- essential reporting materials (for example, intervention description, outcome definition, effect estimate, or measure of uncertainty) are unavailable after reasonable recovery attempts; or
- duplication means the same trial population and contrast are already represented by a more complete record.

Exclusion from the main synthesis will not necessarily mean removal from the feasibility assessment. Studies excluded because of missing reporting or coding will be retained in the screening dataset so that their absence can inform recommendations for archive development.

Component Eligibility

A component will be eligible for individual estimation where:

- it falls within the agreed taxonomy scope;
- its coding has been completed and quality checked;
- it has a non-zero weight in at least three evaluations;
- it shows sufficient variation across intervention arms; and
- Its weights are not identical to those of another component.

The threshold of three evaluations is a minimum feasibility threshold rather than a guarantee of reliable estimation. Precision, covariance, leverage and model stability will also be considered. Components that do not meet the individual-analysis threshold may:

- contribute to a broader dimension score;
- be combined with closely related components where conceptually and empirically justified;
- be represented descriptively; or
- be excluded from quantitative component estimation.

Outcome Eligibility

An outcome will be eligible where it:

- measures an eligible pupil outcome, including attainment or, where sufficient comparable data are available, outcomes such as attendance, behaviour, self-regulation or other relevant pupil-level outcomes;
- is identified as primary or otherwise directly relevant to the selected domain;
- was measured after intervention exposure;
- permits a treatment–comparator contrast;
- can be converted to the common effect metric; and
- is based on a sufficiently defined analysis sample.

Where several versions of an outcome are reported, preference will ordinarily be given to:

1. the prespecified primary outcome;

2. the prespecified primary follow-up (post-intervention time point);
3. the evaluator's adjusted intention-to-treat estimate;
4. the measure with the clearest validity and completeness; and
5. the estimate most consistent with EEF statistical guidance.

Research Stages

Stage 1: Data Access and Feasibility

The core components coding handbook will be reviewed to establish the intended definition and application of each eligible component and to identify any ambiguities relevant to quantitative analysis. The research team will review component and dimension definitions; document coding rules for absent, minor and major components; examine worked examples and borderline cases; identify changes in definitions across coding versions; distinguish confirmed absence from incomplete coding; assess whether the scale is applied to intervention arms or whole programmes; and agree how comparator conditions will be represented.

The team will map the organisation of the evaluation dataset and undertake an initial data clean to standardise identifiers and labels; identify duplicate records; verify treatment allocation; check sample sizes against reports; examine missingness; reconcile outcome names; identify impossible or inconsistent values; map follow-up points; and flag discrepancies between reports and the working dataset. They will then assess whether each intervention arm has been coded with sufficient completeness and consistency to enter the primary analysis. For each candidate primary outcome, the study will assess whether the following are available or recoverable:

- outcome definition;
- effect estimate;
- effect direction;
- confidence interval or standard error;
- sample size;
- analysis population;
- clustering structure;
- adjustment variables; and
- follow-up point.

This activity will produce a coding-completeness matrix, a list of coding recovery priorities, data-quality flags and an estimate of the proportion of potentially eligible studies ready for analysis.

Where outcome or coding information is incomplete, the team will search for final reports and supplementary publications. A retrieval log will record which materials were sought, located and

used. Information recovered from reports will be entered into a structured extraction form. Extracted values will be independently checked where they determine eligibility or effect-size derivation. The infill dataset will distinguish clearly between directly reported values, calculated values, values reproduced from pupil-level data, inferred labels and unresolved missing information. For each eligible contrast, the team will identify the effect estimate that most closely represents the intended post-intervention attainment effect. Where the evaluator reports a suitable effect size and confidence interval, these values will normally be used. Where necessary, the team will calculate a standardised mean difference; apply a small-sample correction; derive a standard error from a confidence interval; derive uncertainty from reported test statistics; use sample information to reconstruct variance; or extract from the sensitivity analysis reports obtained via the EEF archive. The derivation will reflect clustering and the analysis design wherever possible. Estimates that cannot be placed on a compatible scale without strong assumptions will not be combined. The linked outcome and coding datasets will undergo a second round of data cleaning to check for duplicate contrasts, mismatched arm identifiers, inconsistent effect directions, implausible variances, missing component codes, incomplete provenance, inconsistencies between trial counts and arm counts and errors in subject or outcome classification.

At this point, the study will have reached ‘Decision Gate 1’². This will determine whether sufficient evidence exists to progress to formal component harmonisation and modelling by addressing four principal questions:

1. Is component coding sufficiently complete?
2. Are eligible primary outcomes available or recoverable?
3. Can effect-size precision be obtained?
4. Is there preliminary evidence that a viable component network may exist?

Possible pathways are:

- **Continue:** sufficient evidence exists to proceed.
- **Exclude individual studies:** the dataset remains viable without them.
- **Restrict the subject or intervention scope:** only particular domains are ready.
- **Terminate formal cNMA feasibility:** the dataset cannot currently support defensible estimation.

Stage 2: Component Coding and Harmonisation

The team will calculate, for every component:

² More detailed description of the workflows and decisions for each study stage are provided in Appendix A.

- the number of evaluations with a non-zero code;
- the number of intervention arms with a non-zero code;
- the number of minor and major codes;
- the distribution across subjects and phases;
- precision-weighted representation; and
- occurrence within comparator arms.

Only components with non-zero weight in at least three evaluations will be considered for individual component estimation. Higher practical thresholds may be applied where estimates are unstable. A component frequency distribution figure will present components ordered by the number of evaluations and arms in which they occur, distinguishing minor and major prominence and marking the minimum individual-analysis threshold. The study will examine whether components almost always occur together (i.e., rarely occur independently); have identical coding profiles; are strongly correlated in prominence; or are confounded with subject or trial characteristics. The analysis will use correlation matrices, co-occurrence tables, cross-tabulations and visual inspection of the arm-by-component matrix. Heat maps of the arm-by-component matrix and component co-occurrence patterns will be used to illustrate component frequency, overlap and clustering. High correlation will not automatically justify collapsing components; it will instead trigger joint empirical and conceptual review. A component may be collapsed with another where:

1. the components are conceptually part of a coherent higher-order construct;
2. the original reports do not reliably distinguish them;
3. the coding profiles are highly overlapping;
4. separate estimation would be unstable or uninterpretable; and
5. sensitivity analyses can compare the collapsed and original representations.

A component dictionary will be created to record the original taxonomy code and label, operational definition, coding scale, eligible studies and arms, observed frequencies, common co-occurring components, ambiguities, collapsed or higher-order groupings, quality flags and final analytical treatment. The dictionary will be version controlled and supplied as part of the technical documentation. At this point, the research team will share a frozen dataset with EEF before commencing the principal models. This will lead to ‘Decision Gate 2’, to determine whether the final dataset is ready for modelling.

Stage 3: Prototype Component-Focused Synthesis and cNMA

A preliminary meta-analysis will describe the overall attainment evidence, quantify heterogeneity and establish the variance structure needed for subsequent component analyses. Effect sizes will be analysed using a multilevel random-effects model. Eligible outcomes will be nested within evaluations so that correlated results from the same trial are

not treated as independent. The initial model will contain no component predictors. Depending on feasibility findings, it may be fitted across eligible attainment subjects; with subject included as a moderator; or separately by subject domain. This analysis will be used to assess effect-size comparability, influential studies, residual heterogeneity, distribution of residuals, sensitivity to alternative variance estimators, dependence assumptions and the effect of retaining multiple outcomes.

Next, each sufficiently common component will be examined separately using component-wise models. The component's prominence score will be used as the predictor, subject to the findings of the scale-assessment stage. The component-wise estimate will represent the association between greater prominence of that component and the intervention effect across the eligible evidence base. It will not be interpreted as the isolated causal effect of adding the component to an otherwise identical programme. Weighted least-squares or equivalent multilevel meta-regression models will account for sampling precision, dependence among outcomes from the same trial and between-study heterogeneity.

Related components will also be grouped within their taxonomy dimensions to explore dimension-wise models. A study- or arm-level dimension score will be calculated from the component weights belonging to that dimension, initially using the mean of available component prominence scores. Alternative dimension scoring may be considered where simple averaging would be misleading. Any alternative will be documented before the primary dimension model is fitted. Dimension-level analyses provide a more stable but less granular representation of intervention content and may remain interpretable where individual components are too sparse or correlated.

Before moving to an additive cNMA, the evidence network will be examined for:

- connectedness;
- isolated interventions;
- bridge³ interventions;
- bridge components;
- disconnected sub-networks;
- component configurations represented by only one study;
- overlap between intervention and comparator content; and
- subject-specific fragmentation.

Where the full network is disconnected, the study may analyse defensible connected sub-networks separately; restrict the model to a better-connected subject domain; use dimension-

³ Bridging: refers to the connectivity of intervention and/or component within the evidence network.

level rather than component-level representations; or stop short of formal cNMA. Where model stability permits, Bayesian versions of the overall, component and dimension models will be fitted to provide probability-based summaries. Thresholds of 0.2 and 0.5 standard deviation units will be reviewed because they may represent relatively demanding thresholds in educational trials. Weakly informative priors will generally be preferred to completely non-informative priors where they improve estimation without dominating the observed evidence. Prior choices and prior-sensitivity analyses will be reported.

Where connectedness and data sufficiency criteria are satisfied, an additive cNMA model will represent each intervention as a bundle of components relative to a reference condition, such as business-as-usual. The additive model assumes that the expected effect of an intervention bundle can be represented by the combined contributions of its coded components. This assumption will be treated as an analytical approximation rather than a substantive claim that components always operate independently. The model will assess:

- consistency of the network;
- fit relative to simpler models;
- residual heterogeneity;
- component collinearity;
- influential configurations;
- uncertainty in component estimates; and
- sensitivity to alternative coding and network restrictions.

Component estimates will be described as indicative marginal associations under the model assumptions. They will not be presented as universally transferable causal effects. Selected combinations may be explored where they occur frequently enough; there is variation in their joint and separate occurrence; there is a substantive rationale; the number of parameters is proportionate to the evidence; and the additive model suggests that interaction testing is warranted. Interactions will be limited and prespecified wherever possible. The study will avoid an unrestricted search across all possible component combinations. Meta-Analytic Classification and Regression Trees will be used to examine whether combinations of components or dimensions partition studies into groups with different effect sizes. The method will account for between-study heterogeneity and will use cross-validation or pruning procedures to reduce overfitting. Findings will be treated as hypothesis-generating.

Moderator analyses will be undertaken only where category sizes and model degrees of freedom are adequate. Where data are sufficiently complete, potential moderators will include:

- subject area;
- school phase or pupil age;
- trial design;

- intervention type;
- delivery setting;
- Free School Meal eligibility;
- school-level disadvantage;
- prior attainment; and
- low-attainer status.

Where disadvantage estimates are available only for a subset of trials, the analysis will explicitly assess selection and comparability. Failure to conduct a defensible equity analysis will be reported as an archive-readiness finding. Where disadvantage-focused cNMA or subgroup analysis is not currently feasible, the study will identify the conditions that would need to be met to enable such analyses in the future. This will include assessing whether a sufficient number of studies report comparable disadvantage-related estimates, whether there is adequate variation in intervention components within this subset of studies, and whether the resulting evidence network would be sufficiently connected to support component-level estimation. The study will use these findings to identify priorities for future commissioning, evaluator reporting and archive development that could strengthen the evidence base for subsequent equity-focused synthesis.

Finally, sensitivity analyses will be utilised to examine:

- one primary outcome per study;
- alternative component prominence coding;
- alternative component collapsing decisions;
- exclusion of weakly specified intervention arms;
- exclusion of high-risk or highly influential studies;
- subject-specific rather than pooled models;
- restriction to connected sub-networks;
- alternative effect-size derivations;
- alternative assumptions about within-study outcome correlation;
- exclusion or separate treatment of non-randomised designs;
- alternative prior distributions; and
- alternative thresholds for rare components.

Before findings are translated for wider audiences, the research team will prepare an interpretation memo for EEF, outlining what the models estimate; assumptions required for

interpretation; strengths and weaknesses of the network; estimates judged sufficiently stable; findings that should remain exploratory; sensitivity of conclusions; limitations of disadvantage analyses; and language that should and should not be used in reporting. This will help inform ‘Decision Gate 3’, to determine the level at which results may be interpreted. Four principal outcomes are possible:

1. Full component interpretation

Component and selected combination estimates are sufficiently stable for cautious substantive interpretation.

2. Restricted interpretation

Only a limited subset of components or a specific connected sub-network is interpreted.

3. Dimension-only interpretation

Individual components are too sparse or correlated, but broader dimensions support cautious interpretation.

4. Feasibility findings only

The EEF trial data does not support defensible substantive component estimates; reporting focuses on readiness, limitations and future requirements.

Data analysis

This section provides a consolidated reference for the statistical methods embedded within Stage 3. The final choice and complexity of each model will depend on the outcomes of Decision Gates 1 and 2. All analysis will be conducted using reproducible scripts, principally in R. EEF’s Statistical Analysis Guidance will be followed where applicable. Departures required by the secondary synthesis design will be discussed with the EEF Evaluation Manager and documented.

Effect-Size Metric

For projects where ESs are available but confidence intervals (CIs) are missing, both will be extracted from sensitivity analysis reports obtained via the EEF archive. Where CIs remain unavailable, sample sizes will be retrieved and used to compute CIs for Hedges' g , applying a small-sample correction factor $J = 1 - \frac{3}{4(n_0 + n_1 - 2) - 1}$ following the approach of Borenstein et al. (2009).

$$CI = g \pm 1.96 \sqrt{\left(J^2 \frac{n_0 + n_1}{n_0 n_1} + \frac{g^2}{2(n_0 + n_1)} \right)}$$

Standard errors for each trial, which are not directly available in the dataset, will be derived from the CIs. Trials featuring two primary outcomes will be handled by treating each outcome

as a separate row, thereby ensuring all primary outcomes are captured within a single unified dataset.

$$S_k = \frac{UB_k - LB_k}{2 \times 1.96}, \text{ with } UB_k = \text{upper bound and } LB_k = \text{lower bound of the CI for study } k.$$

Multilevel Random-Effects Meta-analysis

Multilevel meta-analysis will be conducted using a random-effects model, implemented via the `rma.mv` function from the `metafor` package in R (Viechtbauer, 2010). Reported effect sizes from each trial will serve as the outcome variable. A random effect at the study (trial) level will be specified to account for clustering within studies reporting multiple outcomes. The model will initially be fitted without component-level information in order to obtain an estimate of between-study variance, using the restricted maximum likelihood (REML) approach. This estimate will subsequently inform both the component-wise and dimension-wise analyses. No fixed-effect covariates will be included in this preliminary stage.

Component-Wise Analysis

To evaluate the effect of individual study components, a component-wise meta-analysis will be carried out. Only components with at least three outcomes receiving non-zero weights will be included, ensuring sufficient data for reliable estimation. For each component, the pooled ES will be estimated using a weighted least squares (WLS) approach, with model weights accounting for the variance-covariance structure and component-specific weights. This framework is well suited to the data structure, given the presence of multiple primary outcomes per trial, which introduces heteroscedasticity and dependence across estimates.

$$c_{jk}$$

$= \{1 \text{ if component } j \text{ is present as major in trial } k \quad 0.5 \text{ if component } j \text{ is present as minor in trial } k \quad 0 \text{ if component } j \text{ is absent from trial } k$

Dimension-Wise Analysis

To assess broader intervention categories, component weights will be aggregated within each dimension. Each dimension's weight will be computed as the mean of the constituent component weights for a given study. Dimension-level information will be recorded in a matrix format, and the aggregated weight for each dimension-study combination will be calculated accordingly. The same WLS framework used in the component-wise analysis will then be applied at the dimension level.

Bayesian Posterior Probability

A Bayesian meta-analysis will be conducted to estimate the overall SMD across studies, explicitly accounting for both within-trial and between-trial variability. This approach will also accommodate component-specific weights that modulate each study's contribution to the analysis. Observed ESs will be modelled using a multivariate normal distribution, with a non-informative prior placed on the overall effect. For component-specific analyses, study weights will be incorporated through a precision matrix transformation. Posterior inference will

encompass the mean, standard deviation, and 95% credible intervals for the ES, as well as posterior probabilities that the effect exceeds pre-specified thresholds. These results will be reported for both the overall effect and for each individual component or dimension.

Model Selection Approaches: glmulti and glmmLasso

Two complementary modelling approaches will be used to determine which components and contextual factors appear most influential. First, automated model selection within a meta-analytic framework will be performed using the glmulti package (Calcagno & de Mazancourt, 2010), which evaluates all possible predictor combinations against information criteria such as AIC and BIC. This will support identification of parsimonious models and influential moderators while accounting for uncertainty through model averaging. As a complementary method, regularised multivariate regression will be applied using the glmmLasso package, which extends Lasso penalisation to generalised linear mixed models. This approach is particularly valuable in high-dimensional settings or where predictors are correlated, as it performs variable selection by shrinking uninformative coefficients towards zero, thereby enhancing interpretability and predictive accuracy. Following guidance from Schmid et al. (2020), the number of components or dimensions included in any single meta-regression model will be capped to maintain a minimum of ten studies per variable. With 54 studies in the dataset, this will result in a maximum of six predictors per model.

Meta-Analytic Classification and Regression Trees: metacart

To explore which combinations of components or dimensions may account for heterogeneity in ESs, the Meta-Analytic Classification and Regression Trees (Meta-CART) approach will be employed. This recursive partitioning technique, adapted specifically for meta-analysis, will be used to identify subgroups of studies with differing treatment effects based on moderator combinations. A random-effects framework will be adopted within the Meta-CART procedure to account for between-study variability, with the analysis performed using the REMrt function from the metacart R package (Li et al., 2020).

Prototype Component Network Meta-Analysis (cNMA)

Where feasibility criteria relating to network connectedness, component frequency, component variation and model stability are satisfied, an additive Component Network Meta-Analysis (cNMA) will be undertaken. In this framework, each intervention is represented as a combination of its constituent components rather than as a single intervention package. The model estimates the average association between each component and intervention effectiveness, while accounting for the fact that components occur in different combinations across interventions. The primary analysis will use an additive cNMA model in which the expected effect of an intervention is represented as the sum of the contributions of its coded components.

Model diagnostics will assess network connectedness, residual heterogeneity, component collinearity, influential intervention configurations and sensitivity to alternative coding assumptions. Where assumptions are not satisfied, analyses may be restricted to connected subnetworks, broader component dimensions, or alternative synthesis approaches.

Component estimates will be interpreted as indicative marginal associations under the assumptions of the model and not as independent causal effects of adding a component to an otherwise identical intervention. Findings will therefore be interpreted alongside sensitivity analyses, network diagnostics and evidence regarding component co-occurrence.

Caveats and potential insights of cNMA

Component network meta-analysis (cNMA) is particularly suited to the present setting, where interventions are largely distinct but share components. By exploiting the overlap in component composition across interventions. This shared component structure is what cNMA relies on. However, the interpretation of these estimates depends on the structure of the component network. Components that occur frequently across a range of different intervention combinations are more amenable to estimation, whereas components that are rare or tend to occur together with the same other components may be difficult to disentangle. The analysis also relies on the assumptions underlying the cNMA model, including consistency of the evidence and the adequacy of the assumed relationship between components and intervention effects. Overall, this study will be able to describe which components are likely to be associated with stronger effects, conditional on the other components present. It will also be able to describe how components can be ranked against each other. It will not be able to say that a specific component causes a given effect when added to a new intervention. It will also not be able to say that two interventions sharing a component would perform in a similar way, since the intervention is not based on a single component.

Ethics

This study is a secondary analysis of existing Education Endowment Foundation (EEF)-funded trial data, evaluation reports, and core component coding. No new data will be collected, no participants will be recruited, and there will be no direct contact with pupils, teachers, school leaders, schools or other stakeholders. The project does not involve any intervention, randomisation or alteration to educational provision and therefore presents minimal ethical risk. Analyses will focus on intervention characteristics, patterns across studies and the readiness of the archive to support more sophisticated evidence synthesis. Findings will therefore be reported at an aggregate level and will not be used to judge the performance of individual schools, educators or programmes. Because the project seeks to identify patterns across multiple evaluations, particular care will be taken to ensure that findings are interpreted responsibly. Component-level estimates will be presented as indicative associations within a body of evidence rather than as definitive causal effects of individual intervention components. All findings will be interpreted within the methodological assumptions and limitations of the study, with clear distinction between confirmatory analyses, exploratory analyses and feasibility findings.

The study also recognises the importance of responsible reporting in relation to educational disadvantage. Where analyses involving socio-economic disadvantage are undertaken, findings will be interpreted within an equity-focused framework consistent with EEF's mission. Care will be taken to avoid deficit-based interpretations of pupil characteristics or overgeneralisation

from incomplete evidence. Any limitations arising from inconsistent reporting of disadvantage-related variables across evaluations will be explicitly acknowledged.

Ethical approval will be sought from the appropriate Durham University Research Ethics Committee where required under institutional policy. As the study involves secondary analysis of existing de-identified research data with no participant recruitment or additional data collection, it is anticipated that the project may be eligible for proportionate review or exemption from full ethical review. The appropriate determination will be made through Durham University's ethics governance procedures before substantive analysis begins. Where formal ethical review is not required, written confirmation of the exemption or institutional determination will be retained within the project records. The study will not commence until all institutional governance requirements, EEF approvals and data access arrangements have been completed. Artificial intelligence (AI) platforms will be used to support the reviewing and editing of written outputs produced by the research team. Specifically, AI tools will be used to check grammar, spelling, and sentence structure, and to improve the clarity and flow of writing.

Data protection

The study will be conducted in accordance with the UK General Data Protection Regulation (UK GDPR), the Data Protection Act 2018, EEF data-sharing agreements and all applicable institutional information governance policies. The research team will use only the data required to answer the research questions and undertake the planned feasibility analyses. Data processing will adhere to the principles of lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability. Only authorised members of the research team who require access for the purposes of the study will be permitted to access EEF datasets.

The study will process personal data solely for scientific research purposes under Article 6(1)(e) of the UK GDPR (performance of a task carried out in the public interest), together with the scientific research provisions of Article 89 and Schedule 1 of the Data Protection Act 2018, where applicable. The research team does not anticipate processing special category personal data directly. Where datasets contain variables relating to characteristics such as ethnicity, disability or socio-economic disadvantage that fall within the definition of special category data, these will only be processed where necessary for the approved research purposes and under the safeguards required by Article 9(2)(j) UK GDPR and Schedule 1, Part 1, Paragraph 4 of the Data Protection Act 2018 (research purposes). Processing will be limited to variables necessary for the approved analyses and will remain subject to the conditions specified within the EEF Data Sharing Agreement.

All analyses will be undertaken using de-identified or pseudonymised datasets. The research team will not attempt to identify individual pupils, teachers, schools or trial participants, nor will any analyses be undertaken for that purpose. Outputs will be reported only in aggregate form. Disclosure control procedures will be applied before publication to ensure that no individual, school or organisation can be identified, particularly where subgroup analyses involve small numbers of studies or settings.

Access permissions will be restricted to authorised project staff and protected through institutional authentication procedures. Data will not be transferred to unsecured devices or external systems without explicit approval and appropriate safeguards. Data handling procedures will comply with all requirements specified by EEF's data access arrangements.

The study will implement comprehensive quality assurance procedures throughout data management and analysis. These will include verification of extracted information against original evaluation reports; reproducible data cleaning and transformation scripts; version-controlled derived datasets; documented data linkage procedures; a complete data provenance record for all derived variables; a component dictionary recording all coding decisions; and an audit trail documenting all amendments made following the formal decision gates.

Research data will be retained only for the period specified within Durham University's Research Data Management Policy, the EEF Data Sharing Agreement and any associated data access conditions. At the conclusion of the project, all datasets and derived files containing controlled information will be securely deleted or archived in accordance with EEF's requirements and institutional data retention policies. Any retained analytical outputs or reproducible scripts will exclude identifiable information and will remain subject to the permissions governing the original archive data.

Personnel

- Dr Emma Dobson (WhatWorked Education) will lead the study overall and take responsibility for design, interpretation, reporting and alignment with EEF priorities
- Dr Akansha Singh (Durham University) will lead data preparation and quality assurance
- Rahil Khawaja (WhatWorked Education) will lead the component modelling and reproducible statistical workflow
- Germaine Uwimpuhwe (Durham University, WhatWorked Education) will support component coding, data familiarisation and disadvantage-related feasibility work
- Professor Steve Higgins (Durham University, WhatWorked Education) will provide senior methodological oversight and continuity with EEF's wider synthesis standards.

Risks

Risk	Impact	Likelihood	Planned mitigation
Incomplete or inconsistent component coding across studies	High	Low	Implement a formal data freeze; include only fully coded studies in the primary analysis;

			maintain a cumulative update file for partially coded studies; document all coding decisions in a component dictionary.
Insufficient variation in component combinations to support component-level estimation	High	Medium	Conduct feasibility assessment before modelling; collapse conceptually related components where justified; report feasibility findings rather than forcing component estimates where evidence is insufficient.
Poor network connectedness preventing a credible cNMA	High	Low	Assess network structure and overlap; restrict analyses to defensible subnetworks where appropriate; pivot to feasibility analysis and recommendations if cNMA assumptions are not met.
Rare components generating unstable estimates	Medium	Medium	Analyse individual components only when they have non-zero weight in at least three study outcomes; otherwise combine into broader dimensions or exclude from component-level modelling.

Missing confidence intervals, standard errors, or other precision information	Medium	Medium	Recover missing information from reports where possible; use established meta-analytic methods to derive standard errors from available data; conduct sensitivity analyses.
Dependence between multiple outcomes reported within the same trial	Medium	Medium	Use multilevel modelling to account for within-study clustering; conduct sensitivity analyses using a single outcome per study.
Heterogeneity in outcome measures across studies	Low	High	Convert outcomes to a common effect-size metric; conduct subject-specific analyses where pooling across domains is not justified.
Project unable to answer substantive component questions because of dataset limitations	Medium	Low	Treat feasibility as a primary objective; document limitations systematically; produce recommendations for future commissioning, reporting and archive development.
Strong correlations between components leading to multicollinearity	Medium	Medium	Examine component co-occurrence patterns; combine highly overlapping components where theoretically and empirically justified; use regularised

			methods in exploratory analyses.
Findings interpreted as causal effects of individual components rather than associations	Medium	Low	Clearly state that findings are observational and exploratory; emphasise that components operate within broader intervention contexts; avoid prescriptive recommendations based solely on component estimates.
Over-interpretation of subgroup findings relating to disadvantage	Medium	Low	Treat subgroup analyses as exploratory unless supported by sufficient evidence; report uncertainty transparently; avoid definitive claims where data are sparse.
Potential disclosure of identifiable information	High	Low	Use only de-identified or pseudonymised data; restrict access to authorised researchers; report results only in aggregate form; comply with EEF data-sharing agreements and UK GDPR requirements.
Data security breach or unauthorised access	High	Low	Store data on secure institutional systems with access controls; limit access to named project staff; follow institutional

			information governance and data protection policies.
Selective reporting of favourable component findings	Low	Low	Pre-specify core analyses in the study plan; distinguish confirmatory and exploratory analyses; report all major modelling decisions and sensitivity analyses transparently.

Timeline

Dates	Activity	Staff responsible/leading
June 2026	- Project initiation meeting with EEF - Team roles agreed - Analysis plan drafted - Feasibility criteria refined	ED
August 2026	- Study protocol finalised - Analysis workflow completed - EEF data request submitted - Review of component coding handbook - Data familiarisation - Initial data cleaning - Feasibility assessment of component coding and primary outcomes	ED, RK, GU
September 2026	- Source additional study reports - Structured data extraction - Dataset completion - Effect-size harmonisation	AS, RK, GU
October 2026	Final data cleaning Decision Point 1: Final dataset agreed	Team

	Component frequency and correlation analyses Component dictionary (Version 1) Analysis dataset frozen and approved	
November 2026	• Baseline multilevel meta-analysis • Preliminary component network construction • Network connectedness diagnostics	RK, GU
December 2026	• Bayesian component modelling • Prototype additive component Network Meta-Analysis (cNMA) • Component interaction modelling	RK, GU
January 2026	• Meta-CART analyses • Moderator and disadvantage feasibility analyses • Sensitivity analyses	RK, GU
February 2027	• CNMA interpretation memo • Decision Point 2: Go / No-Go for interpretation and translation • Draft findings summary	ED, AS
March 2027	• Draft EEF briefing documents • EEF briefing meeting • Agree reporting structure and dissemination plan	ED
April-May 2027	• Draft full technical report • Draft technical appendix • Draft component dictionary (final version) • Internal quality assurance and statistical review	AS, GU
June 2027	• Peer review revisions • Incorporate EEF feedback • Finalise analyses and interpretation	ED
July 2027	• Prepare final report package • Final technical appendix • Final briefing documents • Knowledge mobilisation materials	ED

August- September 2027	• Final EEF report submitted • Final EEF briefing documents submitted • Technical appendix and analysis scripts submitted	ED
---------------------------------------	---	----

References

Clark, R. E., & Saxberg, B. (2012). The “active ingredients” approach to the development and testing of evidence-based instruction by instructional designers. *Educational Technology*, 52(5), 20–25.

<https://eric.ed.gov/?id=EJ1000524>

Education Endowment Foundation. (2024). *A school’s guide to implementation* (3rd ed.).

<https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/implementation>

Education Endowment Foundation (2025) *2025 Impact Report*. London: Education Endowment Foundation.

Education Endowment Foundation (2026) *Teaching and Learning Toolkit*. London: Education Endowment Foundation.

Ha, C. (2023). How to compare multiple intervention effects: Educational research using network meta-analysis. *International Journal of Research & Method in Education*, 46(1), 98–112. <https://doi.org/10.1080/1743727X.2022.2035352>

Mertens, E., Deković, M., Leijten, P., van Londen, M., & Reitz, E. (2020). Components of school-based interventions stimulating students’ intrapersonal and interpersonal domains: A meta-analysis.

Clinical Child and Family Psychology Review, 23(4), 605–631. [https://doi.org/10.1007/s10567-020-](https://doi.org/10.1007/s10567-020-00328-y)

[00328-y](https://doi.org/10.1007/s10567-020-00328-y)

Peng, P., Wang, W., Filderman, M. J., Zhang, W., & Lin, L. (2024). The active ingredient in reading comprehension strategy intervention for struggling readers: A Bayesian network meta-analysis. *Review of Educational Research*, 94(1), 210–240. <https://doi.org/10.3102/00346543231171345>

Peng, P., Wang, W., Lin, L., Liu, Y., Yan, X., Tan, Y., Zhang, Z., Zhang, W., & Huang, Y. (2025). A

network meta-analysis of multi-component reading interventions for students with reading difficulties: Active-ingredient vs. ingredient-interaction? Educational Psychology Review, 37, Article

85. <https://doi.org/10.1007/s10648-025-10063-x>

Pellegrini, M., Pigott, T. D., Chubb, C. S., Pruitt, N., & Scarbrough, H. F. (2025). Exploring the methodological quality of education intervention meta-analyses: A meta-review. Review of Educational Research. Advance online publication. <https://doi.org/10.3102/00346543251389938>

Petropoulou, M., Efthimiou, O., Rücker, G., Schwarzer, G., Furukawa, T. A., Pompoli, A., & Lesniak,

M. (2021). A review of methods for addressing components of interventions in meta-analysis. PLOS

ONE, 16(2), e0246631. <https://doi.org/10.1371/journal.pone.0246631>

Rücker, G., Petropoulou, M., & Schwarzer, G. (2020). Network meta-analysis of multicomponent

interventions. Biometrical Journal, 62(3), 808–821. <https://doi.org/10.1002/bimj.201800167>

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., & Bond, M. & Koryakina, A. (2023) EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. EPPI Centre, UCL Social Research Institute, University College London

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. BMJ, 374, Article n2061. <https://doi.org/10.1136/bmj.n2061>

Tsokani, S., Seitidis, G., & Mavridis, D. (2023). Component network meta-analysis in a nutshell.

BMJ Evidence-Based Medicine, 28(3), 183–186. <https://doi.org/10.1136/bmjebm-2021-111906>

Uwimpuhwe, G., Kang, S., Singh, A., Sayari, M., Akhter, N., Coolen-Maturi, T., & Einbeck, J. (2025). IPD meta-analysis: The impact of EEF-funded trials on the educational attainment of pupils with

English as an additional language (2011–2024) Education Endowment Foundation Report

https://d2tic4wvo1iusb.cloudfront.net/production/documents/tenders/WS4-EAL_IPD-meta-analysis_v.1.0.0.pdf?v=1768330855

Veroniki, A. A., Wolfe, D. M., Hutton, B., Schwarzer, G., McIsaac, D. I., Straus, S. E., Jackson, D., & Tricco, A. C. (2026). Analysing complex interventions using component network meta-analysis. *BMJ*, 393, Article e083075. <https://doi.org/10.1136/bmj-2024-083075>

Waddell, S., & Sharples, J. (2020). Developing a behavioural approach to knowledge mobilisation: Reflections for the What Works Network. Early Intervention Foundation and Education Endowment Foundation.

Appendix A

Study Workflows

Stage 1 Workflow

Study and arm identified

↓

Coding record available?

→ No: refer for coding or exclude from primary analysis

↓ Yes

All relevant dimensions reviewed?

→ No: classify as partial

↓ Yes

Zeros confirmed as absence?

→ No: seek clarification

↓ Yes

Report supports coding?

→ No: quality flag or exclusion

↓ Yes

Include in candidate analysis dataset

Primary outcome identified

↓

Suitable effect size and uncertainty reported?

→ Yes: extract and verify

→ No

↓

Sufficient report statistics available?

→ Yes: derive effect and standard error

→ No

↓

EEF trial dataset or technical material available?

→ Yes: recreate or recover estimate

→ No

↓

→ No: exclude from quantitative synthesis and retain as feasibility finding

Original report and archive data

↓

Identify preferred estimand and outcome

↓

Reported compatible effect size?

→ Yes: verify direction, scale and precision

→ No: derive from adjusted estimate or summary statistics

↓

Apply standardisation and small-sample correction where required

↓

Derive sampling variance or standard error

↓

Apply clustering/multi-arm adjustments

↓

Quality check

↓

Harmonised effect-size record

Coding sufficiently complete?

↓

Primary outcomes available?

↓

Precision recoverable?

↓

Preliminary variation and connectedness plausible?

↓

Yes: proceed to Stage 2

Study-specific failure: exclude affected records

Systemic failure: report feasibility findings and stop formal modelling

Stage 2 workflow

Components rare or highly correlated?

↓

Conceptually coherent?

→ No: retain separately, omit or report descriptively

↓ Yes

Can reports distinguish them reliably?

→ Yes: prefer original coding and test stability

→ No

↓

Would collapsing improve estimability without obscuring meaning?

→ No: omit individual estimates

→ Yes: create higher-order component

↓

Compare original and collapsed specifications in sensitivity analysis

Eligibility complete

- Coding quality checked
- Outcomes harmonised
- Component dictionary agreed
- Rare/overlapping component decisions documented
- Network scope provisionally defined

↓

Independent quality assurance

↓

EEF review

↓

Approve dataset and freeze / Revise and recheck / Restrict analytical scope

Stage 3 Workflow

Frozen arm-by-component matrix

↓

Apply frequency and variation threshold

↓

Fit separate weighted component model

↓

Assess uncertainty, heterogeneity and influence

↓

Repeat using binary or alternative prominence coding

↓

Classify estimate as interpretable, unstable or descriptive only

Original component codes

↓

Group components by taxonomy dimension

↓

Calculate arm-level dimension score

↓

Check score distribution and variation

↓

Fit weighted dimension model

↓

Compare with component-level findings

Network connected?

↓

Component estimates stable and sufficiently precise?

↓

Robust to coding and sample sensitivity analyses?

↓

Yes: full or restricted component interpretation

Component failure but dimensions viable: dimension-only interpretation

Network or data failure: feasibility findings only