Education
Endowment
Foundation

The impact of Feedback on student attainment: a systematic review

August 2021

Mark Newman (EPPI Centre, UCL Institute of Education)

Irene Kwan (EPPI Centre, UCL Institute of Education)

Karen Schucan Bird (EPPI Centre, UCL Institute of Education)

Hui-Teng Hoo (Nanyang Technological University, Singapore)

UCL
Institute of Education

EPPI Centre
Evidence for
Policy & Practice

Please cite this report as follows: Newman, M., Kwan, I., Schucan Bird, K., Hoo, H.T. (2021), *The impact of Feedback on student attainment: a systematic review,* London: Education Endowment Foundation

The EEF Guidance Report ***Teacher Feedback to Improve Pupil Learning*** is available at: https://educationendowmentfoundation.org.uk/public/files/Publications/Feedback/Teacher_Feedback_to_Improve_Pupil_Learning.pdf

Acknowledgements

The authors would like to thank the peer reviewers, Professor Steve Higgins (Durham University), Ian Shemilt and James Thomas for support with MAG searching, and Zak Ghouze and Sergio Graziosi for support with EPPI-Reviewer.

ISBN: 978-1-911605-28-7

Table of contents

List of tables

List of figures

Abstract

Meta-syntheses have reported positive impacts of feedback for student achievement at different stages of education and have been influential in establishing feedback as an effective strategy to support student learning. However, these syntheses combine studies of a variety of different feedback approaches, combine studies where feedback is one of a number of intervention components and have several methodological limitations (for example, the lack of quality appraisal of the included studies). There is also still more research needed to investigate the impact of different types of feedback on different students in different settings.

Objective

This systematic review was conducted at the request of the Education Endowment Foundation to provide more precise estimates of the impact of different types of feedback in different contexts for different learners aged between 5 and 18. The review analysis sought to explore potential variations in the impact of feedback through subgroup analysis of the characteristics of the feedback, the educational setting, the learners and the subject. This review provides evidence that can be used to support the development of guidance for teachers and schools about feedback practices.

Methods design

A systematic review was undertaken in two stages. First, a systematic map identified and characterised a subset of studies that investigated the attainment impacts of feedback. Second, an in-depth review comprising of a meta-analysis was performed to answer the review questions about the impact of interventions that comprised of feedback only and to explore the variety of characteristics that may influence the impact of feedback.

Methods search

We used the Microsoft Academic Graph (MAG) dataset hosted in EPPI-Reviewer to conduct a semantic network analysis to identify records related to a set of pre-identified study references. The MAG search identified 23,725 potential studies for screening.

Methods study selection

Studies were selected using a set of pre-specified selection criterion. Semi-automated priority screening was used to screen the title and abstract of studies using bespoke systematic review software EPPI-Reviewer. The title and abstract screening was stopped after 3,028 studies and 745 were identified for full-text screening. Reviewers carried out a moderation exercise, all screening a selection of the same titles to develop consistency of screening. Thereafter, single reviewer screening was used with referral for a second reviewer opinion in cases of uncertainty.

Methods data collection

Studies were coded using a bespoke data extraction tool developed by the EEF Database Project. Study quality was assessed using a bespoke risk of bias assessment adapted from the ROBINS-I tool. The review team undertook a moderation exercise coding the same set of studies to develop consistency. Thereafter, single reviewer coding was used, based on the full text with referral for a second opinion in cases of uncertainty.

Methods synthesis

Data from the studies was used to calculate standardised effect sizes (Standardised Mean Difference- Hedge's *g*). Effect sizes from each study were combined to produce a pooled estimate of effect using Random Effects Meta-analysis. Statistical Heterogeneity tests were carried out for each synthesis. Sensitivity analysis was carried out for assessed study quality. Subgroup analysis was completed using meta-analysis to explore outcomes according to the different characteristics of feedback, context and subjects.

Main results

The full text screening identified 304 studies to include in the initial systematic map, of which 171 studies investigated feedback only. After applying final selection criteria, 43 papers with 51 studies published in and after the year 2000 were included. The 51 studies had approximately 14,400 students. Forty studies were experiments with random allocation to groups and 11 were prospective quantitative experimental design studies. The overall ecological validity was assessed as moderate to high in 40 studies and the overall risk of bias assessed as low to moderate in 44 studies.

The interventions took place in curriculum subjects including literacy, mathematics, science, social studies, and languages, and tested other cognitive outcomes. The source of feedback included teacher, researcher, digital, or automated means. Feedback to individual students is reported in 48 studies and feedback to group or class is reported in four studies. Feedback took the form of spoken verbal, non-verbal, written verbal, and written non-verbal. Different studies investigated feedback that took place immediately after the task, during the task and up to one week after the task (delayed feedback). Most of the feedback interventions gave the learner feedback about the outcome and the process/strategy. Some provided feedback on outcome only and two provided feedback about process /strategy only.

On the main research question, the pooled estimate of effect of synthesis of all studies with a low or moderate risk of bias indicated that students who received feedback had better performance than students who did not receive feedback or experienced usual practice (g = 0.17, 95% C.I. 0.09 to 0.25). However, there is statistically significant heterogeneity between these studies (I^2 = 44%, Test for Heterogeneity: Q(df = 37) = 65.92, p = 0.002), which suggests that this may not be a useful indicator of the general impact of feedback on attainment when compared to no feedback or usual practice.

The heterogeneity analysis suggested considerable heterogeneity between studies in the main synthesis and all the subgroup synthesis, and in the majority of the cases the heterogeneity is statistically significant. This means caution is required when considering the results of the synthesis. The results of the subgroup synthesis suggest that a variety of student and context factors may have an effect on the impact of feedback.

Conclusions

The results of the review may be considered broadly consistent with claims made on the basis of previous synthesis and meta-synthesis, suggesting that feedback interventions, on average, have a positive impact on attainment when compared to no feedback or usual practice. The limitations in the study reports and the comparatively small number of studies within each subgroup synthesis meant that the review was not able to provide very much more certainty about the factors that affect variation in the impact of single component feedback interventions within different contexts and with different students. More research is needed in this area to consider what may moderate the impact of feedback.

However, the findings further support the conclusion made by previous studies that feedback, on average, has a positive impact on attainment; moreover, this is based on a more precise and robust analysis than previous syntheses. This suggests that feedback may have a role to play in raising attainment alongside other effective interventions.

Findings were further interpreted by a panel of expert practitioners and academics to produce the EEF's *Teacher feedback to improve pupil learning* guidance report.

1. Background and review rationale

Feedback can be defined as information communicated to the learner that is intended to modify the learner's thinking or behaviour for the purpose of improving learning.[1] Meta-syntheses have reported positive impacts of feedback, with effect sizes ranging from *d* = 0.70 to *d* = 0.79 for student achievement at different stages of education[2] and have been influential in establishing feedback as highly effective with regards to student learning. For example, the EEF Teaching and Learning Toolkit meta-synthesis suggests that feedback may have 'very high' impact (equivalent to eight months' additional progress) for relatively low cost.[3]

However, caution is necessary when interpreting the findings of these meta-syntheses for a number of reasons. Firstly, the average effect size reported in the EEF Toolkit is based on combining the estimates from existing meta-analyses of individual studies, which may contain limitations of various kinds (see the list below for examples) that may mean that average effect sizes identified are overestimates. Second, some studies included in syntheses (such as Kluger and DeNisi's meta-analysis[4]) suggest that some feedback interventions may, in fact, negatively impact pupils. Third, previous meta-syntheses have not explored in detail the impact of potential moderating factors, such as different types of feedback. As Ekecrantz has argued, there is still a need to better understand how and under what circumstances teacher feedback on student performance promotes learning as well as to question the generalised claim (that feedback improves attainment) itself.[5]

For example, a recent meta-analysis that re-analysed studies included in the original synthesis by Hattie and Timperley[6] revised down the average effect size from the estimates of the effects of feedback from their originally published Standardised Mean Difference of *d* = 0.79 to *d* = 0.48.[7] In the revised meta-analysis, 17% of the effect sizes from individual studies were negative. The confidence interval ranged from *d* = 0.48 to *d* = 0.62, and the authors found a wide range of effect sizes. Different moderators were also investigated to explore the impact of different characteristics of context and feedback. Whilst this meta-analysis offers improvements over previous meta-syntheses, it has a number of limitations, including:

- It only included studies drawn from 36 existing meta-analyses, the most recent of which was published in 2015. Eligible studies published after 2015 or not included in these meta-analyses would not have been included.
- All comparative study designs were included. Less robust study designs may have overestimated the positive effect of feedback.
- There was no reported study quality assessment/moderation or sensitivity analysis, which may have led to an overestimation of the pooled effect sizes.
- The meta-analyses included studies with high levels of heterogeneity, $I^2 = 80\%$ or more (in the main and moderator analysis). This suggests that the synthesis may be combining studies/comparing feedback practices inappropriately.
- The meta-analysis did not consider all potentially relevant moderating factors. It may also be the case that the impact of feedback depends on factors other than those analysed, including the ability of the learner, the learning context, and/or the frequency, duration, timing, and type of feedback.

This systematic review was conducted at the request of the EEF to try and provide more accurate and precise estimates of the impact of different types of feedback in different schooling contexts. The review examines the impact of single component feedback, in different contexts, and for different learners with a greater degree of granularity and precision than is currently available via the EEF Teaching and Learning Toolkit strand on 'Feedback'. For EEF, the purpose of the systematic review is to provide evidence that can be used to inform guidance for teachers and schools about effective feedback practices.

[1] Shute V.J. (2007). *Focus on Formative Feedback*. Research Report RR-07-11. Princeton NJ. Education and Testing Service.
[2] Hattie, J. (2009). *Visible Learning: A Synthesis of 800+ Meta-Analyses on Achievement*. London: Routledge; Hattie, J. and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res*. 77, 81–112; Hattie, J. and Zierer, K. (2019). *Visible Learning Insights*. London: Routledge.
[3] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/feedback/
[4] Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.
[5] Ekecrantz S. (2015). Feedback and student learning—A critical review of research. *Utbildning & Larande* 9(2) pp15-32.
[6] Hattie, J. and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res*. 77, 81–112.
[7] Wisniewski, B., Zierer, K. and Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Front. Psychol*. 10:3087.

The systematic review methods and processes were developed and carried out conterminously with the EEF Database project with a view to facilitating the future use of the produced resources and supporting the ongoing work of the Database project.

1.1 Domain being studied: Feedback approaches

This review focuses on interventions that provide feedback from teachers to learners in mainstream educational settings. Feedback is defined in accordance with the EEF toolkit definition:[8]

> *'Feedback is information given to the learner and/or teacher about the learner's performance relative to learning goals or outcomes. It should aim to produce (and be capable of) producing improvement in students' learning. Feedback redirects or refocuses either the teacher's or the learner's actions to achieve a goal, by aligning effort and activity with an outcome. It can be about the output of the activity, the process of the activity, the student's management of their learning or self-regulation, or them as individuals. This feedback can be verbal or written or can be given through tests or via digital technology. It can come from a teacher or someone taking a teaching role, or from 'peers'.'*

This initial broad definition, whilst conceptually coherent, does create challenges both in practice for teachers and in terms of identifying and distinguishing between practices when considering research evidence. For example, what is the difference between small group learning and 'peer feedback'? It seems perfectly reasonable to assume that small group learning must contain conversations between students about their work and the task they have been asked to complete and thus is 'feedback'. However, in practice, this may not be what teachers think of as 'feedback' and in the research literature, 'small group learning' is investigated both as a unique pedagogical strategy and as a component of a number of other pedagogical strategies.

As the development of the understanding of the scope of the review evolved, the working definition of feedback for the review became modified practically through the exclusion of certain categories of intervention, even though they may contain an element of feedback practice. The inclusion criteria in the methods section outlines the revised definition that the review team used.

1.2 Conceptual framework/Theory of Change

There are several ways in which feedback is conceptualised as improving learner performance—i.e. as a Theory of Change. The 'Feedback' strand in the EEF Teaching and Learning Toolkit draws most explicitly on the conceptualisation of Hattie and Timperley's (2007) model. This model emphasises the importance of systems of feedback where the teacher provides feedback to the specific needs of individual students. The searching processes used in this review are consistent with this model as the studies used in the Feedback strand of the EEF Teaching and Learning Toolkit were used to 'seed' the search. However, they did not preclude the inclusion of studies that may draw on other 'models' of feedback which, though similar to Hattie and Timperley (2007), may be argued to place more emphasis on, for example: developing learner self-regulation (Nicole and Macfarlane-Dick, 2006); students' intrinsic motivation (Dweck, 2016); and/or are subject specific—for example, 'Thinking Mathematically' (Mason, Burton and Stacey, 2010). The coding tools used in the review were informed by the model (in terms of coding about the source and content of the feedback; see Appendix 3).

1.3 Review design

A systematic review approach was used to investigate the research questions. The review was undertaken in two stages. First, a systematic map identified and described the feedback characteristics of a subset of studies that investigated the attainment impacts of feedback. The map was used to make decisions about focusing the analysis in the second in-depth systematic review stage. At the second stage an in-depth review, including meta-analysis, was performed on a subset of the studies identified in the map to answer the review questions and explore the variety of intervention and context characteristics that may influence the impact of feedback.

[8] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/feedback/technical-appendix/

This systematic review was designed to complement the work of the EEF Database project. The EEF Database project is currently undertaking a programme to extract and code the individual studies from the meta-synthesis used in the EEF Teaching and Learning Toolkit.[9] The search strategy used in this review was 'seeded' from studies identified as being about 'feedback' in the database, and this systematic review used the coding tools developed by the Database team (see Appendix 3). The studies newly identified in this review will be subsequently included in the EEF Database.

This systematic review was also designed to provide additional research evidence for use in guidance on feedback developed for schools produced by the EEF, and therefore to fit with a particular time window for the review's production. The results of the meta-analyses were presented to an advisory panel of academics and teaching practitioners, who used the results, their own expertise, a review of practice undertaken by the University of Oxford,[10] and conceptual models (such as Hattie and Timperley) to draft recommendations for practice.

[9] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/
[10] Elliott, V. *et al* (2020). *Feedback in Action: A review of practice in English schools*. Department of Education, University of Oxford, Education Endowment Foundation.

2. Objectives

2.1 Systematic map research question

What are the characteristics of the research using counterfactual designs measuring the attainment impacts of feedback interventions/approaches in mainstream schools?

2.2 Systematic review research question

What is the difference in attainment of learners aged 3–18, receiving a single component feedback intervention/approach in comparison to learners receiving 'the usual treatment' (with regard to feedback practices in the setting)/no feedback?

Given the large number of studies identified, a pragmatic decision was taken based on the initial mapping of the literature at stage 1, that in order to complete the review within a given time frame and resources (September 2020 to March 2021), the in-depth review would focus on studies published post-2000, in which the feedback was the only intervention component, the sources of feedback were teacher, researcher and/or digital/automated feedback, and which only focussed on learners aged between 5–18 years old. Thus the research question for the completed in-depth review is:

What is the difference in attainment of learners, aged 5–18, receiving a single component feedback intervention/approach from a teacher/researcher/digital/automated source in comparison to learners receiving 'the usual teaching'/no feedback?

The review analysis explored through subgroup analysis potential variations in the impact of feedback on attainment through the following factors:

- the source of feedback (e.g. teacher, researcher, digital/automated);
- whether feedback is given to the individual student or to a group (e.g. individual, class);
- how the feedback is delivered (e.g. verbal, written);
- when the feedback is provided (e.g. prior, during, immediate, delayed (short), delayed (long));
- the content of the feedback (e.g. about outcome, process/ strategies) ;
- the characteristics of the educational setting (phase of schooling); and
- characteristics of the subject (e.g. maths, science, literacy).

The review had initially intended to answer additional questions; however, it did not identify enough evidence to address questions about:

- the tone of the feedback (positive, negative, neutral);
- providing feedback on correct answers or incorrect answers; or
- the impact of feedback on learners with different characteristics—e.g. age, gender, disadvantage, level of prior attainment.

3. Methods

The full protocol for the review can be found on the EEF website.[11]

3.1 Inclusion and exclusion criteria for the review

The inclusion criteria for the first stage of the review are set out below in Table 1. These selection criteria are those used in the EEF Database project. The criterion for 'feedback intervention' was developed for this project based on the EEF Database project definition of feedback above. There are no restrictions on the eligibility of studies to be included in the review beyond those described in the table—i.e. empirical research studies published in any format from anywhere in the world investigating any kind of feedback can be included, providing all other criteria are met.

Table 1: First stage systematic map selection criteria

Criteria	Included	Excluded
Population	The majority of the sample (>50%) on which the analysis is based are learners or pupils aged between 3–18 (further education or junior college students are be included where their study is for school level qualifications).	The majority of the sample is post-secondary education; in higher education; adults; infants under 3; other students over 18.
Intervention	*An educational intervention or approach, recognisable as feedback that aims to help the learner improve their performance: (I) Source: Feedback can be provided by a teacher or person acting in the teaching role (such as teaching assistant), parent/carer or other family members, or peers. Feedback can be digital or otherwise automated or generated by the learner. (II) Form: Feedback can take the form of spoken, written or non-verbal statements. (III) Kind: Feedback can focus on the learner's academic performance/outcome, the process, the learner's strategies/approach or about the learner. Feedback includes praise and rewards.	Intervention or approach is not recognisable feedback: (I)Consists of only feedback on behaviour. (II) Student performance data given only to the teacher. (III) The study/intervention is Mastery Learning. (IV) The study intervention is Tutoring. (V) The study intervention is a type labelled as 'learning strategy'. (VI) The study intervention is aimed at developing metacognition/self-regulation.
Setting	The intervention or approach is undertaken in a mainstream educational setting or environment for the learners involved, such as a nursery or school or a typical setting (e.g. an outdoor field centre or museum).	(I) Laboratory studies: Children are removed from classroom or school to specially created environments (both physical and virtual). (II) The setting is EFL/ESL learning outside the UK.
Comparison	Receiving 'treatment' as usual, no feedback or an alternative intervention.	No comparison.
Study design	A valid (see exclusion criteria) counterfactual comparison between those receiving the feedback intervention or approach and those not receiving it.	Single group and single subject designs where there is no control for maturation or growth.

[11] https://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Systematic_Review_of_Feedback._M_Newman._Dec_2020b._Protocol.pdf

Criteria	Included	Excluded
Outcomes	Assessment of educational or cognitive attainment/achievement which reports quantitative results from testing of attainment/achievement or learning outcomes, such as by standardised tests, other appropriate curriculum assessments, school examinations, or appropriate cognitive measures.	No quantitative outcomes measured. Purely qualitative outcomes.
Language	English only	Not published in English
Publication date	Post 1960**	Prior to 1960

**Review specific based on the EEF Database definition of feedback given above.*

*** The EEF Teaching and Learning Toolkit Database currently does not contain any studies before 1960. On this basis we have selected this cut-off date for selection.*

3.2 Search strategy for identification of studies

Our initial search strategy included five strands:

- an automated electronic search using Microsoft Academic Graph (MAG);
- a conventional search of the *ProQuest Dissertations and Theses Global* database;
- forwards and backwards citation searches;
- related publications searches; and
- contacting experts.

The results of the MAG database search and initial screening yielded a high number of potential study includes (see further details below). Therefore, in order to complete the review in the set timeline, we had to adopt the revised strategy using only the MAG database.

We used a semi-automated study identification workflow, powered by the MAG dataset and hosted in EPPI-Reviewer.[12,13] The MAG dataset currently comprises 240 million bibliographic records of research articles from across science, connected in large network graph of conceptual and citation relationships. MAG records include abstracts and (often multiple) links to online full-text sources, when available. We used MAG to conduct a semantic network analysis to identify records related to a set of pre-identified study references.

The 'SEED' source used comprised of three sets of 'MAG Matched' records:

- all studies included in meta-analysis that are used in six strands of the EEF Teaching and Learning Toolkit (n = 2066 records);
- all studies included in meta-analysis in the EEF Teaching and Learning Toolkit feedback strand (n = 1025 records); and
- a corpus of n = 144 unique study reports that were selected by the EEF Database team from the above group as eligible for this review.

Semantic network analysis was then used to identify related MAG records in 'one-hop' ('proximal') or 'two-hop' ('extended') networks citation and/or 'related publications' relationship with one or more of the 'seed' records.[14]

[12] Shemilt I. and Thomas J. MAG-Net-ise it! How the use of Microsoft Academic Graph with machine learning classifiers can revolutionise study identification for systematic reviews. Oral paper accepted for presentation at the 26th Cochrane Colloquium, Santiago, Chile, 22–25 October 2019.

[13] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews 4:5. doi:10.1186/2046-4053-4-5

[14] Shemilt, I. and Thomas, J. MAG-Net-ise it! How the use of Microsoft Academic Graph with machine learning classifiers can revolutionise study identification for systematic reviews. Oral paper accepted for presentation at the 26th Cochrane Colloquium, Santiago, Chile, 22–25 October 2019.

3.3 Screening (study selection)

A screening training and moderation exercise was completed whereby the EPPI-Centre team 'rescreened' a random selection of the studies included and excluded by the EEF database team at Durham. Screening was undertaken by all members of the EPPI-Centre review team. Each study was screened by a single team member initially. A study may have been rescreened by a second team member in case of a selection query and/or at a later stage in the review process.

The MAG search identified 23,725 potential studies for screening. Manual screening of records retrieved from the MAG dataset was conducted using 'priority screening' mode in EPPI-Reviewer. 'Priority screening' mode utilises 'active learning', which involves periodic automatic reprioritisation of the rank-ordered lists of 'new' candidate records by a machine learning classifier, based on all preceding title and abstract eligibility screening decisions made by the researchers (also 'seeded' by our corpus of 'known includes') in each workflow.[15] The retrospective simulation study estimated that approximately 5,000 of these (i.e. the first 5,000 in priority order) would need to be screened to identify all the studies meeting the review selection criteria.

We also monitored the screening using 'screening progress' record in EPPI-Reviewer, to inform a pragmatic decision about when to truncate screening (within available resources). In consultation with the stakeholders, the review team managed the dynamic process of completing the review within a defined deadline.

3.4 Data extraction

Studies were coded using the EEF's Database 'Main', 'Effect Size' and 'Feedback coding frames (see Appendix 3). This coding was carried using the EPPI-reviewer systematic review software tool. Where an individual paper reported more than a single study, each study was coded separately and recorded individually in any relevant analyses. The review team undertook a coding moderation exercise prior to coding where all of the team coded the same studies and compared results. Thereafter studies were coded by one team member and referred to a second team member where there were any queries.

3.5 Stage 2: In-depth review

The full text screening initially identified 304[16] studies to include in the initial systematic map. The first stage of coding coded the studies for whether or not the intervention was feedback (or variations of feedback) only or feedback and other components. The second stage of coding for the complete systematic map was carried out on the 171 studies that investigated feedback only interventions. The studies were coded on the following characteristics:

- What was the educational setting?
- What was the source of the feedback?
- Who was the feedback directed to?
- What form did the feedback take?
- When did the feedback happen?
- What kind of feedback was provided?

Given the large number of studies, a pragmatic decision was taken in order to complete the review within a given time frame and resources. The in-depth review focused on feedback only studies published post-2000, in which the sources of feedback are teacher, researcher and/or digital/automated feedback.

The research question for the in-depth review was:

[15] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews 4:5. doi:10.1186/2046-4053-4-5

[16] The descriptive map was produced under dynamic conditions during the review process to inform the focus of the second stage in-depth review. A number of studies that were initially included in the map were subsequently excluded from the map and in-depth review after further scrutiny of the paper revealed that they did not meet a review inclusion criteria. Other studies were added to the map/review after the initial map report as they were subsequently identified in the coding process.

What is the difference in attainment of learners, aged 5–18, receiving a feedback only intervention/approach from a teacher/researcher/digital/automated source in comparison to learners receiving 'the usual treatment' (with regard to feedback practices in the setting)/no feedback or an alternative approach?

3.5.1 Stage 2 selection criteria

Following the focusing of the research question, a further selection process was undertaken on studies that met the first stage screening criteria to select studies for the in-depth review based on the following second stage criteria:

- Feedback is the only component of the intervention being investigated
- The source of feedback is either the teacher, researcher, or digital/automated.
- The study takes place in mainstream educational setting among 5 to 18 year olds.
- The study was published after 2000.

3.6 Selecting outcomes and calculating effect sizes

The outcomes specified as of interest for the review were educational attainment, which is defined as some kind of curriculum-related test or assessment (45 studies), or where this was not measured in the study, a measure of non-curriculum-based test of cognition (six studies). Where attainment outcome measures were present, all were data extracted and cognitive measures were not coded even if present. The first focus of outcome data extraction was to code descriptive or statistical data that could be used to calculate a standardised effect size such as Hedges *g*, e.g. Means, Standard Deviations, Group size, F value, P value, T value, Proportions. If study authors reported a standardised effect size then this was used. Where study outcomes are only reported for separate groups (e.g. for males and females), mean and standard deviation for a combined group were calculated using Cochrane guidance.[17] In study outcomes that were measured as reduction in negative outcome (e.g. errors), these were recoded to match direction of effect for positively framed measures. Where data for the calculation of outcomes was not reported by study authors, record was made of study author conclusions about the result for that outcome. Standardised effect sizes (Hedges *g)* were calculated using the EPPI Reviewer[18] or the Campbell Collaboration Effect Size calculator.[19] The one study that reported binary outcomes was also translated to Hedges *g* using the Campbell Collaboration Effect Size calculator.

3.7 Study quality assessment

The use of the pre-existing EEF database coding tools for this review required the development of a bespoke study quality assessment tool that utilised the information already coded using the EEF database tools. The development of the study quality assessment tool was shaped by two concerns that are relevant to review users. Firstly, about attributing study outcome to the effect of the feedback intervention, and secondly, about the applicability of the results to the context of mainstream UK schools. The review has been designed to optimise both possibility of making claims about the impact of feedback and to maximise the potential relevance of the evidence to mainstream schools through both the search process and the selection criteria used. However, given the diversity of studies that could be included, there was still a need to provide further information and judgement about a study's quality and relevance.

The review only included studies in which the researchers had created conditions to support a causal claim (i.e. a comparison between feedback and no feedback/usual practice). However, even with this condition, it is still necessary to judge whether the comparison represents a fair test. In the research field the problem of attributing causal impact is considered in terms of threats to validity or bias. Therefore, we developed the tool with reference to various factors that may influence the outcome of a study and thus be 'risks of bias'. Given the high prevalence of quasi-experimental

[17] https://handbook-5-1.cochrane.org/chapter_7/table_7_7_a_formulae_for_combining_groups.htm.

[18] Effect Size calculations and meta-analysis functions are based on the 'metafor' package in R. Viechtbauer W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1-48. Additonal sources used for functions are Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R. (2009). Subgroup analyses. In: Introduction to Meta-Analysis. John Wiley & Sons, Ltd, pages 59-86; and Deeks, J.J., Douglas, A.G. and Bradburn, M.J. (2001). Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger, M., Davey Smith, G.; Altman, D.G. Systematic Reviews in Health Care: Meta-analysis in Context. London: BMJ Publishing Group.

[19] https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html.

studies in the education field, we used the ROBINS-I tool[20] as a point of reference to construct a bespoke risk of bias assessment tool based on the coding questions available in the EEF data extraction tool. The study quality assessment tool produces an overall risk of bias rating for each study—low, moderate, or serious risk of bias. In general, the greater the risk of bias the less confident we would be about a causal attribution claim in a study. In terms of impact on study outcomes, we would expect to see larger positive effects in higher risk studies and vice versa.

The issue of study relevance is sometimes referred to as Ecological Validity. This is essentially a question of 'would the same results be achieved in a different setting?' This is rather difficult to judge given the complexity and variation in settings both in the original study and in any potential setting of application. The review was designed to identify and select studies that are potentially relevant through the focus on studies in school settings. The review takes the perspective that beyond this the question of relevance is most reasonably judged by experts in the context. Therefore, the assessment of ecological validity is limited to two elements: 'Who was responsible for teaching at the point of delivery?' and 'What was the source of feedback?'

The Study Quality Assessment tool can found in Appendix 4. The review team undertook a coding moderation exercise prior to undertaking the study quality assessment where all members of the team coded the same studies and compared results. Thereafter, studies were coded for study quality assessment by one team member.

3.8 Data synthesis

Quantitative synthesis using statistical meta-analysis was carried using the following procedures:

3.8.1 Selection of outcome measures for inclusion in meta-analysis

Where a study reports more than one outcome, this could be for a number of reasons—for example, different measures of the same outcome, a science test with multiple parts, groups exposed to different intervention characteristics, and/or different curriculum subjects tested. Every relevant outcome (i.e. that met the inclusion criteria) was coded. An important principle of meta-analysis is that the same subjects cannot appear more than once in the same meta-analysis. So it is highly unlikely that more than one outcome from a study will be included in the same meta-analysis. The following rules were used when selecting outcomes in these circumstances:

- Select the outcome appropriate for the synthesis question—e.g. if the question is about digital feedback, select an outcome from a digital feedback group compared to control.
- Use post-test only outcomes
- Select (or create) an outcome for combined groups (where results are reported in subgroups).
- Where there is more than one effect size recorded in a study for a particular outcome, use the effect size closest to zero whether positive or negative.[21]

In addition to the above for sub group synthesis

- If the outcome is measured in a general assessment and curriculum subject, then select that curriculum subject for the synthesis (e.g. maths).
- Where there is an immediate and a delayed post-intervention test use as appropriate to the synthesis,

3.8.2 Meta-analysis

The meta-analysis combined standardised effect sizes from each study (Standardised Mean Difference (SMD) Hedges *g*) to compute an overall point estimate of effect. The interpretation of SMD has two elements: the direction and size of effect. The point of 'no effect' (no difference between groups) is indicated by the value $g = 0$. Values less than zero indicate that the control (no feedback) group had a better outcome than the intervention (feedback) group.

[20] Sterne, J.A.C., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., Viswanathan, M., Henry, D., Altman, D.G., Ansari, M.T., Boutron, I., Carpenter, J.R., Chan, A.W., Churchill, R., Deeks, J.J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y.K., Pigott, T.D., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F., Higgins, J.P.T. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. BMJ 2016; 355; i4919; doi: 10.1136/bmj.i4919.

[21] The review team felt that where a single study produced results with different effect sizes, then at the very least this was indicative of the outcome being sensitive to factors within the study. Therefore a cautious approach was preferable when selecting effect sizes for inclusion in a synthesis from such a study.

Values greater than zero indicate that the intervention group had a better outcome than the control group. The larger the effect size (positive or negative) the bigger the difference in outcome between the groups.

The analysis also includes an estimate of the precision of the point estimate in the form of 95% confidence intervals (C.I). For practical purposes, this can be thought of as the probable range in which the 'true' result lies. The narrower this range, the more accurately the point estimate of effect is as an indicator of the 'true' effect size. A key issue for interpretation is whether the 95% C.I range crosses the value $g = 0$ (no effect). If it does then the interpretation is that we are not confident of excluding the opposite effect to that indicated by the point estimate.

Effect sizes from individual studies were combined using Random Effects Model Meta-analysis. Each synthesis included a statistical assessment of heterogeneity between studies. The I^2 statistic provides a value between 1% and 100%, with 100% being high. The higher the value the greater the statistical heterogeneity between studies. There will always be some heterogeneity between studies. The statistic is an indicator that signals the degree to which there might be 'real' heterogeneity between studies that is impacting the outcomes and which may mean that studies are not sufficiently similar to make the pooled estimate of effect size a useful or valid indicator of the general impact of feedback. There are many potential causes of real study heterogeneity, one of which could be study design, so a sensitivity analysis using the risk of bias assessment was completed for each synthesis where relevant. Other study characteristics may also affect study outcomes—for example, the characteristics of the sample, settings or feedback—and these were explored through the subgroup analysis.

4 Search results

The search identified 23,725 potential studies for screening. The screening was carried out dynamically and simultaneously through all stages of the review with a view to ensuring that the workload of review processes could be managed within the required review deadlines. This meant that the title and abstract screening was stopped after 3,028 studies had been screened and 745 potentially includable studies had been identified for full text screening.

During the screening, the review team identified that many of the interventions in the studies appeared to have actions in addition to feedback. The components in addition to feedback varied in the different studies but included amongst other actions: instruction of various kinds, guided practice, inclusion techniques, peer feedback (in addition to teacher feedback), and others. Therefore the first stage of coding identified whether or not the intervention was feedback (or variations of feedback) only or feedback and other components. The second stage of coding of feedback characteristics was carried out on the 171 studies that investigated feedback only. After applying the final selection criteria, the in-depth review included 51 eligible studies reported in 43 published papers.

The flow of studies is reported in the diagram in Appendix 1. The dynamic screening process and the involvement of two teams in the screening process (the database team and the review team) meant that studies continued to be excluded throughout the review process as the review team looked at papers in more detail. Similarly, multiple studies within the same papers were identified and screened at different points in the process of the review. Studies excluded or added at later stages of the review were not retrospectively recoded and therefore the data for the number of studies is not precise at all stages of the review. The numbers where this is the case are shown in the boxes in red in Appendix 1.

5 Results of effectiveness review

5.1 Definitions

The feedback practices that were included in the in-depth review had to be:

- from the teacher/researcher/digital or other technology to the student;
- delivered to 5–18 year olds;
- feedback about process/strategy or outcome;
- the only component of the intervention investigated in the study (i.e the 'test' is feedback compared to no feedback or usual practice); and
- reported in studies conducted in 2000 or after.

Some examples of the practices investigated in the studies are given in Table 2.

Table 2: Examples of feedback practices from included studies

- Curriculum-Based Measurement Written Expression (CBM-WE) probes are brief, timed (four-minute) assessments that look at a student's mastery of writing mechanics and conventions. The student is given a 'story starter'—a brief introductory story stem that serves as a stimulus for the student to create his or her own writing sample. Fourth grade students in the intervention group were provided both with (a) feedback from their teachers regarding their performance on CBM-WE probes and (b) new weekly goals (Alitto *et al*, 2016).
- Students in a mainstream secondary school in North East England undertook a cognitive ability test on two occasions. In one condition, students received item-specific accuracy feedback while in the other (standard condition) no feedback was provided (Beckmann; Beckmann and Elliott, 2009).
- A computer tutor that offers a supportive context for students to practice summary writing, guiding them through successive cycles of revising with feedback on the content of their writing. Automatic evaluation of the content of student summaries is enabled by Latent Semantic Analysis (LSA) (Franzke *et al*, 2005).
- In the intervention group, before starting the teaching unit, the teachers received an overview of their students' prior knowledge of Pythagoras as assessed in the pretest. The teachers assessed students' performance at the end of each phase at three predefined points in time (in the 5th, 8th, and 11th lessons) and provided students with written process-oriented feedback in the following lesson using the diagnostic and feedback tool developed (Rakoczy; Pinger and Hochweber, 2018).

5.2 Description of the evidence base

We identified 51 studies, published in or after 2000, to be included for the review. Five studies (Brosvic *et al,* 2006—Experiment 1a; Brosvic *et al*, 2006—Experiment 1b; Brosvic *et al*, 2006—Experiment 2; Dihoff *et al*, 2005—Experiment 1; Golke, Dörfler and Artelt, 2015—Experiment 1) did not provide usable data to compute effect sizes and thus could not be included in the meta-analyses. The remaining 46 studies involved approximately 14,400 participants. Details of each study are presented in the table of characteristics and study quality in Appendix 2.

The descriptive characteristics of the evidence base of included studies are given in tables 3 to 17 below. The number of studies referred to in the tables may differ from that used in the synthesis reports in the following section because not all studies reported data to calculate effect sizes and/or where synthesis included only studies with a low or moderate risk of bias. The number of the studies in the systematic map (from stage 1) is given for the characteristics coded at that stage.

Table 3: Characteristics of included studies—Year of publication

Year of publication	No. of studies
2000–2005	10
2006–2010	11
2011–2015	13
2016–2020	17

Table 4: Characteristics of included studies—Country where study completed

Country	No. of studies
UK	3
US	30
Belgium	1
Germany	5
Indonesia	1
Latvia	1
The Netherlands	2
Nigeria	2
Slovakia	1
Spain	3
Switzerland	1
Taiwan	1

The selection criterion for inclusion in the study required that the attainment of a group of students receiving feedback was compared to a groups of students not receiving feedback/usual practice. This meant that only comparative study designs were included in the review. These studies were coded as either experiments with random allocation to groups (Randomised Controlled Trial) or Prospective Quantitative Experimental Designs, as shown in Table 5 below.

Table 5: Characteristics of included studies—Study design

Study design	No. of studies
Randomised Controlled Trial	40
Prospective Quantitative Experimental design	11

Table 6: Characteristics of included studies—Educational Setting

Educational settings	No. of studies
Nursery school/pre-school	2*
Primary/elementary school	24
Middle school	7
Secondary/high school	18

*participants UK primary age

Table 7: Characteristics of included studies—Age of study participants

Age (*not mutually exclusive*)	No. of studies
4	1
5	3
6	4
7	6
8	12
9	10
10	8
11	7
12	12
13	12
14	13
15	7
16	2
17	1
No information provided	8

Table seven above shows the ages of participants in the studies. This is the age of students as provided by the authors and/or where reviewers could work out the age based on information about the school year group ages in the educational system of the country where the study took place.

The studies were coded for gender/sex of participants as described by the authors. Where the information was provided, all studies included both male and female participants. Study outcome data (required for calculating effect sizes) were not reported separately for males and females.

Table 8: Study characteristics—Gender/sex of participants

Gender/sex	No. of studies
Mixed gender	45
No information provided	6

Table 9: Characteristics of studies—Curriculum subjects

Curriculum subjects tested (*not mutually exclusive*)	No. of studies
Literacy (total)	23
Literacy: reading comprehension	14
Literacy: decoding/phonics	2
Literacy: spelling	2
Literacy: reading other	2
Literacy: speaking and listening/oral language	2
Literacy: writing	11
Mathematics	17
Science	7
Curriculum: social studies	1
Languages	2
Others/cognitive outcomes	6

Table 10: Characteristics of included studies—Source of feedback

Source of feedback (*not mutually exclusive*)	No. of studies in review	No. of studies in map
Teacher	14	32
Researcher	18	73
Digital or automated	31	78

Table 11: Characteristics of included studies—Feedback directed to

Feedback directed to (*not mutually exclusive*)	No. of studies in review	No. of studies in map
Individual pupil	48	169
General (group or class)	4	8

Feedback can be communicated in different ways. This is coded as form of feedback shown in Table 12. Spoken verbal refers to feedback provided in spoken form. Non-verbal refers to feedback communicated physically, other than with words, such a through body language, gesture, or other non-verbal means, such as extended wait time. Written verbal refers to where written comments are provided, either handwritten or digitally. Written, non-verbal refers to feedback in the form of tick or check marks, or with symbols or icons (this includes marked tests or test results).

Table 12: Characteristics of included studies—Form of feedback

Form of feedback (*not mutually exclusive*)	No. of studies in review	No. of studies in map
Spoken verbal	22	67
Non-verbal	0	6
Written verbal	27	68
Written, non-verbal	21	68

Table 13: Characteristics of included studies—When was feedback given?

When feedback happened (*not mutually exclusive*)	No. of studies	No. of studies in map
During the task	17	62
Immediately after task	30	107
Delayed (short—up to 1 week after task)	14	31
Delayed (long—more than 1 week after task)	1	3

Table 14: Characteristics of included studies—Kind of feedback given?

Kind of feedback* (*not mutually exclusive*)	No. of studies	No. of studies in map
About the outcome	49	164
About the process of the task	13	41
About the learner's strategies or approach	9	19

*See the synthesis by kind of feedback for further discussion of these categories.

Table 15: Characteristics of included studies—Emotional tone of feedback

Emotional tone of the feedback (*not mutually exclusive*)	No. of studies	No. of studies in map
Positive	2	20
Neutral	50	161
Negative	1	5

Each study was assessed using the ecological validity tool (see Appendix 4 for details). As already noted, the review selection criteria included requirements that support ecological validity (e.g. must be in mainstream school age groups). The results of the ecological validity assessment in Table 16 should be viewed in that context.

Table 16: Characteristics of included studies—Overall ecological validity

Overall ecological validity	No. of studies
High & High = High	24
High & Moderate = Moderate	16
Moderate & Moderate = Moderate	11
	51 total

Table 17 shows the results of the overall risk of bias analysis for all the studies. The method of assessing the risk of bias is described in the method section above and the tool. The assessment is based on the information reported in the studies on the dimensions in the assessment tool (see Appendix 4).

Table 17: Included study characteristics—Overall risk of bias assessment

Overall risk of bias	No. of studies
Low risk of bias	4
Moderate risk of bias	40
Serious risk of bias	7
	51 total

The synthesis results in Table 18 below show that the greater the assessed risk of bias, the larger the pooled estimate of effect and the greater the statistical heterogeneity of the studies. This is what you would anticipate based on the dimensions assessed in the tool.

Table 18: Synthesis results of studies within groups by risk of bias

Risk of bias assessment	**Pooled Effect size *g* (95%C.I)**	**Heterogeneity**
Low (n = 4)	0.07 (0.00 to 0.14)	I^2 = 0%. Test for Heterogeneity: Q(df = 3) = 1.01, p = 0.79
Moderate (n = 35)*	0.20 (0.10 to 0.30)	I^2 = 51%. Test for Heterogeneity: Q(df = 34) = 68.76, p = 0.0004
Serious (n = 7)	0.62 (0.24 to 0.99)	I^2 = 92%. Test for Heterogeneity: Q(df = 6) = 71.52, p = <0.0001

*Only studies with data to calculate an effect size.

Another study design issue that might influence synthesis outcomes and study heterogeneity is the nature of the comparison being made. We attempted to code for whether a study compared feedback to 'usual teaching' or 'active control' (control for novelty or new treatment). This information was not available in all studies. This element of study design is not assessed in the risk of bias tool. Table 19 below shows the results of synthesis of studies in these two groups. The pooled estimate of effect in each group is not markedly different and neither are the levels of heterogeneity.

Table 19: Synthesis results of studies grouped by type of comparison group

Comparison group received	**Pooled effect size (95% C.I)**	**Heterogeneity**
Usual teaching (20 studies)	*g* = 0.14 (0.03–0.25)	I^2 = 54%. Test for Heterogeneity: Q(df = 19) = 41, p = 0.002
Active control (19 studies)	*g* = 0.22 (0.09–0.34)	I^2 = 41%. Test for Heterogeneity: Q(df = 18) = 30.57, p = 0.032

We did not identify any studies which reported providing feedback on correct answers or incorrect answers. There are some studies that provide information about the socioeconomic status of sample participants (for example, percentage eligible for free school meals). However, these studies did not present any subgroup data analysis in these categories. Some authors make comments about the results in these groups but the data was not presented in the study. We did not identify any studies which conducted subgroup analysis relating to prior attainment level on students.

5.3 What is the impact of feedback compared to no feedback or usual practice on student attainment?

There were five studies identified as meeting the review selection criteria but which did not report the data needed to calculate an effect size. The author reported outcomes from these studies are shown in Table 20 below.

Table 20: Included studies for which no data to calculate effect sizes reported

Study	Author reported result
Brosvic *et al* (2006). Experiment 1a	Significant positive effects were found in the groups that received feedback when compared to the no feedback groups.
Brosvic *et al* (2006). Experiment 1b	Significant positive effects were found in the groups that received feedback when compared to the no feedback groups.
Brosvic *et al* (2006). Experiment 2	Significant positive effects were found in the groups that received feedback when compared to the no feedback groups.
Dihoff *et al* (2005). Experiment 1	Significant positive effects were found in the groups that received feedback when compared to the no feedback groups.
Golke, Dörfler & Artelt (2015). Experiment 1	No significant difference between feedback and no feedback group in literacy categories of text comprehension.

Figure 1 below is a forest plot showing the result of each included study (the point estimate) as a Standardised Mean Difference (Hedges *g*) and the pooled estimate of effect resulting from combining the individual study results using a Random Effects meta-analysis (the diamond at the bottom of the plot). A number of papers published the results of more than one study (for example, Allitto *et al*, 2016). Where these studies involved completely distinct participants, they are included in the review as separate studies. Hence the same publication citation but not the same study may appear twice in the same synthesis.

When interpreting the results, an effect size greater than zero indicates that outcomes in the feedback group were better than in the non-feedback/usual practice group. The 'whiskers' each size of the point estimate of effect are the 95% confidence interval. If the upper or lower confidence interval crosses the line of no effect (g=0) then we cannot exclude the possibility that the true effect may be opposite to that indicated by the point estimate.

There is considerable statistical heterogeneity between the studies (I^2 = 76%; Test for Heterogeneity: Q(df = 45) = 187.95, p < 0.0001). A higher I^2 value combined with a statistically significant test for heterogeneity suggests that the pooled estimate of effect may not be a useful indicator of the general effect of feedback on attainment.

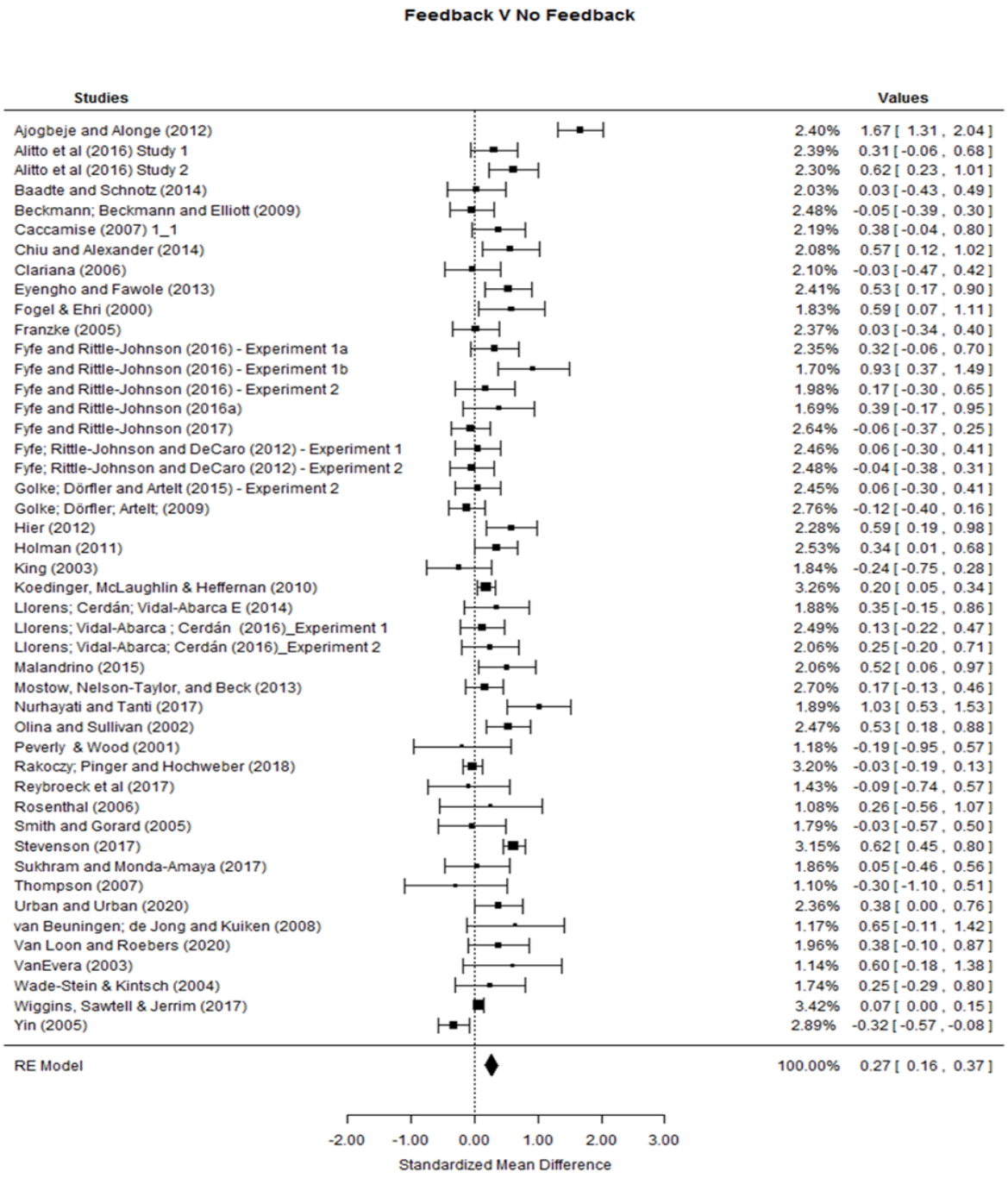
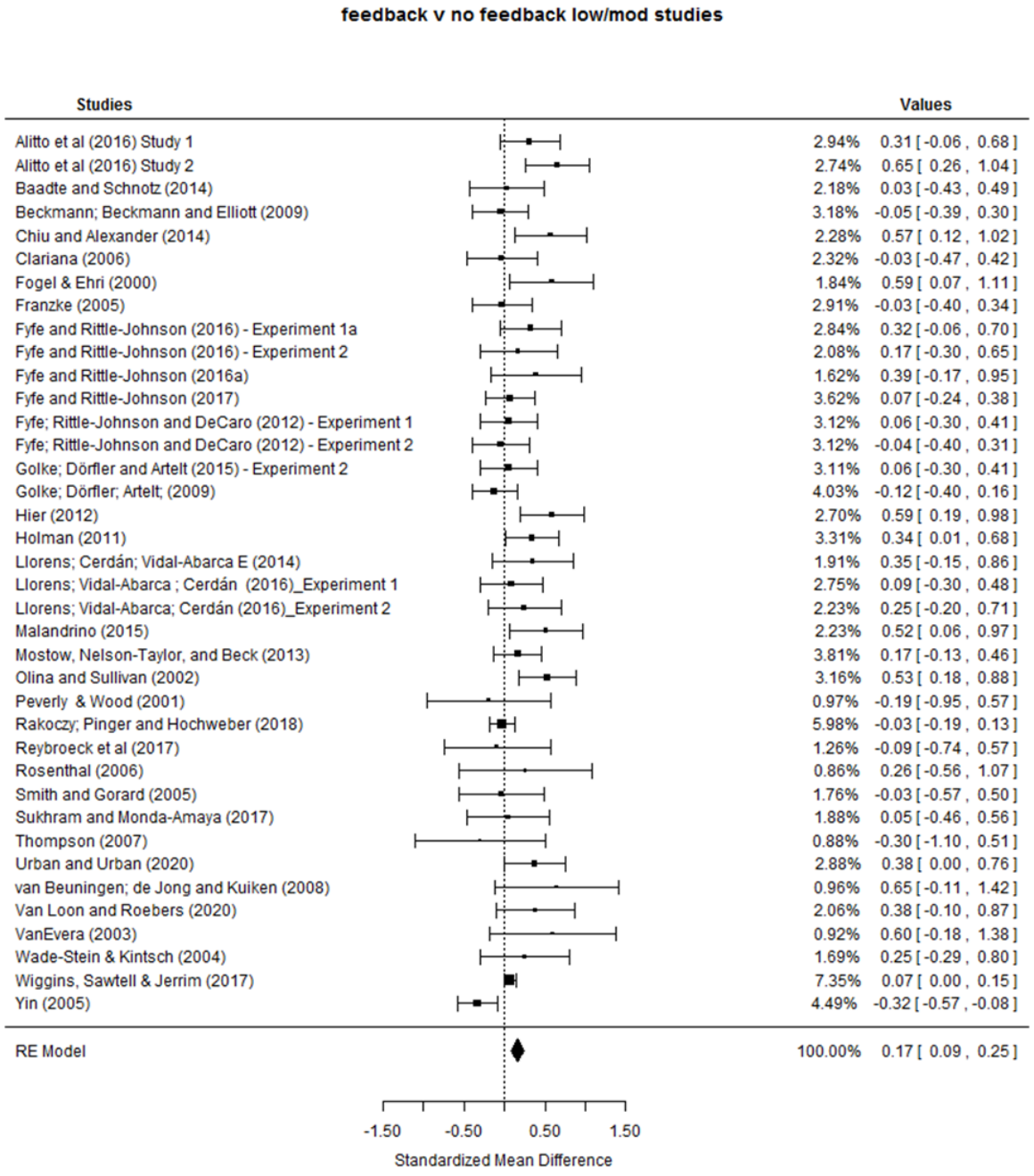


Figure 1: Synthesis: Feedback compared to no feedback or usual practice—All studies

Differences in study design may contribute to heterogeneity between studies. Furthermore, a pooled estimate of effect synthesised from studies with a lower risk of bias may represent a more valid estimate of impact as these studies will have fewer threats to validity than studies with a high risk of bias. Figure 2 below is a forest plot for all studies with a low or moderate risk of bias (ROB) assessment. The pooled estimate of effect indicates that students who received feedback had better performance than students who did not receive feedback (g = 0.17; 95% C.I 0.09 to 0.25). The 95% confidence interval does not cross the line of no effect and therefore the opposite effect can be exluded. However there is statistical heterogeneity between these studies (I^2 = 44%, Test for Heterogeneity: Q(df = 37) = 65.92, p = 0.002), suggesting that this may not be a useful indicator of the general impact of feedback on attainment.

Figure 2: Synthesis: Feedback compared to no feedback or usual practice—Low and moderate risk of bias studies



The review also investigated a number of sub questions about a variety of factors that may theoretically influence the impact of feedback. These questions were investigated through subgroup analysis reported in the following sections. For all subgroups analysis, the synthesis compares feedback to no feedback or usual practice.

5.4 Impact of feedback in different curriculum subjects

5.4.1 Literacy

There are 23 studies in which feedback was investigated in the curriculum subject of literacy. Figure 3 is a forest plot showing the synthesis of 23 studies including all the sub-categories measured. The pooled estimate of effect indicates that the students receiving feedback performed better than students who did not receive feedback (g = 0.22, 95% C.I, 0.12 to 0.31) and the 95% confidence interval excludes the opposite effect. There is no statistically significant heterogeneity (I^2 = 32% Test for Heterogeneity: Q(df = 22) = 32.32, p = 0.07) and therefore this may be a useful indicator of the impact of feedback in the curriculum subject of literacy.

One study (Golke, Dörfler and Artelt, 2015, Experiment 1) provided no usable data to compute an effect size. The authors stated that there was no significant difference in outcome between the feedback and no feedback group on literacy outcomes.

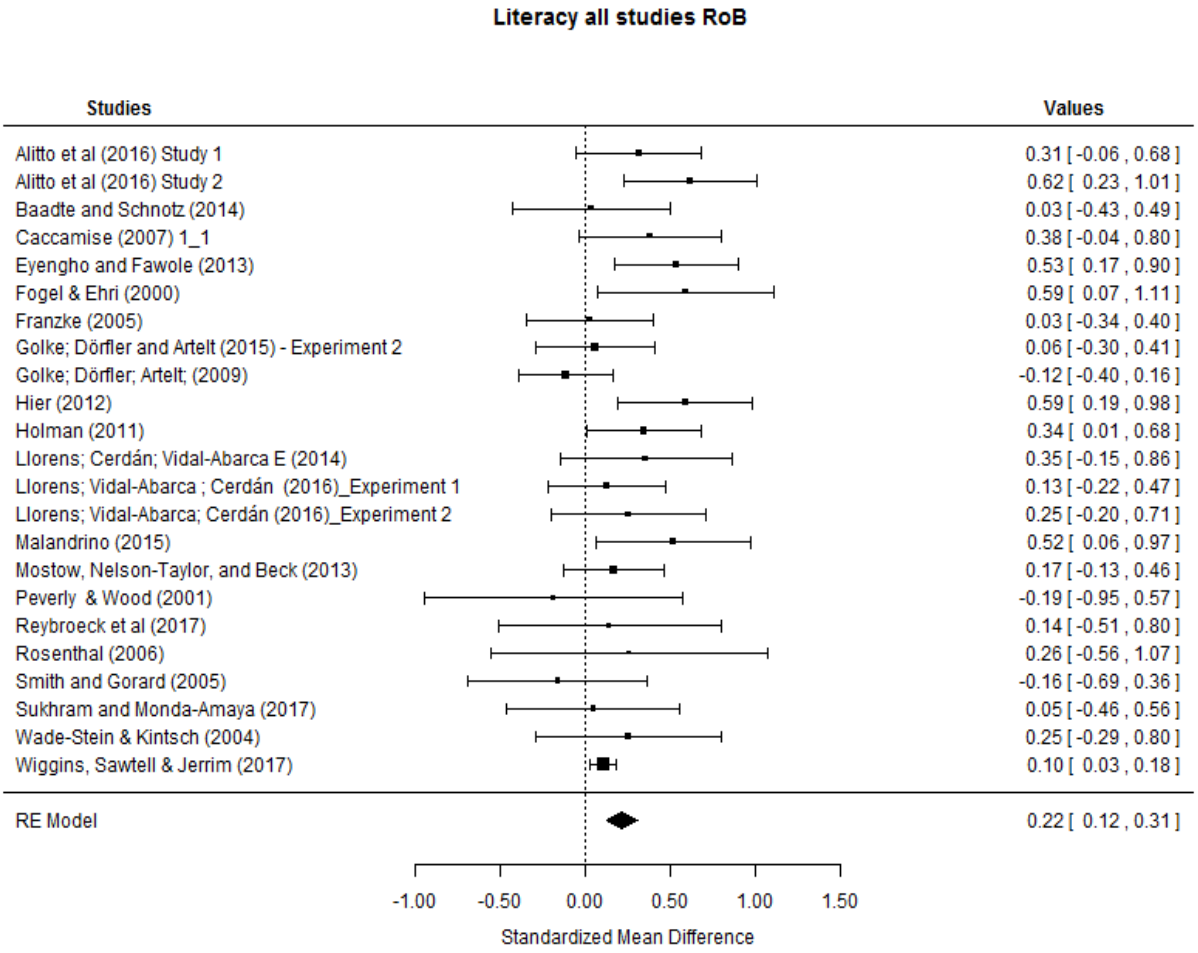


Figure 3: Synthesis: Curriculum subject literacy—All studies

As shown in Figure 4, limiting the synthesis to the 21 studies of low and moderate risk of bias reduces the heterogeneity (I^2 = 26%, Test for Heterogeneity: Q(df = 20) = 28.85, p = 0.13). The direction of effect continues to favour feedback and exclude the opposite effect (*g* = 0.19, 95% C.I 0.09 to 0.28).

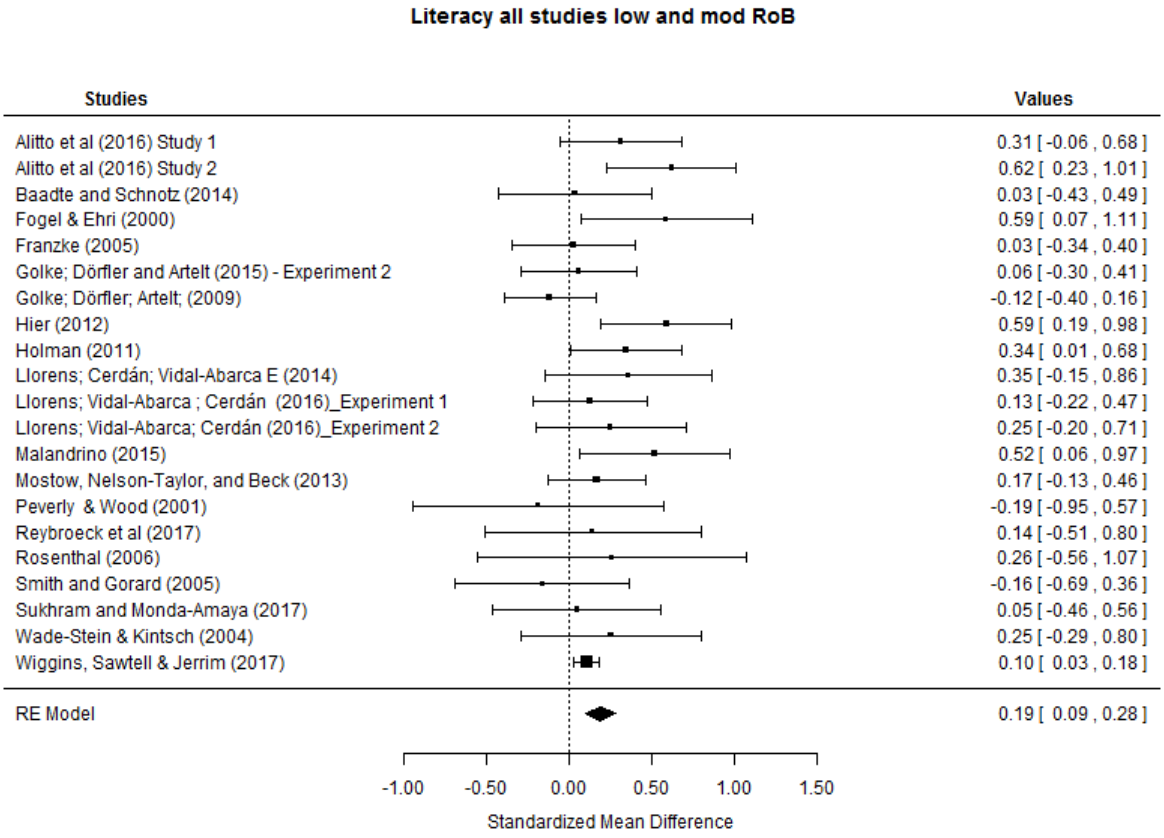


Figure 4: Synthesis: Curriculum subject literacy—Low or moderate risk of bias studies only

5.4.2 Mathematics

There are four studies (Brosvic *et al*, 2006, Experiment 1a 2006; Brosvic *et al*, 2006, Experiment 1b 2006; Brosvic *et al*, 2006, Experiment 2 2006; Dihoff *et al*, 2005, Experiment 1 2005) which did not provide useful data to compute effect sizes. The respective authors stated that significant positive effects in mathematics were found in the groups that received feedback when compared to the no feedback groups.

There are 17 studies which assessed the effect of feedback in the maths curriculum. Figure 5 is a forest plot showing the synthesis of all studies where the curriculum subject was mathematics. The pooled estimate of effect indicates that the students receiving feedback performed better than students who did not receive feedback (g = 0.25, 95% C.I, 0.06 to 0.45) but there is statistical heterogeneity (I^2 = 86%, Test for Heterogeneity: Q(df = 12) = 88.68, p < 0.0001).

.

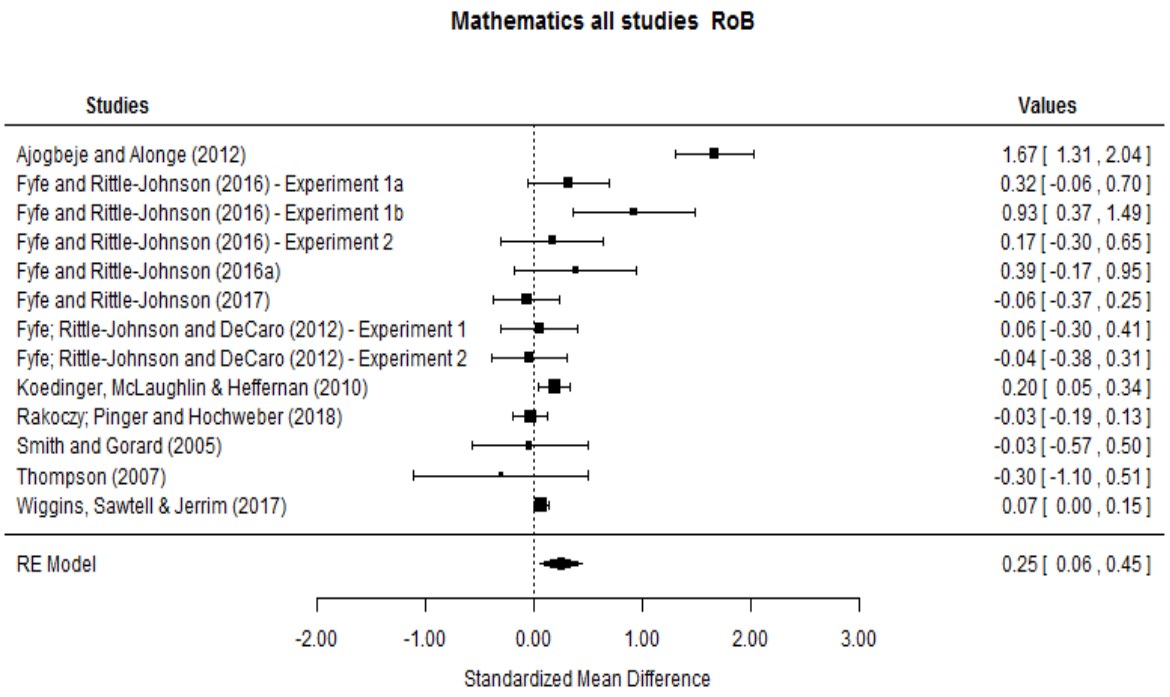


Figure 5: Synthesis: Curriculum subject mathematics—All studies

Synthesis of only the 15 studies of low and moderate risk of bias does not have statistically significant heterogeneity (I^2 = 36%, Test for Heterogeneity: Q(df = 10) = 15.65, p = 0.11). Figure 6 shows a pooled estimate of effect favouring feedback of *g* = 0.08 but the 95% confidence interval crosses the line of no effect, therefore we cannot exclude the opposite effect.

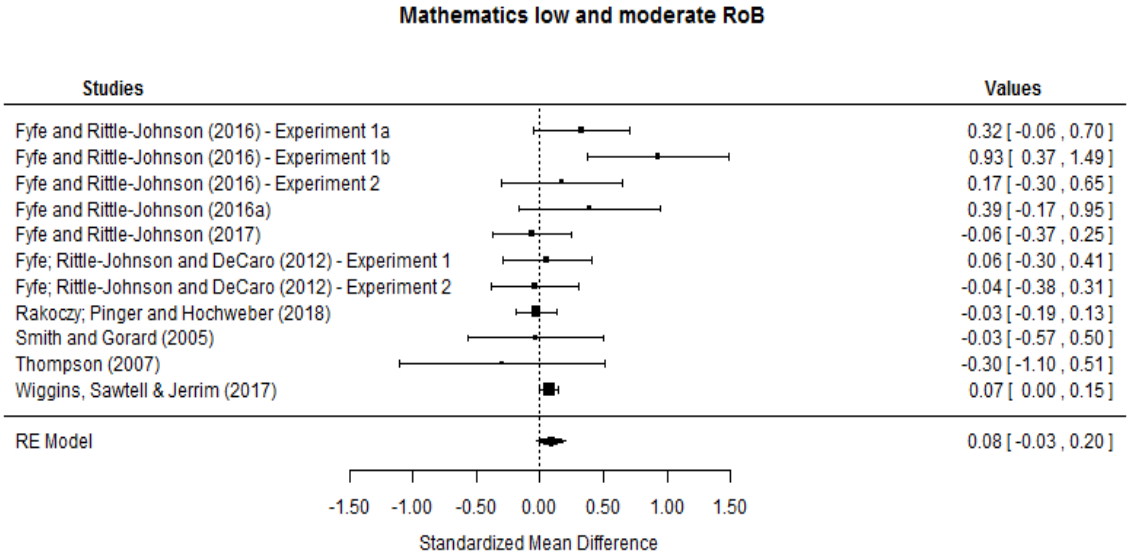


Figure 6: Synthesis: Curriculum subject mathematics—Low and moderate risk of bias studies only

5.4.3 Curriculum subjects: Science

There are seven studies which investigated the effect of feedback in the science curriculum. Figure 7 is a forest plot showing the synthesis of all studies where the curriculum subject was science. The analysis indicates substantial statistical heterogeneity (I^2 = 80%, Test for Heterogeneity: Q(df = 6) = 30.57, p < 0.0001).

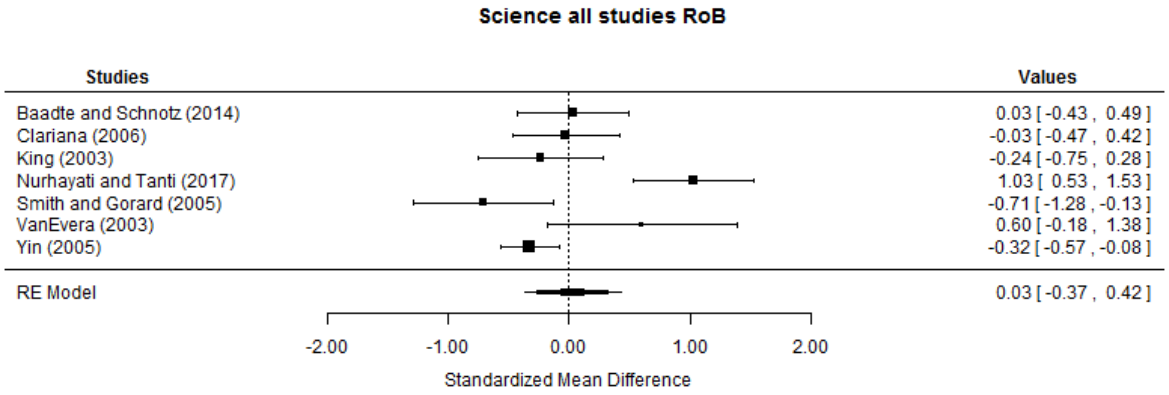


Figure 7: Synthesis: Curriculum subject science—All studies

Limiting the synthesis to five studies of low and moderate risk of bias reduces the heterogeneity (I^2 = 57%, Test for Heterogeneity: Q(df = 4) = 9.47, p = 0.05) but it remains statistically significant. As shown in Figure 8, the pooled estimate of effect (*g* = -0.15) indicates that students who received feedback had a worse outcome than students who did not receive feedback. However, the 95% confidence interval crosses the line of no effect and so we cannot confidently exclude the opposite effect.

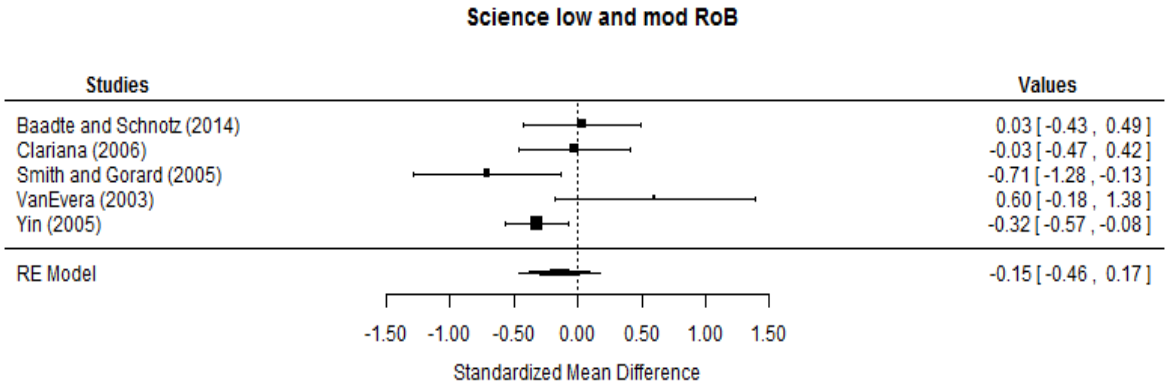


Figure 8: Synthesis: Curriculum subject science—Low or moderate risk of bias studies only

The results of the synthesis in the different curriculum areas cannot be compared directly. There may be many other variables that are differently affecting impact in the studies in these groups apart from the 'curriculum subject'. The statistical heterogeneity amongst the studies in the science curriculum area in particular means the results are difficult to interpret. Nevertheless, the different results in the pooled estimate of effect in the three different curriculum areas would seem worthy of further investigation.

5.5 Impact of feedback by age: Synthesis in UK key stages

Information about the age of participants was coded in the review. This was either stated in the study report or deduced by the reviewers from the details provided (for example, year group). There are eight studies for which it was not possible to ascertain the age of participants. The study participants were typically within a single school year group and thus contained children within a two year age range. Therefore there is considerable overlap in ages between studies in different year groups. Furthermore, most studies are international and thus not conducted in contexts where the UK key stage system operates. We have therefore used a modified version of the UK key stage age ranges in the synthesis to minimise the overlap between studies in the different key stages. The students in the studies are in the age range indicated in each of the key stages.

5.5.1 Key Stage 1 (ages 5–7)

The source of feedback in the studies in this key stage was either researcher or digital/automated, and the form of the feedback was both verbal and written. There is statistically significant heterogeneity between the studies (I^2 = 57%, Test for Heterogeneity: Q(df = 8) = 18.4504, p = 0.02). There is not statistically significant heterogeneity between the low moderate risk of bias studies (I^2 = 37%, Test for Heterogeneity: Q(df = 7) = 11.15, p = 0.1324). The pooled estimate of effect, which indicates that performance was better in the group that received feedback shown in Figure 7 (*g* = 0.34, 95% C.I 0.15 to 0.52), may therefore be a useful indicator of the impact of feedback compared to no feedback or usual practice in Key Stage 1.

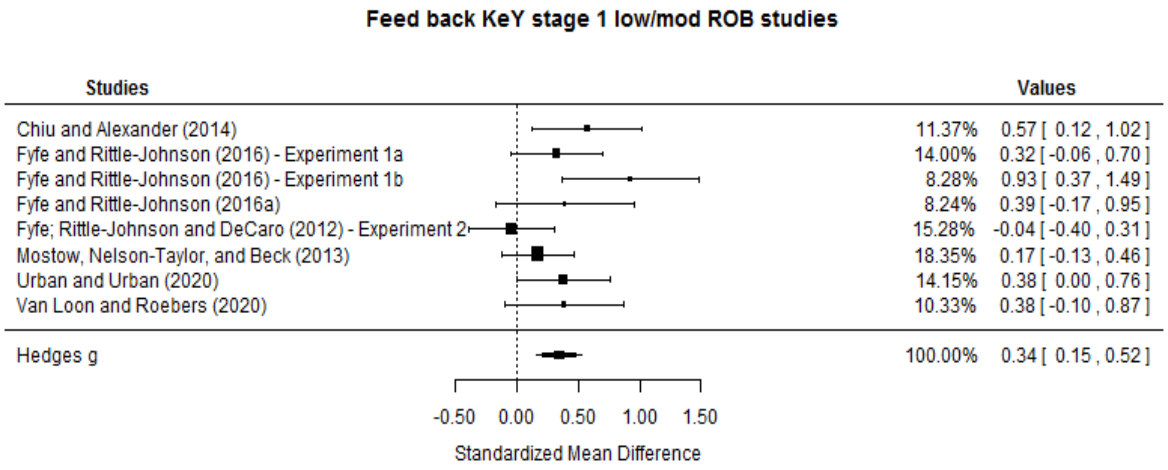


Figure 9: Synthesis: Key Stage 1—Low or moderate risk of bias studies

5.5.2 Key Stage 2 (ages 8–11)

We have included studies where the age range of students was 8–11 rather than the 7–11 used in the UK system. There is statistically significant heterogeneity between the studies that included participants in the Key Stage 2 age range (I^2 = 62%, Test for Heterogeneity: Q(df = 18) = 47.73, p = 0.0002). This suggests that pooled estimate of effect shown in Figure 10 (*g* = 0.20, 95% C.I 0.07 to 0.33) may not be a useful indicator of the impact of feedback compared to no feedback or usual practice in the Key Stage 2 age range.

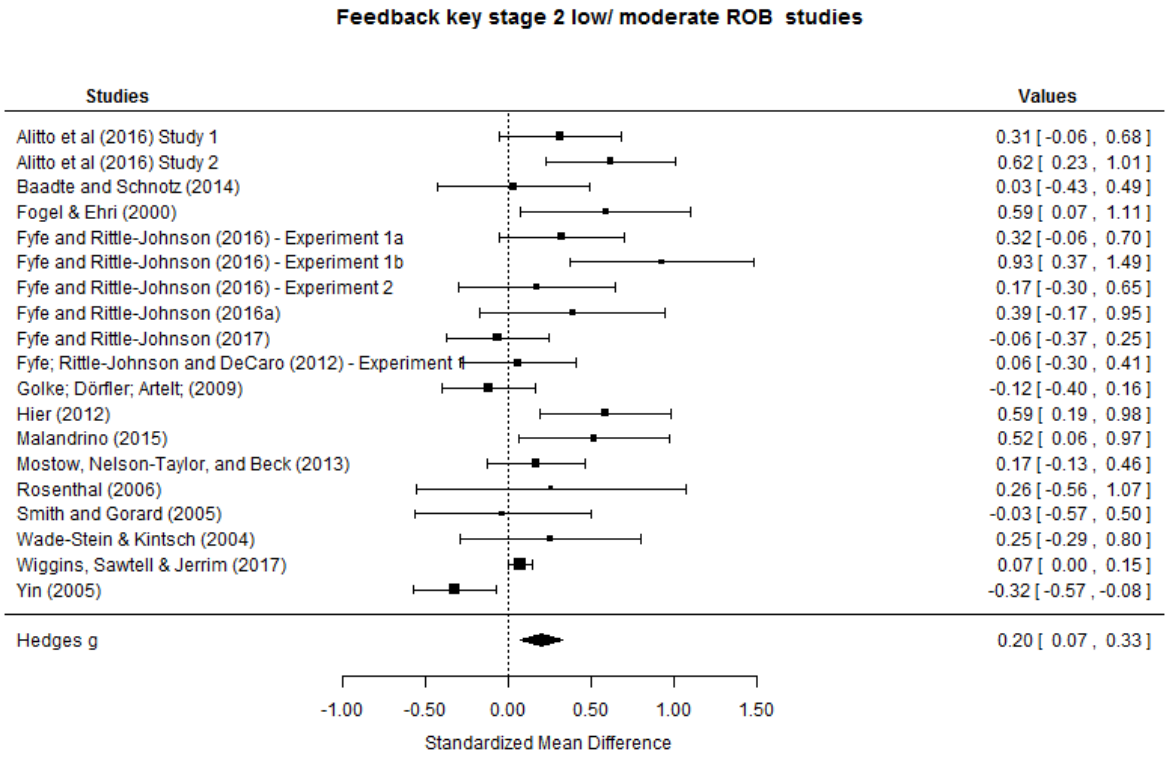


Figure 10: Synthesis: Key Stage 2—Low or moderate risk of bias studies

5.5.3 Key Stage 3 (ages 12–14)

We have included studies where the age range of students was 12–14 rather than the 11–14 used in the UK system. There is statistically significant heterogeneity between the studies in this group (I^2 = 55%, Test for Heterogeneity: Q(df = 18) = 40.16).

There is not statistically significant heterogeneity of the studies with a low or moderate risk of bias assessment (I^2 = 30%, Test for Heterogeneity: Q(df = 15) = 21.53, p = 0.12). As shown in Figure 11, the pooled estimate of effect indicates that students who received feedback performed better than students who did not (*g* = 0.05, 95% C.I -0.07 to 0.18). However the 95% confidence interval crosses the line of no effect and so we cannot be confident of excluding the opposite effect.

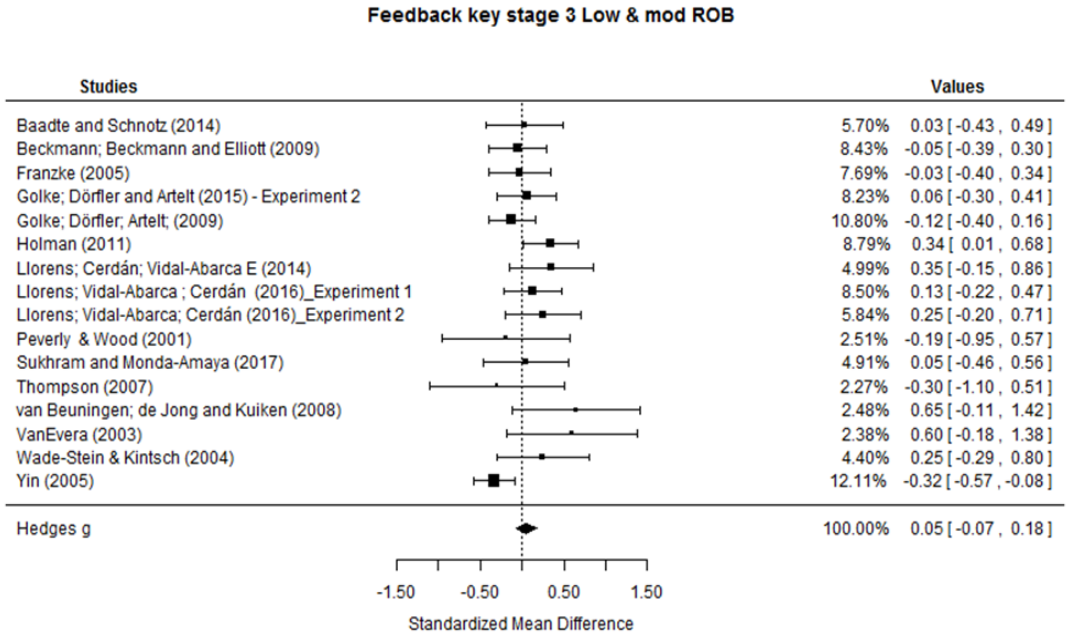


Figure 11: Synthesis: Key Stage 3—Low or moderate risk of bias studies

5.5.4 Key Stage 4 (age 15–16)

We have included studies where the age range of students was 15–16 rather than the 14–16 used in the UK key stage system. The studies with participants at Key Stage 4 do not have not statistically significant heterogeneity (I^2 = 0%, Test for Heterogeneity: Q(df = 6) = 4.14, p = 0.66). There is one study with a serious risk of bias assessment in this group. A synthesis without this study (see Figure 12) gives a pooled estimate of *g* = -0.04, 95% C.I -0.17 to 0.09. The group of studies is not statistically heterogenous (I^2 = 0%, Test for Heterogeneity: Q(df = 5) = 0.58, p = 0.99). The pooled estimate of effect indicates that students who received feedback performed worse than students who did not receive feedback. However, as the 95% confidence interval crosses the line of no effect, we cannot exclude the opposite effect.

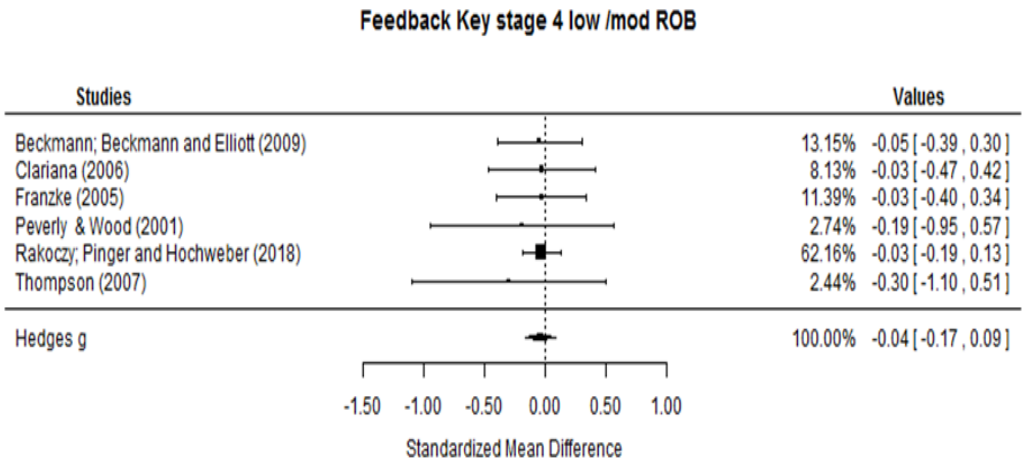


Figure 12: Synthesis: Key Stage 4—Low or moderate risk of bias studies

Care is required when comparing the synthesis results between the key stage age groups, but it is interesting to note that for the low or moderate risk of bias studies, synthesis was not statistically heterogenous in three of the four key stage age groups (Key Stages 1, 3 and 4). The synthesis results in these three groups were also different. In Key

Stage 1 the individual study results were all positive with one exception, and the pooled estimate of effect was also positive with the largest effect size found across any of the syntheses we completed (*g* = 0.34, 95% C.I 0.15 to 0.52). We should perhaps note however that most of the studies in key stage 1 were carried out by the same group of researchers. In Key Stage 4 the individual study results were all negative as is the pooled estimate of effect (*g* = -0.04, 95% C.I -0.17 to 0.09). These findings may suggest that age (particularly at the youngest and oldest end of the school age spectrum) may be a factor in influencing the impact of feedback.

5.6 Impact of feedback: Educational setting

5.6.1 Primary schools

Twenty six studies were conducted in the primary school setting (elementary schools equivalent in the US), including two in a preschool setting where the children were aged 5–6 (Chiu, 2014; Urban, 2020).

There is statistically significant heterogeneity between the studies (I^2=68%, Test for Heterogeneity: Q(df = 21) = 67.04, p = 0.0001), suggesting that this pooled estimate of effect shown in Figure 13 (*g* = 0.30, 95% C.I 0.18 to 0.43) may not be a particularly useful indicator of the impact of feedback in the primary school setting.

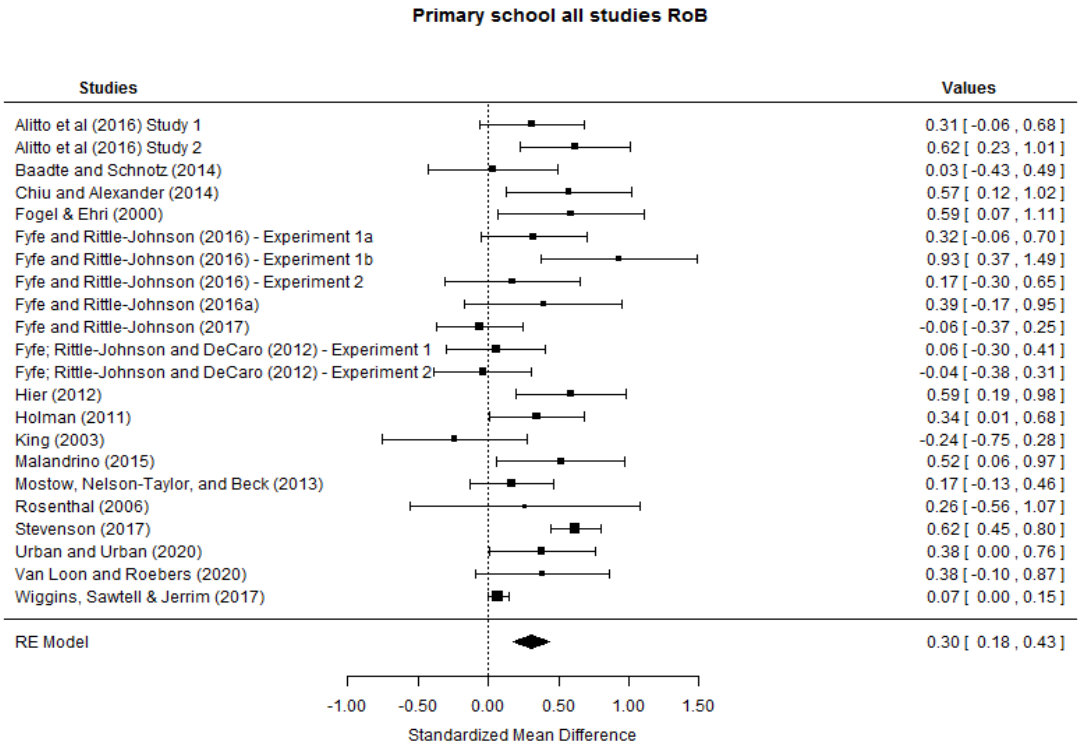


Figure 13: Synthesis: School setting, primary—All studies

The synthesis of studies with low or moderate risk of bias (see Figure 14) has statistically significant heterogeneity (I^2 = 52%, Test for Heterogeneity: Q(df = 19) = 40.10, p = 0.003), suggesting that the point estimate (*g* = 0.29, 95% C.I 0.18 to 0.43), may not be a useful indicator of the effect of feedback in the primary school setting.

Four studies (Brosvic *et al,* 2006, Experiment 1a, Experiment 1b, Experiment 2; Dihoff *et al*, 2005, Experiment 1) in primary school settings did not provide useful data to compute effect sizes. The respective authors stated that significant positive effects were found in the groups that received feedback when compared to the no feedback groups.

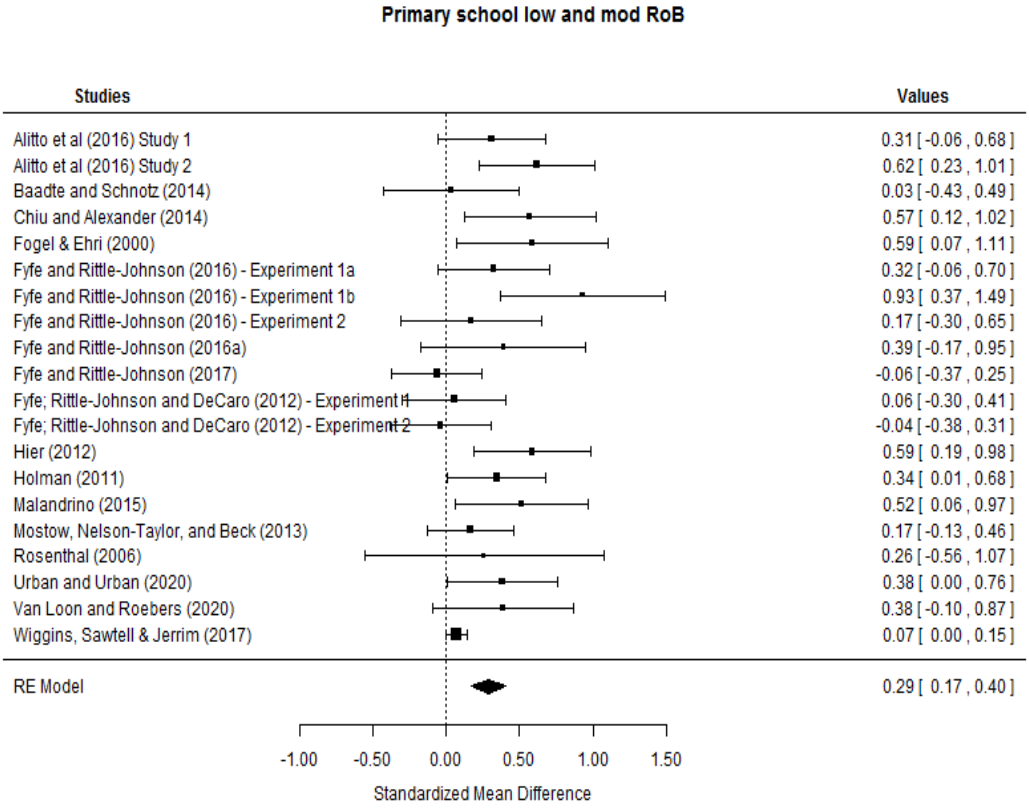


Figure 14: Synthesis: School setting, primary—Low or moderate risk of bias studies

5.6.2 Secondary schools

One study (Golke, Dörfler and Artelt, 2015, Experiment 1) in the secondary school setting did not provide usable data to compute an effect size. The authors reported that there was no significant difference in effect between the feedback and the no feedback groups in the subject of literacy in secondary setting.

There are 25 studies that assessed feedback in secondary school settings (including middle and high school). The synthesis of all studies (Figure 15) has statistically significant heterogeneity (I^2 = 81%, Test for Heterogeneity: Q(df = 23) = 120.19, p < 0.0001).

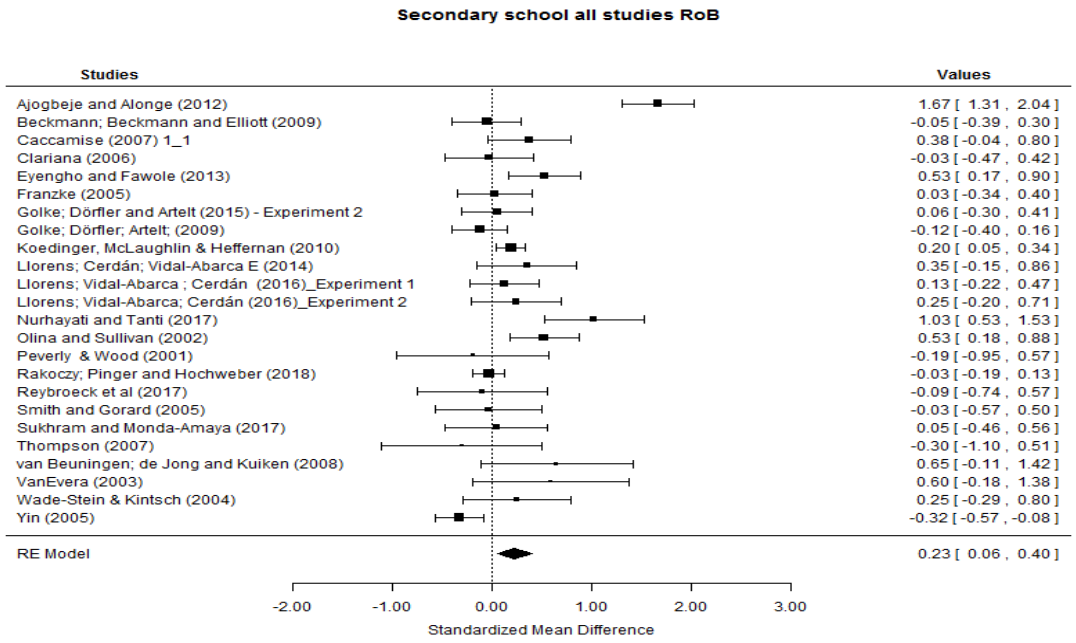


Figure 15: Synthesis: School setting, secondary—All studies

When the synthesis is limited to the 20 studies of low or moderate risk of bias (Figure 16), there was no statistically significant heterogeneity (I^2 = 32%, Q(df = 18) = 26.53, p = 0.08). The pooled estimate of effect (*g* = 0.05, 95% C.I 0.07 to 0.16) indicates that the students who received feedback performed better than the students who did not receive feedback. However, the 95% confidence interval crosses the line of no effect, meaning that we cannot exclude the opposite effect.

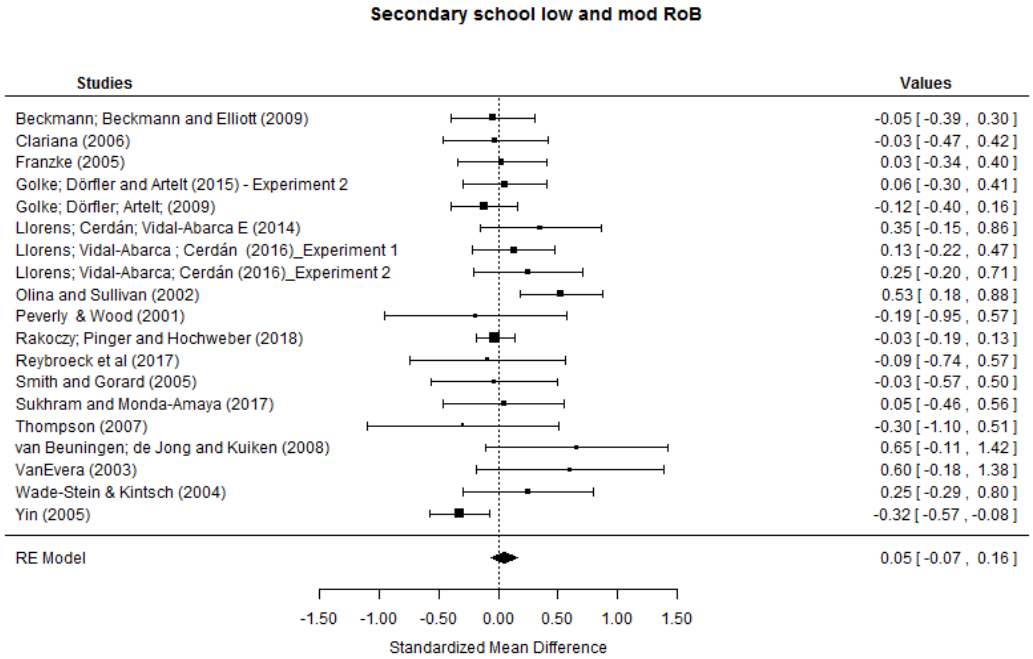


Figure 16: Synthesis: School setting, secondary—Low or moderate risk of bias studies only

5.7 Impact of feedback: Source of feedback

There are a number of ways in which feedback could be delivered to students. The review focused on three categories of feedback: feedback from the teacher, feedback from the researcher, and digital or automated feedback. Some study reports were not always entirely clear about whether the source of the feedback was the researcher or the teacher and in some studies (n=2) it appeared to be both. In some studies it appeared that the feedback was both automated in some way and reported by the teacher/researcher (n = 10).

5.7.1 Source of feedback: Teacher

There are 14 studies where the source of feedback is the teacher, with ten providing data for the calculation of effect sizes. All of these studies were moderate or high risk of bias. Figure 17 shows the results of all the studies where the teacher is the source of feedback. There was statistically significant heterogeneity between the studies (I^2 = 81%, Test for Heterogeneity: Q(df = 9) = 45.1705, p < 0.0001).

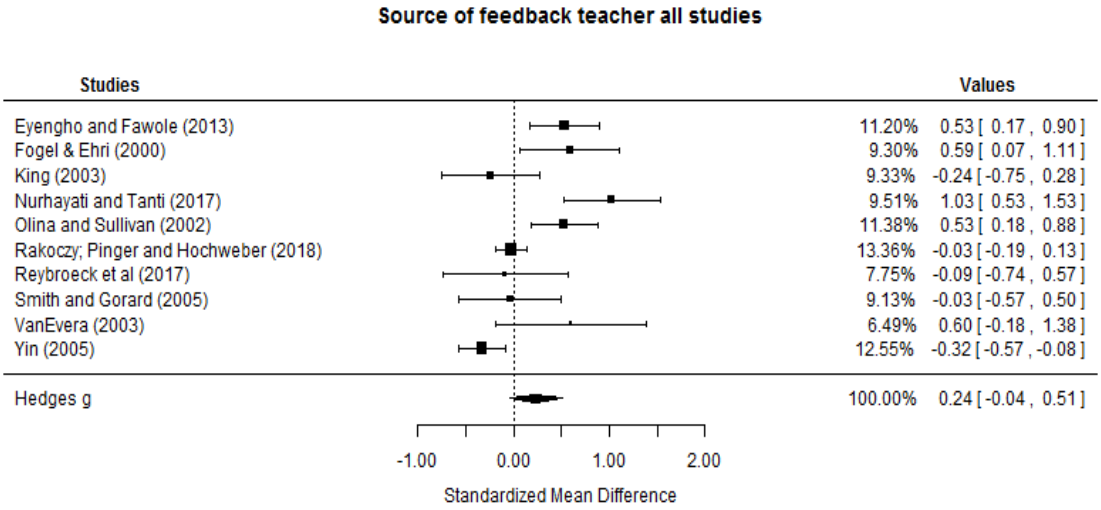


Figure 17: Synthesis: Source of feedback, teacher—All studies

Limiting the synthesis to moderate risk of bias studies, the pooled estimate of effect (*g* = 0.13, 95% C.I -0.15 to 0.51) in Figure 16 indicates that the students who received feedback from the teacher performed better than those who did not receive the feedback intervention. The 95% confidence interval crosses the line of no effect, therefore we cannot exclude the opposite effect. The statistically significant heterogeneity (I^2 = 74%, Test for Heterogeneity: Q(df = 6) = 25.32, p = 0.0007) suggests that the pooled estimate may not be a useful indicator of the general effect of teacher feedback.

There were four moderate risk of bias studies with no data to calculate effect sizes for teacher feedback. In Brosvic *et al* (2006), the authors report that all three studies and all outcomes favoured the feedback intervention group and were statistically significant (moderate risk of bias). In Dihoff *et al* (2005, Experiment 1), the authors report that all outcomes favoured the feedback intervention group and were statistically significant.

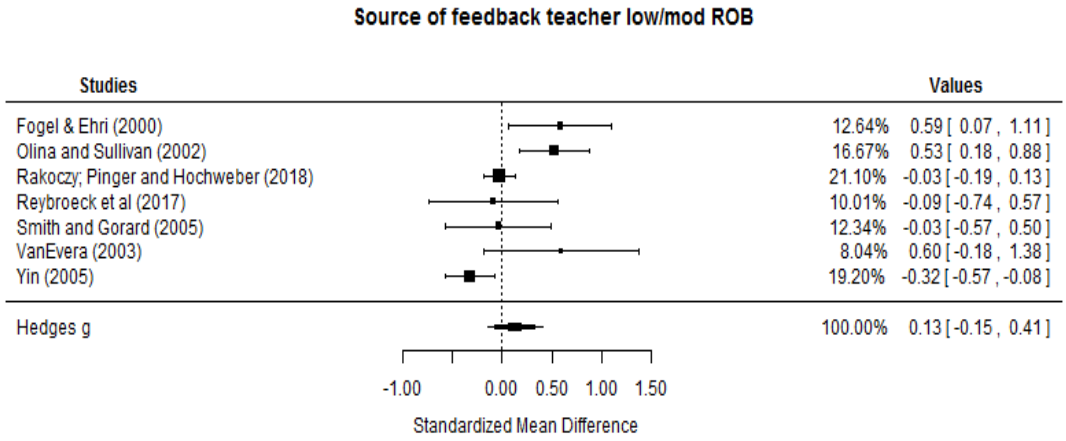


Figure 18: Synthesis: Source of feedback, teacher—Low or moderate risk of bias studies only

5.7.2 Researcher

There were 18 studies where the source of feedback was the researcher. There is statistically significant heterogeneity between the studies in Figure 19 (I^2 = 78%, Test for Heterogeneity: Q(df = 17) = 77.88, p < 0001).

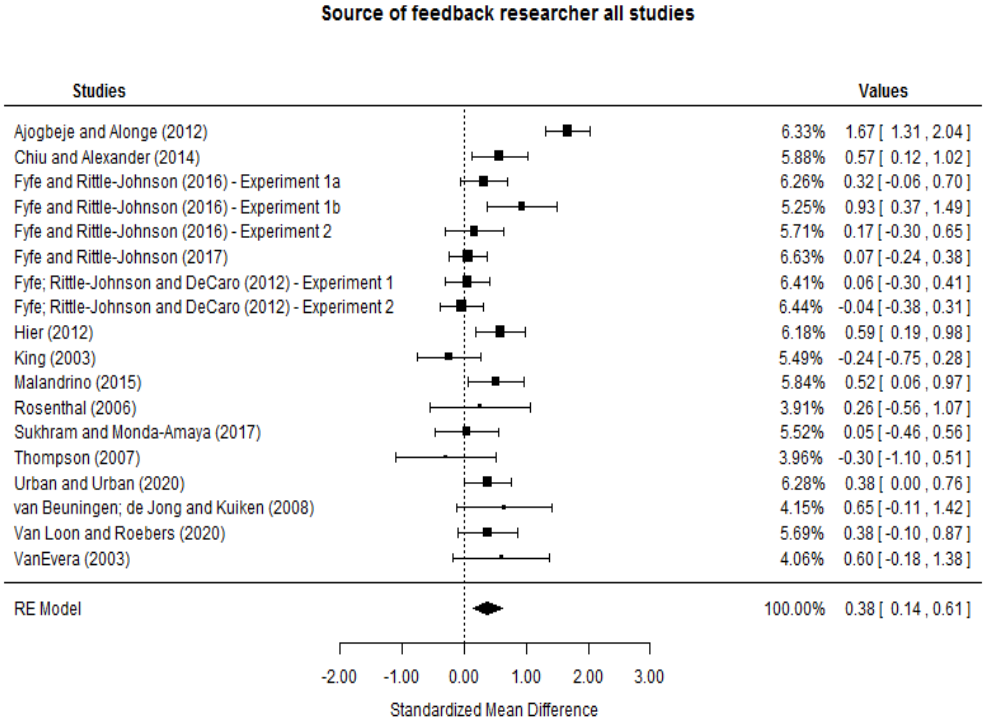


Figure 19: Synthesis: Source of feedback, researcher—All studies

Limiting the synthesis to studies with a low or moderate risk of bias (see Figure 20) reduces the statistical heterogeneity but it remains statistically significant (I^2 = 61%, Test for Heterogeneity: Q(df=21) = 54.12, p < 0.0001). This suggests that the pooled estimate of effect shown (*g* = 0.30, 95% C.I 0.16 to 0.44) may not be a useful general indicator of the effect of feedback provided by a researcher.

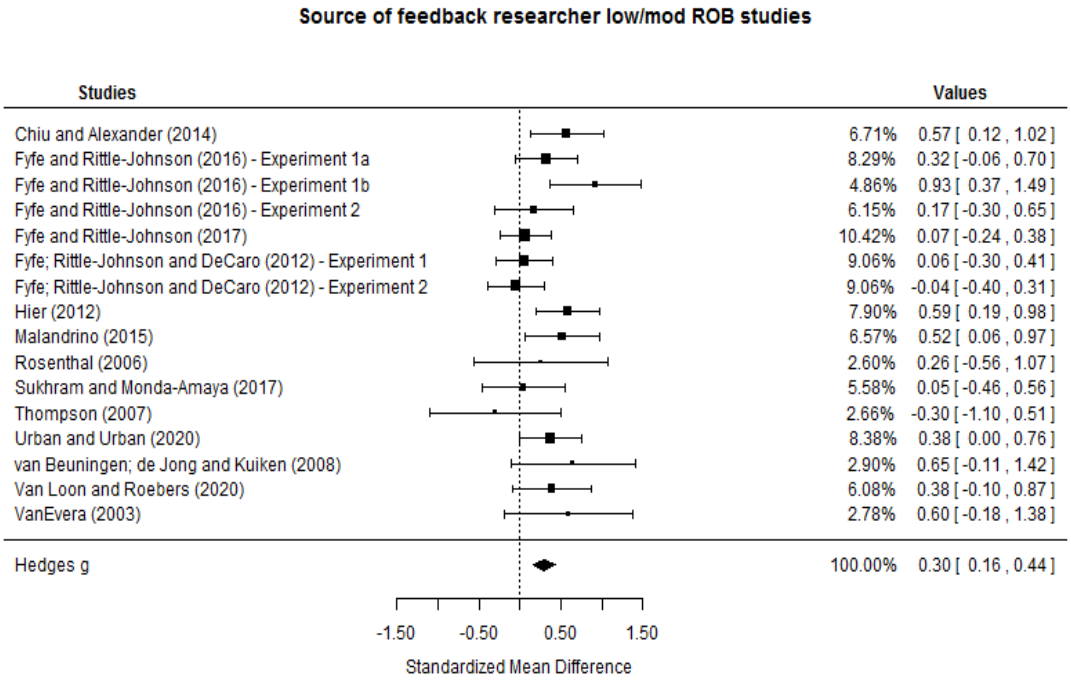


Figure 20: Synthesis: Source of feedback, researcher—Low or moderate risk of bias studies

5.7.3 Researcher or teacher

The studies in which the feedback was provided by a researcher could be argued to be testing a particular feedback technique with the intention of providing a model for teachers to use. It is therefore reasonable to combine the studies where the source of feedback was a teacher or a researcher and consider the results as source of feedback from 'a person'.

Figure 21 shows a synthesis of all studies where the feedback is from a teacher or researcher, where the study has a low or moderate risk of bias. The pooled estimate of effect favours feedback (g = 0.25, 95% C.I 0.10 to 0.41) and the confidence interval excludes the opposite effect. However, the 22 studies had statistically significant heterogeneity (I^2 = 61%, Test for Heterogeneity: Q(df = 21) = 54.12, p < 0.0001), suggesting that the pooled estimate of effect may not be a useful general indicator of the impact of feedback from a person.

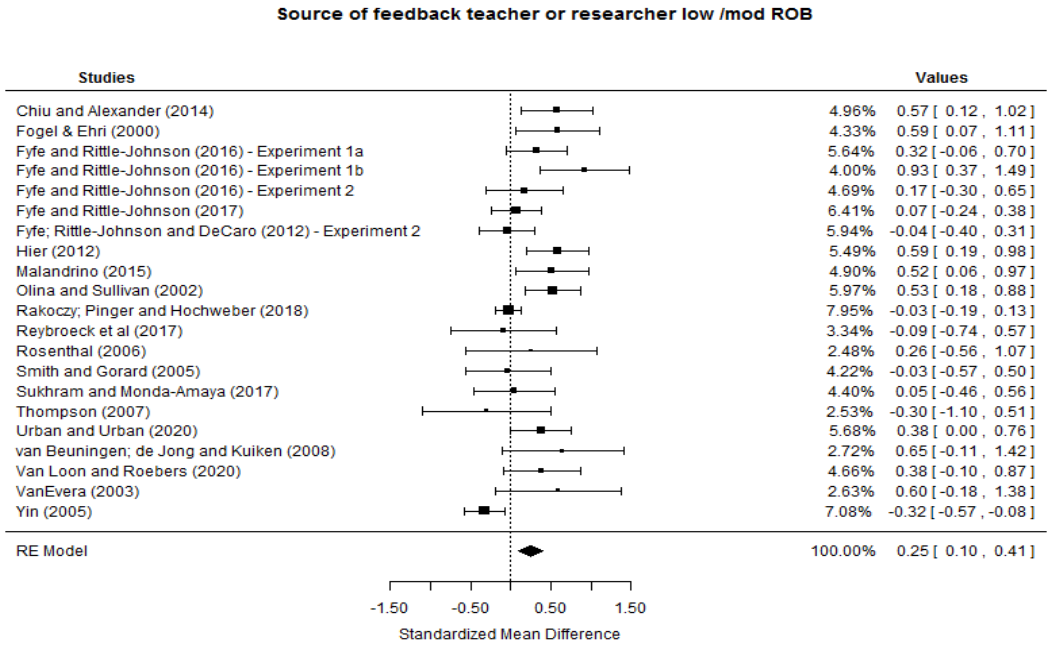


Figure 21: Synthesis: Source of feedback, teacher or researcher—Low or moderate risk of bias studies

5.7.4 Digital or automated feedback

There were 31 studies in which the feedback was provided by digital or automated methods. The non-digital automated methods were used when the students completed some kind of test-like task and then were given some kind of 'reveal' card, which when used revealed the correct answer to the student. Figure 22 shows the synthesis of all of these studies. There is statistical heterogeneity between these studies (I^2 = 63%, Test for Heterogeneity: Q(df = 25) = 67, p < 0.0001).

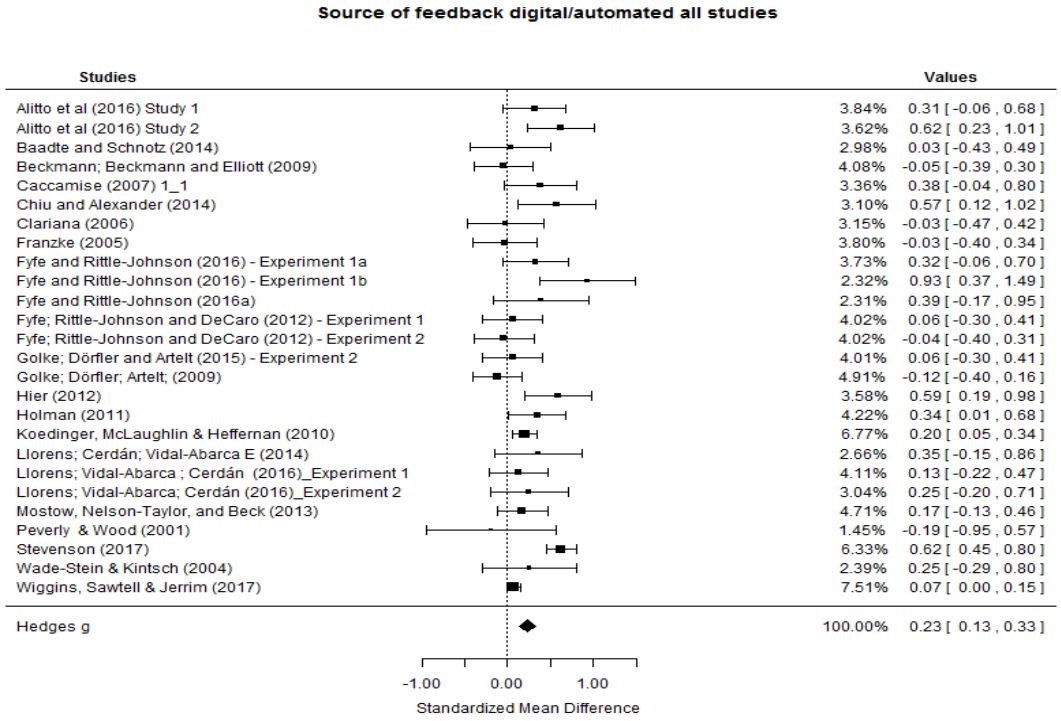


Figure 22: Synthesis: Source of feedback, digital/automated—All studies

The 23 low or moderate risk of bias studies where the source of feedback is digital or automated had statistically significant heterogeneity (I^2 = 42%, Test for Heterogeneity: Q(df = 22) = 38.11, p = 0.02), suggesting that the positive pooled estimate of effect shown in Figure 23 (*g* = 0.19, 95% C.I 0.09 to 0.28), may not be a useful indicator of the impact of feedback from a digital or automated source.

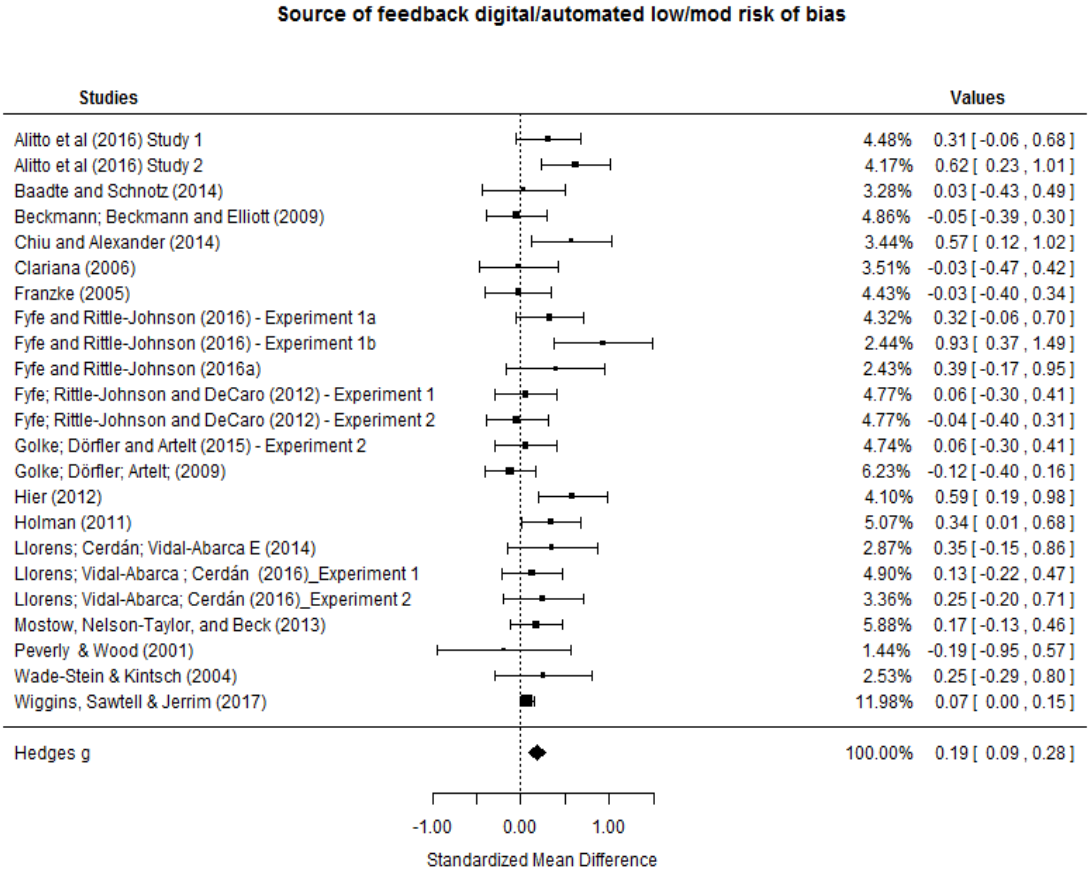


Figure 23: Synthesis: Source of feedback, digital or automated—Low or moderate risk of bias studies

5.8 Impact of feedback: Target of feedback

Feedback can be provided to either individual students or groups of students. The majority of studies in the review investigated the outcome of feedback provided to individual students.

5.8.1 Individual students

There were five studies with no data to compute effect sizes, all with a moderate risk of bias (Brosvic *et al*, 2006 (3 studies); Dihoff *et al*, 2005; Golke, Dörfler and Artelt 2015,Experiment 1)). In Dihoff *et al* (2005) and Brosvic *et al* (2006), the authors state all outcomes favoured the feedback intervention group and are statistically significant. In Golke, Dörfler and Artelt (2015,Experiment 1), the authors state no statistically significant difference between feedback and non-feedback groups on all outcomes.

Figure 24 shows the results of the meta-analysis of studies where feedback is provided to individual students. There was statistically significant heterogeneity between the studies (I^2 = 75%; Test for Heterogeneity: Q(df = 42) = 162.41, p < 0.0001).

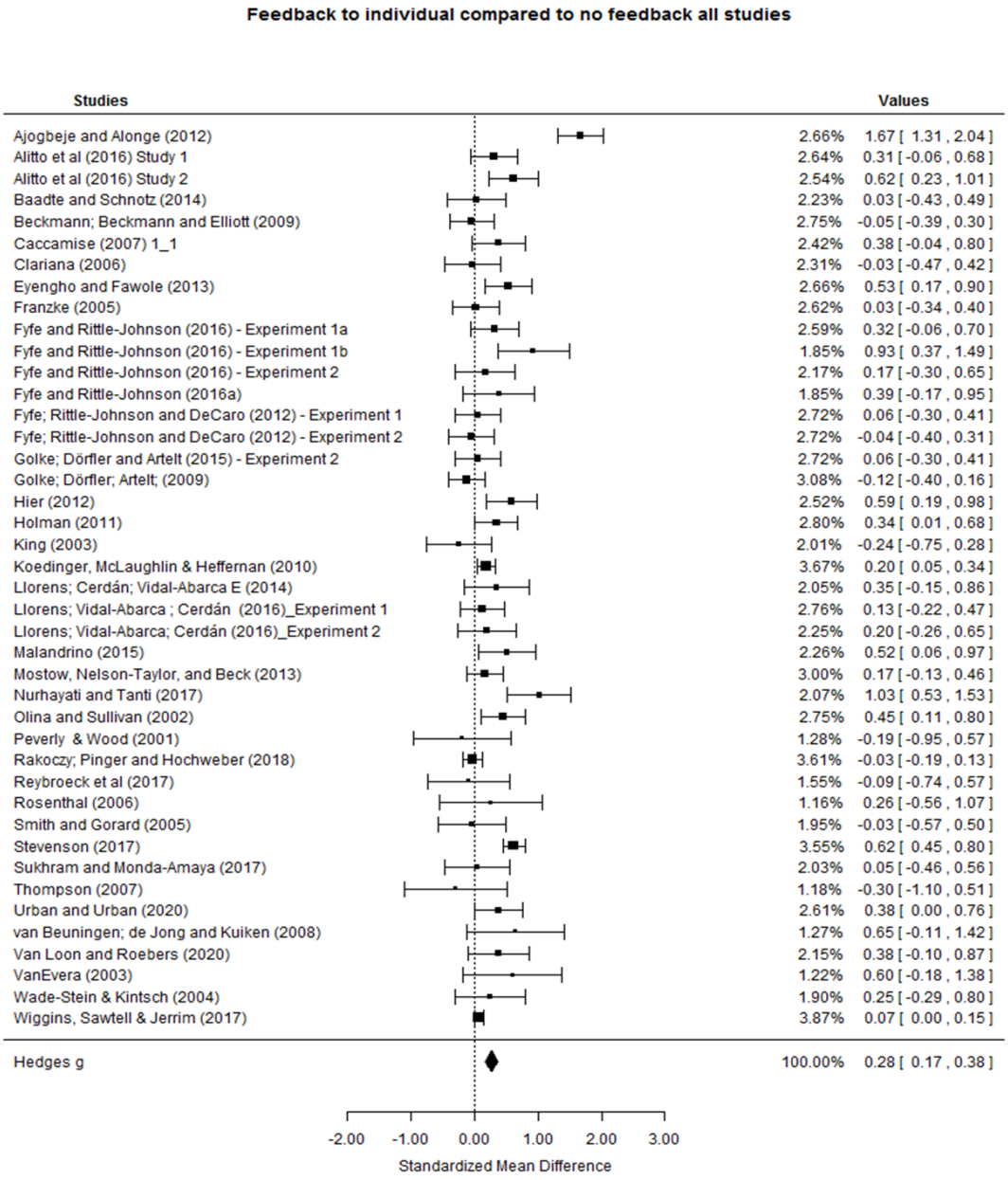


Figure 24: Synthesis: Target of feedback, individual students—All studies

Limiting the synthesis to studies with a low or moderate risk of bias reduces the statistical heterogeneity but it remains statistically significant (I^2 = 33%; Test for Heterogeneity: Q(df = 35) = 52.13, p = 0.03). The pooled estimate of effect shown Figure 25 (*g* = 0.18, 95% C.I. 0.10 to 0.26) may not be a useful indicator of the impact of feedback given to individual students.

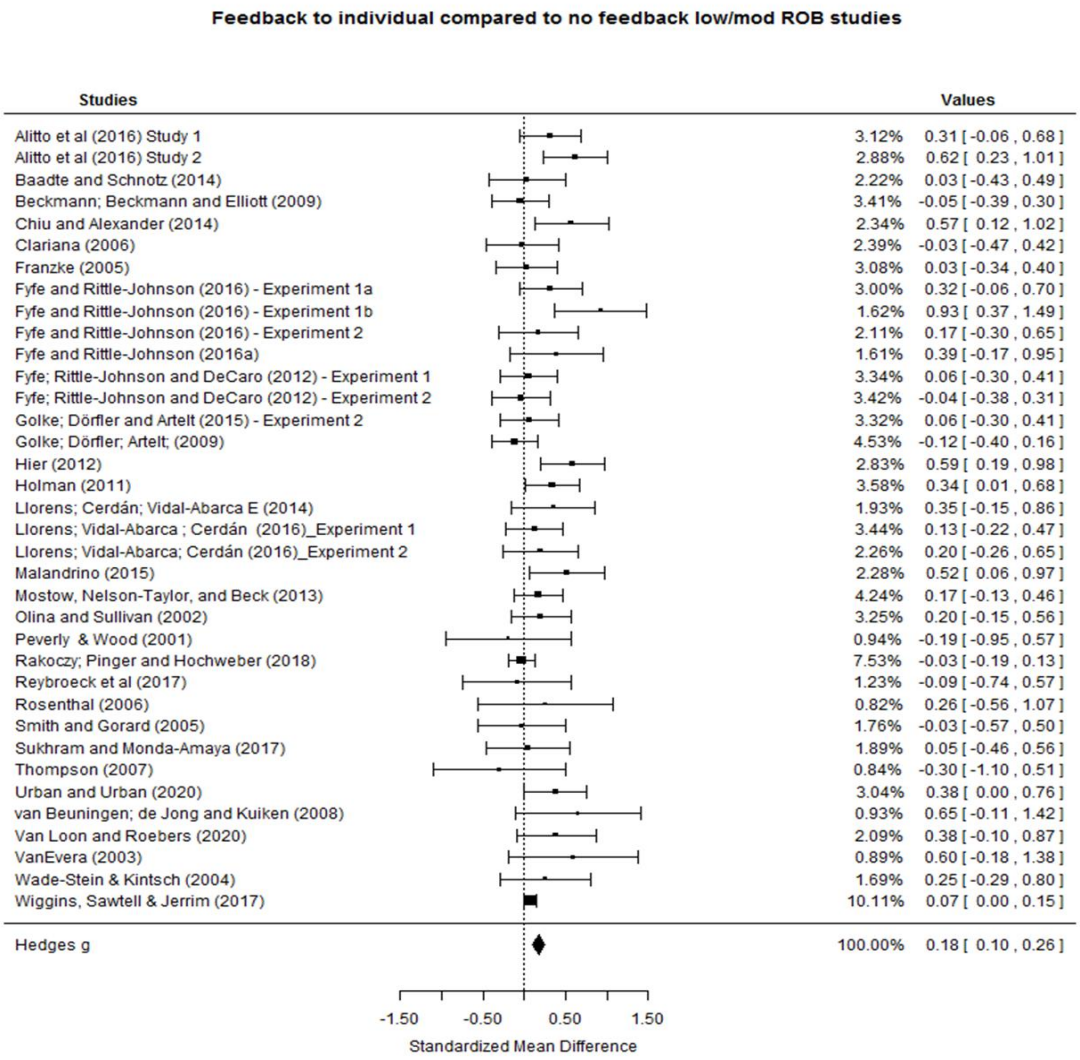


Figure 25: Synthesis: Target of feedback, individual students—Low moderate risk of bias studies

5.8.2 Group or whole class

There are four studies where feedback was given to a group or whole class. All but one study were moderate risk of bias. Figure 26 shows the results of the meta-analysis of all studies where feedback is provided to a group or class of students. There is a statistically significant heterogeneity between the studies (I^2 = 96%, Test for Heterogeneity: Q(df = 3) = 84.11, p < 0.0001).

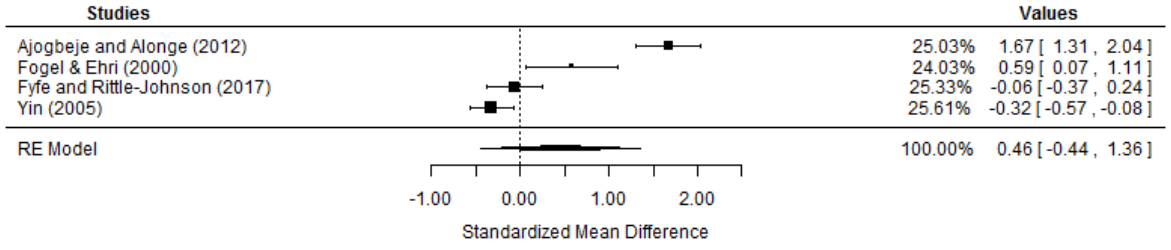


Figure 26: Synthesis: Target of feedback, group or whole class—All studies

Limiting the synthesis to studies with a low or moderate risk of bias reduces the statistical heterogeneity but it remains statistically significant (I^2 = 80%; Test for Heterogeneity: Q(df = 2) = 9.89, p = 0.007). This suggests the pooled estimate of effect shown in Figure 27 (*g* = 0.01, 95% C.I -0.42 to 0.45) is not likely to be a useful general indicator of the effect of providing feedback to groups compared to no feedback or usual practice.

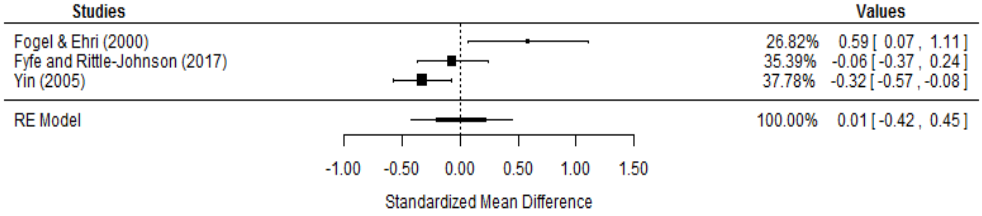


Figure 27: Synthesis: Target of feedback, group or whole class—Low or moderate risk of bias studies

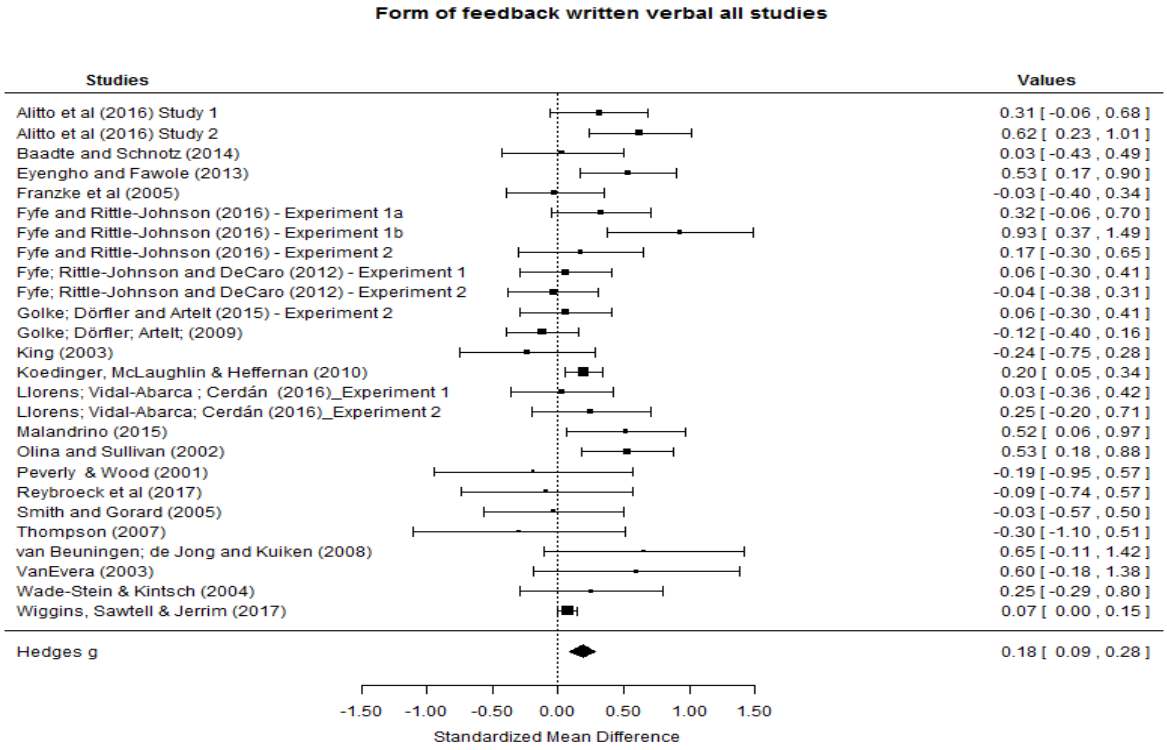
5.9 Impact of feedback: Form of feedback

Feedback can include written words (written verbal) and/or the use of written symbols, numbers or text (written non-verbal). Feedback can also be provided verbally. In some studies, combinations of feedback are provided and thus the number of studies in the synthesis are not mutually exclusive.

5.9.1 Written verbal feedback (text)

There are 27 studies that assessed the effect of feedback provided as written words. Figure 28 is a forest plot showing the synthesis of these studies with a pooled estimate of effect (*g* = 0.18, 95% C.I 0.09 to 0.28), but given the statistically significant heterogeneity (I^2 = 45%, Test for Heterogeneity: Q(df = 25) = 45.11, p = 0.008), the pooled estimate of effect may not be a useful indicator of the impact of feedback provided in written verbal form.

Figure 28: Synthesis: Form of feedback, written verbal text—All studies



Synthesis of the 24 studies of low and moderate risk of bias has statistically significant heterogeneity (I^2 = 41%, Test for Heterogeneity: Q(df = 22) = 37.38, p = 0.02). The pooled estimate of effect (*g* = 0.18, 95% C.I 0.07 to 0.28) shown

in Figure 29 therefore may not be a useful general indicator of the impact of written verbal feedback compared to no feedback or usual practice.

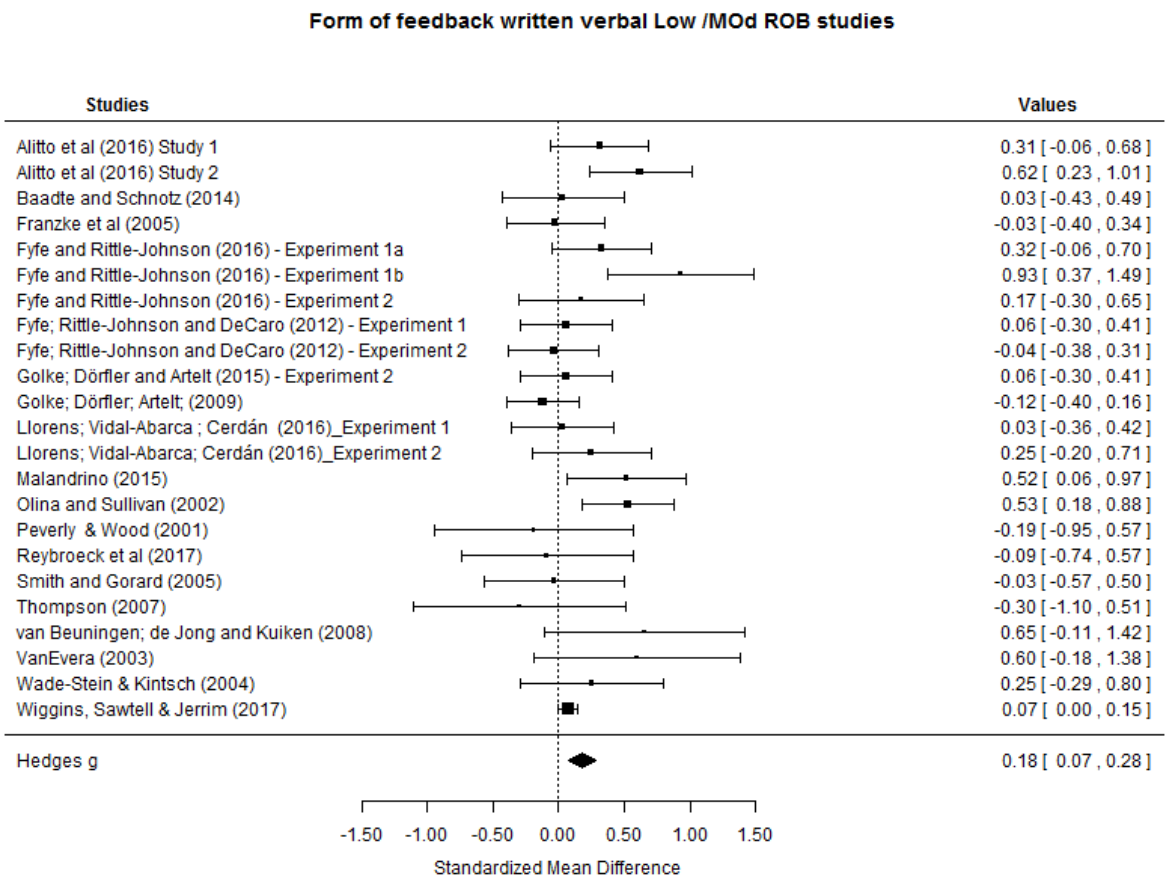


Figure 29: Synthesis: Form of feedback, written verbal text—Low or moderate risk of bias studies

One study (Golke, Dörfler and Artelt, 2015, Experiment 1) did not provide usable data to compute effect size. The authors reported that there was no significant difference in effect between the feedback and the no feedback groups.

5.9.2 Written non-verbal feedback (not using words)

There are 21 studies that assessed the effect of feedback provided in written form without using words. Figure 30 is a forest plot showing the synthesis of these studies. There is statistically significant heterogeneity between the studies (I^2 = 62%, Test for Heterogeneity: Q(df = 17) = 45.51, p = 0.0002). Two studies were judged to be at serious risk of bias. Limiting the synthesis to the 21 studies of low and moderate risk of bias reduced the heterogeneity between the studies (I^2 = 41%, Test for Heterogeneity: Q(df = 15) = 25.49, p = 0.04). However, it remains statistically significant and therefore the pooled estimate of effect shown in Figure 31 (*g* = 0.23, 95% C.I 0.10 to 0.35) may not be a useful indicator of the general impact of non-verbal written feedback.

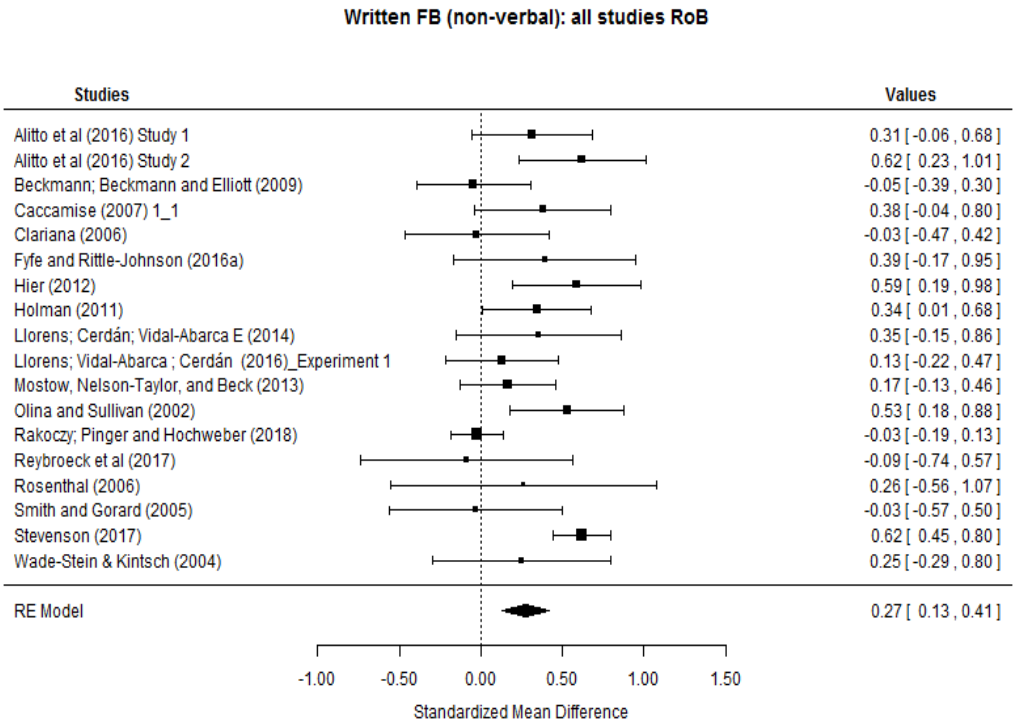


Figure 30: Synthesis: Form of feedback, written non-verbal—All studies

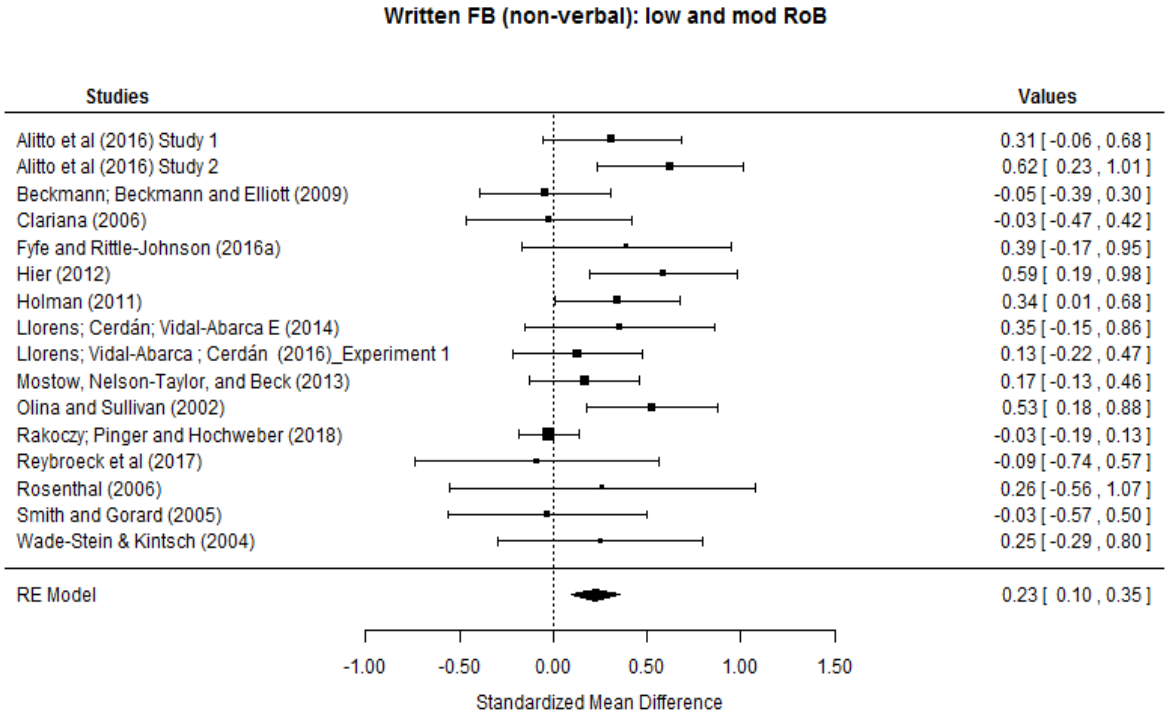


Figure 31: Synthesis: Form of feedback, written non-verbal—Low or moderate risk of bias studies

Three studies (Brossvic *et al*, 2006—Experiment 1a; Brossvic *et al*, 2006—Experiment 1b; Dihoff *et al* 2005—Experiment 1) did not provide useful data to compute effect sizes. The respective authors stated that significant

positive effects in mathematics were found in the groups that received feedback when compared to the no feedback groups.

5.9.3 Type and source of feedback

Further subgroup analysis was undertaken to explore the effect of combining the source and type of feedback (the results of which are shown in Table 21 below). The categories are not mutually exclusive—i.e. a single study could appear in more than one category. The teacher written, researcher written and the combined teacher and researcher written feedback synthesis groups of studies had statistically significant heterogeneity. The pooled estimate of effect is consistent across the synthesis with the exception of the 'Researcher non text written' feedback category, where the pooled estimate of effect is *g* = 0.52 (95% C.I. 0.17 to 0.58). There were only two studies in this category, and in both cases the feedback was a written 'score' given to students by the researchers. The results of these syntheses do not appear to suggest that who provides what type of written feedback differentially effects the impact of feedback.

Table 21: Syntheses—Types and sources of outcome combined

Outcome (n-studies)	Heterogeneity	Effect size *g*	(95% C.I)
Teacher written feedback (verbal and non-verbal), low and mod ROB studies (5)*	I^2=60% Test for Heterogeneity: Q(df=4) = 10.13, p-val=0.04	0.17	-0.13 to 0.47
Teacher written feedback (non-verbal), all studies (moderate ROB) (4)*	I^2=64% Test for Heterogeneity: Q(df=3) = 8.33, p-val=0.04	0.11	-0.19 to 0.43
Teacher written verbal feedback, mod/low ROB studies (4)	I^2=41% Test for Heterogeneity: Q(df=3) = 5.04, p-val=0.17	0.27	-0.08 to 0.62
Researcher written (verbal and non-verbal) feedback, mod/low ROB (8)	I^2=28% Test for Heterogeneity: Q(df=7) = 9.76, p-val=0.20	0.28	0.08 to 0.49
Researcher written verbal, low/mod ROB (8)	I^2=50% Test for Heterogeneity: Q(df=7) = 13.93, p-val=0.05	0.27	0.03 to 0.51
Researcher written non-verbal (low mod only) (2)	I^2=0%	0.52	0.17 to 0.58
Teacher or researcher written feedback, mod/low ROB, all studies (15)*	I^2=54% Test for Heterogeneity: Q(df=14) = 31.10, p-val=0.005	0.26	0.09 to 0.43

*Four studies did not report data to calculate effect sizes; all report group received feedback performed better than group that did not receive feedback and the results are statistically significant.

5.9.4 Verbal feedback

There are 22 studies that evaluated spoken feedback. One study was low risk of bias, and all others were moderate or serious risk of bias. There were four studies where there was no data to compute effect sizes (Brosvic *et al*, 2006, Experiment 1a, Experiment 1b, Experiment 2; Dihoff *et al*, 2005, Experiment 1). All studies reported the outcome favoured the feedback intervention group and was statistically significant (moderate risk of bias).

Figure 32 shows the results of the meta-analysis of 18 studies for which effect sizes could be calculated. The overall point estimate of effect may not be an accurate indicator of effect due to statistically significant heterogeneity between studies (I^2 = 86%, Q Test (df = 17) = 125.25, p < 0.0001).

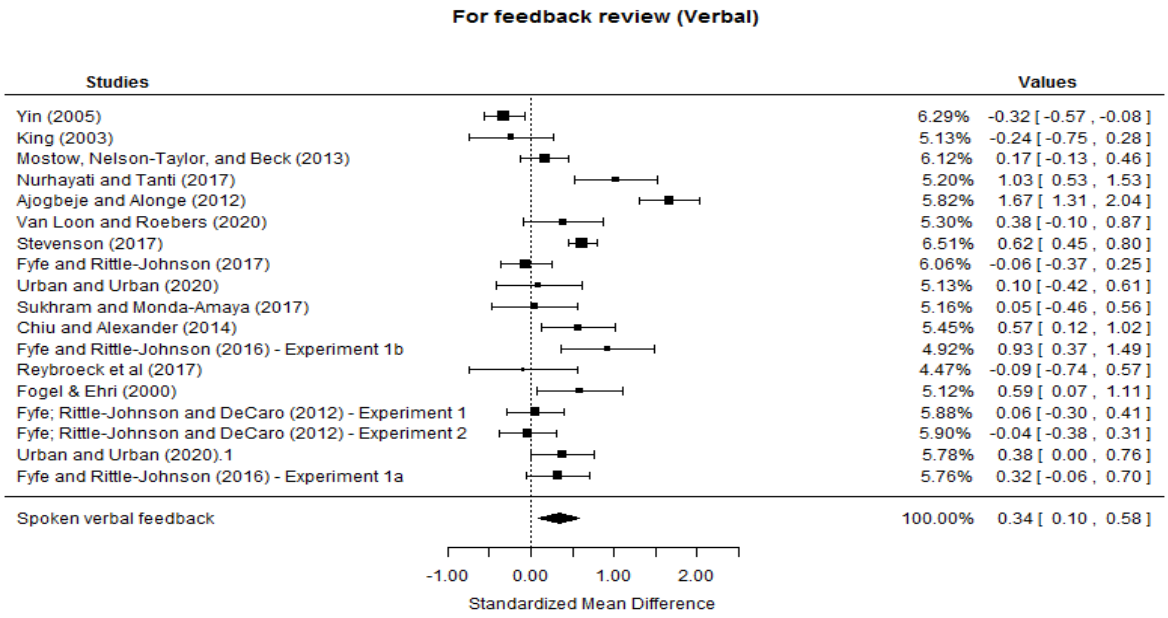


Figure 32: Synthesis: Form of feedback, verbal—All studies

Limiting the meta-analysis to studies with low or moderate risk of bias (n = 14) reduces heterogeneity (I^2 = 62%, Q(df = 13) = 34.37, p = 0.001), but it remains statistically significant. This suggest that the pooled estimate of effect shown in Figure 33 (*g* = 0.19, 95% C.I 0.01 to 0.36) may not be a useful indicator of the general impact of verbal feedback.

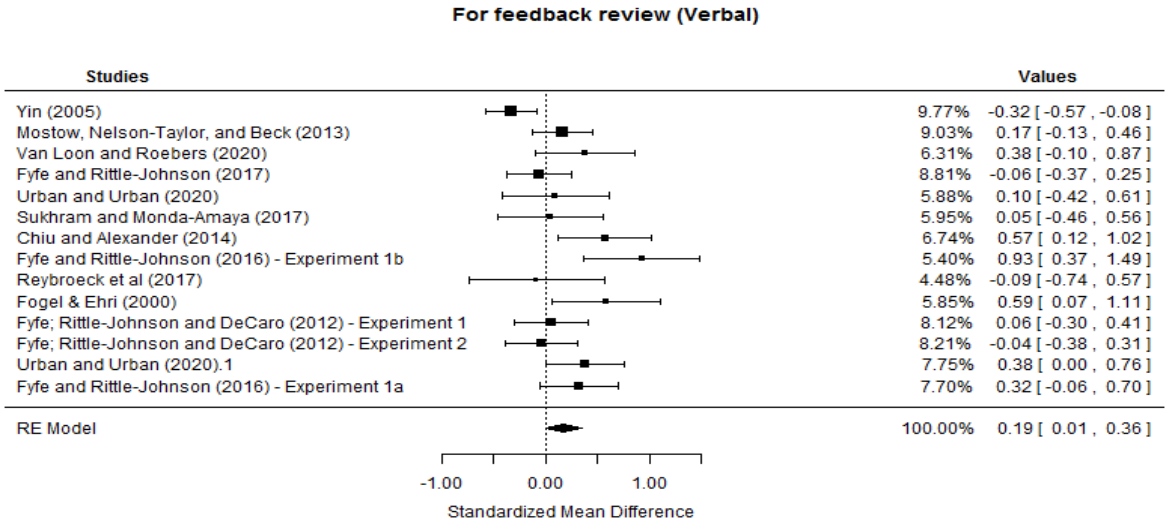


Figure 33: Synthesis: Form of feedback, verbal—Low or moderate risk of bias studies

5.10 Impact of feedback: Timing of feedback

Feedback can be provided immediately after task, during task, or delayed for a short period of time (more than one day and up to a week) after the task.

5.10.1 Feedback immediately after task

Three studies in this group did not report data to calculate effect sizes. Brossvic *et al* (2006)(two studies), Dihoff *et al* (2005) report that for all three studies, the outcomes favoured the feedback intervention group and are statistically significant..

Figure 34 shows the results of the meta-analysis of studies where immediate feedback was provided after task. There is statistically significant heterogeneity between the studies (I^2 = 71%, Test for Heterogeneity: Q(df = 25) = 89, p < 0.0001).

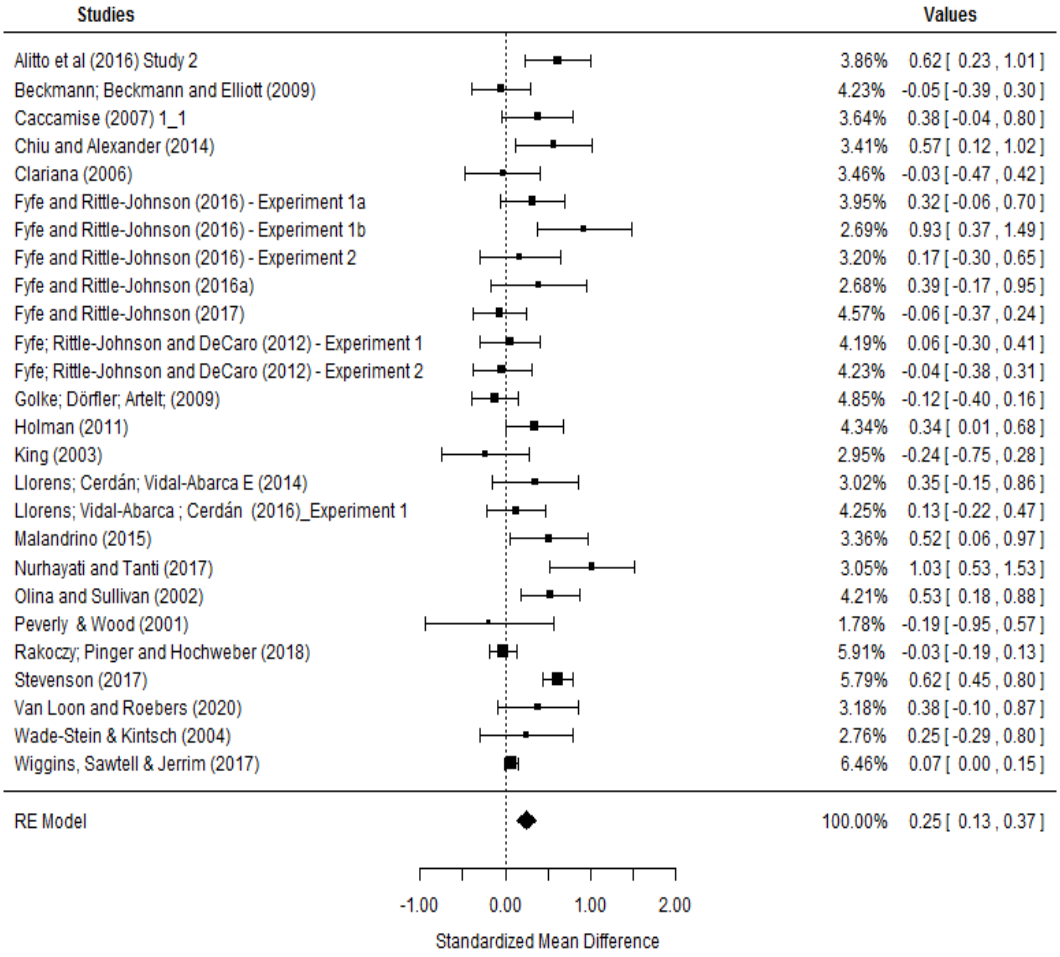


Figure 34: Synthesis: Timing of feedback, immediately after the task—All studies

As shown in Figure 35, limiting the synthesis to studies with a low or moderate risk of bias (n = 22) reduces the statistical heterogeneity, which remains statistically significant (I^2 = 52%, Test for Heterogeneity: Q(df = 21) = 44.21, p = 0.002). The pooled estimate of effect (*g* = 0.19, 95% C.I. 0.09 to 0.29) may not be a useful indicator of the impact of immediate feedback compared to no feedback or usual practice.

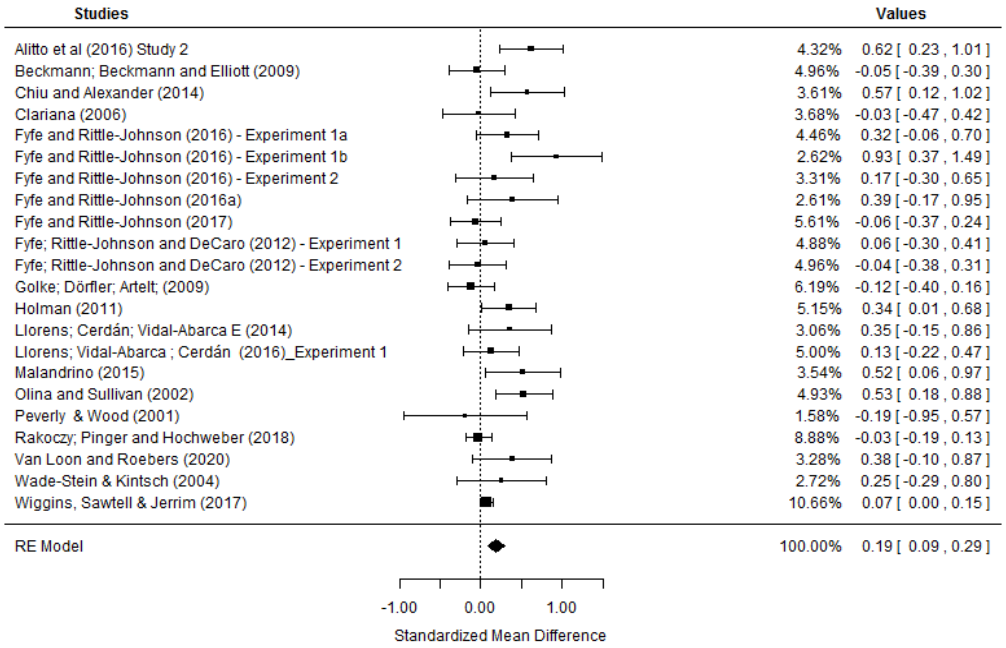


Figure 35: Synthesis: Timing of feedback, immediately after the task—Low or moderate risk of bias studies

5.10.2 Feedback during task

Figure 36 shows the results of the meta-analysis of all studies where feedback was given during task. There is statistically significant heterogeneity between the studies (I^2 = 69%, Test for Heterogeneity: Q(df = 15) = 48.67, p < 0.0001).

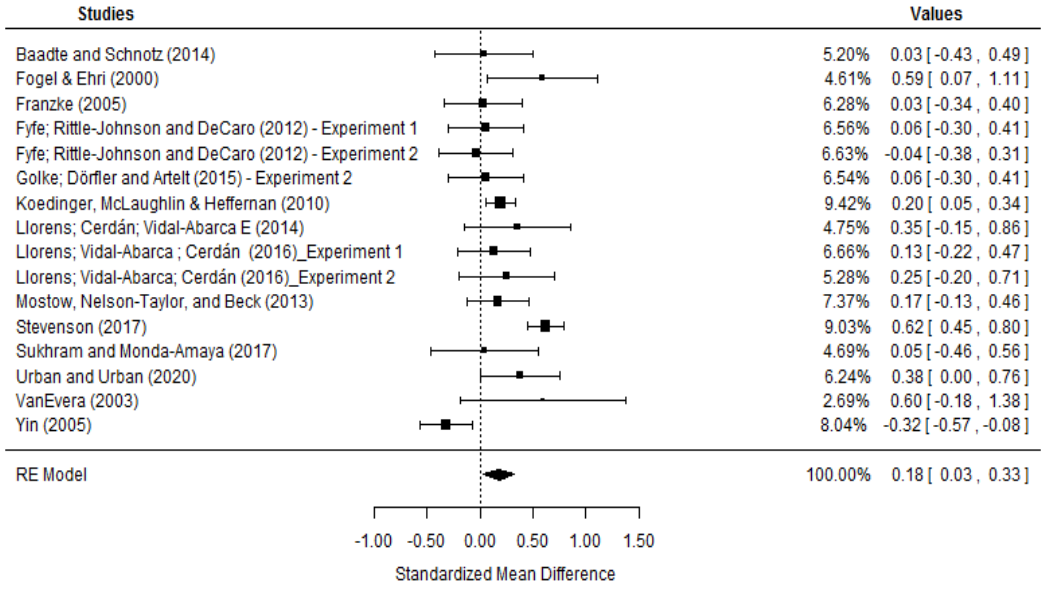


Figure 36: Synthesis: Timing of feedback, during the task—All studies

Figure 37 illustrates that limiting the synthesis to studies with low or moderate risk of bias reduces the statistical heterogeneity between the studies (I^2 = 37%, Test for Heterogeneity: Q(df = 13) = 20.72, p = 0.08). This is therefore likely to be a useful general indicator of the impact of providing feedback during task compared to no feedback or usual practice. The pooled estimate of effect (*g* = 0.11, 95% C.I. -0.02 to 0.24) indicates that feedback given during task leads to improved outcomes when compared to no feedback or usual practice. The confidence interval crosses the line of no effect and therefore we cannot be confident excluding the opposite effect.

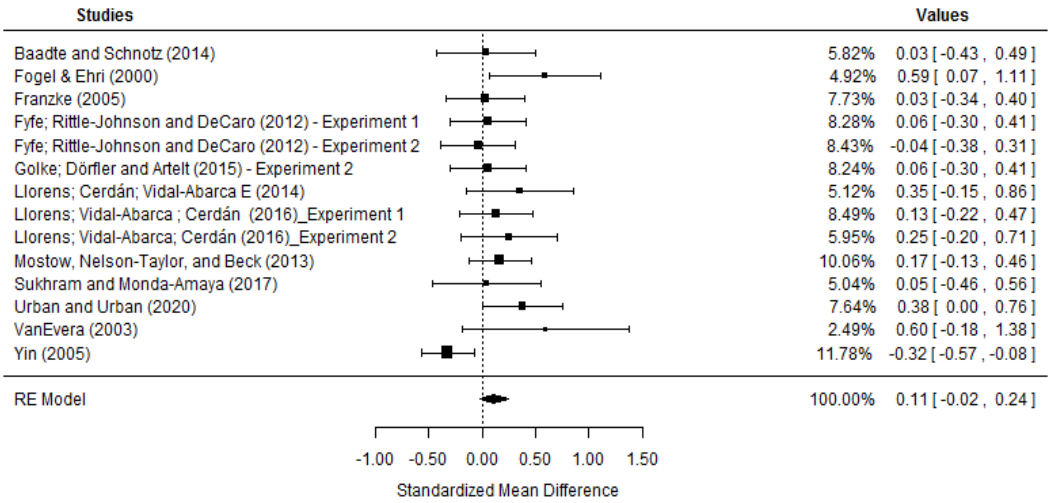


Figure 37: Synthesis: Timing of feedback, during the task—Low or moderate risk of bias studies

One study (moderate risk of bias) for which there was no data to compute effect sizes (Golke, Dörfler and Artelt, 2015, Experiment 1) reported no statistically significant difference between groups provided with feedback during the task and non-feedback groups on all outcomes.

5.10.3 Feedback delayed shortly after task (more than one day and up to a week)

Figure 38 shows the results of the meta-analysis of all studies where delayed feedback was provided. In these studies, the feedback was given between a day and a week after the 'learning task' had been completed by the students. There is a statistically significant heterogeneity between the studies (I^2 = 85%, Test for Heterogeneity: Q(df = 10) = 70.62, p < 0.0001).

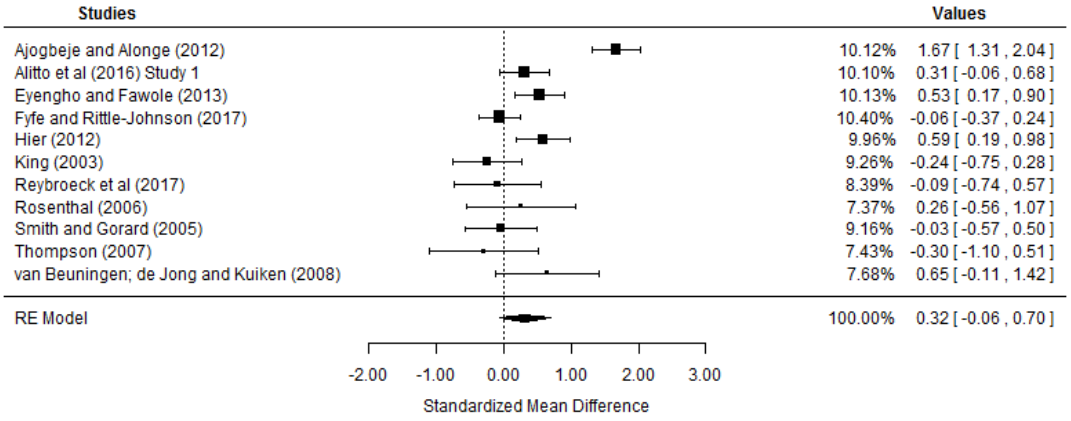


Figure 38: Synthesis: Timing of feedback, shortly delayed after task—All studies

As shown in Figure 39, limiting the synthesis to studies with a low or moderate risk of bias (n = 8) reduces the statistical heterogeneity, which is not statistically significant (I^2 = 37%; Test for Heterogeneity: Q(df = 7) = 11.04, p = 0.13). The pooled estimate of effect (*g* = 0.18, 95% C.I. -0.05 to 0.41) indicates that delayed feedback given after task leads to improved outcomes when compared to no feedback or usual practice. The confidence interval crosses the line of no effect and therefore we cannot be confident excluding the opposite effect

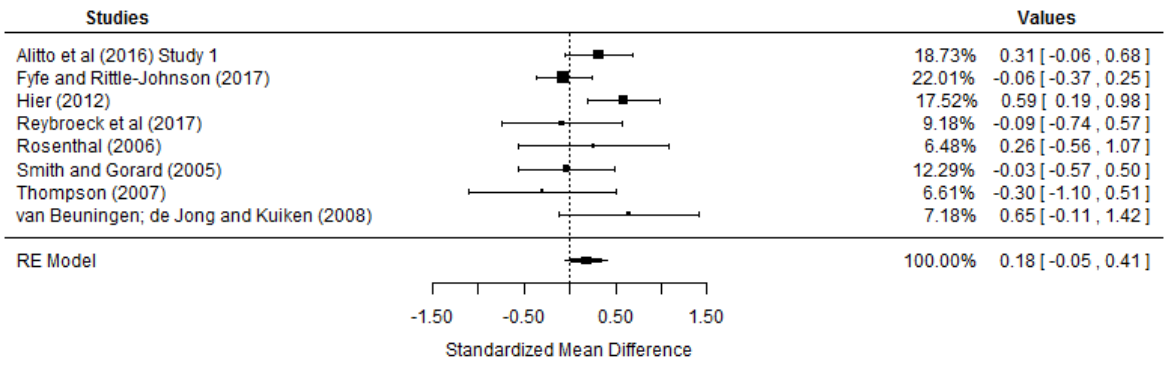


Figure 39: Synthesis: Timing of feedback, shortly delayed after task—Low or moderate risk of bias studies

5.11 Impact of feedback: Kind of feedback

The content of feedback can vary. The review coding attempted to distinguish between feedback content about the outcome or completed task (for example, scores, grades, correct/incorrect); feedback about the process of the task (for example, how the task or activity is or should be undertaken); and where the feedback is about the learners' strategies or approaches (for example, prompts to support learners' self-regulation). These coding categories are derived from Hattie and Timperley's (2007) feedback model (with 'outcome' feedback resembling the 'task level' feedback discussed in their review).

However, the descriptions of type of feedback provided lacked detail in many cases and rarely used these terms, requiring the reviewers to make judgements about which category the kind of feedback provided in the study fitted into. Whilst it was usually clear when feedback on outcome was provided, the limitations of the study descriptions means that it may be possible that some studies coded as 'outcome only' did have some additional elements of feedback. It also became clear at the in-depth review stage that it was very difficult in practice to consistently distinguish between feedback on process and feedback on strategy, based on the descriptions provided in the studies. These two categories were therefore combined for the purpose of synthesis.

Outcome feedback was included in 49 out of 51 studies in the review. However, in many of these studies, the feedback also included feedback on process or strategy. There were only two studies in which the feedback was process/strategy only.

5.11.1 Feedback on outcome only

In 32 studies, the feedback type was outcome only. There are four studies (Brossvic *et al*, 2006—Experiment 1a; Brossvic *et al*, 2006—Experiment 1b; Brossvic *et al*, 2006—Experiment 2; Dihoff *et al*, 2005—Experiment 1) in which the feedback was outcome only, which did not provide useful data to compute effect sizes. The authors state that all outcomes favoured the group receiving feedback and was statistically significant.

Figure 40 shows a synthesis of all studies where the feedback type was outcome only. There is statistically significant heterogeneity between the studies (I^2 = 45%, Test for Heterogeneity: Q(df = 27) = 49.56, p = 0.005).

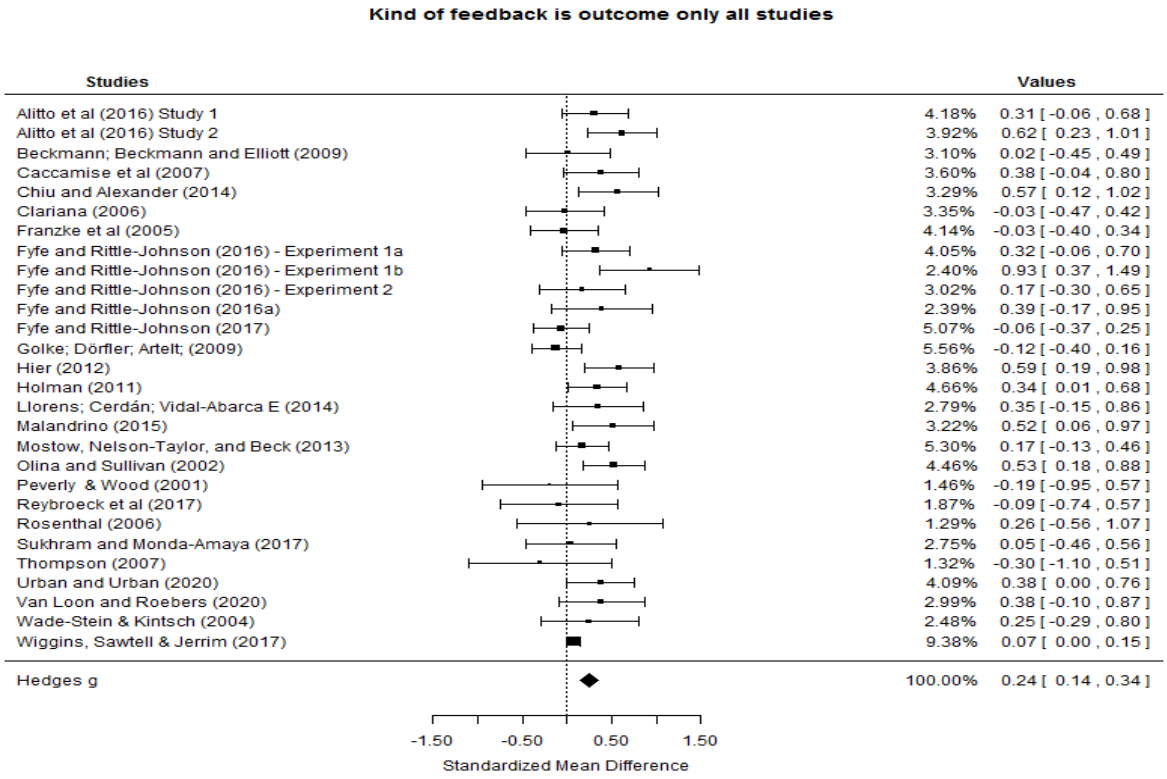


Figure 40: Synthesis: Kind of feedback, outcome only—All studies

The synthesis without the one study with a high risk of bias assessment (Caccamise *et al*, 2007) has statistically significant heterogeneity (I^2 = 47%, Test for Heterogeneity: Q(df = 26) = 49.05, p = 0.004). The pooled estimate of effect shown in Figure 41 (*g* = 0.24, 95% C.I 0.14 to 0.34) is possibly not a useful indicator of the general impact of outcome only feedback.

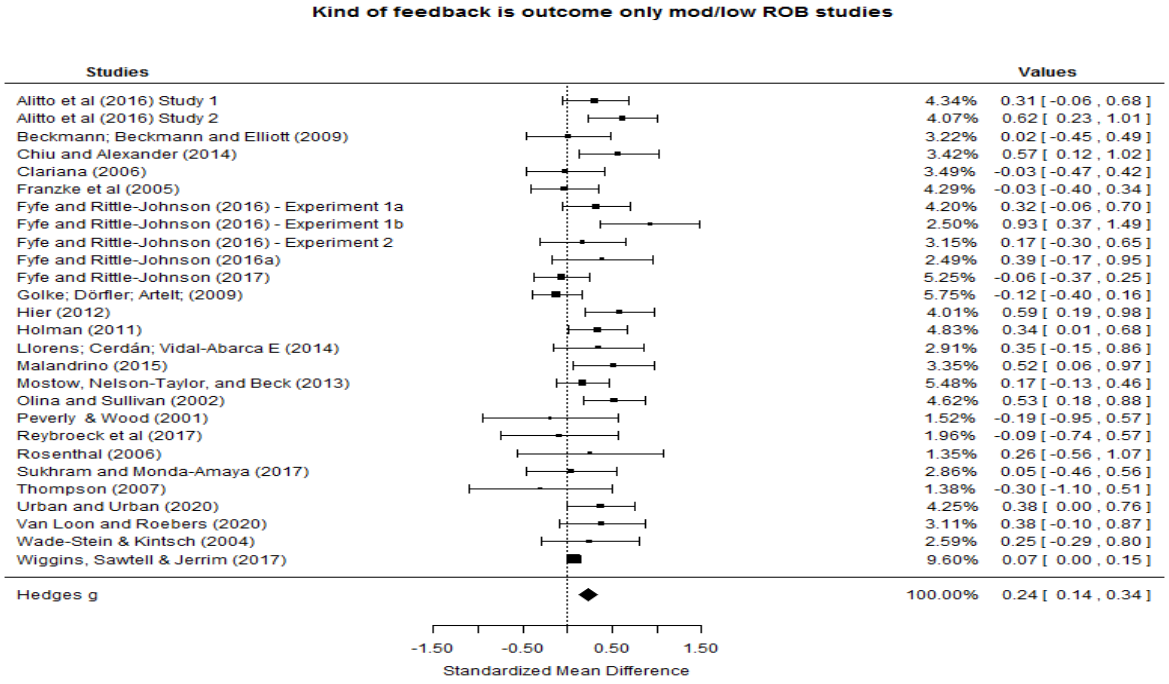


Figure 41: Synthesis: Kind of feedback, outcome only—Low or moderate risk of bias studies

5.11.2 Feedback on process or strategy

There are only two studies in which the feedback was process/strategy only (King, 2003; Rakoczy, Pinger and Hochweber, 2018). One has a high risk of bias assessment (King 2003) and therefore these two studies were not synthesised. The outcomes in both studies favoured the group that did not receive feedback, but the 95% confidence interval did not exclude the opposite effect in either study.

5.11.3 Feedback on both outcome and process/strategy

Sixteen of the studies provided feedback on both outcome and process/strategy. The synthesis of these studies is shown in figure 42. There is statistically significant heterogeneity between these studies (I^2 = 87%, Test for Heterogeneity: Q(df = 15) = 116.62, p < 0.0001).

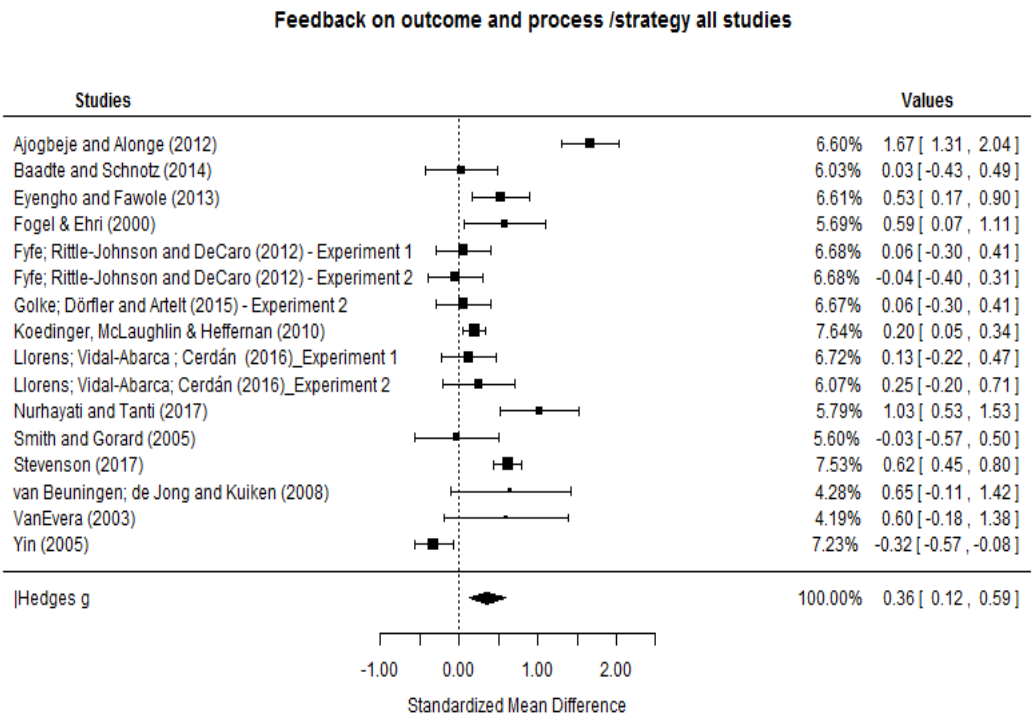


Figure 42: Synthesis: Kind of feedback, outcome and process/strategy—All studies

Restricting the synthesis to studies with a low or moderate risk of bias assessment reduces the heterogeneity between the studies (I^2 = 45%, Test for Heterogeneity: Q(df = 10) = 18.34, p = 0.05). The pooled estimate of effect shown in Figure 43 (*g* = 0.09, 95% C.I -0.08 to 0.26) indicates that the group that received feedback on outcomes and process had a better outcome. However, the 95% confidence interval crosses the line of no effect and therefore we cannot exclude the opposite effect.

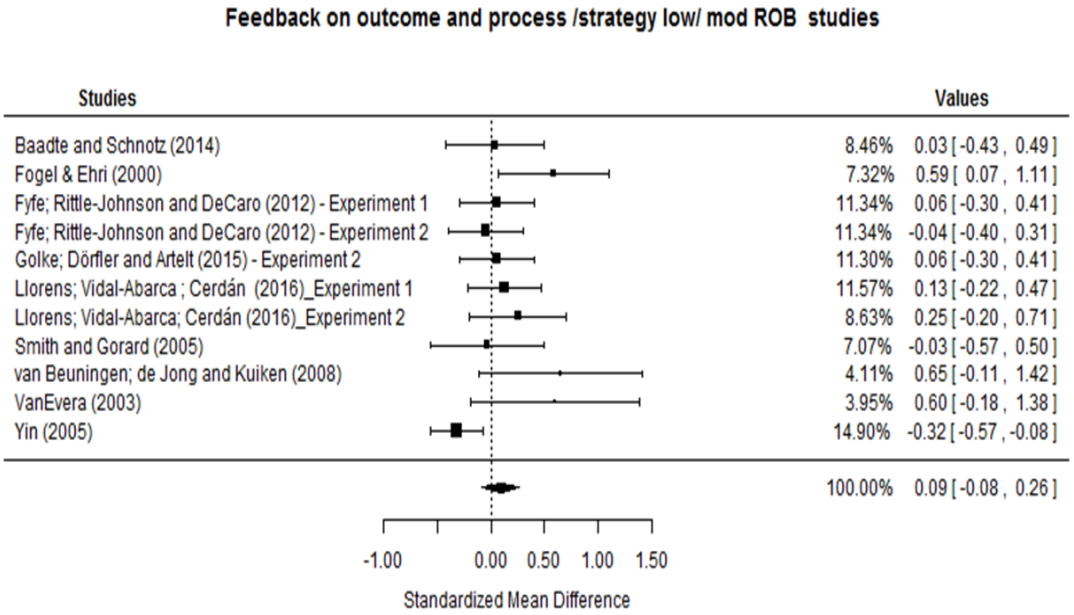


Figure 43: Synthesis: Kind of feedback, outcome and process/strategy—Low or moderate risk of bias studies

6 Applicability and gaps of the evidence base

The review design ensured that only studies carried out in mainstream educational settings that measured attainment outcomes were included. The studies are international in scope. The review selection criteria limited the focus of investigation to studies where the 'active' practice investigated was feedback only. There are a wider range of educational practices in which 'feedback' may be included as an element.

The pragmatic requirements of the review meant that the screening was incomplete and that the review focused only on studies published after 2000. Therefore there may be other published research investigating of the impact of feedback on attainment in mainstream educational settings that has not been either identified or included in the review.

7 Overall evidence statement

The results of the review provide evidence to suggest that, on average, single component 'feedback only' interventions lead to better attainment outcomes for students in mainstream education, when compared to no feedback or usual practice (*g* = 0.17, 95% C.I 0.09 to 0.25, low/moderate risk of bias studies). However, the statistical analysis found considerable unexplained heterogeneity in the main and subgroup analysis. Furthermore, there are also studies where the results showed that students who received feedback had a worse outcome than those who either received no feedback or usual practice. This may indicate that that not all 'feedback only' interventions are effective in improving attainment in all contexts.

Caution is required when interpreting all of the results of the subgroup analysis, given the degree of heterogeneity between studies and the lack of direct comparisons between studies or statistical moderator analysis. There are quite possibly factors other than the characteristics investigated in this review that are systematically different between these studies.

The results of the subgroup analysis do, in some cases, appear indicative of some kind of systemic variation in impact. The results of feedback studies in literacy appear to favour feedback compared to no feedback or usual practice, whereas in mathematics and science, the results are more equivocal. The results appeared to favour feedback when compared to no feedback at primary level, particularly at Key Stage 1, but were more equivocal at secondary level. The positive impact of feedback of digital/automated feedback appears slightly more clear than it is for feedback from a person (either teacher or researcher).

The results for feedback of outcome only and feedback on outcome and process/strategy are different. However, this difference should not be interpreted as if this were a direct comparison between the two kinds of feedback. There was also statistically significant heterogeneity between the low/moderate risk of bias outcome studies in these syntheses. The outcome feedback in the studies was of the form correct/incorrect or a grade or score of some kind, which falls into the Hattie and Timperley category 'feedback on the task'. The review results are therefore arguably consistent with the Hattie and Timperley model in that they state that feedback on the task can be successful when given immediately and when aligned with task definition. However, we might anticipate based on the same model that feedback on both outcome and process/strategy might have a stronger effect than feedback on outcome alone. While, on average, the effect was still positive versus no feedback/usual practice, the effect of outcome *and* process/strategy feedback (*g* = 0.09, 95% C.I -0.08 to 0.25) was not as large as feedback on outcome alone *(g* = 0.24, 95% C.I 0.14 to 0.34). However, as discussed above, coding for the kind of feedback based on the information provided in studies was challenging. Whilst the reviewers were able to code for 'outcome feedback', the detail of process/strategy feedback was often less clear. It may be the case that the process/strategy feedback in these studies was more limited than that envisaged in the Hattie and Timperley model.

Table 22: Summary of findings

	Specific subject outcomes	Impact Hedges *g* (95% C.I)	Number, location and design of studies; Number of participants (if available)	Impact heterogeneity
Feedback compared to no feedback or usual practice	**All subjects**	All studies 0.27 (0.16, 0.25) Low/moderate ROB (44 studies) 0.17 (0.09–0.25)	UK 3, USA 30, Belgium 1, Germany 5, Indonesia 1, Latvia 1, The Netherlands 2, Nigeria 2, Slovakia 1, Spain 3, Switzerland 1, Taiwan 1. Randomised Controlled Trial = 40. Prospective Quantitative Experimental design = 11. Data from Approximately 14,400 students.	I^2=76%, Test for Heterogeneity: Q (df=45) = 187.95, p=0.0001 I^2 = 44%, Test for Heterogeneity: Q(df = 37) = 65.91 p-val = 0.0024)
Curriculum subjects tested				
Curriculum subjects	**Literacy** All studies	0.22 (0.12, 0.31)	13 studies from US, 3 from Germany, 3 from Spain, 2 from UK, 1 from Belgium, 1 from Nigeria --- 12 RCTs, 5 cluster RCTs, 1 multisite RCT, 5 quasi-experimental designs --- Data from 9,288 pupils*	No significant heterogeneity (I^2=31.9%, p=0.07)
	Literacy Low/Mod ROB studies	0.19 (0.09, 0.28)	12 studies from US, 3 from Germany, 3 from Spain, 2 from UK, 1 from Belgium --- 12 RCTs, 5 cluster RCTs, 1 multisite RCT, 3 quasi-experimental designs --- Data from 8,849 pupils*	No significant heterogeneity (I^2=25.5%, p=0.14)
	Mathematics All studies	0.25 (0.06, 0.45)	9 studies from US, 2 from UK, 1 from Germany, 1 from Nigeria --- 8 RCTs, 2 cluster RCTs, 3 quasi-experimental designs --- Data from 9,552 pupils*	Significant heterogeneity (I^2=86.5%, p<0.0001)
	Maths Low/Mod ROB studies	0.08 (–0.03, 0.20)	8 studies from US, 2 from UK, 1 from Germany --- 8 RCTs, 2 cluster RCTs, 1 quasi-experimental design --- Data from 7,968 pupils*	No significant heterogeneity (I^2=36.1%, p=0.11)
	Science All studies	0.03 (–0.37, 0.42)	4 studies from US, 1 from UK, 1 from Germany, 1 from Indonesia --- 2 RCTs, 2 cluster RCTs, 3 quasi-experimental designs --- Data from 741 pupils	Significant heterogeneity (I^2=80.4%, p<0.0001)
	Science Low/Mod ROB studies	-0.15 (–0.46, 0.17)	3 studies from US, 1 from UK, 1 from Germany ---	Heterogeneity (I^2=57.8%, p=0.05)

			2 RCTs, 2 cluster RCTs, 1 quasi-experimental design --- Data from 606 pupils	
	Combined subjects outcomes (*not mutually exclusive*)	**Overall Impact SMD (95% C.I)**	**Number, location and design of studies; Number of participants (if available)**	**Impact heterogeneity**
Key stages				
Key Stage 1 (aged 5–7 years) Low/Mod ROB studies	Mathematics (N=4), literacy (N=1), and cognitive outcomes (N=3)	0.34 (0.15, 0.52)	5 studies from US, 1 from Taiwan, 1 from Slovakia, 1 from Switzerland --- 7 RCTs, 1 quasi-experimental design --- Data from 702 pupils*	No significant heterogeneity (I^2=37%, p=0.13)
Key Stage 2 (aged 8–11 years) Low/Mod ROB studies	Literacy (N=11), mathematics (N=8), science (N=3), social studies (N=2)	0.20 (0.07, 0.33)	15 studies from US, 2 from Germany, 2 from UK, 1 from The Netherlands --- 12 RCTs, 4 cluster RCTs, 3 quasi-experimental designs --- Data from 8,540 pupils*	Significant heterogeneity (I^2=62%, p=0.0002)
Key Stage 3 (aged 12–14 years) Low/Mod ROB studies	Literacy (N=10), science (N=3), mathematics (N=1), language (N=1), social studies (N=1), cognitive outcomes (N=1)	0.05 (–0.07, 0.19)	8 studies from US, 3 from Germany, 3 from Spain, 1 from The Netherlands, 1 from UK --- 11 RCTs, 1 multisite RCT, 3 cluster RCTs, 1 quasi-experimental design --- Data from 1,875 pupils*	No significant heterogeneity (I^2=30%, p=0.12)
Key Stage 4 (aged 15–16 years) Low/Mod ROB studies	Literacy (N=2), mathematics (N=2), science (N=1), cognitive outcomes (N=1)	–0.04 (–0.17, 0.09)	4 studies from US, 1 from Germany, 1 from UK --- 4 RCTs, 1 multisite RCT, 1 cluster RCT --- Data from 1,024 pupils	Significant heterogeneity (I^2=0%, p=0.99)
Educational setting				
Primary schools All studies	Literacy (N=9), mathematics (N=8), science (N=2), cognitive outcomes (N=4), social studies (N=1)	0.30 (0.18, 0.43)	16 studies from US, 1 from Germany, 1 from UK, 1 from Taiwan, 1 from Slovakia, 1 from Switzerland, 1 from The Netherlands --- 15 RCTs, 4 cluster RCTs, 3 quasi-experimental designs --- Data from 9,527 pupils*	Significant heterogeneity (I^2=68.68%, p<0.0001)
Primary schools Low/Mod ROB studies	Literacy (N=9), mathematics (N=8), science (N=1), cognitive outcomes (N=3), social studies (N=1)	0.29 (0.17, 0.40)	15 studies from US, 1 from Germany, 1 from UK, 1 from Taiwan, 1 from Slovakia, 1 from Switzerland --- 15 RCTs, 4 cluster RCTs, 1 quasi-experimental design --- Data from 8,463 pupils*	Significant heterogeneity (I^2=52.6%, p=0.003)

Secondary schools All studies	Literacy (N=13), science (N=5), mathematics (N=5), language (N=2) cognitive outcomes (N=2)	0.23 (0.06, 0.40)	10 studies from US, 3 from Germany, 3 from Spain, 2 from UK, 1 from Latvia, 1 from The Netherlands, 1 from Belgium, 2 from Nigeria, I from Indonesia --- 12 RCTs, 5 cluster RCTs, 7 quasi-experimental designs --- Data from 4,857 pupils*	Significant heterogeneity (I^2=80.9%, p<0.0001)
Secondary schools Low/Mod ROB studies	Literacy (N=11), science (N=4), mathematics (N=3), language (N=2) cognitive outcomes (N=2)	0.05 (–0.07, 0.16)	8 studies from US, 3 from Germany, 3 from Spain, 2 from UK, 1 from Latvia, 1 from The Netherlands, 1 from Belgium --- 12 RCTs, 5 cluster RCTs, 2 quasi-experimental designs --- Data from 2,764 pupils*	No significant heterogeneity (I^2=32.2%, p=0.088)
Source of feedback				
Teacher All studies	Literacy (N=4), science (N=5), mathematics (N=2), language (N=1) cognitive outcomes (N=1)	0.24 (–0.04, 0.51)	4 studies from US, 1 from UK, 1 from Germany, 1 from Belgium, 1 from Latvia, 1 from the Indonesia, 1 from Nigeria --- 6 cluster RCTs, 4 quasi-experimental designs --- Data from 1,778 pupils	Significant heterogeneity (I^2=81%, p<0.0001)
Teacher Low/Mod ROB studies	Literacy (N=2), science (N=3), mathematics (N=2), language (N=1) cognitive outcomes (N=1)	0.13 (–0.15, 0.41)	3 studies from US, 1 from UK, 1 from Germany, 1 from Belgium, 1 from Latvia, --- 6 cluster RCTs, 1 quasi-experimental design --- Data from 1,447 pupils	Significant heterogeneity (I^2=74%, p=0.0007)
Researcher All studies	Literacy (N=4), science (N=2), mathematics (N=8), language (N=1) cognitive outcomes (N=3)	0.38 (0.14, 0.61)	13 studies from US, 1 from Taiwan, 1 from Slovakia, 1 from The Netherlands, 1 from Switzerland, 1 from Nigeria --- 14 RCTs, 2 cluster RCTs, 2 quasi-experimental designs --- Data from 1,654 pupils*	Significant heterogeneity (I^2=78%, p<0.0001)
Researcher Low/Mod ROB studies	Literacy (N=4), science (N=1), mathematics (N=7), language (N=1) cognitive outcomes (N=3)	0.30 (0.16, 0.44)	12 studies from US, 1 from Taiwan, 1 from Slovakia, 1 from The Netherlands, 1 from Switzerland --- 14 RCTs, 2 cluster RCTs, 2 quasi-experimental designs --- Data from 1,349 pupils*	Significant heterogeneity (I^2=61%, p<0.0001)
Teacher/researcher Low/Mod ROB studies	Literacy (N=6), science (N=3), mathematics (N=9), language (N=2) cognitive outcomes (N=4)	0.25 (0.1, 0.41)	14 studies from US, 1 from Taiwan, 1 from UK, 1 from Slovakia, 1 from The Netherlands, 1 from Switzerland, 1 from Latvia, 1 from Germany, 1 from Belgium --- 15 RCTs, 6 cluster RCTs, 1 quasi-experimental design --- Data from 2,635 pupils*	Significant heterogeneity (I^2=61%, p<0.0001)

Digital/automated All studies	Literacy (N=15), science (N=2), mathematics (N=7), social studies (N=1), cognitive outcomes (N=3)	0.23 (0.13, 0.33)	16 studies from US, 2 from UK, 3 from Germany, 3 from Spain, 1 from Taiwan, 1 from The Netherlands --- 19 RCTs, 2 cluster RCTs, 5 quasi-experimental designs --- Data from 11,497 pupils*	Significant heterogeneity (I^2=63%, p<0.0001)
Digital/automated Low/Mod ROB studies	Literacy (N=14), science (N=2), mathematics (N=6), social studies (N=1), cognitive outcomes (N=2)	0.19 (0.09, 0.28)	14 studies from US, 2 from UK, 3 from Germany, 3 from Spain, 1 from Taiwan --- 19 RCTs, 2 cluster RCTs, 2 quasi-experimental designs --- Data from 8,911 pupils*	No significant heterogeneity (I^2=42%, p=0.02)
Feedback directed to				
Individual pupil All studies	Literacy (N=21), science (N=5), mathematics (N=11), social studies (N=1), cognitive outcomes (N=5)	0.28 (0.17, 0.38)	23 studies from US, 4 from Germany, 3 from UK, 3 from Spain, 2 from The Netherlands, 1 from Belgium, 1 from Latvia, 1 from Slovakia, 1 from Switzerland, 2 from Nigeria, 1 from Indonesia, 1 from Taiwan --- 26 RCTs, 6 cluster RCTs, 1 multisite RCT, 10 quasi-experimental designs --- Data from 13,801 pupils*	Significant heterogeneity (I^2=75%, p<0.0001)
Individual pupil Low/Mod ROB studies	Literacy (N=19), science (N=2), mathematics (N=9), social studies (N=1), cognitive outcomes (N=4)	0.18(0.10, 0.26)	20 studies from US, 4 from Germany, 3 from UK, 3 from Spain, 1 from The Netherlands, 1 from Belgium, 1 from Latvia, 1 from Slovakia, 1 from Switzerland, 1 from Taiwan --- 26 RCTs, 6 cluster RCTs, 1 multisite RCT, 3 quasi-experimental designs --- Data from 10,644 pupils*	No significant heterogeneity (I^2=33%, p=0.03)
Group All studies	Literacy (N=1), mathematics (N=2), science (N=1)	0.46 (–0.44, 1.36)	3 studies from US, 1 from Nigeria --- 1 RCT, 2 cluster RCTs, 1 quasi-experimental design --- Data from 823 pupils	Significant heterogeneity (I^2=96.4%, p<0.0001)
Group Low/Mod ROB studies	Literacy (N=1), mathematics (N=1), science (N=1)	0.01 (–0.42, 0.45)	3 studies from US --- 1 RCT, 2 cluster RCTs --- Data from 583 pupils	Significant heterogeneity (I^2=80%, p=0.007)
Form of feedback				
Written verbal All studies	Literacy (N=14), mathematics (N=8), science (N=5), social science (N=1), cognitive outcomes (N=1)	0.18(0.09, 0.28)	15 studies from US, 3 from Germany, 2 from UK, 2 from Spain, 1 from The Netherlands, 1 from Belgium, 1 from Latvia, 1 from Nigeria --- 17 RCTs, 3 cluster RCTs, 6 quasi-experimental designs ---	Significant heterogeneity (I^2=45%, p=0.008)

			Data from 10,416 pupils*	
Written verbal Low/Mod ROB studies	Literacy (N=13), mathematics (N=7), science (N=4), social science (N=1), cognitive outcomes (N=1)	0.18 (0.07, 0.28)	13 studies from US, 3 from Germany, 2 from UK, 2 from Spain, 1 from The Netherlands, 1 from Belgium, 1 from Latvia --- 17 RCTs, 3 cluster RCTs, 3 quasi-experimental designs --- Data from 8,811 pupils*	Significant heterogeneity (I^2=41%, p=0.02)
Written non-verbal All studies	Literacy (N=12), mathematics (N=3), science (N=2), language (N=1), cognitive outcomes (N=3)	0.27 (0.13, 0.41)	10 studies from US, 2 from UK, 2 from Spain, 1 from Germany, 1 from The Netherlands, 1 from Belgium, 1 from Latvia --- 8 RCTs, 4 cluster RCTs, 6 quasi-experimental designs --- Data from 3,552 pupils*	Significant heterogeneity (I^2=62%, p=0.0002)
Written non-verbal Low/Mod ROB studies	Literacy (N=11), mathematics (N=3), science (N=2), language (N=1), cognitive outcomes (N=2)	0.23 (0.10, 0.35)	9 studies from US, 2 from UK, 2 from Spain, 1 from Germany, 1 from Belgium, 1 from Latvia --- 9 RCTs, 4 cluster RCTs, 4 quasi-experimental designs --- Data from 2,310 pupils*	Significant heterogeneity (I^2=41.2%, p=0.04)
Verbal feedback All studies	Literacy (N=8) mathematics (N= 11) science (N=3) cognitive (N=4)	0.34 (0.10 to 0.58)	15 studies from USA, 1 each from Belgium, Indonesia, The Netherlands, Nigeria, Slovakia, Switzerland, Taiwan --- Study design, 15 RCTs, 3 Cluster RCT, 5 quasi-experimental designs --- Data from 2,088 pupils	Significant heterogeneity (I^2=86%, p<0.0001).
Verbal feedback Low/Mod ROB studies	Literacy (N=8) mathematics (N= 9) science (N=1) cognitive (N=3)	0.19 (0.01 to 0.36)	13 studies from USA, 1 each from Belgium, Slovakia, Switzerland, Taiwan -- Study design, 14 RCTs, 3 Cluster RCT, 1 quasi-experimental design -- Data from 669 pupils	Significant heterogeneity (I^2=62%, p=0.001)
Timing of feedback				
Immediate All studies	Literacy (N=10), mathematics (N=9), science (N=4), cognitive outcomes (N=5)	0.25 (0.13, 0.37)	15 studies from US, 2 from UK, 2 from Spain, 2 from Germany, 1 from Switzerland, 1 from Latvia, 1 from The Netherlands, 1 from Taiwan, 1 from Indonesia --- 16 RCTs, 4 cluster RCTs, 6 quasi-experimental designs --- Data from 10,672 pupils*	Significant heterogeneity (I^2=71.9%, p<0.0001)
Immediate Low/Mod ROB studies	Literacy (N=9), mathematics (N=9), science	0.19 (0.09, 0.29)	13 studies from US, 2 from UK, 2 from Spain, 2 from Germany, 1 from	Significant heterogeneity

	(N=2), cognitive outcomes (N=4)		Switzerland, 1 from Latvia, 1 from Taiwan --- 16 RCTs, 4 cluster RCTs, 2 quasi-experimental designs --- Data from 9,295 pupils*	(I^2=52.5%, p=0.0022)
During All studies	Literacy (N=8), mathematics (N=3), science (N=3), social studies (N=1), cognitive outcomes (N=2)	0.18 (0.03, 0.33)	9 studies from US, 3 from Spain, 2 from Germany, 1 from The Netherlands, 1 from Slovakia --- 9 RCTs, 1 multisite RCT, 3 cluster RCTs, 3 quasi-experimental designs --- Data from 4,099 pupils*	Significant heterogeneity (I^2=69.2%, p<0.0001)
During Low/Mod ROB studies	Literacy (N=8), mathematics (N=2), science (N=3), social studies (N=1), cognitive outcomes (N=1)	0.11 (–0.02, 0.24)	8 studies from US, 3 from Spain, 2 from Germany, 1 from Slovakia --- 9 RCTs, 1 multisite RCT, 3 cluster RCTs, 1 quasi-experimental design --- Data from 1,756 pupils*	No significant heterogeneity (I^2=37%, p=0.079)
Short delay All studies	Literacy (N=6), mathematics (N=4), science (N=2), language (N=2)	0.32 (–0.06, 0.70)	6 studies from US, 2 from Nigeria, 1 from UK, 1 from Belgium, 1 from The Netherlands --- 6 RCTs, 1 cluster RCT, 4 quasi-experimental designs --- Data from 1348 pupils	Significant heterogeneity (I^2=85%, p<0.0001)
Short delay Low/Mod ROB studies	Literacy (N=5), mathematics (N=3), science (N=1), language (N=2)	0.18 (–0.05, 0.41)	5 studies from US, 1 from UK, 1 from Belgium, 1 from The Netherlands --- 6 RCTs, 1 cluster RCT, 1 quasi-experimental design --- Data from 847 pupils	No significant heterogeneity (I^2=36%, p=0.137)
Kind of feedback				
Feedback type: outcomes only All studies	Literacy (N=16), mathematics (N=7), science (N=1), cognitive outcomes (N=5)	0.24 (0.14, 0.34)	19 studies from US, 2 from UK, 1 from Belgium, 1 from Germany, 1 from Spain, 1 from Latvia, 1 from Switzerland, 1 from Slovakia, 1 from Taiwan --- 19 RCTs, 1 multisite RCT, 4 cluster RCTs, 4 quasi-experimental designs --- Data from 9,401 pupils*	Significant heterogeneity (I^2=45%, p=0.005)
Outcomes only Low/Mod ROB studies	Literacy (N=15), mathematics (N=7), science (N=1), cognitive outcomes (N=5)	0.24 (0.14, 0.34)	18 studies from US, 2 from UK, 1 from Belgium, 1 from Germany, 1 from Spain, 1 from Latvia, 1 from Switzerland, 1 from Slovakia, 1 from Taiwan --- 19 RCTs, 1 multisite RCT, 4 cluster RCTs, 3 quasi-experimental designs --- Data from 9,158 pupils*	Significant heterogeneity (I^2=47%, p=0.004)

Feedback type: outcome and process/strategy All studies	Literacy (N=6), mathematics (N=5), science (N=5), language (N=2), social studies (N=1), cognitive outcomes (N=1)	0.36 (0.12, 0.59)	6 studies from US, 2 from Germany, 2 from The Netherlands, 2 from Spain, 2 from Nigeria, 1 from UK, 1 from Indonesia --- 7 RCTs, 3 cluster RCTs, 6 quasi experimental-designs --- Data from 4,198 pupils*	Significant heterogeneity (I^2=87%, p<0.0001)
Outcome and process/strategy Low/Mod ROB studies	Literacy (N=5), mathematics (N=3), science (N=4), language (N=2), social studies (N=1)	0.09 (–0.08, 0.26)	5 studies from US, 2 from Germany, 2 from Spain, 1 from The Netherlands, 1 from UK --- 7 RCTs, 3 cluster RCTs, 1 quasi-experimental design --- Data from 1,349 pupils*	No significant heterogeneity (I^2=45%, p=0.05)

C.I= Confidence Interval; ROB = Risk of Bias

*Data likely to include some double counting due to uncertainty of sample size described in multiple trials within the published papers.

8. Agreements and disagreements with other reviews

The findings of this review are not straightforward to compare directly with other systematic reviews because of the decisions made about the scope and process of this review. As noted in the introduction to this report, a recent meta-analysis of 435 studies of feedback produced a weighted average effect size of *d* = 0.55 (95% C.I *d* = 0.48 to *d* = 0.62) and 17% of the effect sizes from individual studies were negative. There was also considerable variance in the weighted average effect size across the characteristics explored.[22] In one of the most comprehensive historical reviews and meta-analysis of feedback, Kluger and DeNisi[23] found a weighted effect of feedback of *d* = 0.41, but in over 38% of studies the effects were negative. In this review, 24% of the studies with a low or moderate risk of bias were negative and the point of the weighted average effect size of these studies was *g* = 0.17 (95% C.I 0.09 to 0.25).

There may be a number of reasons why the results of these reviews appear to be different, including (as noted above) the focus of the reviews and the selection of the studies. It is clear that these two previous reviews included studies from a wider range of contexts, including higher education and business, a wider range of quasi-experimental study designs, and a wider range of actions under the umbrella 'feedback'. In this review, it is not necessarily clear to what extent the 'feedback only' interventions investigated meet with the feedback 'theories' put forward in either of the other two reviews. But as Kluger and DeNisi point out, the broad scope of what is considered 'feedback' makes the testing of any model or practice practically difficult.

9. Implications for policy and practice

It is difficult to draw clear policy and practice implications from the results of the review. The perspective of the review team is that drawing implications of the results of a review to any particular set of policy and practice contexts requires detailed practical knowledge of the conditions of the context into which findings are being translated, and is therefore best done by users in those contexts. This was done by the EEF's Guidance Report process, where an expert advisory panel (consisting of expert academics and practitioners) interpreted the meta-analyses presented here, in addition to scrutinising individual studies and a review of practice to produce recommendations[24].

In terms of more general reflections, the overall synthesis results suggest that feedback does, on average, have a positive impact on attainment when compared to no feedback or usual practice. The size of the impact of feedback identified in the synthesis carried out in this review is not of the scale identified by Kluger and De Nisi (1986), Wisniewski, Zierer and Hattie (2020) or in the EEF feedback strand of the EEF toolkit.[25]

Furthermore, there is considerable heterogeneity amongst the studies. This may suggest that caution is required in making strong claims about the implications of this review for practice. The heterogeneity between studies found across all of the subgroup analysis meant the review was not able to provide particularly clear evidence about the factors that affect the impact of feedback on attainment. It might be argued that the review results suggest that the factors that affect the impact of feedback may include others that have not been identified in this review or combinations of factors that it was not possible to investigate.

10. Implications for research

The operating parameters for the review processes meant that (i) not all of the studies identified as potentially relevant could be screened, and (ii) only studies of 'feedback only' published after 2000 were included in the review. This means that there are potentially more studies of 'feedback to be identified and also studies that have already been identified that should be scrutinised in more detail to identify potential gaps in the research evidence base about the impact of feedback practices on student attainment in mainstream education.

However, in taking forward either primary or secondary research evaluating the impact of feedback, consideration will need to be given to the boundaries of interventions labelled as 'feedback'. In terms of practical application, what are

[22] Wisniewski, B., Zierer, K. and Hattie, J. (2020). 'The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research', *Front. Psychol*. 10:3087.
[23] Kluger, A. and DeNisi, A. (1986). 'The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory', *Psychological Bulletin*, Vol. II9, No. 2, 254–284

[24] https://educationendowmentfoundation.org.uk/public/files/Publications/Feedback/Teacher_Feedback_to_Improve_Pupil_Learning.pdf
[25] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/feedback/

the specific points of practice that demark ‘feedback’ from another educational interventions that might be used by a practitioner, such as ‘mastery learning’? Based on the experience of the review, there are practical differences between interventions labelled as ‘feedback’. These differences make claims about impact rather fuzzy in terms of interpreting their practical application.

In taking forward either primary or secondary research to investigate the impact of practices labelled as ‘feedback’, it seems likely that greater practical clarity will be necessary in delimiting the boundaries of each type of feedback intervention.

11. Limitations

The review only searched Microsoft Academic Graph (MAG) as a source. This may mean that relevant studies were not identified. Our initial indications from the pilot searches on MAG suggested that theses/dissertations may not be identified. The final list of included studies does however include six theses.

The review included only studies of the impact of feedback as a single component intervention published after 2000. The screening of identified studies was also stopped before the optimally identified ‘stopping moment’. It is therefore possible that other studies that have investigated the impact of feedback on attainment in mainstream school settings were not identified and/or selected into the review.

The review findings have been expressed with a degree of caution that is appropriate to the processes used and results obtained in the review. However, the synthesis included studies that were assessed as having a moderate risk of bias, which may mean that even the modest claims made about the impact of feedback are potentially optimistic. There is also heterogeneity between the studies as indicated by the statistical heterogeneity analysis. There are differing views about how to interpret statistically significant heterogeneity. There will always be some heterogeneity between studies. The position taken in the reporting of the findings of this review is that the presence of statistically significant heterogeneity means that the pooled estimate may not be a useful indicator of the general effect of single component feedback. This interpretation is based on the I^2 measure and the statistical significance of the test for heterogeneity, as this seems the most transparent and systematic approach to adopt.

12. Team

Dr Mark Newman (EPPI-Centre, UCL Institute of Education): Principle investigator, team leader, lead author of the systematic review .

Dr Karen Schucan Bird (EPPI-Centre, UCL Institute of Education): Co-investigator, co author systematic review.

Irene Kwan (EPPI-Centre, UCL Institute of Education): Co-investigator, co author systematic review.

Dr Mary Richardson (Dept of Curriculum Pedagogy and Assessment, UCL Institute of Education): Co-investigator, led and authored the scoping review.

Dr Hui-Teng Hoo (Nanyang Technological University, Singapore): development associate, co-author systematic review .

Ian Shemilt (EPPI-Centre, UCL Institute of Education): led the development of the MAG workflows.

Conflicts of interest

The team members are all members of staff at University College London. Dr Hui-Teng Hoo is a member of staff at Nanyang Technological University, Singapore. They are not in receipt of personal or research funding from any third parties relevant to this review or topic. They are not in receipt of any other funding from the EEF. Dr Hui-Teng Hoo has published academic research on the topic of feedback. EPPI-Centre collaborates with Microsoft on the developmental use of MAG for systematic reviews but receives no funding from Microsoft.

13. References of included studies

N = 51 reported in 43 published papers

Ajogbeje, O.J., and Alonge, M.F. (2012). Effect of Feedback and Remediation on Students' Achievement in Junior Secondary School Mathematics (use in feedback review). International Education Studies, 5(5), pp.153–162.

Alitto, J., Malecki, C.K., Coyle, S. and Santuzzi, A. (2016). Examining the effects of adult and peer mediated goal setting and feedback interventions for writing: Two studies. Journal of School Psychology, 56, pp.89–109.

Baadte, C., and Schnotz, W. (2014). Feedback Effects on Performance, Motivation and Mood: Are They Moderated by the Learner's Self-Concept? Scandinavian Journal of Educational Research, 58(5), pp.570–591.

Beckmann, N., Beckmann, J.F., and Elliott, J.G. (2009). Self-Confidence and Performance Goal Orientation Interactively Predict Performance in a Reasoning Test with Accuracy Feedback. Learning And Individual Differences, 19(2), pp.277–282.

Brosvic, G.M., Dihoff, R.E., Epstein, M.L. and Cook, M.L. (2006). Feedback Facilitates the Acquisition and Retention of Numerical Fact Series by Elementary School Students with Mathematics Learning Disabilities—Experiment 1a. Psychological Record, 56(1), pp.35–54.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E. and Kintsch, W. (2007). Guided practice in technology-based summary writing. In: McNamara, D.S. ed., Reading Comprehension Strategies: Theory, interventions, and technologies. Mahwah, N.J.: Erlbaum, pp.375–396.

Chiu, S. and Alexander, P.A. (2014). Young Children's Analogical Reasoning The Role of Immediate Feedback. Journal Of Psychoeducational Assessment, 32(5), pp.417–428.

Clariana, R.B. and Koul, R. (2006). The Effects of Different Forms of Feedback on Fuzzy and Verbatim Memory of Science Principles. British Journal of Educational Psychology, 76(2), pp.259–270.

Dihoff, R.E., Brosvic, G.M., Epstein, M.L. and Cook, M.J. (2005). Adjunctive Role for Immediate Feedback in the Acquisition and Retention of Mathematical Fact Series by Elementary School Students Classified with Mild Mental Retardation—Experiment 1. Psychological Record, 55(1), pp.39–66.

Eyengho, T. and Fawole, O. (2013). Effectiveness of indirect and direct metalinguistic error correction techniques on the essays of senior secondary school students in South Western Nigeria. Educational Research Review, 8(17), pp.1613–1620.

Fogel, H. and Ehri, L.C. (2000). Teaching Elementary Students Who Speak Black English Vernacular to Write in Standard English: Effects of Dialect Transformation Practice. Contemporary Educational Psychology, 25(2), pp.212–235.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N. and Dooley, S. (2005). Summary Street: Computer support for comprehension and writing. Journal of Educational Computing Research, 33(1), pp.53–80.

Fyfe, E.R. and Rittle-Johnson, B. (2016). Feedback Both Helps and Hinders Learning: The Causal Role of Prior Knowledge—Experiment 1a. Journal Of Educational Psychology, 108(1), pp.82–97.

Fyfe, E.R. and Rittle-Johnson, B. (2016). The benefits of computer-generated feedback for mathematics problem solving. Journal Of Experimental Child Psychology, 147, pp.140–151.

Fyfe, E.R. and Rittle-Johnson, B. (2017). Mathematics practice without feedback: A desirable difficulty in a classroom setting. Instructional Science, 45(2), pp.177–194.

Fyfe, E.R., Rittle-Johnson, B. and DeCaro, M.S. (2012). The Effects of Feedback During Exploratory Mathematics Problem Solving: Prior Knowledge Matters. Journal Of Educational Psychology, 104(4), pp.1094–1108.

Golke, S., Dörfler, T. and Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. Learning And Instruction, 39, pp.123–136.

Golke, S., Dörfler, T. and Artelt, C. (2009). The effects of accuracy feedback during a text comprehension test. Educational and Child Psychology, 26, pp.30–39.

Hier, B. (2012). Generality of treatment effects: Evaluating elementary-aged students' abilities to generalize and maintain fluency gains of a performance feedback writing intervention. Masters Dissertation Thesis. University of Syracuse. USA.

Holman, L. (2011). Automated writing evaluation program's effects on student writing achievement. Doctoral Dissertation. Tennessee State University. USA.

King, M. (2003). The effects of formative assessment on student self-regulation, motivational beliefs and achievement in elementary science. PhD Thesis. George Mason University. USA.

Koedinger, K.R., McLaughlin, E.A. and Heffernan, N.T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. Educational Computing Research, 43(4), pp.489–510.

Llorens, A.C., Cerdán, R. and Vidal-Abarca, E. (2014). Adaptive formative feedback to improve strategic search decisions in task-oriented reading. Journal of Computer Assisted Learning, 30(3), pp.233–251.

Llorens, A.C., Vidal-Abarca, E. and Cerdán, R. (2016). Formative feedback to transfer self-regulation of task- oriented reading strategies. Journal of Computer Assisted Learning, 32(4), pp.314–331.

Malandrino, R.D. (2015). Generalization Programming and the Instructional Hierarchy: A Performance Feedback Intervention in Writing. Dissertation. Syracuse University. USA.

Mostow, J., Nelson-Taylor, J. and Beck, J.E. (2013). Computer-Guided Oral Reading versus Independent Practice: Comparison of Sustained Silent Reading to an Automated Reading Tutor That Listens. Journal of Educational Computing Research, 49(2), pp.249–276.

Nurhayati, S. and Tanti, I.A. (2017). The Influence of Giving Direct Corrective Feedback on Big Task toward Student’s Learning Result. 2, pp.42–48.

Olina, Z. and Sullivan, H.Z. (2002). Effects of Classroom Evaluation Strategies on Student Achievement and Attitudes. Educational Technology Research And Development, 50(3), pp.61–75.

Peverly, S.T. and Wood, R. (2001). The Effects of Adjunct Questions and Feedback on Improving the Reading Comprehension Skills of Learning-Disabled Adolescents. Contemporary Educational Psychology, 26(1), pp.25–43.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B. and Besser, M. (2018). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. Learning And Instruction, 60(1), pp.154–165.

Reybroeck, M.V., Penneman, J., Vidick, C. and Galand, B. (2017). Progressive treatment and self-assessment: effects on students’ automatisation of grammatical spelling and self-efficacy beliefs. Reading And Writing, 30(9), pp.1965–1985.

Rosenthal, B.D. (2006). Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency. PhD Thesis. Syracuse University. USA.

Smith, E. and Gorard, S. (2005). 'They don't give us our marks': The role of formative feedback in student progress. Assessment In Education Principles Policy & Practice, 12(1), pp.21–38.

Stevenson, C.E. (2017). Role of Working Memory and Strategy-Use in Feedback Effects on children's Progression in Analogy Solving: an Explanatory Item Response Theory Account. 27(3), pp.393–418.

Sukhram, D. and Monda-Amaya, L.E. (2017). The effects of oral repeated reading with and without corrective feedback on middle school struggling readers. British Journal Of Special Education, 44(1), pp.95–111.

Thompson, D.J.B. (2007). Effects of evaluative feedback on math self-efficacy, grade self-efficacy, and math achievement of ninth grade algebra students: a longitudinal approach. PhD Thesis. University of Louisville. USA.

Urban, K. and Urban, M. (2020). Effects of performance feedback and repeated experience on self-evaluation accuracy in high- and low-performing preschool children. European Journal Of Psychology Of Education, pp.1–16.

Van Beuningen, C.G., de Jong, N.H. and Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. Itl International Journal Of Applied Linguistics, 156(156), pp.279–296.

Van Loon, M.H. and Roebers, C.M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. Early Childhood Research Quarterly, 53, pp.301–313.

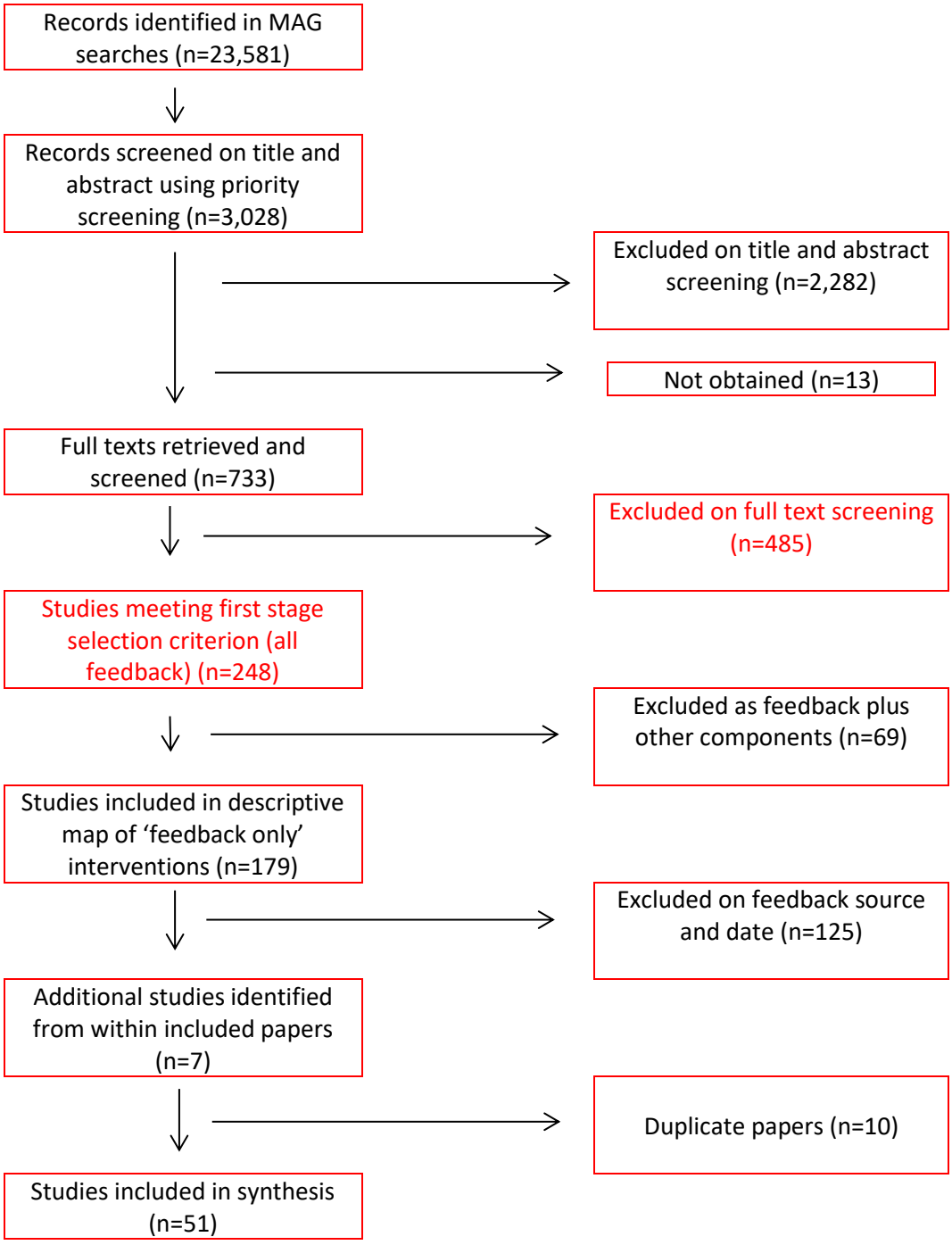
VanEvera, W.C. (2003). Achievement and motivation in the middle school science classroom: The effects of formative assessment feedback. PHD Thesis. George Mason University. USA.

Wade-Wade-Stein, D. and Kintsch, E. (2004). Summary Street: Interactive computer support for writing. Cognition and Instruction, 22(3) pp.333–362.

Wiggins, M., Sawtell, M. and Jerrim, J. (2017). Learner Response System: Evaluation report and executive summary. London. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Learner_Response_System.pdf

Yin, Y. (2005). The influence of formative assessments on student motivation, achievement, and conceptual change. PhD Thesis. Stanford University. USA.

Appendix 1: Flow of studies through the review



Appendix 2: Table of characteristics of included studies

	Author/year Study ID Title	Country/design	Participants/ educational setting/curriculum subject	Feedback characteristics	Main results	Study quality
1.	**Ajogbeje and Alonge (2012)** 50079413 *Effect of Feedback and Remediation on Students' Achievement in Junior Secondary School Mathematics (use in feedback review)*	**Country** • Nigeria **Study design** • Prospective QED	**Population** • Students (N=240) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Mathematics	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil • Group **Form of feedback** • Spoken verbal • Non-verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of the feedback** • Neutral	**Post-test effect sizes** • Mathematics (SMD=1.67[SE=0.19])	**Overall ecological validity** • Moderate **Overall risk of bias** • Serious
2.	**Alitto *et al* (2016), Study 1** 50078937	**Country** • USA	**Population** • Students (N=114) **Age** • 9–10 years	**Source of feedback** • Digital or automated	**Post-test effect sizes** • Literacy Writing (1) (SMD=0.49[SE=0.19])	**Overall ecological validity** • High

	Examining the effects of adult and peer mediated goal setting and feedback interventions for writing: Two studies (Study 1)	**Study design** • Individual RCT	**Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy (4 tests)	**Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Positive • Neutral	• Literacy Writing (2) (SMD=0.31[SE= 0.19]) • Literacy Writing (3) (SMD=0.35[SE= 0.19]) • Literacy Writing (4)*	**Overall risk of bias** • Moderate
3.	**Alitto *et al* (2016), Study 2** 55547946 *Examining the effects of adult and peer mediated goal setting and feedback interventions for writing: Two studies (Study 2)*	**Country** • USA **Study design** • Prospective QED	**Population** • Students (N=106) **Age** • 10–11 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy (4 tests)	**Source of the feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome	**Post-test effect sizes** • Literacy Writing (1) (SMD=0.62[SE=0.20]) • Literacy Writing (2) (SMD=0.75[SE= 0.20]) • Literacy Writing (3) (SMD=0.70[SE= 0.20]) • Literacy Writing (3) (SMD=0.65[SE= 0.20])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

				Emotional tone of feedback • Neutral		
4.	**Baadte and Schnotz (2014)** 50085016 *Feedback Effects on Performance, Motivation and Mood: Are They Moderated by the Learner's Self-Concept?—Updated (use for feedback review)*	**Country** • Germany **Study design** • Individual RCT	**Population** • Students (N=72) **Age** • 10–12 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Science/social studies combined	**Source of the feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Science (SMD=0.03[SE=0.24])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
5.	**Beckmann, Beckmann and Elliott (2009)** 50081348 *Self-Confidence and Performance Goal Orientation Interactively Predict*	**Country** • UK **Study design** • Individual RCT	**Population** • Students (N=105) **Age** • 13–15 years **Gender** • Mixed gender **Educational setting)** • Secondary/high school	**Source of the feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened**	**Post-test effect sizes** • Cognitive (1) (SMD=0.02[SE=0.24]) • Cognitive (2) SMD=–0.31[SE=0.27]) • Cognitive (3) (SMD=–0.05[SE=0.18])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

	Performance in a Reasoning Test with Accuracy Feedback		**Curriculum subjects tested** • Cognitive reasoning (3 tests)	• Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback?** • Neutral		
6.	**Brosvic, Dihoff, Epstein and Cook—Experiment 1a (2006)** 50079146 *Feedback Facilitates the Acquisition and Retention of Numerical Fact Series by Elementary School Students with Mathematics Learning Disabilities—Experiment 1a*	**Country** • USA **Study design** • Individual RCT	**Population** • Students with a learning disability in mathematics (MLD) (N=40) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics	**Source of the feedback** • Teacher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written, non-verbal **When feedback happened** • Immediate • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Maths*	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
7.	**Brosvic, Dihoff, Epstein and Cook—**	**Country** • USA	**Population** • Students normally achieving (NA) (N=40) **Age**	**Source of the feedback** • Teacher • Digital or automated	**Post-test effect sizes** • Maths*	**Overall ecological validity** • High

	Experiment 1b (2006) 54977848 *Feedback Facilitates the Acquisition and Retention of Numerical Fact Series by Elementary School Students with Mathematics Learning Disabilities—Experiment 1b*	**Study design** • Individual RCT	Not reported **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics	**Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written, non-verbal **When feedback happened** • Immediate • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		**Overall risk of bias** • Moderate
8.	**Brosvic, Dihoff, Epstein and Cook—Experiment 3 (2006)** 54978775 *Feedback Facilitates the Acquisition and Retention of Numerical Fact Series by Elementary School Students with Mathematics Learning*	**Country** • USA **Study design** • Individual RCT	**Population** • Students with a learning disability in mathematics (MLD) (N=40) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics	**Source of the feedback** • Teacher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal **When feedback happened** • Immediate • Delayed (short) **Kind of feedback provided** • About the outcome	**Post-test effect sizes** • Maths*	**Overall ecological validity** • High **Overall risk of bias** • Moderate

	Disabilities—Experiment 3			**Emotional tone of feedback** • Neutral		
9.	**Caccamise (2007) 1_1** 37092575 *Guided practice in technology-based summary writing*	**Country** • USA **Study design** • Prospective QED	**Population** • Students (N=243) **Age** • 12–15 years **Gender** • Not reported **Educational setting** • Middle school **Curriculum subjects tested** • Literacy: writing	**Source of the feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (SMD=0.38[SE=0.21])	**Overall ecological validity** • Moderate **Overall risk of bias** • Serious
10.	**Chiu and Alexander (2014)** 50101990 *Young Children's Analogical Reasoning, The Role of Immediate Feedback*	**Country** • Taiwan **Study design** • Individual RCT	**Population** • Students (N=80) **Age** • 5 years **Gender** • Mixed gender **Educational setting** • Nursery school/pre-school	**Source of the feedback** • Researcher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Non-verbal **When feedback happened**	**Post-test effect sizes** • Cognitive (SMD=0.57[SE=0.23])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

			Curriculum subjects tested Cognitive reasoning	• Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		
11.	**Clariana (2006)** 46888078 *The Effects of Different Forms of Feedback on Fuzzy and Verbatim Memory of Science Principles*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=82) **Age** • 15–17 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Science (4 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Science (1) (SMD=0.33[SE=0.23]) • Science (2) (SMD=–0.03[SE=0.23]) • Science (3) (SMD=0.24[SE=0.23]) • Science (4) (SMD=–0.11[SE=0.23])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
12.	**Dihoff, Brosvic, Epstein and Cook—Experiment 1 (2005)** 50079830	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=16) **Age** • 10.5 years **Gender** • Mixed gender	**Source of feedback** • Teacher • Digital or automated **Feedback directed to** • Individual pupil	**Post-test effect sizes** • Maths (1)* • Maths (2)*	**Overall ecological validity** • High **Overall risk of bias**

	Adjunctive Role for Immediate Feedback in the Acquisition and Retention of Mathematical Fact Series by Elementary School Students Classified with Mild Mental Retardation—Experiment 1		**Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (2 tests)	**Form of feedback** • Spoken verbal • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral		• Moderate
13.	**Eyengho and Fawole (2013)** 50095529 *Effectiveness of indirect and direct metalinguistic error correction techniques on the essays of senior secondary school students in South Western Nigeria*	**Country** • Nigeria **Study design** • Prospective QED	**Population** • Students (N=196) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: writing (2 tests)	**Source of feedback** • Teacher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback**	**Post-test effect sizes** • Literacy (1) (SMD=0.53[SE=0.19]) • Literacy (2) (SMD=0.64[SE=0.19])	**Overall ecological validity** • Moderate **Overall risk of bias** • Serious

				• Neutral		
14.	**Fogel and Ehri (2000)** 50084152 *Teaching Elementary Students Who Speak Black English Vernacular to Write in Standard English: Effects of Dialect Transformation Practice*	**Country** • USA **Study design** • Cluster RCT	**Population** • Students (N=60) **Age** • 8–10 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy: writing (5 tests)	**Source of feedback** • Teacher **Feedback directed** • Group **Form of feedback** • Spoken verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral • Negative	**Post-test effect sizes** • Literacy (1) (SMD=1.00[SE=0.27]) • Literacy (2) (SMD=0.72[SE=0.27]) • Literacy (3) (SMD=0.78[SE=0.27]) • Literacy (4) (SMD=0.59[SE=0.26]) • Literacy (5) (SMD=0.91[SE=0.27])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
15.	**Franzke and Kintsch (2005)** 37092578 *Summary Street: Computer support for comprehension and writing*	**Ccountry** • USA **Study design** • Multisite RCT	**Population** • Students (N=121) **Age** • 13–15 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested**	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Non-verbal **When feedback happened**	**Post-test effect sizes** • Literacy (1) (SMD=0.31[SE=0.19]) • Literacy (2) (SMD=—0.22[SE=0.19]) • Literacy (3) (SMD=0.15[SE=0.19]) • Literacy (4) (SMD=0.03[SE=0.19]) • Literacy (5) (SMD=—0.03[SE=0.19]) • Literacy (6)	**Overall ecological validity** High **Overall risk of bias** • Moderate

			• Literacy: writing (7 tests)	• During the task **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	(SMD=0.25[SE=0.19]) • Literacy (7) (SMD=0.08[SE=0.19])	
16.	**Fyfe and Rittle-Johnson (2016)—Experiment 1a** 50079852 *Feedback Both Helps and Hinders Learning: The Causal Role of Prior Knowledge—Experiment 1a*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=112) **Age** • 7–9 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (3 tests)	**Source of feedback** • Researcher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Maths (1) (SMD=–0.60[SE=0.28]) • Maths (2) (SMD=4.83[SE=0.55]) • Maths (3) (SMD=0.32[SE=0.19])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
17.	**Fyfe and Rittle-Johnson (2016)—Experiment 1b**	**Country** • USA	**Population** • Students (N=112) **Age** • 7–9 years **Gender**	**Source of feedback** • Researcher • Digital or automated	**Post-test effect sizes** • Maths (1) (SMD=0.93[SE=0.28]) • Maths (2)*	**Overall ecological validity** • Moderate

	54240106 *Feedback Both Helps and Hinders Learning: The Causal Role of Prior Knowledge—Experiment 1b*	**Study design** • Individual RCT	• Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (2 tests)	**Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		**Overall risk of bias** • Moderate
18.	**Fyfe and Rittle-Johnson (2016)—Experiment 2** 54235415 *Feedback Both Helps and Hinders Learning: The Causal Role of Prior Knowledge—Experiment 2*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=113) **Age** • 8–9 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (4 tests)	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback**	**Post-test effect sizes** • Maths (1) (SMD=0.17[SE=0.24]) • Maths (2) (SMD=–0.41[SE=0.25]) • Maths (3) (SMD=–0.50[SE=0.25]) • Maths (4) (SMD=–0.56[SE=0.25])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

				• Neutral		
19.	**Fyfe and Rittle-Johnson (2016a)** 50079801 *The benefits of computer-generated feedback for mathematics problem solving*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=77) **Age** • 7–9 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (2 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Maths (Immediate) (SMD=0.67[SE=0.29]) • Maths (Summative) (SMD=0.39[SE=0.29])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
20.	**Fyfe and Rittle-Johnson (2017)** 50083136 *Mathematics practice without feedback: A desirable difficulty in a classroom setting*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=243) **Age** • 8 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (4 tests)	**Source of feedback** • Researcher **Feedback directed to** • Group **Form of feedback** • Spoken verbal **When feedback happened** • Immediate • Delayed (short)	**Post-test effect sizes** • Maths (1) (SMD=0.11[SE=0.16]) • Maths (2) (SMD=0.07[SE=0.16]) • Maths (3) (SMD=–0.06[SE=0.16]) • Maths (s4) (SMD=–0.06[SE=0.16])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

				Kind of feedback provided • About the outcome **Emotional tone of feedback** • Neutral		
21.	**Fyfe, Rittle-Johnson and DeCaro (2012)—Experiment 1** 50079651 *The Effects of Feedback During Exploratory Mathematics Problem Solving: Prior Knowledge Matters—Experiment 1*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=93) **Age** • 8 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (11 tests)	**Source of feedback** • Researcher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • During the task • Immediate **Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Maths (1) (SMD=0.39[SE=0.26]) • Maths (2) (SMD=–0.34[SE=0.26]) • Maths (3) (SMD=0.20[SE=0.27]) • Maths (4) (SMD=–0.80[SE=0.28]) • Maths (5) (SMD=0.12[SE=0.25]) • Maths (6) (SMD=–0.40[SE=0.26]) • Maths (7) (SMD=0.12[SE=0.27]) • Maths (8) (SMD=–1.47[SE=0.30]) • Maths (9)* • Maths (10) (SMD=0.06[SE=0.18]) • Maths (11) (SMD=–0.21[SE=0.19])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
22.	**Fyfe, Rittle-Johnson and DeCaro (2012)—Experiment 2**	**Country** • USA	**Population** • Students (N=101) **Age** • 7 years	**Source of feedback** • Researcher • Digital or automated	**Post-test effect sizes** • Maths (1) (SMD=0.40[SE=0.25]) • Maths (2)	**Overall ecological validity** • Moderate

	54124473 *The Effects of Feedback During Exploratory Mathematics Problem Solving: Prior Knowledge Matters—Experiment 2*	**Study design** • Individual RCT	**Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Mathematics (11 tests)	**Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • During the task • Immediate **Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	(SMD=–0.51[SE=0.25]) • Maths (3) (SMD=0.53[SE=0.26]) • Maths (4) (SMD=–0.54[SE=0.26]) • Maths (5) (SMD=0.26[SE=0.25]) • Maths (6) (SMD=–0.65[SE=0.26]) • Maths (7) (SMD=0.18[SE=0.25]) • Maths (8) (SMD=–0.76[SE=0.26]) • Maths (9)* • Maths (10) (SMD=−0.04[SE=0.18]) • Maths (11) (SMD=–0.04[SE=0.18])	**Overall risk of bias** • Moderate
23.	**Golke, Dörfler and Artelt (2009)** 46888085 *The effects of accuracy feedback during a text comprehension test*	**Country** • Germany **Study design** • Individual RCT	**Population** • Students (N=198) **Age** • 11–12 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome	**Post-test effect sizes** • Literacy (Reading) (SMD=–0.12[SE=0.14])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

				Emotional tone of feedback • Neutral		
24.	**Golke, Dörfler and Artelt (2015)—Experiment 1** 54732130 *The impact of elaborated feedback on text comprehension within a computer-based assessment—Experiment 1*	**Country** • Germany **Study design** • Individual RCT	**Population** • Students (N=566) **Age** • 12 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy*	**Overall ecological validity** • High **Overall risk of bias** • Moderate
25.	**Golke, Dörfler and Artelt (2015)—Experiment 2** 50082195 *The impact of elaborated feedback on text comprehension within a computer-based*	**Country** • Germany **Study design** • Individual RCT	**Population** • Students (N=251) **Age** • 12 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading (2	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task	**Post-test effect sizes** • Literacy (1) (SMD=0.06[SE=0.18]) • Literacy (2) (SMD=0.36[SE=0.18])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

	assessment—Experiment 2		tests)	**Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral		
26.	**Hier (2012)** 50079304 *Generality of treatment effects: Evaluating elementary-aged students' abilities to generalize and maintain fluency gains of a performance feedback writing intervention*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=103) **Age** • 8–9 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy: writing (2 tests)	**Source of feedback** • Researcher • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=0.59[SE=0.20]) • Literacy (2) (SMD=0.29[SE=0.20])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
27.	**Holman (2011)** 37092584 *Automated writing evaluation*	**Country** • USA **Study design** • Cluster RCT	**Population** • Students (N=160) **Age** • 13–14 years **Gender** • Mixed gender	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil	**Post-test effect sizes** • Literacy (SMD=0.34[SE=0.17])	**Overall ecological validity** • High **Overall risk of bias**

	program's effects on student writing achievement		**Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy: writing	**Form of feedback** • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		• Moderate
28.	**King (2003)** 37092606 *The effects of formative assessment on student self-regulation, motivational beliefs and achievement in elementary science*	**Country** • USA **Study design** • Prospective QED	**Population** • Students (N=65) **Age** • 10–11 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Science	**Source of feedback** • Teacher • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal **When feedback happened** • Immediate • Delayed (short) **Kind of feedback provided** • About the process of the task • About the learner's strategies or approach	**Post-test effect sizes** • Science (SMD=–0.24[SE=0.26])	**Overall ecological validity** • Moderate **Overall risk of bias** • Serious

				Emotional tone of feedback • Neutral		
29.	**Koedinger, McLaughlin and Heffernan (2010)** 37092607 *A quasi-experimental evaluation of an on-line formative assessment and tutoring system*	**Country** • USA **Study design** • Prospective QED	**Population** • Students (N=1344) **Age** • 12–13 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested** • Mathematics	**Source of the feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of the feedback** • Neutral	**Post-test effect sizes** • Maths (SMD=0.20[SE=0.07])	**Overall ecological validity** • High **Overall risk of bias** • Serious
30.	**Llorens, Cerdán and Vidal-Abarca (2014)** 46888095 *Adaptive formative feedback to improve strategic search decisions*	**Country** • Spain **Study design** • Individual RCT	**Population** • Students (N=92) **Age** • 12–14 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects**	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form feedback** • Written, non-verbal **When feedback happened**	**Post-test effect sizes** • Literacy (1) (SMD=0.68[SE=0.27]) • Literacy (2) (SMD=0.35[SE=0.26])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

	in task-oriented reading—Updated		**tested** • Literacy: reading (2 tests)	• During the task • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		
31.	**Llorens, Vidal-Abarca and Cerdán (2016)—Experiment 1** 46888096 *Formative feedback to transfer self-regulation of task-oriented reading strategies—Experiment 1*	**Country** • Spain **Study design** • Individual RCT	**Population** • Students (N=142) **Age** • 12–14 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading (5 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • During the task • Immediate **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=0.03[SE=0.20]) • Literacy (2) (SMD=0.09[SE=0.20]) • Literacy (3) (SMD=0.44[SE=0.23]) • Literacy (4) (SMD=0.16[SE=0.23]) • Literacy (5) (SMD=0.13[SE=0.18])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

32.	**Llorens, Vidal-Abarca and Cerdán (2016)—Experiment 2** 49106831 *Formative feedback to transfer self-regulation of task-oriented reading strategies—Experiment 2*	**Country** • Spain **Study design** • Individual RCT	**Population** • Students (N=112) **Age** • 12–14 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading (3 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=0.88[SE=0.24]) • Literacy (2) (SMD=0.25[SE=0.23]) • Literacy (3) (SMD=0.36[SE=0.23])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
33.	**Malandrino (2015)** 50082272 *Generalization Programming and the Instructional Hierarchy: A Performance Feedback Intervention in Writing*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=116) **Age** • 8 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy: writing	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Immediate **Kind of feedback provided**	**Post-test effect sizes** • Literacy (SMD=0.52[SE=0.23])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

				• About the outcome **Emotional tone of feedback** • Neutral		
34.	**Mostow, Nelson-Taylor and Beck (2013)** 46888304 *Computer-Guided Oral Reading versus Independent Practice: Comparison of Sustained Silent Reading to an Automated Reading Tutor That Listens*	**In which Country** • USA **Study design** • Prospective QED	**Population** • Students (N=193) **Age** • 6–10 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Literacy: reading/spelling (6 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written, non-verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=0.10[SE=0.15]) • Literacy (2) (SMD=0.17[SE=0.15]) • Literacy (3) (SMD=0.41[SE=0.15]) • Literacy (4) (SMD=0.73[SE=0.15]) • Literacy (5) (SMD=0.37[SE=0.15]) • Literacy (6) (SMD=0.17[SE=0.15])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
35.	**Nurhayati and Tanti (2017)** 50098095 *The Influence of Giving Direct Corrective Feddback on Big Task toward*	**In which Country** • Indonesia **Study design** • Prospective QED	**Population** • Students (N=70) **Age** • 14 years **Gender** • Not reported **Educational setting** • Secondary/high school	**Source of feedback** • Teacher **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal	**Post-test effect sizes** • Science (SMD=1.03[SE=0.26])	**Overall ecological validity** • High **Overall risk of bias** • Serious

	Student's Learning Result		**Curriculum subjects tested** • Science	**When feedback happened** • Immediate **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral		
36.	**Olina and Sullivan (2002)** 50081169 *Effects of Classroom Evaluation Strategies on Student Achievement and Attitudes*	**Country** • Latvia **Study design** • Cluster RCT	**Population** • Students (N=189) **Age** • Not reported **Gender** • Not reported **Educational setting** • Secondary/high school **Curriculum subjects tested** • Other curriculum test/cognitive (3 tests)	**Source of feedback** • Teacher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Others (1) (SMD=0.45[SE=0.18]) • Others (2) (SMD=0.53[SE=0.18]) • Cognitive (3) (SMD=0.20[SE=0.18])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

37.	**Peverly and Wood (2001)** 47269862 *The Effects of Adjunct Questions and Feedback on Improving the Reading Comprehension Skills of Learning-Disabled Adolescents*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=50) **Age** • 14–16 years **Gender** • Not reported **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: reading (2 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=–0.19[SE=0.39]) • Literacy (2) (SMD=2.01[SE=0.48])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
38.	**Rakoczy, Pinger and Hochweber (2018)** 50080103 *Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy*	**Country** • Germany **Study design** • Cluster RCT	**Population** • Students (N=620) **Age** • 15 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested** • Mathematics	**Source of feedback** • Teacher **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the process of the task	**Post-test effect sizes** • Maths (SMD=–0.03[SE=0.08])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

				Emotional tone of feedback • Neutral		
39.	**Reybroeck and Penneman (2017)** 50079005 *Progressive treatment and self-assessment: effects on students' automatisation of grammatical spelling and self-efficacy beliefs*	**Country** • Belgium **Study design** • Individual RCT • Cluster RCT	**Population** • Students (N=126) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy: writing/spelling (3 tests)	**Source of feedback** • Teacher • Self **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written verbal • Written, non-verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=–0.09[SE=0.33]) • Literacy (2) (SMD=–0.26[SE=0.33]) • Literacy (3) (SMD=0.14[SE=0.33])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
40.	**Rosenthal (2006)** 37092595 *Improving elementary-age children's writing fluency: A comparison of improvement*	**Country** • USA **Study design** • Cluster RCT	**Population** • Students (N=45) **Age** • 8–9 years **Gender** • Mixed gender **Educational setting** • Primary/elementary	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Written, non-verbal	**Post-test effect sizes** • Literacy (1) (SMD=0.39[SE=0.42]) • Literacy (2) (SMD=0.26[SE=0.42]) • Literacy (3) (SMD=0.47[SE=0.40]) • Literacy (4) (SMD=0.52[SE=0.40])	**Overall ecological validity** •Moderate **Overall risk of bias** • Low

	based on performance feedback frequency		school **Curriculum subjects tested** • Literacy: reading/writing (4 tests)	**When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		
41.	**Smith and Gorard (2005)** 50079102 *'They don't give us our marks': The role of formative feedback in student progress*	**Country** • UK **Study design** • Prospective QED	**Population** • Students (N=104) **Age** • Not reported **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Literacy • Mathematics • Science • Languages (Welsh)	**Source of feedback** • Teacher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • Delayed (short) • Delayed (long) **Kind of feedback provided** • About the outcome • About the process of the task **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (SMD=–0.16[SE=0.27]) • Maths (SMD=–0.03[SE=0.27]) • Science (SMD=–0.71[SE=0.29]) • Language (SMD=–1.20[SE=0.43])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

42.	**Stevenson (2017)** 50080874 *Role of Working Memory and Strategy-Use in Feedback Effects on children's Progression in Analogy Solving: An Explanatory Item Response Theory Account*	**Country** • The Netherlands **Study design** • Prospective QED	**Population** • Students (N=999) **Age** • 4–8 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Cognitive reasoning	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal • Written, non-verbal **When feedback happened** • During the task • Immediate **Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Cognitive (SMD=0.62[SE=0.09])	**Overall ecological validity** • Moderate **Overall risk of bias** • Serious
43.	**Sukhram and Monda-Amaya (2017)** 50079125 *The effects of oral repeated reading with and without corrective feedback on middle school struggling readers*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=60) **Age** • 12–14 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested** • Literacy (3 tests)	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal **When feedback happened** • During the task	**Post-test effect sizes** • Literacy (1) (SMD=0.08[SE=0.02]) • Literacy (2) (SMD=0.15[SE=0.26]) • Literacy (3) (SMD=0.05[SE=0.26])	**Overall ecological validity** • Moderate **Overall risk of bias** • Low

				Kind of feedback provided • About the outcome **Emotional tone of feedback** • Neutral		
44.	**Thompson (2007)** 50080408 *Effects of evaluative feedback on math self-efficacy, grade self-efficacy, and math achievement of ninth grade algebra students: a longitudinal approach*	**Country** • USA **Study design** • Individual RCT	**Population** • Students (N=46) **Age** • 13–15 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects** • Mathematics (2 tests)	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Maths (1) (SMD=−0.30[SE=0.41]) • Maths (2) (SMD=0.32[SE=0.43])	**Overall ecological validity** • Moderate **Overall risk of bias** • Low
45.	**Urban and Urban (2020)** 50084250 *Effects of performance feedback and repeated*	**Country** • Slovakia **Study design** • Individual RCT	**Population** • Students (N=111) **Age** • 6 years **Gender** • Mixed gender **Educational setting**	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal	**Post-test effect sizes** • Cognitive (1) (SMD=0.86[SE=0.29]) • Cognitive (2) (SMD=0.10[SE=0.26]) • Cognitive (3) (SMD=0.38[SE=0.19])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

	experience on self-evaluation accuracy in high- and low-performing preschool children		• Nursery school/pre-school **Curriculum subjects tested** • Cognitive reasoning (3 tests)	**When feedback happened** • During the task **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral		
46.	**van Beuningen, de Jong and Kuiken (2008)** 50088090 *The effect of direct and indirect corrective feedback on L2 learners' written accuracy*	**Country** • The Netherlands **Study design** • Individual RCT	**Population** • Students (N=66) **Age** • 14 years **Gender** • Not reported **Educational setting** • Secondary/high school **Curriculum subjects tested** • Languages (4 tests)	**Source of feedback** • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • Delayed (short) **Kind of feedback provided** • About the outcome • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Language (1) (SMD=1.12[SE=0.41]) • Language (2) (SMD=0.65[SE=0.39]) • Language (3) (SMD=0.84[SE=0.37]) • Language (4) (SMD=0.67[SE=0.37])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate
47.	**van Loon and Roebers (2020)**	**Country** • Switzerland	**Population** • Students (N=105) **Age**	**Source of feedback** • Researcher	**Post-test effect sizes** • Cognitive (1) (SMD=0.82[SE=0.25])	**Overall ecological validity**

	50088442 *Using feedback to improve monitoring judgment accuracy in kindergarten children*	**Study design** • Individual RCT	• 5 years **Gender** • Mixed gender **Educational setting** • Primary/elementary school **Curriculum subjects tested** • Cognitive (2 tests)	**Feedback directed to** • Individual pupil **Form of feedback** • Spoken verbal **When feedback happen** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of the feedback** • Neutral	• Cognitive (2) (SMD=0.38[SE=0.25])	• Moderate **Overall risk of bias** • Moderate
48.	**VanEvera (2003)** 37092614 *Achievement and motivation in the middle school science classroom: The effects of formative assessment feedback*	**Country** • USA **Study design** • Cluster RCT	**Population** • Students (N=68) **Age** • 13–14 years **Gender** • Mixed gender **Educational setting** • Secondary/high school **Curriculum subjects tested** • Science	**Source of feedback** • Teacher • Researcher **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the process of the task • About the learner's strategies or approach	**Post-test effect sizes** • Science (SMD=0.60[SE=0.40])	**Overall ecological validity** • Moderate **Overall risk of bias** • Moderate

				• About the person **Emotional tone of feedback** • Positive		
49.	**Wade-Stein and Kintsch (2004)** 37092600 *Summary Street: Interactive computer support for writing*	**Country** • USA **Study design** • Prospective QED	**Population** • Students (N=52) **Age** • 11–12 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested** • Literacy: writing (2 tests)	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil **Form of feedback** • Written verbal • Written, non-verbal **When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Literacy (1) (SMD=0.85[SE=0.29]) • Literacy (2) (SMD=0.25[SE=0.28])	**Overall ecological validity** • High **Overall risk of bias** • Moderate
50.	**Wiggins, Sawtell and Jerrim (2017)** 38296697 *Learner Response System: Evaluation report*	**Country** • UK **Study design** • Cluster RCT	**Population** • Students (N=6572) **Age** • 9–11 years **Gender** • Mixed gender **Educational setting** • Primary/elementary	**Source of feedback** • Digital or automated **Feedback directed to** • Individual pupil • Teacher **Form of feedback** • Written verbal	**Post-test effect sizes** • Maths (1) (SMD=0.07[SE=0.04]) • Maths (2) (SMD=–0.05[SE=0.04]) • Maths (3) (SMD=0.06[SE=0.05]) • Maths (4) (SMD=–0.07[SE=0.05]) • Literacy (1)	**Overall ecological validity** • High **Overall risk of bias** • Low

	and executive summary		school **Curriculum subjects tested** • Mathematics (4 tests) • Literacy: reading (4 tests)	**When feedback happened** • Immediate **Kind of feedback provided** • About the outcome **Emotional tone of feedback** • Neutral	(SMD=0.10[SE=0.04]) • Literacy (2) (SMD=0.01[SE=0.04]) • Literacy (3) (SMD=0.06[SE=0.05]) • Literacy (4) (SMD=0.01[SE=0.05])	
51.	**Yin (2005)** 37092616 *The influence of formative assessments on student motivation, achievement, and conceptual change*	**Country** • US **Study design** • Cluster RCT	**Population** • Students (N=280) **Age** • 11–13 years **Gender** • Mixed gender **Educational setting** • Middle school **Curriculum subjects tested** • Science	**Source of feedback** • Teacher **Feedback directed to** • Group **Form of feedback** • Spoken verbal **When feedback happened** • During the task **Kind of feedback provided** • About the outcome • About the process of the task • About the learner's strategies or approach **Emotional tone of feedback** • Neutral	**Post-test effect sizes** • Science (SMD=–0.32[SE=0.13])	**Overall ecological validity** • High **Overall risk of bias** • Moderate

***No usable data to compute effect sizes; SMD = Standard Mean Difference; SE = Standard Error**

Appendix 3: EEF feedback review—Data extraction tool

This is the data extraction tool used in the EEF feedback review. It is comprised of the EEF database extraction tools (main, subject specific, outcome and study quality assessment tool) put together in a single document.

Section 1: What is the publication type?

- Journal article
  A report published in a peer-reviewed journal with an ISSN.
- Dissertation or thesis
  A report of a study in a dissertation or thesis submitted as all or part of the assessment for a higher degree.
- Technical report
  An unpublished report, technical report or document providing details of a research study or studies without an ISSN or ISBN. (EEF evaluation reports are classified as technical reports.)
- Book or book chapter
  A report of a research study published in a book or book chapter with an ISBN.
- Conference paper
  A report of a study presented at a research conference and subsequently made more widely available.
  NB Peer-reviewed conference proceedings with an ISBN should still be classified as a conference paper.
- Other (Please specify)
  A report not classifiable according to the categories above (e.g. a website). Please add further details in the notes field.

Section 2: What is the research design and which methods were used?

- What is the intervention name?
  Provide the name of the intervention, programme or approach as given in the report.
- How is the intervention described?
  Brief summary of the intervention as provided in the report(s). Please include the rationale for impact on learning if given.
- What are the intervention objectives?
  Please provide the specific objectives or aims of the intervention, programme or approach as provided in the report.
- Is there more than one treatment group?
  Does the research design include more than one arm or contrast so that more than one estimate of the impact of the intervention or approach can be made from a different comparison group or version of the intervention?
  - Yes (Please specify)
    Highlight in the text (or use the info box) to describe the design and specify the other interventions or comparisons relative to the main intervention group.
  - No
  - Not specified or N/A

- How were participants assigned?
  How were the participants assigned or allocated to their group (i.e. treatment and control)?

- Random (please specify)
  Select this code where the report describes the participants' allocation to their group as random or pseudo-random (computer generated). Please highlight in the text or add information to the info box about the randomisation details.
- Non-random, but matched
  No randomisation, but matched at allocation prospectively to balance on attainment (or on attainment and other variables).
- Non-random, not matched prior to treatment
  No random allocation and not matched prior to treatment. The nature and extent of any group differences in attainment at baseline is described and then accounted for in the analysis of impact (retrospective matching).
- Unclear
  Please only select this code if there are no details about control and intervention allocation or if the information is so unclear as to prevent a reasonable inference.
- Not assigned—naturally occurring sample
  This is where researchers take advantage of a situation where a comparison can be made between groups from changes that either are planned or have already happened which will give an estimate of the impact of the intervention or approach of interest.
  - Retrospective Quasi-Experimental Design (QED)
    Where an experiment is created from a naturally occurring situation and two groups (or more) are compared to give an estimate of impact.
  - Regression discontinuity
    This is a type is a quasi-experimental pre-test/post-test design that identifies the causal effects of an intervention or approach by assigning a cutoff or threshold above or below which an intervention is assigned (e.g. policy change where smaller classes are introduced in a district or a test is used to allocate students to additional support). By comparing results close to but either side of the threshold, it is possible to estimate effect.

- What was the level of assignment?
  At which level was the assignment to intervention and control group conducted?
  - Individual
    The assignment was at the level of the individual student or pupil. No account was taken of class or school. All of the individual participants were included as a single group for allocation or randomisation.
  - Class
    The class or usual teaching group of the students was the level at which the intervention or approach was allocated. Intact classes were allocated or assigned to the intervention or approach (taking no account of school).
  - School—cluster
    The school was the level of assignment and all pupils in a single school are allocated to the same grouping (i.e. a single school would not include both intervention and control).
  - School—multi-site
    The school is the level of assignment, but each school contains both intervention and control groups. The design allows a within-school comparison to be made.
  - Region or district
    The region or district is the level at which the assignment is made.
  - Not provided/not available
    A description of the level of allocation is not provided or available in the report.
  - Not applicable

- How realistic was the study?
  Was the intervention implemented under 'real world' conditions? Factors to consider in assessing the 'ecological validity' include where the intervention took place (usual educational setting for educational approaches of this kind) and who taught or led the intervention with the pupils (e.g. did it involve usual teachers or other education professionals).
  - High ecological validity
    Select this code where the intervention or approach seems realistic for schools or teachers to adopt.
    Any adaptations to enable the research to be conducted do not appear to affect the validity of the findings and implications for schools. Studies which take place in schools and are taught by the usual teachers or staff have high ecological validity.
  - Low ecological validity
    Select this code where the intervention or approach does not seems realistic or practical for schools or teachers to adopt. Studies which take place in laboratory settings and are only taught by researchers have low ecological validity.
  - Unclear
    Select this code where there are no details about where the intervention took place or who was responsible for its delivery and it is not possible to infer sufficient details to make a judgement about the ecological validity of the study.

Section 3 Where did the study take place?

- In which country/countries was the study carried out? (Select ALL that apply)
  Countries which are recognised as sovereign states by the United Nations. If you think there is a country missing please ask!
  - UK (Select all that apply)
    - England
    - Northern Ireland
    - Scotland
    - Wales
  - USA
  - Afghanistan
  - Albania
  - Argentina
  - Angola
  - Armenia
  - Austria
  - Australia
  - Azerbaijan
  - Bahamas, The
  - Bahrain
  - Bangladesh
  - Belarus
  - Barbados
  - Belize
  - Belgium
  - Benin
  - Bhutan
  - Bosnia and Herzegovina
  - Botswana

- Brazil
- Bolivia
- Brunei Darussalam
- Burkina Faso
- Bulgaria
- Cabo Verde
- Cambodia
- Canada
- Cameroon
- Central African Republic
- Chad
- Chile
- Colombia
- Congo
- Costa Rica
- Côte d'Ivoire / Ivory Coast
- Croatia
- China
  If just Hong Kong, use Hong King code only, NOT China
- Cuba
- Cyprus
- Denmark
- Czech Republic
- Dominican Republic
- Egypt
- Ecuador
- El Salvador
- Equatorial Guinea
- Estonia
- Eritrea
- Ethiopia
- Finland
- Fiji
- France
- Gabon
- Georgia
- Gambia, The
- Germany
- Greece
- Ghana
- Guatemala
- Grenada
- Guinea-Bissau
- Guinea
- Guyana
- Haiti
- Honduras
- Hong Kong (see China)
- Hungary
- Iceland

- Indonesia
- India
- Iran
- Iraq
- Ireland
- Italy
- Israel
- Jamaica
- Japan
- Jordan
- Kenya
- Kazakhstan
- Kuwait
- Kiribati
- Lao (or Laos)
  Lao People's Democratic Republic
- Kyrgyzstan
- Latvia
- Lebanon
- Liberia
- Lesotho
- Libya
- Liechtenstein
- Luxembourg
- Lithuania
- Madagascar
- Macedonia
- Malaysia
- Malawi
- Mali
- Maldives
- Malta
- Marshall Islands
- Mauritania
- Mauritius
- Micronesia
- Mexico
- Moldova
- Mongolia
- Mozambique
- Namibia
- Myanmar (Burma)
- Nepal
- Nauru
- The Netherlands
- New Zealand
- Nicaragua
- Nigeria
- Niger
- Pakistan

- Norway
- Palau
- Panama
- Papua New Guinea
- Peru
- Philippines
- Poland
- Puerto Rico (US dependency)
- Portugal
- Qatar
- Romania
- Rwanda
- Russia
- Saint Kitts and Nevis
- Saint Lucia
- Saint Vincent and the Grenadines
- San Marino
- Samoa
- Saudi Arabia
- São Tomé and Príncipe
- Serbia
- Senegal
- Seychelles
- Sierra Leone
- Slovakia
- Singapore
- Slovenia
- Solomon Islands
- South Africa
- Somalia
- South Korea / Republic of Korea
- South Sudan
- Sri Lanka
- Spain
- Sudan
- Suriname
- Swaziland / Eswatini
- Sweden
- Switzerland
- Taiwan
- Syria
- Tanzania
- Tajikistan
- Thailand
- Timor-Leste
- Togo
- Tonga
- Tunisia
- Trinidad and Tobago
- Turkey

- Turkmenistan
- Tuvalu
- Ukraine
- Uganda
- United Arab Emirates
- Uruguay
- Uzbekistan
- Vanuatu
- Venezuela
- Vietnam
- West Indies (Use for Caribbean colonial dependencies)
  Cayman Islands (United Kingdom)
  Anguilla (United Kingdom)
  Antigua and Barbuda
  Aruba (Netherlands)
  Bonaire (Netherlands)
  British Virgin Islands (United Kingdom)
  Curaçao (Netherlands)
  Guadeloupe (France)
  Martinique (France)
  Montserrat (United Kingdom)
  Nueva Esparta (Venezuela)
  Saba (Netherlands)
  Saint Barthélemy (France)
  Saint-Martin (France)
  Sint Eustatius (Netherlands)
  Sint Maarten (Netherlands)
  United States Virgin Islands (United States)
  Federal Dependencies of Venezuela (Venezuela)
  Turks and Caicos Islands (United Kingdom)
- Yemen
- Zambia
- Zimbabwe

- Is there more specific information about the location?
  Further information on where the study took part (e.g. city, district, urban, suburban, rural etc.) as provided by the study.
  - Specific to the location or place
    Information about the specific place where the research was undertaken (e.g. name of the city, state, city or region)
  - Information about the type of location
    Information about what kind of location (e.g. urban, rural, suburban).
  - No information provided
    Please use this code if there is no further information about the specific location (place name) or the type of location (e.g. urban/ rural).
- What is the educational setting (Select ALL that apply)
  What is the type of educational setting that the students attend which is the focus of the intervention or approach?
  - Nursery school/pre-school
    A separate nursery school or pre-school setting or a nursery or early years class in a primary school.
    The focus is on the type of setting or educational provision.

- Primary/elementary school
  A school for children of normal school age (depending on the jurisdiction).
  The focus is on the type of school or setting. Pupils will typically be between the ages of 5 and 11.
- Middle school
  An intermediate school provided in some jurisdictions for pupils between their primary (or elementary) and secondary educational stages.
- Secondary/high school
  A school for older pupils, after primary or elementary education (and after middle school where provided). Pupils will usually be between the ages of 11 and 18.
- Residential/boarding school
  A school where pupils reside as well as study; boarding either by week or over a term.
- Independent/private school
- Home
- Further education/junior or community college
  A formal educational setting for older secondary pupils. Students will usually be 16 or older, but still studying for school-level, vocational or professional qualifications (i.e. not higher education or leading to a Bachelor's degree).
- Other educational setting (please specify)
  An educational setting which cannot be classified under one of the other definitions. Please provide details of the educational setting as given in the study (e.g. field centre, museum classroom, concert or rehearsal hall, public theatre, workplace training, etc.).
- Outdoor adventure setting
  Educational activities taking place outdoors, such as Outward Bound courses, sailing and kayaking or canoeing, camping, climbing or courses based at an outdoor education centre.
  All studies classified under the Toolkit strand 'Outdoor adventure learning' should be included.
  Field studies centres where the activities focus solely on school subjects like Geography or Biology should not be included (please use 'Other' for these and specify the type of setting).
- No information provided

Section 4 What is the sample of the study?

- What is the overall sample analysed?
  What is the total number of participants in the data analysed (both intervention and control/comparison)? Please add additional details in the notes.
- What is the gender of the students?
  Please indicate the gender of the total sample.
  - Female only
  - Male only
  - Mixed gender
    Provide the percentage or number of female pupils in the study. Please highlight the section or add details of where this can be found in the report.
  - No information provided
- What is the age of the students? (Select ALL that apply)
  Please provide additional information if available (e.g. grade level(s), mean age, or mean and standard deviation).

    - 3
    - 4
    - 5
    - 6
    - 7
    - 8
    - 9
    - 10
    - 11
    - 12
    - 13
    - 14
    - 15
    - 16
    - 17
    - 18
    - No information provided
- What is the proportion of low SES/FSM students in the sample?
  What proportion of the students in the study are receiving free school meals (FSM) or reduced price lunches or are identified as being from a low socio-economic status? If possible, record this as a percentage. Please highlight or add further details as reported in the study.
    - FSM or low SES student percentage
      Please add the percentage of pupils in the sample who are receiving free school meals (FSM) or reduced price lunches or are identified as being from a low socio-economic status background.
    - Further information about FSM or SES in the study sample.
      Please highlight any details provided in the study about the socio-economic status of the students involved in the research (such as eligibility for free or reduced price school meals or lunches).
    - No SES/FSM information provided
      Select this option if there is no information about the socio-economic status of the students involved in the research (such as eligibility for free or reduced price school meals or lunches).

Section 5: What was involved in the intervention?
Details about the intervention, approach or policy being evaluated.

- What type of organisation was responsible for providing the intervention?
  Please indicate what kind of organisation was responsible for the provision or management and organisation of the intervention?
    - School or group of schools
    - Charity or voluntary organisation
    - University/researcher design
    - Local education authority or district
      Local education authority or district (government or public funding)
    - Private or commercial company
    - Other (please provide details)
- Was training for the intervention provided?
  Was training provided to the delivery team as part of the preparation and support for the intervention? If so, who provided it?

- Yes (Please specify)
  Please highlight the text or add details to the info box as provided in the report.
- No
- Unclear/not specified

- Who is the focus of the intervention? (Select ALL that apply)
  Who is the main focus of the intervention study? Although the interest of the Toolkit is on student outcomes, the focus of behavioural change may be on others in educational settings, such as teachers or parents. NB All interventions must report outcomes on student's attainment.
  - Students
    The main focus of the intervention is on the behaviours, interactions or activities of the students or pupils. Others may be involved (such as in training to deliver or implement a new approach), but the main aim is to change students' activities, behaviours and interactions to improve educational outcomes.
  - Teachers
    The main focus of the intervention is on the teachers and their behaviours, interactions and activities. Although the final outcome may be to improve students' attainment, the focus and study aims focus on the teachers as a clear or explicit part of the rationale.
  - Teaching assistants
    The focus of the intervention includes teaching assistants or teacher's aides (and/or other para-professionals) and their behaviours, interactions and activities. Although the final outcome may be to improve students' attainment, the focus and study aims involve teaching assistants as part of the process.
  - Other education practitioners
  - Non-teaching staff
    The main focus of the intervention is on the non-teaching staff in schools and their behaviours, interactions and activities. This includes all staff who would not normally have a teaching role (e.g. administrative staff, lunchtime supervisors, facilities management etc.). Although the final outcome may be to improve students' attainment, the focus and study aims include the non-teaching staff as part of the rationale.
  - Senior management
    The main focus of the intervention is on the senior management in schools (e.g. headteachers, deputy head teachers, heads of department) and their behaviours, interactions and activities. Although the final outcome may be to improve students' attainment, the focus and study aims include the senior management as part of the rationale.

  - Parents
    Parents or carers of students in the educational settings involved are involved because of their parental or caring responsibilities.
  - Other (Please specify)
- What is the intervention teaching approach? (Select ALL that apply)
  What was the main teaching or learning approach used for an intervention session?
  - Large group/class teaching (6+)
    A large group (more than 6 students) with a teacher or supporter of the intervention, typically in a classroom setting.
  - Small group/intensive support (3–5)
    Intensive small group provision by a teacher, teaching assistant or other supporter of the intervention in small group setting (3–5 participants in a group), sometimes in a separate teaching space or classroom.

    - Paired learning
      Two pupils either working together, or peer teaching each other.
    - One-to-one
      One-to-one instruction where the teacher is not a peer, but a teacher, teaching assistant, volunteer or other education professional.
    - Student alone (self-administered)
      Pupils or students working through study materials independently and/or unsupervised.
    - Other (Explain in notes)
- Were any of the following involved in the intervention or approach?
    - Digital technology
      The main approach depends on the use of digital technology (e.g. tablets, laptops, software, internet) by pupils or teachers (e.g. interactive whiteboards).
        - Yes
        - No
    - Parents or community volunteers
      Parents or community volunteers working with their children (or other pupils).
        - Yes
        - No
- When did the intervention take place? (Select ALL that apply)
  When was the intervention delivered?
    - During regular school hours
      The intervention or approach takes place completely or mainly during regular school hours.
    - Before/after school
      The intervention or approach takes place completely or mainly before or immediately after normal school hours. This should mainly apply to activities taking place on school or normal educational settings.
    - Evenings and/or weekends
      Where the intervention or approach takes place during evenings or weekends. Activities which take place immediately after school and at school (or in the same educational setting) should not be included.
    - Summer/holiday period
      Where the educational activity takes place as additional time in what would normally be a holiday period (e.g. summer holidays or other vacation times).
    - Other (please specify)
    - Unclear/ not specified
      Use this code where there are no details provided of when the intervention was delivered and where the information provided does not allow a reasonable inference to be made about timing.
      The usual inference for most interventions where the timing is not specified will be 'During regular school hours'. If this inference cannot reasonably be made please indicate in the notes the details in the report which produce the ambiguity or lack of clarity.
- Who was responsible for the teaching at the point of delivery? (Select ALL that apply)
  Please provide details (e.g. staff involved, training level provided, number/proportions of staff).
  This should focus on the experience of pupils, rather than any initial training and support.
    - Research staff
      Select this code where the intervention or approach was delivered largely or exclusively by researchers or the research team.

- Class teachers
  Select this code when the intervention or approach was taught or delivered by professional teachers as part of their usual teaching or wider professional activity.
- Teaching assistants
  Select this code where the majority of the teaching or delivery of the intervention is undertaken by teaching assistants (or teacher's aides, para-professionals, auxiliary teachers, nursery nurses in early years settings and other cognate terms). These will be staff usually employed by a school, but without a full teaching qualification.
- Other school staff
  Staff employed by the school, but neither teachers nor teaching assistants (or those in similar paid roles). It includes administrative staff, lunch-time supervisors, facilities staff.
- External teachers
  Teachers or other professional educational staff hired or employed by the research team or the delivery organisation.
- Parents/carers
  Parents or carers whose main relationship with the intervention is through their parental or caring responsibilities. This includes where parents are working with their own children, or working with other children in the school or educational setting that their own children attend.
- Lay persons/volunteers
  Adults (over 18 years) involved as volunteers or undertaking unpaid work who provide the majority of the support to pupils or lead in the delivery of the intervention to students.
- Peers
  Other students or pupils at the same school or educational setting as the intervention group; or at another local school (e.g. secondary students tutoring pupils at their own or their peers' primary schools). Peers will normally be of similar age and socio-economic or cultural background.
  University students tutoring primary school pupils would not be classified as 'peers'.
- Digital technology
  Include digital technology where the technology has a role in the educational activity, such as where automated feedback or marking is provided, or where it provides an explicit teaching role (intelligent tutoring or the use of explanatory videos) or where differentiated activities are offered or allocated automatically to learners. Incidental use of technology which is usually involved in the normal teaching and learning activities of the intervention group should not be included as this has already been recorded.
- Unclear/not specified
  Use this code where there are no details provided of who or how the intervention was delivered or where the information provided does not allow a reasonable inference to be made.

- What was the duration of the intervention? (Please add to info box and specify units)
  Duration of the intervention or approach (from beginning to end). Please specify units (e.g. months, weeks, days). This may differ from the duration of the research project or evaluation which could involve pre- and post-testing periods.
- What was the frequency of the intervention?
  What is the frequency of the intervention (as delivered)? e.g. daily, twice weekly, weekly monthly.
- What is the length of intervention sessions?
  What is the length in minutes of a typical session?

- Are implementation details and/or fidelity details provided?
  Are details provided about how successfully the intervention was implemented or taken up? Please indicate what type of information by selecting the appropriate checkbox and highlighting relevant text in the report.
  - Qualitative
    Please select if qualitative details about the intervention or approach are provided, such as describing any issues or challenges about implementation, or comments on the training and/ orimplementation process.
  - Quantitative
    Please select if quantitative details about implementation are provided, such as number of schools or teachers trained, or number of sessions attended.
  - No implementation details provided
    No details about the implementation process are provided.
- Are the costs reported?
  Are there any financial costs or details reported?
  - Yes (Please add details)
    If this option is selected, please add details as provide in the report(s).
  - No
- Who undertook the outcome evaluation?
  Here we are interested in how independent the evaluation was.
  - The developer
    This is the usual option and should be selected unless the information is unclear or confusing. This is where the researcher or developer evaluated their own programme or approach.
  - A different organisation paid by developer
    The development team is different from the evaluation team but it is commissioned directly by the developer or researcher who developed the intervention approaches.
  - An organisation commissioned independently to evaluate
    The research team is different from the evaluation team and commissioned independently (e.g. EEF reports).
  - Unclear/not stated
    There is insufficient information about the status of the evaluation research to indicate or infer how independent the evaluation is.
  - Is this an EEF evaluation?
    If the evaluation was funded by the Education Endowment Foundation please select.

Section 6: What kind of primary outcomes are reported?

- What kind of tests were used? (Select ALL that apply)
  What type(s) of test(s) were used to measure the intervention outcomes on learning at pupil/student level?
  - Standardised test (Please specify)
    A standardised test is administered and scored in a consistent way. The properties of the test are established through piloting on a group to determine the mean and spread of the scores for a particular target group. Standardised tests are usually named and the properties published.
    Please add the name of the test(s) used, a brief description and any details reported.
  - Researcher-developed test (Please add details)
    A test developed or designed for a specific research project. Please add any details as provided in the report(s).

- School-developed test (Please add details)
  A test or examination developed and used by a school or schools involved in the research as part of their usual assessment approach. Please add any details as provided in the report(s).
- National test or examination (Please specify)
  A test or examination used in regional or national evaluations of student and school performance. These may be optional or compulsory, but are organised and/or administered by the regional or national education administration in a particular jurisdiction.
- International tests (Please specify)
  Tests used for international comparisons of student performance (e.g. PISA, TIMMS, PIRLS etc.). Please specify the name of the test.

- Curriculum subjects tested (Select ALL that apply)
  If the outcomes relate to the subjects of the school curriculum outcomes, record which subjects are included.
  - Literacy (first language)
    Aspects of literacy including speaking and listening, reading and writing. Include study of literature when this is first language study.
    - Reading comprehension
      This may include aspects such as main idea identification and passage comprehension. When a test provides different outcomes, e.g. TOWRE (Test of Word Reading Efficacy) provides word attack, word identification, & passage comprehension, choose passage comprehension as main outcome.
    - Decoding/phonics
      These measures gave a focus on recognising letters and making the correct sounds associated with the letters or letter combinations. They may be referred to as phonological or phonemic awareness.
    - Spelling
      Where the focus is on the correct spelling of words.
    - Reading other
      E.g. phonics, reading fluency, vocabulary comprehension (receptive vocabulary). When a test provides different outcomes, e.g. TOWRE (Test of Word Reading Efficacy) provides word attack, word identification, & passage comprehension, choose passage comprehension as main outcome.
    - Speaking and listening/oral language
      Speaking and listening or oral language and communication outcomes, including vocabulary use (productive spoken vocabulary).
    - Writing
      A test of written language including quality, quantity and written vocabulary (range).
  - Mathematics
    All aspects of mathematics including number and numerical operations, shape and space (geometry), algebra, data-handling etc.
  - Science
    All general science subjects including physics, chemistry, biology as well as specific subjects such as ecology or astronomy.
  - Social studies
    Either integrated social studies courses or programmes or separate curriculum areas of social studies (e.g. history, geography, civics, sociology, economics or anthropology).

- Arts
  Expressive and performing arts, including music, art, drama, drawing, painting, sculpture and the decorative arts.
- Languages
  Where the aim is to develop communicative or literacy capability in a language other than the first language or usual language of instruction in the school.
- Other curriculum test
  Please provide a description of the outcome as reported where it is a test of a school curriculum subject not included in the categories above (e.g. music, art, classics).

- In addition to the primary educational attainment outcome, are there other outcomes reported?
  - Yes
  - No
- If yes, which other outcomes are reported?
  - Cognitive outcomes measured (Please specify)
    If non-curricular cognitive outcomes are measured, please indicate and specify the outcomes (e.g. reasoning, memory, intelligence, etc.). Include the name of the test where possible (e.g. Raven's Matrices, Stanford–Binet Intelligence Scales etc.).
  - Other types of student outcomes (Please specify)
    E.g. attendance, measures of behaviour, health status, non-cognitive attitudes/dispositions, etc. as assessed through a test or a survey.
  - Other participants' (i.e. not students) outcomes (Please specify)
    If outcomes are measured and reported for other participants involved in the research (such as teachers or parents), please note which participants and which outcomes have been measured, e.g. parental participation.

Feedback v.02 October 2018

Feedback is information given to the learner and/or the teacher about the learner's performance relative to learning goals. It should aim towards (and be capable of producing) improvement in students' learning. Feedback redirects or refocuses either the teacher's or the learner's actions to achieve a goal, by aligning effort and activity with an outcome. It can be about the learning activity itself, about the process of activity, about the student's management of their learning or self-regulation or (the least effective) about them as individuals. This feedback can be verbal or written and can be delivered by a person or via technology.

- **What was the source of the feedback?**
  - Teacher
  - Teaching assistant
  - Volunteer
  - Parent(s) or other relatives
    Parent(s), carer(s) or guardian(s). Also use for other family members (such as grandparents or siblings).
  - Researcher
  - Peer (same age/class)
  - Peer (group)
    Feedback from more than one same age pupil (e.g. when feedback is formalised in collaborative learning).
  - Peer (older)
  - Digital or automated
    Feedback from a computer or other digital device (e.g. mobile phone, website or programme) where there is some automation involved.
  - Other non-human
    Such as from a worked example or where answers are checked after the task has been completed.
  - Self
    Only use this code when checking or self-assessment is strategic and self-regulated (such as applying a checking algorithm or mnemonic).
  - Other (please specify)
    Please add notes about the source for this category, as described in the study.

- **Who was the feedback directed to?**
  This will almost always be to pupils, but may be to the teacher. If to the teacher, then there should be some explicit model of further feedback to change subsequent pupil behaviours or performance.
  - Individual pupil
  - General (group or class)
    Where the feedback is not specific to an individual learner, please indicate.
  - Teacher
    Only select this code when this is explicitly part of the model of feedback in the research study.

- **What form did the feedback take?** (Select one)
  This focuses on how the feedback was communicated. Choose the main feedback approach if there is more than one.
  - Spoken verbal
    Feedback provided in spoken form, this includes audio recorded comments.

- Non-verbal
  Where feedback was communicated physically other than with words, such a through body language, gesture or other non-verbal means, such as extended wait time.
- Written verbal
  Where written comments are provided, either handwritten or digitally.
- Written, non-verbal
  Such as tick or check marks, or with symbols or icons (this includes marked tests or test results).

- **When did the feedback happen?** (Select one)
  Choose the option which best describes the feedback timing.
  - Prior to the task
    Sometimes described as 'feedforward', this is where pupils are primed with information before undertaking a task (e.g. students complete test and get positive, negative results regardless of actual score and then their performance on a following test is measured).
  - During the task
    Where the feedback is contemporaneous with the task or part of the task.
  - Immediate
    Where the feedback was provided immediately or shortly after the activity was completed (such as at the end of the task, or later the same day.
  - Delayed (short)
    Where the feedback occurred more than one day and up to a week after the task or activity.
  - Delayed (long)
    Where the feedback occurred more than a week after the task or activity.

- **What kind of feedback was provided?**
  - About the outcome
    Where the feedback was about the outcome or completed task (e.g. correct or incorrect).
    - Correct
      Where feedback was about the correct answers or responses.
    - Incorrect
      Where feedback focused on the incorrect answers or responses.

  - About the process of the task
    Where the feedback is about how the task or activity is currently being, or should be, undertaken (process rather than outcome).
  - About the learner's strategies or approach
    Where the feedback was to support the learner's own regulation or control of what they were doing (i.e. metacognition and/or self-regulation), often in the form of prompts or cues.
  - About the person
    Feedback directed at the individual or self, such as ‘good boy’ or ‘clever girl’.

- **What was the emotional tone of the feedback?**
  Select the most appropriate description for the emotional tone of the feedback. Select more than one only where this is explicitly part of the design, otherwise select the best overall description, based on how it is described in the study.
  - Positive

- Neutral
  Where the feedback was designed or perceived to be neutral in tone.
- Negative
  This is where the feedback is deliberately designed to be discouraging. It should not be used for feedback about incorrect responses or results.

EEF Toolkit effect size data extraction v1.0 October 2019 [Standard]

Data extraction tool to support meta-analysis of the impact data from included studies. Updated October 2019.

- **Section 1: What are the details of the study design?**
  - What was the study design?
    What type of study design is used for the evaluation of impact?
    - Individual RCT
      An experimental design where individual participants are the unit of randomisation and no provision is made for clustering in the design or analysis.
    - Cluster RCT
      An experimental design where school or class is the unit of randomisation (i.e. all pupils in the same school are in same group and where classes are randomised between schools. The school-level variance should be assigned to either intervention or control in the analysis.
    - Multisite RCT
      An experimental design where both control and intervention pupils may be in the same class or school (within school/class) so that in the analysis the school or class level variance should be shared between intervention and control groups.
    - Prospective QED
      A quasi-experimental design which is planned in advance. There may be a prospective allocation, but the design may also take advantage of a naturally occurring experiment. There is often some matching but no randomisation.
    - Retrospective QED
      A post-hoc natural experiment where matching and/or equivalence is achieved through the design and/or analysis. There is no attempt to manage or control the intervention or phenomenon under investigation.
    - Interrupted time series QED
      A design where the same group is treated as control and comparison, e.g. ABAB and the counterfactual is created over time.
    - Regression discontinuity with randomisation
      Prospective regression discontinuity design where participants around the cut off are randomised to treatment or control.
    - Regression discontinuity—not randomised
      RD with non-random allocation (prospective matching to create equivalence).
    - Regression continuity—naturally occurring
      Regression (dis) continuity design naturally occurring—retrospective matching. Exploits or manipulates a naturally occurring discontinuity to explore the causal effect of an educational intervention or approach. Regression discontinuity designs elicits the causal effects of interventions by assigning a cut off or threshold above or below which an intervention is assigned.

  - What is the number of schools involved in the study?
    - What is the number of schools involved in the intervention group(s)?
      Please provide the number of schools involved in the intervention or versions of the intervention. Please only enter numeric data in the info box.
    - What is the number of schools involved in the control or comparison group?
      Please provide the number of schools involved in the control group. Please only enter numeric data in the info box.

- What is the total number of schools involved?
  Please record the total number of schools involved in the study. This will be the sum of intervention and control schools in a cluster randomised trial, but in a multisite trial, where there are control and intervention pupils in each school, it may be the same as for intervention/control. Please only enter numeric data in the info box.
- Not provided/unclear/not applicable
  Please indicate if the number of schools involved is not provided, is unclear, or is not applicable (such as in an Outdoor Education study).

- What is the number of classes involved?
  - What is the total number of classes involved in the intervention group?
    Please provide the number of classes involved in the intervention or versions of the intervention. Please only enter numeric data in the info box.
  - What is the total number of classes involved in the control or comparison group?
    Please provide the number of classes involved in the control group. Please only enter numeric data in the info box.
  - What is the total number of classes involved?
    Please record the total number of classes involved in the study. Please only enter numeric data in the info box.
  - Not provided/unclear/not applicable
    Please indicate if the number of classes involved is not provided, is unclear, or is not applicable (such as in an Outdoor Education study).

- Are details of randomisation provided? [Not selectable (no checkbox)]
  - Not applicable
    Please select if the study is not described as a randomised design (e.g. quasi-experimental or naturally occurring experiment).
  - No/unclear
    Please select if the study is described as randomised but no details are provided or these details are unclear. If the details are unclear, please highlight the relevant section of the report.

- **Section 2: How is the sample described?**
  Information about the sample size, groups and comparability.
  - What is the sample size for the intervention group?
    Record the initial or assigned sample size for the treatment group in the notes. Please enter numeric data only in the info box. This should be either the main counterfactual comparison of the intervention or approach for the Toolkit from this study, or the first reported.
  - What is the sample size for the control group?
    Record the initial or assigned sample size for the control group in the notes. Please enter numeric data only in the info box.
  - *What is the sample size for the second intervention group?
    *Record the initial or assigned sample size for a second or alternative treatment group in the notes (*if there is one). This should be an equally valid comparison of the intervention*

or approach for the Toolkit as the first intervention group reported above. Please enter numeric data only in the info box.

- *What is the sample size for the third intervention group?
  *Record the initial or assigned sample size for a third or different treatment group in the notes (*if there is one). This should be an equally valid comparison of the intervention or approach for the Toolkit as the other intervention groups reported above. Please enter numeric data only in the info box.*
- Does the study report any group differences at baseline?
  Is there quantitative information about the similarity of treatment and control groups at the beginning of the intervention?
  - Yes
    Please select if there is information provided about how comparable the intervention and control groups are at the beginning of the study in terms of the analysis. Please also highlight the relevant section of the text where this is possible.
  - No/unclear
    Please select this option if there is no information about the baseline comparability of the groups or if this is unclear. If there is information, but it is unclear, please highlight the relevant section of the study, where this is possible.
- Is comparability taken into account in the analysis?
  Are covariates in treatment and control groups assessed, and, if unbalanced, controlled in adjusted analysis?
  - Yes
  - No
  - Unclear or details not provided
- Is attrition or drop-out reported?
  If the sample recruited differs from the sample analysed, are the reasons for this reported? Please include details of attrition or drop-out or any pupils excluded from the analysis.
  - Yes
  - No
  - Unclear (please add notes)
    Please check this option if the amount of attrition is unclear. Please also add notes about attrition if there is information about different groups or outcomes.
- What is the attrition in the treatment group?
  Number of drop-outs in the intervention group as a percentage of the n of the intervention group. Please enter numeric data only in the info box.
- Are the variables used for comparability reported?
  Does the study state which variables are used to assess the comparability of the treatment and control groups?
  - Yes
  - No
  - N/A
  - If yes, which variables are used for comparability?
    Select the variables considered in assessment of similarity, e.g. prior attainment, age, gender, SES, special educational needs, ethnicity.
    - Educational attainment
      A measure of either direct (e.g. reading comprehension) or indirect (reasoning) educational performance or capability.
    - Gender
    - Socio-economic status
    - Special educational needs

- Other (please specify)
- What is the total or overall percentage attrition?
  Please report the percentage of drop-outs or overall attrition in the whole sample. This is the number of drop-outs divided by the initial sample x 100. Or you can calculate as the (initial sample minus the analysed sample) divided by the initial sample times 100. ((N-n)/N) x 100. Please add the % sign (e.g. 15.8%). For more information see: https://ies.ed.gov/ncee/wwc/Docs/OnlineTraining/wwc_training_m2.pdf
- Is clustering accounted for in the analysis?
  Does analysis take account of clustering? E.g. regression with school or cluster or MLM (multi-level modelling) or HLM (hierarchical linear modelling)?
  - Yes
  - No
  - Unclear

- **Section: 3 Outcome details**
  - Outcomes
    - Are descriptive statistics reported for the primary outcome?
      - Yes
        1. If yes, please add for the intervention* group
           *Descriptive statistics for the intervention group. *If there is more than one intervention group please add this below.*
           - Number (n)
             What is the number for the intervention group in the data analysed for this outcome? Add numeric data only to the info box.
           - Pre-test mean
             Please record the pre-test mean (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
           - Pre-test standard deviation
             Please record the pre-test standard deviation (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
           - Post-test mean
             Please report the post-test mean for the intervention group (if provided) for this outcome. Add numeric data only to the info box.
           - Post-test standard deviation
             Please record the post-test standard deviation for the intervention group for this outcome (if provided). Add numeric data only to the info box.
           - Gain score mean (if reported)
             Please add the gain score (pre-test to post-test) mean for the intervention group. Add numeric data only to the info box.
           - Gain score standard deviation (if reported)
             Please add the gain score (pre-test to post-test) standard deviation for the intervention group. Add numeric data only to the info box.
           - Any other information?
             Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)), or use to add notes about the numeric data in the categories above.
        2. If yes please add for the control group
           Descriptive statistics for the intervention group

- Number (n)
  What is the number for the control group in the data analysed for this outcome? Add numeric data only to the info box.
- Pre-test mean
  Please record the pre-test mean (if provided) for the control group for this outcome. Add numeric data only to the info box.
- Pre-test standard deviation
  Please record the pre-test standard deviation (if provided) for the control group for this outcome. Add numeric data only to the info box.
- Post-test mean
  Please report the post-test mean for this outcome for the control group (if provided) for this outcome.
- Post-test standard deviation
  Please record the post-test standard deviation for the control group for this outcome (if provided).
- Gain score mean (if reported)
  Add numeric data only to the info box.
- Gain score standard deviation (if reported)
  Add numeric data only to the info box.
- Any other information?
  Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)).

3. If yes, please add for a second intervention* group (if needed)
   Descriptive statistics for a second intervention group, if needed.
   - Number (n)
     What is the number for the intervention group in the data analysed for this outcome? Add numeric data only to the info box.
   - Pre-test mean
     Please record the pre-test mean (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
   - Pre-test standard deviation
     Please record the pre-test standard deviation (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
   - Post-test mean
     Please report the post-test mean for the intervention group (if provided) for this outcome. Add numeric data only to the info box.
   - Post-test standard deviation
     Please record the post-test standard deviation for the intervention group for this outcome (if provided). Add numeric data only to the info box.
   - Gain score mean (if reported)
     Please add the gain score (pre-test to post-test) mean for a second intervention group (if needed). Add numeric data only to the info box.
   - Gain score standard deviation (if reported)
     Please add the gain score (pre-test to post-test) standard deviation for a second intervention group (if needed). Add numeric data only to the info box.
   - Any other information?
     Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)), or use to add notes about the numeric data in the categories above.

- If needed, please add for the control group
  Descriptive statistics for the second control group (if needed and if different from the primary outcome control)
  - Number (n)
    What is the number for the control group in the data analysed for this outcome? Add numeric data only to the info box.
  - Pre-test mean
    Please record the pre-test mean (if provided) for the control group for this outcome. Add numeric data only to the info box.
  - Pre-test standard deviation
    Please record the pre-test standard deviation (if provided) for the control group for this outcome. Add numeric data only to the info box.
  - Post-test mean
    Please report the post-test mean for the control group (if provided) for this outcome.
  - Post-test standard deviation
    Please record the post-test standard deviation for the control group for this outcome (if provided).
  - Gain score mean (if reported)
    Please add the gain score (pre-test to post-test) mean for this group (if needed). Add numeric data only to the info box.
  - Gain score standard deviation (if reported)
    Please add the gain score (pre-test to post test) standard deviation for this group (if needed). Add numeric data only to the info box.
  - Any other information?
    Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)).

4. If yes, please add for a third intervention* group (if needed)
   Descriptive statistics for a third intervention group, if needed.
   - Number (n)
     What is the number for the intervention group in the data analysed for this outcome? Add numeric data only to the info box.
   - Pre-test mean
     Please record the pre-test mean (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
   - Pre-test standard deviation
     Please record the pre-test standard deviation (if provided) for the intervention group for this outcome. Add numeric data only to the info box.
   - Post-test mean
     Please report the post-test mean for the intervention group (if provided) for this outcome. Add numeric data only to the info box.
   - Post-test standard deviation
     Please record the post-test standard deviation for the intervention group for this outcome (if provided). Add numeric data only to the info box.
   - Gain score mean (if reported)
     Please report the gain score (pre-test to post-test) mean for this outcome for a third intervention group (if needed) for this outcome. Add numeric data only to the info box.
   - Gain score standard deviation (if reported)
     Add numeric data only to the info box.

        - Any other information?
          Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)), or use to add notes about the numeric data in the categories above.
        - If needed please add for a control group
          Descriptive statistics for a third control group (if needed and if different from the primary outcome control)
          - Number (n)
            What is the number for the control group in the data analysed for this outcome? Add numeric data only to the info box.
          - Pre-test mean
            Please record the pre-test mean (if provided) for the control group for this outcome. Add numeric data only to the info box.
          - Pre-test standard deviation
            Please record the pre-test standard deviation (if provided) for the control group for this outcome. Add numeric data only to the info box.
          - Post-test mean
            Please report the post-test mean for the control group (if provided) for this outcome.
          - Post-test standard deviation
            Please record the post-test standard deviation for the control group for this outcome (if provided).
          - Gain score mean (if reported)
            Add numeric data only to the info box.
          - Gain score standard deviation (if reported)
            Add numeric data only to the info box.
          - Any other information?
            Please add any other statistical information reported about this outcome for the intervention group (e.g. standard error (SE)).
    - No
- Is there follow up data?
  Please provide details of any assessment to measure long-lasting effects (e.g. delayed post-test or long term follow up)
    - Yes
    - No
- Primary outcome [Outcome]
  Please indicate the primary outcome and enter additional data using the 'Outcomes' box.
  The primary outcome should be the outcome most relevant to the Toolkit strand(s) in terms of educational impact, such as standardised tests of reading or mathematics (for literacy or mathematics interventions) or national test or examination results. See handbook and supporting resources for further information.
- Secondary outcome(s) [Outcome]
  Please add secondary outcomes in this section where they represent a fair test of the impact of the evaluation at post-test. This should not include delayed or follow up tests, or outcomes used to check the specificity of impact (e.g. a maths test used to control for intervention effect in a literacy intervention) or checking for transfer outcomes.
- SES/FSM outcome [Outcome]
  If a separate effect is reported for low socio-economic status or free or reduced price school meals pupils, please add here.

- Outcome classification
  Outcome classifications for meta-analysis and meta-regressions. Please select all that apply.
  - Sample (select one from this group)
    Outcome classification relating to the sample.
    - Sample: All [Outcome classification code]
      Analysis applied to normal or typical sample of pupils. The whole range of attainment or 'ability' for the educational setting was included in the intervention.
    - Sample: Exceptional [Outcome classification code]
      Students described as gifted and talented or of exceptional 'ability'. Usually those in the top 10 per cent of the distribution.
    - Sample: High achievers [Outcome classification code]
      Classification of the students in the sample in relation to their level of academic attainment. Those described as high attainers or high 'ability'; usually those in the top half or the top third of the distribution (depending on classifications).
    - Sample: Average [Outcome classification code]
      Classification of the students in the sample in relation to their level of academic attainment. Those described as performing at or around average attainment or of average 'ability'; usually those in the middle quartiles (depending on classifications).
    - Sample: Low achievers [Outcome classification code]
      Classification of the students in the sample in relation to their level of academic attainment. Those described as low attainers or low 'ability'; usually those in the bottom half or the bottom third of the distribution (depending on classifications).
  - Test type (select one from this group)
    - Test type: Standardised test [Outcome classification code]
      A standardised test is administered and scored in a consistent way. The properties of the test are established through piloting on a group to determine the mean and spread of the scores for a particular target group. Standardised tests are usually named and the properties published.
    - Test type: Researcher-developed test [Outcome classification code]
      A test developed or designed for a specific research project
    - Test type: National test [Outcome classification code]
      A test or examination used in regional or national evaluations of students and school performance. These may be optional or compulsory, but are organised and/or administered by the regional or national administration in a particular jurisdiction.
    - Test type: School-developed test [Outcome classification code]
      A test or examination developed and used by a school or schools involved in the research as part of their usual assessment approach.
    - Test type: International tests [Outcome classification code]
      Tests used for international comparisons of student performance (e.g. PISA, TIMMS, PIRLS etc.)
  - Effect size calculation (select one from this group)
    What kind of effect size is being reported for this outcome?
    - Post-test unadjusted (select one from this group) [Outcome classification code]
      A simple comparison of the differences between control and intervention groups using only the post-test data, usually from an older randomised controlled trial (RCT) or where baseline equivalence has been established.
    - Post-test adjusted for baseline attainment [Outcome classification code]
      A post-test comparison where a measure of educational attainment at pre-test is

controlled for in the analysis of the impact of the intervention or approach, e.g. ANCOVA, OLS regression.

- Post-test adjusted for baseline attainment AND clustering [Outcome classification code]
  A post-test comparison where a measure of educational attainment at pre-test is controlled for in the analysis of the impact of the intervention or approach and where the estimate is adjusted for clustering at class or school level (e.g. ANCOVA, MLM, OLS regression).
- Pre-post gain [Outcome classification code]
  Outcome assessment based on the difference between an individual's pre-test and post-test scores and the range of these difference (gain score or pre-post analysis).

- Toolkit strand(s) (select at least one Toolkit strand)
  Please select the Toolkit strand or strands which this outcome is evaluating. Each study has usually been classified as appropriate for the Tookit. There will not usually be more than one, but occasionally some outcomes are appropriate measures of more than one approach (such as when a teaching assistant delivers a phonics intervention). If unsure please check with the Tookit team.
  - Toolkit: Arts participation [Outcome classification code]
    Arts participation is defined as involvement in artistic and creative activities, such as dance, drama, music, painting, or sculpture. It can occur either as part of the curriculum or as extra-curricular activity. Participation may be organised as regular weekly or monthly activities, or more intensive programmes such as summer schools or residential courses. Whilst these activities have educational value in themselves, this Toolkit entry focuses on the benefits of arts participation for core academic attainment.
  - Toolkit: Aspiration interventions [Outcome classification code]
    By aspirations we mean the things children and young people hope to achieve for themselves in the future. To meet their aspirations about careers, university, and further education, pupils often require good educational outcomes. Raising aspirations is therefore often believed to incentivise improved attainment.
  - Toolkit: Behaviour interventions [Outcome classification code]
    Behaviour interventions seek to improve attainment by reducing challenging behaviour. This entry covers interventions aimed at reducing a variety of behaviours, from low-level disruption to general anti-social activities, aggression, violence, bullying, and substance abuse. The interventions themselves can be split into three broad categories:
    1. Approaches to developing a positive school ethos or improving discipline across the whole school, which also aim to support greater engagement in learning.
    2. Universal programmes which seek to improve behaviour and generally take place in the classroom.
    3. More specialised programmes which are targeted at students with specific behavioural issues.
  - Toolkit: Block scheduling [Outcome classification code]
    Block scheduling is an approach to school timetabling in secondary schools. It typically means that pupils have fewer classes (4–5) per day, for a longer period of time (70–90 minutes). The three main types of block schedules found in the research are:
    4x4 block scheduling: 4 blocks of extended (80–90 minute) classes each day, covering the same 4 subjects each day. Students take 4 subjects over 1 term, and 4 different subjects in the following term. A/B block scheduling: 3 or 4 blocks

of extended (70–90 minute) classes each day, covering the same 3 or 4 subjects on alternating days. Students take 6 or 8 subjects each term. Hybrid: a hybrid of traditional models and 3/4-class-per-day approaches. Students have 5 classes per day, of between 60 and 90 minutes.

- Toolkit: Built environment [Outcome classification code]
  Changing the physical conditions or built environment of the learning setting, either by moving to a new school building or seeking to improve the structure, air quality, noise, light, or temperature of an existing building or classroom.
- Toolkit: Collaborative learning [Outcome classification code]
  A collaborative (or cooperative) learning approach involves pupils working together on activities or learning tasks in a group small enough for everyone to participate on a collective task that has been clearly assigned. Pupils in the group may work on separate tasks contributing to a common overall outcome, or work together on a shared task.
  Some collaborative learning approaches put mixed ability teams or groups to work in competition with each other in order to drive more effective collaboration. There is a very wide range of approaches to collaborative and cooperative learning involving different kinds of organisation and tasks. Peer tutoring can also be considered as a type of collaborative learning, but in the Toolkit it is reviewed as a separate topic.
- Toolkit: Digital technology [Outcome classification code]
  The use of digital technologies to support learning. Approaches in this area are very varied, but a simple split can be made between:
  Programmes for students, where learners use technology in problem solving or more open-ended learning, and
  Technology for teachers such as interactive whiteboards or learning platforms which may be used by the teachers, or where the technology may provide instruction more directly.
- Toolkit: Early years intervention [Outcome classification code]
  Early years or early childhood interventions are approaches that aim to ensure that young children have educationally based pre-school or nursery experiences which prepare for school and academic success, usually through additional nursery or pre-school provision. Many of the researched programmes and approaches focus on disadvantaged children. Some also offer parental support. The research summarised here looks at general or multi-component programmes and approaches.
- Toolkit: Extending school time [Outcome classification code]
  This summary focuses on extending core teaching and learning time in schools and the use of targeted before- and after-school programmes. Other approaches to increasing learning time are included in other sections of the Toolkit, such as Homework, Early years intervention and Summer schools.
  The research focuses on three main approaches to extending teaching and learning time in schools:
  extending the length of the school year;
  extending the length of the school day; and
  providing additional time for targeted groups of pupils, particularly disadvantaged or low-attaining pupils, either before or after school.
- Toolkit: Feedback [Outcome classification code]
  Feedback is information given to the learner and/or the teacher about the learner's performance relative to learning goals. It should aim towards (and be capable of producing) improvement in students' learning. Feedback redirects or refocuses either the teacher's or the learner's actions to achieve a goal, by

aligning effort and activity with an outcome. It can be about the learning activity itself, about the process of activity, about the student's management of their learning or self-regulation or (the least effective) about them as individuals. This feedback can be verbal, written, or can be given through tests or via digital technology. It can come from a teacher or someone taking a teaching role, or from peers.

- Toolkit: Homework [Outcome classification code]
  Homework refers to tasks given to pupils by their teachers to be completed outside of usual lessons. Common homework activities in primary schools tend to be reading or practising spelling and number facts, but may also include more extended activities to develop inquiry skills or more directed and focused work such as revision for tests which is more similar to homework set in secondary schools. Other homework activities may include reading or preparing for work to be done in class, or practising and completing tasks or activities already taught or started in lessons, as well as revision for exams.
- Toolkit: Individualised instruction [Outcome classification code]
  Individualised instruction involves different tasks for each learner and support at the individual level. It is based on the idea that all learners have different needs, and that therefore an approach that is personally tailored—particularly in terms of the activities that pupils undertake and the pace at which they progress through the curriculum—will be more effective. Various models of individualised instruction have been tried over the years in education, particularly in subjects like mathematics where pupils can have individual sets of activities which they complete, often largely independently. More recently, digital technologies have been employed to facilitate individual activities and feedback.
- Toolkit: Learning styles [Outcome classification code]
  The idea underpinning learning styles is that individuals all have a particular approach to or style of learning. The theory is that learning will therefore be more effective or more efficient if pupils are taught using the specific style or approach that has been identified as their learning 'style'. For example, pupils categorised as having a 'listening' learning style could be taught more through storytelling and discussion and less through traditional written exercises.
- Toolkit: Mastery learning [Outcome classification code]
  Mastery learning breaks subject matter and learning content into units with clearly specified objectives which are pursued until they are achieved. Learners work through each block of content in a series of sequential steps.
  Students must demonstrate a high level of success on tests, typically at about the 80% level, before progressing to new content. Mastery learning can be contrasted with other approaches which require pupils to move through the curriculum at a pre-determined pace. Teachers seek to avoid unnecessary repetition by regularly assessing knowledge and skills. Those who do not reach the required level are provided with additional tuition, peer support, small group discussions, or homework so that they can reach the expected level.
- Toolkit: Metacognition and self-regulation [Outcome classification code]
  Metacognition and self-regulation approaches aim to help pupils think about their own learning more explicitly, often by teaching them specific strategies for planning, monitoring and evaluating their learning. Interventions are usually designed to give pupils a repertoire of strategies to choose from and the skills to select the most suitable strategy for a given learning task.
  Self-regulated learning can be broken into three essential components: cognition—the mental process involved in knowing, understanding, and learning;

metacognition—often defined as 'learning to learn'; and
motivation—willingness to engage our metacognitive and cognitive skills.

- Toolkit: Mentoring [Outcome classification code]
  Mentoring in education involves pairing young people with an older peer or volunteer, who acts as a positive role model. In general, mentoring aims to build confidence, develop resilience and character, or raise aspirations, rather than to deliver specific academic skills or knowledge.
  Mentors typically build relationships with young people by meeting with them one to one for about an hour a week over a sustained period, either during school, at the end of the school day, or at weekends.
  Activities vary between different mentoring programmes, sometimes including direct academic support with homework or other school tasks. For programmes focused primarily on direct academic support see 'One to one tuition' and 'Peer tutoring'.
  Mentoring has increasingly been offered to young people who are deemed to be hard to reach or at risk of educational failure or exclusion.
- Toolkit: One to one tuition [Outcome classification code]
  One to one tuition involves a teacher, teaching assistant or other adult giving a pupil intensive individual support. It may happen outside of normal lessons as additional teaching—for example as part of extending school time or a summer school—or as a replacement for other lessons.
- Toolkit: Oral language interventions [Outcome classification code]
  Oral language interventions emphasise the importance of spoken language and verbal interaction in the classroom.
  They are based on the idea that comprehension and reading skills benefit from explicit discussion of either the content or processes of learning, or both. Oral language approaches include:
  Targeted reading aloud and discussing books with young children;
  Explicitly extending pupils' spoken vocabulary; and
  The use of structured questioning to develop reading comprehension. All of the approaches reviewed in this section support learners' articulation of ideas and spoken expression, such as Thinking Together or Philosophy for Children. Oral language interventions therefore have some similarity to approaches based on metacognition, which make talk about learning explicit in classrooms, and to Collaborative Learning approaches, which promote pupils' talk and interaction in groups.
- Toolkit: Outdoor adventure learning [Outcome classification code]
  Outdoor adventure learning typically involves outdoor experiences, such as climbing or mountaineering; survival, ropes or assault courses; or outdoor sports, such as orienteering, sailing and canoeing. These can be organised as intensive residential courses or shorter courses run in schools or local outdoor centres.
  Adventure education usually involves collaborative learning experiences with a high level of physical (and often emotional) challenge. Practical problem-solving, explicit reflection and discussion of thinking and emotion (see also 'Metacognition and self-regulation') may also be involved.
  Adventure learning interventions typically do not include a formal academic component, so this summary does not include forest schools or field trips.
- Toolkit: Parental engagement [Outcome classification code]
  We define parental engagement as the involvement of parents in supporting their children's academic learning. It includes:
  1. approaches and programmes which aim to develop parental skills such as literacy or IT skills;

2. general approaches which encourage parents to support their children with, for example, reading or homework;
3. the involvement of parents in their children's learning activities; and
4. more intensive programmes for families in crisis.

- Toolkit: Peer tutoring [Outcome classification code]
  Peer tutoring includes a range of approaches in which learners work in pairs or small groups to provide each other with explicit teaching support. In cross-age tutoring, an older learner takes the tutoring role and is paired with a younger tutee or tutees. Peer-assisted learning is a structured approach for mathematics and reading with sessions of 25–35 minutes two or three times a week. In reciprocal peer tutoring, learners alternate between the role of tutor and tutee. The common characteristic is that learners take on responsibility for aspects of teaching and for evaluating their success. Peer assessment involves the peer tutor providing feedback to children relating to their performance and can have different forms such as reinforcing or correcting aspects of learning.
  Peers are defined as other students or pupils at the same school or educational setting as the intervention group, or at another local school (e.g. secondary students tutoring pupils at their own or their peers' primary schools). Peers will normally be of similar age and socio-economic or cultural background.
  University students tutoring primary school pupils would not usually be classified as 'peers'.
- Toolkit: Performance pay [Outcome classification code]
  Performance pay schemes aim to create a direct link between teacher pay or bonuses and the performance of their class in order to incentivise better teaching and so improve outcomes. A distinction can be drawn between awards, where improved performance leads to a higher permanent salary, and payment by results, where teachers get a bonus for higher test scores. Approaches differ in how performance is measured and how closely those measures are linked to outcomes for learners. In some schemes, students' test outcomes are the sole factor used to determine performance pay awards. In others, performance judgements can also include information from lesson observations or feedback from pupils, or be left to the discretion of the headteacher.
- Toolkit: Phonics [Outcome classification code]
  Phonics is an approach to teaching reading, and some aspects of writing, by developing learners' phonemic awareness. This involves the skills of hearing, identifying and using phonemes or sound patterns in English. The aim is to systematically teach learners the relationship between these sounds and the written spelling patterns, or graphemes, which represent them. Phonics emphasises the skills of decoding new words by sounding them out and combining or 'blending' the sound-spelling patterns.
- Toolkit: Reading comprehension strategies [Outcome classification code]
  Reading comprehension strategies focus on the learners' understanding of written text. Pupils are taught a range of techniques which enable them to comprehend the meaning of what they read. These can include: inferring meaning from context; summarising or identifying key points; using graphic or semantic organisers; developing questioning strategies; and monitoring their own comprehension and identifying difficulties themselves (see also 'Metacognition and self-regulation').
- Toolkit: Reducing class size [Outcome classification code]
  As the size of a class or teaching group gets smaller it is suggested that the range of approaches a teacher can employ and the amount of attention each student will receive will increase, thereby improving outcomes for pupils.

- Toolkit: Repeating a year [Outcome classification code]
  Pupils who do not reach a given standard of learning at the end of a year are required to repeat the year by joining a class of younger students the following academic year. This is also known as 'grade retention', 'non-promotion' or 'failing a grade'. For students at secondary school level, repeating a year is usually limited to the particular subject or classes that a student has not passed. Repeating a year is very rare in the UK but is relatively common in the USA where the No Child Left Behind Act (2002) recommended that students be required to demonstrate a set standard of achievement before progressing to the next grade level. Students can also be required to repeat a year in some European countries including Spain, France and Germany. In some countries, such as Finland, pupils can repeat a year in exceptional circumstances, but this decision is made collectively by teachers, parents and the student rather than on the basis of end of year testing.
- Toolkit: School uniform [Outcome classification code]
  Schools identify clothing considered appropriate for pupils to wear in school, and usually specify the style and colour. Schools vary as to how strictly a uniform policy is enforced.
- Toolkit: Setting or streaming [Outcome classification code]
  Pupils with similar levels of current attainment are grouped together either for specific lessons on a regular basis (setting or regrouping), or as a whole class (streaming or tracking). The assumption is that it will be possible to teach more effectively or more efficiently with a narrower range of attainment in a class.
- Toolkit: Small group tuition [Outcome classification code]
  Small group tuition is defined as one teacher or professional educator working with two, three, four, or five pupils. This arrangement enables the teacher to focus exclusively on a small number of learners, usually on their own in a separate classroom or working area. Intensive tuition in small groups is often provided to support lower attaining learners or those who are falling behind, but it can also be used as a more general strategy to ensure effective progress, or to teach challenging topics or skills.
- Toolkit: Social and emotional learning [Outcome classification code]
  Interventions which target social and emotional learning (SEL) seek to improve attainment by improving the social and emotional dimensions of learning, as opposed to focusing directly on the academic or cognitive elements of learning. SEL interventions might focus on the ways in which students work with (and alongside) their peers, teachers, family or community. Three broad categories of SEL interventions can be identified:
  1. Universal programmes which generally take place in the classroom;
  2. More specialised programmes which are targeted at students with particular social or emotional problems;
  3. School-level approaches to developing a positive school ethos, which also aim to support greater engagement in learning.
- Toolkit: Sports participation [Outcome classification code]
  Sports participation interventions engage pupils in sports as a means to increasing educational engagement and attainment. This might be through after-school activities or a programme organised by a local sporting club or association. Sometimes sporting activity is used to encourage young people to engage in additional learning activities, such as football training at a local football club combined with study skills, ICT, literacy or mathematics lessons.
- Toolkit: Summer schools [Outcome classification code]
  Summer schools are lessons or classes during the summer holidays, and are

often designed as catch-up programmes. Some summer schools do not have an academic focus and concentrate on sports or other non-academic activities. Others may have a specific focus, such as pupils at the transition from primary to secondary school, or advanced classes to prepare high-attaining pupils for university.

- Toolkit: Teaching assistants [Outcome classification code]
  Teaching assistants (also known as TAs or classroom support assistants) are adults who support teachers in the classroom. Teaching assistants' duties can vary widely from school to school, ranging from providing administrative and classroom support to providing targeted academic support to individual pupils or small groups.
  Cognate terms: support staff; adult support staff; teaching assistants; associate staff; classroom assistants; classroom support assistant; auxiliary teachers; teacher's aide; education paraprofessional; nursery nurse (in early years' settings).

-
-

- Comparison
  Please do not mark this section. This section is completed in the 'Outcomes specific code' screen.
  - With active control [Comparison]
    i.e. there is control for novelty/an introduced new treatment
  - With business as usual [Comparison]
    i.e. comparison group having usual learning experience
  - With no equivalent teaching [Comparison]
    i.e. additional learning time/no treatment, such as in a summer school intervention or a before or after school club

- Intervention outcome measure
  Type or focus of educational test used to measure the outcome of the impact of the intervention or approach.
  - Literacy: reading comprehension [Intervention]
    E.g. passage comprehension
  - Literacy: decoding/phonics [Intervention]
  - Literacy: spelling [Intervention]
  - Literacy: reading other [Intervention]
    Other reading outcomes (e.g. reading fluency, vocabulary comprehension (receptive vocabulary))
  - Literacy: speaking and listening/oral language [Intervention]
  - Literacy: writing [Intervention]
  - Mathematics [Intervention]
  - Science [Intervention]
  - Social Studies [Intervention]
    E.g. history, geography, economics
  - Arts [Intervention]
    E.g. music, art
  - Languages [Intervention]
    Second or foreign languages, based on the dominant language of instruction in the educational setting.

- Curriculum: other [Intervention]
  Other curriculum outcomes not included in the above options (please specify).
- Combined subjects [Intervention]
  Where the study combines two or more test outcomes from different subjects to provide an overall measure of educational progress (e.g. KS2 English and mathematics or multiple GCSE subjects).
- Cognitive: reasoning [Intervention]
  Tests of verbal, analogical or visual reasoning, including IQ or other 'intelligence' tests.
- Cognitive: other [Intervention]
  Other tests of cognitive performance such as working memory or perception.

Appendix 4: EEF feedback review—Study quality assessment

This is the study quality assessment tool for the EEF feedback review. Use responses from existing coding in the Main (MDE), Effect Size (ESDE), and Review Specific (RS) data extraction tools as specified.

- **Domain 1: Bias in selection/confounding bias**
  This domain assesses the level of confidence we can have that any differences in outcome between the intervention group and the control group can be attributed to the intervention and not to other differences between the characteristics of these groups or the experiences during the study.
  A) How were participants assigned to groups (see MDE Sec2 & ESDE Sec 1)?
    1. Random allocation (details provided)—Low risk
       Use when method of allocation is Random (MDE Sec 2) and details of the randomisation procedure provided (ESDE Sec 1)
    2. Non-random, but matched—Moderate risk
    3. Random allocation (no details provided)—Moderate risk
       Use when no details of method of randomisation are provided
    4. Not random, not matched prior to intervention—Serious risk
    5. Unclear—assume not random not matched—Serious risk

  B) Is comparability taken into account in the analysis (see ESDE Sec 2)?
    1. Yes—Low risk (also use for studies with random allocation)
       Where a study has random allocation code as—Yes
    2. No—Serious risk

- **Domain 2: Bias in the measurement of outcomes**
  How confident can we be that any difference in outcome between the intervention and control group is attributable to the intervention and not to who measured the outcome or how?
  A) Who undertook the outcome evaluation (see MDE sec 5)?
    1. The developer—Moderate risk
    2. A different organisation paid by the developer—Moderate risk
    3. Independent organisation—Low risk
    4. Unclear—assume developer—Moderate risk

  B) What type of test was used to measure the outcome (see MDE section 6)?
    1. Standardised test—Low risk
    2. Researcher-developed test—Moderate risk
    3. National test—Low risk
    4. School-developed test—Moderate risk
    5. International test—Low risk

- **Domain 3: Bias due to missing data**
  How confident can we be that any difference in outcome between the intervention and control is not due to changes in the composition of the groups between baseline and outcome measurement?
    1. How many participants are entered into the study?
       Use number from the description of sample provided by the authors (not results)
    2. How many participants are included in the analysis?
       Use the total number from the outcome data extraction used for the effect size

  A) Is there a difference between the number of participants entered and the number of participants analysed?
  Use your answers to questions above to calculate this. It is the difference between the number of participants entered as described in the methods/sample section of the paper and the number of participants used to calculate the effect sizes (see note below) expressed as a % of the number entered e.g. If sample = 100 and number used in effect size = 90 then difference n = 10 or 10%.

Where the study has more than 1 group (e.g. 2 intervention groups), the total analysed needs to be the total for all groups analysed in the study report.

1. Difference less than 5%—Low risk
2. Difference 5–19%—Moderate risk
3. Difference 20% or more—Serious risk

- **Domain 4: Bias due to selective outcome reporting**
  How confident can we be that any difference in outcome between the intervention and control is attributable to the intervention and not to the selective reporting of outcomes?
  A) Are results reported for all review relevant outcomes that are specified in the methods?
  Look at the attainment/cognitive outcomes that the authors say are used in the study in the methods section and compare this with the results reported. Are all of the outcomes specified in the methods section that are relevant to the review reported in the results section of the paper?
  e.g. If there are maths and science outcomes specified but only maths outcomes reported, then maths are missing and code = No serious risk
  - Yes—Low risk
  - No—Serious risk (specify those missing)

- **Overall risk of bias**
  Combine the results from Domains 1 to 4 to provide overall estimate of risk of bias.
  1. Low risk of bias
     Not more than 1 moderate risk in any domain
     No serious risks in any domain
  2. Moderate risk of bias
     Not more than 1 serious risk in any domain
     Low or moderate risk of bias in all other domains
  3. Serious risk of bias
     Serious risk of bias in more than one domain

- **Ecological validity**
  How confident can we be that the findings of the study predict the result in real world conditions (mainstream school)?
  A) Who was responsible for teaching at the point of delivery (see MDE Sec 5)?
  1. Research staff—Moderate
  2. Class teachers—High
  3. Teaching assistants—High
  4. Other school staff—High
  5. External teachers—Moderate
  6. Parents/carers—High
  7. Peers—Moderate
  8. Lay person/volunteers—Moderate
  9. Digital—High
  10. Unclear not specified—Moderate

- **B) What was the source of the feedback (see SSDE)?**
  1. Teacher—High
  2. Researcher—Moderate
  3. Digital—High

- **Overall ecological validity**
  Combine the results of the previous questions in the tool.
  1. High & High = High
  2. High & Moderate = Moderate
  3. Moderate & Moderate = Moderate