Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1418

Testing frequency and student achievement: A systematic review

Jens Dietrichson, Julie Kaas Seerup, Anja Bondebjerg Mølgaard, Malene Wallach Kildemoes, Frederikke Lykke Witthöft Schytt, Mikkel Vembye, Elizabeth Bengtsen, Bjørn Christian Ar-leth Viinholt, Morten Kjær Thomsen

School-based testing is widely used for monitoring students' academic progress. Proponents argue that testing ensures accountability and guides teachers and managers, whereas opponents point to adverse consequences such as teaching to the test, and frequent testing creating anxiety and stress. This review examined the effects of interventions that change how frequently primary and second-ary students are tested on measures of student achievement and testing anxiety. The search resulted in 102,451 potentially relevant records. Ninety-three records, nested in 87 studies, met the inclusion criteria, and we included 59 studies in the data synthesis. We found only one study that reported effects on testing anxiety. Almost all interventions involved practice tests that were low-stakes and had a formative purpose. We found statistically significant weighted average effect sizes on academic achievement in interventions where the control group did not receive any practice test: 0.22 (95% CI = [0.09, 0.34]) for between-subject designs and 0.46 (95% CI = [0.29, 0.62]) for within-subject designs. When the control group was also tested, the estimates were consistently positive but not always statistically significant, and sometimes indicated effects that diminished as the control group testing frequency increased. There was substantial heterogeneity in all meta-analyses. In exploratory analyses, practice tests seemed less effective when the learning material was complex, and when the alternative pedagogical method involved more active forms of learning, but more research is needed to pin down the contexts in which testing may not be beneficial for student achievement.

VERSION: March 2026

Suggested citation: Dietrichson, Jens, Julie Kaas Seerup, Anja Bondebjerg Mølgaard, Malene Wallach Kildemoes, Frederikke Lykke Witthöft Schytt, Mikkel Vembye, Elizabeth Bengtsen, Bjørn Christian Ar-leth Viinholt, and Morten Kjær Thomsen. (2026). Testing frequency and student achievement: A systematic review. (EdWorkingPaper: 26-1418). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/jas3-2b83

Testing frequency and student achievement: A systematic review

Jens Dietrichson[1], Julie Kaas Seerup[1], Anja Bondebjerg Mølgaard[1], Malene Wallach Kildemoes[1], Frederikke Lykke Witthöft Schytt[1], Mikkel Vembye[1], Elizabeth Bengtsen[1], Bjørn Christian Arleth Viinholt[2], Morten Kjær Thomsen[3]

[1]VIVE - The Danish Center for Social Science Research
[2]The Royal Danish Library
[3]Leverhulme Centre for Demographic Science, and Nuffield College, University of Oxford, Oxford, UK

Acknowledgements

All meta-analytic data, analysis code, and research materials (including our full coding of each record) will be made available at OSF upon publication. None of the authors has a conflict of interest.

VIVE – The Danish Center for Social Science Research, Denmark provided funding. Morten Thomsen acknowledges funding from the Leverhulme Trust (Grant RC-2018-003) for the Leverhulme Centre for Demographic Science.

We are grateful to Fred Paas for taking the time to answer our queries, to James Pustejovsky and Elizabeth Tipton for helpful discussions, and to seminar participants at the National Agency for Education and Quality for helpful comments and suggestions. We thank Anton Dam, Emilie Erreboe Hougaard, Emilie Mai Anderberg, Frederikke Sissel Blohm, Juliane Esper Ramstedt, and Tanne Ebert Jørgensen for providing excellent research assistance at different stages of this project.

Correspondence concerning this article should be addressed to Jens Dietrichson, Herluf Trolles Gade 11, 1052 Copenhagen K, Denmark. E-mail: jsd@vive.dk.

Abstract

School-based testing is widely used for monitoring students' academic progress. Proponents argue that testing ensures accountability and guides teachers and managers, whereas opponents point to adverse consequences such as teaching to the test, and frequent testing creating anxiety and stress. This review examined the effects of interventions that change how frequently primary and secondary students are tested on measures of student achievement and testing anxiety. The search resulted in 102,451 potentially relevant records. Ninety-three records, nested in 87 studies, met the inclusion criteria, and we included 59 studies in the data synthesis. We found only one study that reported effects on testing anxiety. Almost all interventions involved practice tests that were low-stakes and had a formative purpose. We found statistically significant weighted average effect sizes on academic achievement in interventions where the control group did not receive any practice test: 0.22 (95% CI = [0.09, 0.34]) for between-subject designs and 0.46 (95% CI = [0.29, 0.62]) for within-subject designs. When the control group was also tested, the estimates were consistently positive but not always statistically significant, and sometimes indicated effects that diminished as the control group testing frequency increased. There was substantial heterogeneity in all meta-analyses. In exploratory analyses, practice tests seemed less effective when the learning material was complex, and when the alternative pedagogical method involved more active forms of learning, but more research is needed to pin down the contexts in which testing may not be beneficial for student achievement.

Background

School-based testing is often well suited and widely used for monitoring students' academic progress. However, for almost a century, researchers have taken an interest in the potential moderating qualities of testing students on their academic achievement (Jones, 1923; Keys, 1934; Kulp, 1933; Snedden, 1931; Spitzer, 1939; Turney, 1931). That is, how does exposing students to school-based testing at different frequencies contribute to their academic progress?

In spite of many attempts to answer this question and to formulate guidelines for how to use school-based testing for the benefit of students, educational scholars, policy makers, and teachers have yet to fully understand the potential effects of school-based testing on the academic achievement of students (e.g., Bergbauer et al., 2018; National Research Council, 2011; World Bank, 2017).

In latter years, tests have become associated with the accountability of national educational systems and are now widely used in international comparisons of student achievement. Since the 1970s, scholars have debated this phenomenon, sometimes referred to as 'the global testing culture' or 'the educational reform movement' (National Research Council, 2011; Sahlberg, 2010; Smith, 2016). This trend encompasses an increased political focus on accountability and assessment of learning in educational systems and has fostered both national and international testing schemes, such as the Programme for International Student Assessment (PISA), the progress in International Reading Literacy Study (PIRLS), and the Trends In Mathematics and Science Study (TIMMS). Furthermore, various national educational policies, such as the No Child Left Behind Act in the United States, that emphasised rigorous testing have been implemented.

The testing materials in these aforementioned testing schemes and national policies typically consist of standardised tests, which allow comparisons between separate school systems. With the rapid development of new technologies, the practicalities of testing students become easier to manage. Some would argue that testing students is an inevitable tool in present day educational systems to secure and demonstrate accountability, and guide teachers and managers to create optimal learning environments for students (Buck et al., 2010; Crooks, 1988; National Research Council, 2011). At the same time, there is an ongoing policy debate on the impact of testing on students' academic achievement (e.g., Bergbauer et al., 2018; National Research Council, 2011). Testing has acquired a negative connotation for some practitioners, particularly when it comes to standardised high-stakes, summative testing (Rawson & Dunlosky, 2012). Opponents of standardised testing schemes often refer to adverse consequences such as teachers focusing on 'teaching to the test', and frequent testing fostering distress and emotional difficulties among students (DordiNejad et al., 2011; Organization for Economic Co-operation and Development [OECD], 2017).

Furthermore, testing takes time away from other modes of instruction (Rawson & Dunlovsky, 2012), and implementing large-scale standardised testing, like in the No Child Left Behind policy, is expensive (Azin & Resendez, 2008). Schools may spend as much as 18% of all student time on testing or test preparation (Bruhn et al., 2025; DePaepe et al., 2015). It is therefore important to examine the effects of more frequent testing on student achievement. Also, as there may be adverse effects of testing students more frequently, decisions regarding testing frequency should not only consider academic achievement but also the impact on students' emotional wellbeing.

In accordance with this, the current systematic review aimed to examine how changes in the frequency of school-based tests affect academic achievement (primary outcome measure) and testing-related anxiety (secondary outcome measure) among primary and secondary school students. However, we were not able to conduct analyses of our secondary outcome, testing anxiety, since only one included study reported this outcome (see further discussion in sections Objectives and Results).

The effects of changing the testing frequency may depend on the subject students are tested in, how the tests are conducted, and on student characteristics. Feedback from school-based tests may be more important in some subjects than others (Azmat & Irriberri, 2010), and testing anxiety often differs across subjects (e.g., Wolters & Pintrich, 1998). The age and grade of students may moderate effect sizes (Adesope et al., 2017), for example, because the grade affects how important test information is for teachers, and students react differently to being tested at different ages. As evidenced by the critique against standardised and high-stakes testing, the type of test may matter for both achievement and anxiety. Higher exposure to an increased testing frequency may affect students differently compared to less intensive exposure (Yang et al., 2021). Finally, earlier findings indicate that females are more likely to experience and report higher levels of testing anxiety than males (DordiNejad et al., 2011; OECD, 2017; Segool et al., 2013; von der Embse et al., 2018). In this review, we examined the heterogeneity using both pre-specified moderators and exploratory moderator analyses.

The intervention

We examined the effects of testing frequency alterations on the academic achievement of primary and secondary school students. Whereas specific criteria for inclusion/exclusion of studies, including those concerning the intervention, will be thoroughly described in section Criteria

for considering studies for this review, the present section serves as a short description of the interventions we judged as relevant for this review. Overall, we had six criteria that any intervention had to meet in order to be eligible for inclusion:

First, interventions had to alter the testing frequency for at least one student group and provide information about the benchmark testing frequency in the control group.

Second, to secure the ecological validity of results, we required that interventions had been implemented within a school setting. Hence, laboratory-based interventions were excluded.

Third, interventions had to target either primary or secondary school students. Interventions set in tertiary education, such as university settings, or in pre-school settings, were not included.

Fourth, we required that interventions only manipulated the testing frequency without incorporating any major additional components that may confound the effects of testing. Interventions combining changes in testing frequency (e.g., by introducing curriculum-based measurement) with other components (e.g., peer-assisted learning strategies) were not eligible for inclusion.

Fifth, applied tests had to be recorded on paper, computer, tablet, and so forth. Orally performed tests were included if their results were recorded.

Sixth, intervention effects should not be evaluated by outcome tests that give one group an unfair advantage since this makes it difficult to separate the effect of more frequent testing from more exposure to the instructional material. For example, if the content of the outcome tests was identical to the material used during the intervention, and the control group had not been exposed to this material, we excluded the intervention.

How the intervention might work

Theoretically, testing students may have both beneficial and adverse effects on achievement and anxiety. In favour of increased testing frequencies, numerous researchers have suggested that students achieve more when frequent testing is implemented. Test results may provide teachers with information about a given student's difficulties and strengths, thus serving to qualify personal feedback (Dunlosky et al., 2013; McDaniel et al., 2007; Rawson & Dunlosky, 2012), and allowing for better aimed individual and class-level instruction (Black & Wiliam, 2009). Tests may imply that students receive feedback (with or without teacher mediation) that allows them to spot areas of weaknesses and correct errors (Adesope et al., 2017; Standlee & Popham, 1960). Tests may also help provide structure to a course and point out the most important parts to students (Standlee & Popham, 1960). Furthermore, tests might act as extrinsic motivators, leading students to study harder (Bernatzsky, Cabrera, & Cid, 2017; Standlee & Popham, 1960) and teachers and schools to increase efforts to improve student achievement (Woessmann, 2002).

In addition, it has been suggested that frequent testing administered in the form of practice tests (formative tests) has the potential to improve student retention and thereby learning (Carpenter et al., 2009; Carpenter, 2012; Dunlosky et al., 2013; Gates, 2017; Glover, 1989; Karpicke & Aue, 2015; Rawson & Dunlosky, 2012; Rowland, 2014; Yang et al., 2021). The phenomenon whereby tests can improve learning through enhanced retention is often referred to as the 'testing effect' or 'test-enhanced learning' (Yang et al., 2021). There are two main hypotheses that seek to explain how the testing effect comes into existence. The first hypothesis, 'amount-of-processing', states that exposure to the material is vital for recall: the more exposure,

the higher probability of correct recall. In this perspective, testing is seen as an additional exposure to the material. The second hypothesis, which focuses on the retrieval process of recall, states that retrieving an item from memory will strengthen existing retrieval routes or create new ones, thereby increasing the likelihood of correct recall (Dempster, 1996). That is, the act of processing and retrieving information strengthens recall. Thus, active repetition of the material, for example by taking a recall test, has been shown to yield greater long-term retention of a material than restudying (e.g., Roediger III & Karpicke, 2006a, 2006b).

There may also be disadvantages to testing. Some authors have suggested that the testing effect is more likely to appear when the learning task is of low rather than high complexity (Gates, 2017; Hanham et al., 2017; Kühn, 1914 [cited in van Gog & Sweller, 2015], Leahy et al., 2015; van Gog & Sweller, 2015). Van Gog and Sweller (2015) described the complexity level as being dependent on the skill level of the student, and on the level of element interactivity. A low element interactivity task is characterized by individual elements that can be learned without reference to other elements (e.g., lists of independent word definitions or science facts), and a high element interactivity task is one where elements are related and have to be processed in working memory simultaneously. Solving simple algebraic equations (e.g., solve for a in a/b = c) is an example of a task that is typically complex for novices but becomes less complex as the process is automated and the number of steps and elements is reduced. To learn complex tasks, students must make connections between task elements (van Gog & Sweller, 2015), and build schemas for how the problems should be solved (McLaren et al., 2016). For example, because working memory is capacity constrained (e.g., Cowan, 2001; Oberauer et al., 2018), connections may have to be made between elements to reduce the number of interacting elements and avoid over-

load. Connections and schemas may be easier to form when restudying or studying worked examples than when being tested (Leahy et al., 2015; McLaren et al., 2016; van Gog & Sweller, 2015).

Previous findings support the common sense assumption that increasing the testing frequency indefinitely does not result in indefinite increases in student achievement (Bangert-Drowns et al., 1991). In other words, doing ten tests is not necessarily ten times as good as doing one test; on the contrary, "over-testing" may lead to decreases in student achievement. One reason for this is that testing likely takes time away from instruction, and at some point, additional instruction will be of greater value to student achievement than testing (Rawson & Dunlovsky, 2012). Strengthened extrinsic motivation may lead to an unwanted focus on the tested material (i.e., "teaching to the test") or more generally on tasks that are tested instead of all curriculum elements, which is an extra pertinent risk as schools are "multitasking organisations" (Holmström & Milgrom, 1991).

Being exposed to testing is also associated with a certain amount of stress and potential demotivation (Cheek et al., 2002), and frequent testing may lead to higher levels of testing anxiety. Testing anxiety occurs in situations where one's skills are being evaluated and is most commonly defined in the literature as "a set of phenomenological, physiological, and behavioural responses that accompany concern about the possible negative consequences of failure on an exam" (Zeidner, 1998, p. 17). High levels of testing anxiety can have adverse consequences for students' academic achievement and may have a negative impact on general life satisfaction (OECD, 2017; & Hasson, 2012; von der Embse et al., 2018). It is conceivable that testing frequency has a nonlinear effect on anxiety: moving from zero to one test may increase anxiety, whereas further increases in frequency may familiarise students with taking tests and thus re-

duce anxiety. As we describe in more detail in section Included studies, we found very few studies that measured testing anxiety and fit our inclusion criteria, which meant that we could not conduct a meta-analysis on this outcome.

Potential moderators

Changing the testing frequency may have heterogeneous effects on both achievement and anxiety, which we investigated by testing whether variables related to subject, student, and test characteristics moderated effect sizes. We motivate the choice of these moderators below and define them in section Subgroup analysis and investigation of heterogeneity.

First, turning to the subject variable, it is possible that feedback from school-based tests may be more important in some subjects than others, which could help explain the heterogeneous effects of performance feedback (Azmat & Iriberri, 2010). Feedback may for example be more pertinent in subjects where students are not likely to receive performance information from other sources than schools and teachers (such as science). Also, students may react differently to feedback depending on the perceived importance of the subject. As feedback is one reason why testing may affect achievement, these explanations may also be relevant for changes in testing frequency. Furthermore, students' self-reported levels of testing anxiety differ across subjects (Wolters & Pintrich, 1998), which may directly influence the impact of testing frequency on measures of testing anxiety. Differential anxiety across subjects could also lead to heterogenous effects in student achievement, if it is the case that anxiety affects achievement (von der Embse & Hasson, 2012; von der Embse et al., 2018). We therefore examined heterogeneity of effects across subject categories.

Information derived from tests is more important when teachers know less about their students and when students know less about their own performance (Azmat & Iriberri, 2010;

Raudenbush, 1984). For example, both teachers and students know less about students' achievement level when students have just started school, and teachers know less when students transition between teachers. Also, it may become harder for teachers to keep track of individual students in later grades where classes tend to be larger. Students' emotional reactions to being tested may also depend on their age as well as on how accustomed they are to test-taking. Thus, students' grade level may be a potential source of heterogeneity.

Another student characteristic of potential importance is gender. Earlier findings indicate that female students are more likely to experience and report higher levels of testing anxiety than male students (DordiNejad et al., 2011; OECD, 2017; Segool et al., 2013; von der Embse et al., 2018). In PISA, 64% of girls compared to 47% of boys reported being 'very anxious before a test, even when they are well prepared' (OECD, 2017). This difference motivated us to examine gender as a moderator.

Besides the frequency of testing, the duration of the testing period might matter (Yang et al., 2021). Being exposed to frequent testing during a longer period may compound the benefits of testing but may also lead to, for example, testing fatigue and more anxiety. We therefore included intervention duration as a moderator.

Finally, test characteristics may matter for both achievement and anxiety. High-stakes tests are more likely to strengthen extrinsic motivation compared to low-stakes tests, potentially amplifying both the beneficial and the harmful effects of such motivation. Some studies show that students report significantly higher anxiety levels on standardised high-stakes (summative) tests versus classroom (formative) tests (Segool et al., 2013). Furthermore, some researchers argue that short and frequent practice tests yield more positive outcomes for learning and life satisfaction than long and infrequent summative tests (Dunlosky et al., 2013; Rawson & Dunlosky,

2012). We therefore planned to examine heterogeneity across the effects of high- and low-stakes tests, as well as summative and formative tests, but were unable to do so because only a small number of interventions involved high-stakes or summative tests.

Why it is important to do this review

Four previous reviews examining the effect of testing on student achievement pursued objectives similar to ours and employed meta-analytic methods. Below, we describe these earlier reviews and outline how our review differs from them. We discuss our results in relation to this previous literature (including the magnitude of effects) in the section Agreements and disagreements with other studies or reviews.

Bangert-Drowns et al. (1991) reviewed the effects of frequent testing on test scores. Their meta-analysis revealed an average effect that significantly favoured the experimental groups, that is, the groups that received more frequent testing. The effect size declined considerably when the control group was tested as well (as opposed to receiving zero tests; Bangert-Drowns et al., 1991, p. 94-95). In addition to a considerably longer, up-to-date search period, there are two main differences, which sets our review apart from Bangert-Drowns et al. (1991). Firstly, Bangert-Drowns et al. (1991) focus on students attending secondary school or higher education. Thus, only the secondary school category overlaps between the two reviews. Second, Bangert-Drowns et al. (1991) limited their search to studies from the US, whereas we imposed no geographic restrictions to our literature search.

A recent review by Yang et al. (2021) examined if testing increased learning from elementary school to university and continuing education. They found an overall positive and significant effect in their meta-analysis, which increased with the number of tests in the treatment group. The effect became smaller, albeit still positive and significant, when control groups were

restudying. The effect was positive but clearly smaller and not significantly different from zero when the control group used "elaborative strategies" (e.g., concept mapping, note-taking, summarizing). The most important differences between the inclusion criteria of Yang et al. (2021) and ours are that a) they included interventions in tertiary education; b) they included outcome tests where the control group had clearly less exposure to the tested material; c) they excluded studies in which the control group also received one or more tests. Their effect sizes therefore reflected the contrast between testing and no testing.

The review and meta-analysis of Adesope et al. (2017) focused on the effect of low-stakes practice tests. Overall, they found positive and significant benefits of the tests. The effects were larger when the control group performed a filler task or did not perform any activity other than restudying. The effects were also larger when practice and outcome tests were identical, although still positive and significant when they were different. Similar to Yang et al. (2021), Adesope et al. (2017) reported effect sizes for one and two or more practice tests compared to none, and most studies in their review were laboratory experiments.

Phelps (2012) included both high-stakes and low-stakes tests and a wider range of participants than primary and secondary school students. The effect size closest to the objectives of our review was for a contrast where the treatment group was "tested more frequently than the control group" (p. 34). Thus, neither Adesope et al. (20217) nor Phelps (2012) fully answered our primary research question.

As mentioned, we wanted to examine the effects of testing frequency on testing anxiety but were unable to do so quantitatively because too few studies reported this outcome while also meeting our inclusion/eligibility criteria. Yang et al. (2023) reviewed the effect of practice tests on test anxiety. Their review included studies conducted in both school and laboratory settings,

with no restrictions on student age, grade level, or potential co-interventions. They found 24 studies, only one of which met our inclusion criteria. The meta-analysis indicated that practice tests reduced testing anxiety.

Other researchers who have reviewed related topics concerning test-enhanced learning are Black and Wiliam (2009), Fuchs and Fuchs (2001), Karpicke and Grimaldi (2012), Kingston and Nash (2011), McDaniel et al. (2007), Rawson and Dunlosky (2012), and Rowland (2014). Except for Rowland (2014) and Kingston and Nash (2011), none of the aforementioned reviews conducted a meta-analysis and therefore did not answer our primary research question. Rowland (2014) covered the psychological literature on the testing effect and did not focus on educational contexts. Kingston and Nash (2011) focused only on formative assessment and did not analyse testing frequency.

We believe that this review constitutes a valuable contribution to the testing literature by providing an up-to-date rigorous overview of the current research base regarding the effects of testing frequencies in primary and secondary education. In addition to the differences in inclusion criteria, our search of electronic databases was more encompassing than earlier reviews and we conducted a risk of bias assessment as well as investigations of heterogeneity unseen in the earlier literature. It is our hope that our review will inform teachers, researchers, and policymakers in search for answers on what constitutes an optimal testing frequency in various primary and secondary educational settings. Our primary outcome, students' academic achievement, should be highly relevant to educational stakeholders at all levels, from teachers to national policymakers.

Objectives

The primary research question examined in this review was: What are the effects of different testing frequencies on student achievement?

Our secondary research question was: What are the effects of different testing frequencies on measures of students' testing anxiety?

As mentioned, we could not conduct a meta-analysis corresponding to the secondary research question due to the low number of included studies investigating testing anxiety. Therefore, our examination of testing anxiety is limited to a narrative summary.

We furthermore examined potential moderators. Our tertiary research question asked: How are the effects of different testing frequencies on student achievement and testing anxiety moderated by subject, grade, type of test, duration of the intervention, and gender?

In addition to examining these pre-specified moderators, we examined heterogeneity in a range of exploratory analyses.

Methods

We followed the modernized Campbell's Methodologic Expectations for Campbell Collaboration Intervention Reviews (MECCIR; Aloe et al., 2024) reporting guideline and conducted our analyses in accordance with our pre-registered protocol (Thomsen et al., 2022) to the greatest extent possible.

Criteria for considering studies for this review

Types of studies

We included study designs that contrasted a treatment to a control group (treatment-control design) or two alternative treatments (comparison design). Both randomised controlled trials

(RCT) and quasi-experimental studies (QES) were eligible for inclusion. The treatment group was in all studies a group that was tested more frequently than the control group.

Within the treatment-control design, we defined a control group in terms of 'business as usual', that is, students being exposed to their normal testing regimen. A comparison design implied that all intervention groups were exposed to manipulated testing frequencies in the intervention period – meaning that none experienced their normal testing regimens, including those assigned to no testing, given that this was not their usual practice. Comparison designs also include studies in which the control group instruction did not include any tests but was altered in some other way (e.g., including worked examples, summarizing, etc).

We included both between-subject and within-subject designs. In between-subject designs, different students are assigned to treatment and control/comparison conditions (i.e., each student is assigned to one condition). In within-subject designs, students are their own control group (i.e., each student is assigned to all conditions). As an example, suppose an intervention aiming to teach students science facts uses a within-subject design and two conditions, testing (treatment) and restudy (control). Then each student is assigned to both the testing-condition and the restudy-condition and learn some facts in one condition and the other facts in the other.

Furthermore, to be included in this review, studies needed to assign at least two 'units' (e.g., schools, classes, or students) to the treatment group and at least two units to the control group. We excluded studies assigning only one unit to one or both groups due to the difficulty of separating treatment effects from unit effects.

Types of participants

The eligible population for this review consisted of students attending either primary or secondary school, which in most countries entailed a span from kindergarten until grade 12 (K-

12). In some countries, kindergarten is not a part of the formal school system but rather a form of child care or preschool (e.g., in the United Kingdom, UK). Studies conducted in such settings were excluded from the review, leaving only studies conducted in a formal primary or secondary school settings. We based this decision on the fact that tests of achievement in preschool/child care settings are likely to be incomparable to those applied in a school setting. Furthermore, in countries where kindergarten constitutes a form of child care rather than a school setting, many children do not to attend, making this population different from children in primary and secondary education. For example, in the UK, 29 percent of pre-school aged children do not attend any formal child care (Department for Education, 2018).

As mentioned previously, we also excluded studies with participants enrolled in tertiary education, such as universities. The student population found in higher education is different to the one found in K-12, not only because of age differences but also because higher education is neither obligatory nor close to universal. Course structures, subjects, and curricula typically differ as well. For these reasons, we considered that tertiary settings sufficiently different from primary and secondary school settings to exclude studies set in tertiary education from the review.

Apart from our exclusion of preschool and tertiary education settings, we did not restrict the type of school where an intervention might take place. We included studies carried out in both regular schools, special schools, private schools, and so forth. We also included studies of all types of students, irrespective of their socioeconomic status, potential at-risk status, or mental and/or physical capabilities.

Types of interventions

This review encompassed interventions that manipulated the frequency at which students were tested during an intervention period. We measured the frequency of these “practice tests” as

opposed to “outcome tests”, which are the tests used to measure the effects of the intervention. To be included, studies had to report at least the testing frequency of the control group and ascertain that the treatment group conducted more practice tests than the control group.

Some studies included not only end-of-intervention (or immediate) outcome tests but also delayed or follow-up outcome tests. In these cases, the end-of-intervention test typically covered relevant learning material and thus worked as a practice test in relation to the follow-up outcome test. We therefore counted the end-of-intervention test as a practice test in relation to the effect size based on the follow-up outcome test. For example, suppose an intervention included one practice test in the treatment group and no practice test in the control group, and then one end-of-intervention test outcome test and one follow-up outcome test, which are conducted by both the treatment and control group. In such cases, we calculated two effect sizes, one for each outcome test. The testing frequency for the end-of-intervention effect size was one in the treatment group and zero in the control group, whereas the testing frequency for the follow-up effect size was two in the treatment group and one in the control group.

We did not include “pre-questions” or pre-treatment tests in the testing frequency, that is, tests or quizzes taken before the learning material had been presented. While pre-questions can support learning by helping students and teachers identify prior knowledge and areas needing attention, they operate through different mechanisms than practice tests; consequently, interventions comparing the two types are unlikely to be fully comparable.

To count as a practice test, the results had to be recorded on paper, computer, tablet, and so forth. Orally performed tests were included if their results were recorded. Some interventions included retrieval practice conditions in which students were asked to write down everything they can remember or make free notes about a previously studied topic without any feedback

about what they wrote (e.g., Ritchie et al., 2013; Rowley & McCrudden, 2020). We did not consider such conditions to include practice tests and therefore excluded these studies.

Specific interventions such as progress monitoring or curriculum-based measurement were included as long as there were no significant co-interventions confounding our ability to isolate a testing effect. We assessed this criterion by 1) considering whether there were any co-interventions and 2) judging whether such co-interventions were unequally distributed between treatment and control/comparison groups. That is, if an intervention featured different testing frequencies for the treatment and control/comparison groups, while keeping all co-interventions identical, we considered it possible to isolate a testing effect, as the testing frequency would be the only difference between the groups. We exemplify studies excluded based on the presence of co-interventions in section Excluded studies.

We did not restrict the intervention length, meaning that our analyses covered a relatively broad range of durations, from interventions lasting only one session to interventions lasting a year or more. As part of our data extraction, we recorded intervention length and examined differences between those with longer and shorter durations.

Types of outcome measures

We included studies that test the effect of changing the testing frequency on either our primary outcome (measures of academic achievement), our secondary outcome (measures of testing anxiety), or both.

Primary outcomes. In this review, academic achievement was not restricted to specific subjects, such as math or reading. We included achievement measures within all subjects, including subjects such as history and social and natural sciences. Furthermore, we included both standardised and non-standardised tests, formative and summative tests, as well as low and high

stakes tests, with the intention of performing moderator analyses according to the type of test. However, as noted, we found too few high-stakes or summative tests to conduct the planned analyses.

We did not consider all outcome tests used in the testing frequency literature to be tests of academic achievement. For example, pure memorization tests of non-academic content were excluded (e.g., pictures in Ma et al., 2020, and letter combinations in de Diego-Lázaro, 2024).

Our protocol stated that we would not include studies in which the practice tests used during the intervention were identical to the outcome test used to estimate the effects. The motivation for this exclusion criterion was that identical tests could confer an unfair advantage on the treatment group because they had more exposure to the learning material. That is, the potential effect might not be so much an effect of more frequent testing as an exposure effect. However, certain types of control groups (e.g., restudy conditions) were similarly exposed to the material, implying that they were not at an unfair disadvantage. Furthermore, for certain types of learning material–such as word definitions or history facts–where the learning of one item should not be expected to transfer to other items, testing on non-identical items would be unlikely to capture the intervention effect.

In practice, we therefore judged this criterion by considering two main elements. First, we determined whether practice tests and outcome tests were completely identical. To be identical, the test had to cover the same content or items, and be in the same form. If free recall practice tests covered the same material as a multiple-choice outcome test, we did not consider the tests identical and included the study. If the tests on the other hand contained the same items in the same form but in a different order, we deemed the test identical and excluded. If the test con-

tained both identical and non-identical test items, which were reported separately, we included only the non-identical items.

Secondly, we considered whether the presence of identical tests lead to an unfair advantage for those tested, in the sense that those tested were exposed to the same test items on several occasions, while controls/comparison participants had less exposure to the items. If both treatment and control participants were similarly exposed to the test items – for example, if treated students were tested on the items, while control students studied them – we considered it possible to isolate a testing effect. In such cases, the only difference between the groups was the type of exposure (test vs. study), not the extent of exposure. If an immediate and a delayed outcome test were identical and the immediate test counted as a practice test (see section Types of interventions), we still included the effect size based on the delayed outcome test as both treatment and control received the immediate test.

Secondary outcomes. At the outset of this review, we planned to include measures of testing anxiety as a secondary outcome measure. We considered a wide range of testing anxiety scales to be eligible, examples being the Test Anxiety Inventory (TAI), the Test Anxiety Scale for Children (TASC), and sub-tests from more general social-emotional measures such as the anxiety content subscale from the Behaviour Assessment Scale for Children, Second edition (BASC-2-TA). As previously mentioned, we were unfortunately not able to conduct a meta-analysis of this outcome due to a lack of studies.

Duration of follow-up. We did not restrict the outcomes in terms of the duration of follow-up.

Search methods for identification of studies

We identified relevant studies through searches in electronic databases, hand search in relevant research journals, searches for systematic reviews, grey literature searches, citation-tracking, and contact to international experts. We aimed to retrieve both published, on-going, and unpublished studies, with the goal of securing as close as possible to full coverage of the research field. The searches performed for this review were carried out in two main phases and encompassed both traditional duplicate screening by reviewers and research assistants as well as automated screening procedures, harnessing recent developments within large language model (LLM) technologies. In the following, we describe and document the methods used, with justification of the choices made throughout the search process.

Electronic searches

In the first phase of our search process, we conducted electronic searches in the following bibliographic databases:

- Academic Search Premier (EBSCO-host)
- ERIC (EBSCO-host)
- PsycInfo (EBSCO-host)
- Socindex (EBSCO-host)
- Teacher Reference Center (EBSCO-host)
- EconLit (EBSCO-host)
- Science Citation Index (Web of Science)
- Social Science Citation Index (Web of Science)
- Sociological Abstracts (ProQuest)

These searches were performed in 2020 and updated in 2023.

Search terms. A full description of the search strategy in each database, both for the initial 2020 search and the updated 2023 search, can be found in Supplementary Information Appendix A.

During the screening process, we became concerned that our searches did not provide full coverage of the relevant literature. In order to test this hypothesis, we performed backward citation-tracking on the most related and recent review (Yang et al. (2021)), which yielded a relatively large number of studies not included in our database searches. While we may have been able to secure adequate coverage by just combining our searches from 2020 and 2023 with citation-tracking, hand search, and grey literature searches, as originally planned, we wanted to take further steps to ensure the quality of our searches. This decision was spurred on by the rapid development in AI-tools to support and improve the screening processes of systematic reviews. When we designed our initial search, we had to strike a balance between, on the one hand, ensuring broad coverage and, on the other hand, preventing the number of search hits from becoming unmanageable, given the available manpower and financial resources. This was, and is, common practice in systematic reviews. For this reason, we imposed a "title AND abstract" requirement in the search strategy, as stated in our approved protocol.

Fortunately, we were in a situation in 2024 where recent developments in the use of LLMs and machine learning (ML) classifiers opened up an opportunity for us to manage the vast number of references resulting from the use of a broader "title OR abstract" requirement. Therefore, we conducted additional searches in the three databases generating the most relevant hits in our initial searches: ERIC, APA PsycInfo, and Academic Search Premier (all from the EBSCO Research Databases). Relevant references from the initial search were exclusively from ERIC, APA PsycINFO, Academic Search Premier, and ProQuest. Together, these bibliographical databases contribute to a widespread search of relevant literature for this review. While ERIC and APA PsycINFO are subject-specific to educational and psychological research, respectively, Academic Search Premier and ProQuest have a multidisciplinary focus. Unfortunately, at this time in

2024, we had no institutional access to ProQuest, which meant that additional searches could only be performed in ERIC, APA PsycINFO, and Academic Search Premier.

For the additional searches, we used search terms similar to the original terms, but changed the "title AND abstract" requirement to "title OR abstract". These additional searches, as well as the ones from 2020 and 2023, are all documented in Supplementary Information Appendix A. The result of this broadened search strategy was a new collection of references, including 28,315 references from ERIC, 35,317 references from APA PsycINFO, and 37,823 references from Academic Search Premier. In total, the extra database searches resulted in 101,455 references. Using both a semi-automated duplicate check with a threshold value of 0.85 and a further manual duplicate check, we found that the total number of unique and new references from these searches was 79,114. That is, more than ten times the number of hits found in the original and updated searches combined (6,572).

We describe and discuss the methods we used to handle the screening of the large number of records in detail in Supplementary Information Appendix B. Briefly, we used the title and abstract screening of 6,572 records to train a built-in ML classifier in EPPI Reviewer (Thomas et al., 2023). The classifier returned a ranking of records from least to most likely to be included. We tested a cutoff at 40% probability of inclusion, and two review authors double-screened 50 randomly sampled records with 39% probability of inclusion, none of which were deemed relevant by any of the screeners. We then used AIscreenR (Vembye & Olsen, 2025)–an R package enabling title and abstract screening with OpenAI's GPT API models–to screen the remaining 11,614 records, which had equal or greater than 40% probability of relevance according to the classifier (see Vembye et al., 2025, for extensive testing of AIscreenR's performance). We wrote inclusive prompts, which resulted in 2,239 included records. These were screened by the review

authors, resulting in 321 studies included to full text screening (which were screened using the same procedures as described in section Selection of studies).

Searching other resources

Hand search. We conducted a hand search of the following journals, to make sure that all relevant articles were found. The following five journals were selected on the basis of our initial pilot search, which indicated that they contained the greatest number of relevant articles:

- Assessment in Education: Principles, Policies and Practice
- Journal of Educational Research
- Educational Assessment
- Journal of Educational Psychology
- School Psychology Review

The hand search focused on editions published between 2018 and 2023/2024 to secure recent unpublished articles, which had not yet been indexed in the bibliographic databases.

Search for systematic reviews. We searched for other relevant systematic reviews in the following resources:

- Campbell Systematic Reviews - https://campbellcollaboration.org/
- Cochrane Library - https://www.cochranelibrary.com/
- Centre for Reviews and Dissemination Databases - https://www.crd.york.ac.uk/CRD-Web/
- EPPI-Centre Systematic Reviews - Database of Education Research – https://eppi.ioe.ac.uk/cms/Databases/tabid/185/Default.aspx

Grey literature search. We searched specifically after three types of grey literature: working papers, reports, and dissertations. It should be noted that some of the bibliographic databases also cover grey literature (e.g., ERIC). We searched the following resources for grey literature:

- ProQuest Dissertations & Theses Global (dissertations) (EBSCO-host)
- EBSCO Open Dissertations (dissertations) (EBSCO-host)
- Open Grey (reports, working papers, dissertations) - http://www.opengrey.eu/
- Google Scholar (reports, working papers, dissertations) - https://scholar.google.com/

- Google searches (reports, working papers, dissertations) - https://www.google.com/
- Social Care Online (reports, working papers, dissertations, systematic reviews) – https://www.scie-socialcareonline.org.uk/
- Social Science Research Network (working papers) - https://www.ssrn.com/index.cfm/en/
- Mathematica (reports) - https://www.mathematica.org
- MDRC (reports, working papers) - https://www.mdrc.org/
- Abt Associates (reports) - https://www.abtassociates.com/
- American Institutes for Research (reports) - https://www.air.org/
- WestEd (reports) - https://www.wested.org/
- WeStat (reports) - https://www.westat.com/
- SRI (reports) - https://www.sri.com/

Citation tracking. When relevant studies or reviews were identified, we checked the reference lists of these to check if additional relevant literature had been cited (backward citation tracking). For all relevant reviews we also performed forward citation tracking. Furthermore, we conducted forward and backward citation tracking on all included primary studies.

Contacts to international experts. We contacted three international experts to identify unpublished and ongoing studies, and provided them with the inclusion criteria for the review along with the list of included studies, asking for any other published, unpublished, or ongoing studies relevant to the review. We did not receive any reply.

Data collection and analysis

Selection of studies

The screening process of relevant studies had two stages: (1) screening on title and abstract, and (2) screening on full text. In the screening of title and abstracts identified in the initial and updated searches, we used independent double screening (Polanin et al., 2019; Stoll et al., 2019). In the extra search conducted in 2024, we used the procedure described in section Search strategy (and in Supplementary Information Appendix B). All full texts were screened using independent double screening. The screeners were blind to each other's work until comparing final

judgements. If the two screeners could not agree on the inclusion/exclusion of a specific reference, the reference was sent to one of the review authors for final judgement.

We conducted a pilot-screening for each screening stage and each screener. In the pilot screening of title and abstract, the review team screened and compared 100 references. The review team then discussed and resolved potential disagreements and uncertainties. When the interrater agreement was above 90% in the pilot, the screeners continued to screen the rest of the references. If the interrater agreement was below 90% in the first pilot, the review team members performed a second pilot screening to ensure reliability.

At the full text stage of the screening process, the pilot consisted of 8–10 studies. The pilot procedure at second level was otherwise identical to the process described for first level. The review team met with regular intervals to discuss uncertainties and minimize ‘coders' drift’ (Polanin et al., 2019). Changes to the tool were discussed during the pilot.

During the screening process, none of the review authors or review team members were blind to the authors, journals, or institutions responsible for the publication of eligible studies.

Data extraction and management

At least two members of the review team coded data from the included studies with a coding tool that was piloted and revised beforehand. From all included studies, we coded data on publication characteristics, study characteristics, participant characteristics, intervention characteristics, control/comparison characteristics, and outcome characteristics. Based on this descriptive and numerical coding, one reviewer extracted the data used in the analysis and at least one additional reviewer checked the extraction. If any disagreement or uncertainty emerged during the coding and extraction process, additional reviewers with the appropriate expertise were consulted.

All extracted data are stored electronically using EPPI Reviewer 4 and Microsoft Excel.

Assessment of risk of bias in included studies

Two members of the review team independently assessed the risk of bias for each study and the included study outcomes. The review team members discussed disagreements in their ratings, and if necessary, a third review team member were contacted for final agreement. For included non-randomised studies, we assessed the risk of bias for all included outcomes applying Cochranes ROBINS-I tool (Sterne et al., 2016). For all included randomised studies, we assessed the risk of bias of all outcome measures using a revised version of Cochrane's risk of bias tool, ROB-2 (Eldridge et al., 2016; Sterne et al., 2019). See our protocol for a detailed description of the tools and assessment domains (Thomsen et al., 2022).

Both tools have in common that an overall rating is made on the basis of the domain ratings. A further commonality is that both tools require pre-specification of the effect type that will be assessed. We were most interested in, and believed that most studies would report estimates that were closer to, the effect of starting and adhering to the intervention than the effect of assignment to the intervention.

In the case of an RCT, where there was evidence that the randomisation had gone wrong or was no longer valid, we planned to assess the risk of bias of the outcome measures using ROBINS-I instead of ROB-2. Some included studies contained both RCTs and QESs.

Definition of critical confounders. ROBINS-I dictates that reviewers should define critical confounders relevant to most or all eligible studies at the protocol stage. In the case of this review, we defined the critical confounders as performance at baseline, and gender. Other important confounders may be for example the age and grade of the students, and students' socioeconomic status. If these were unbalanced between treatment and control groups (or comparison

groups), this was reflected in a higher risk-of-bias rating. Identifying critical confounders therefore did not mean that other confounders were not considered. However, we anticipated that most studies would compare students of the same age and in the same grades, while differences in socioeconomic status often would be captured at baseline.

Performance at baseline is generally considered a strong prognostic factor in relation to post-test outcomes (Hedges & Hedberg, 2007). Furthermore, frequent administration of tests might affect students who perform well at baseline differently than students who are struggling academically in terms of their motivation for learning. If high achievers continuously experience success in testing, this will most likely foster feelings of competence, confidence, or relief. In contrast, if students struggling academically continuously experience failures when tested, it may foster feelings of incompetence, shame, and low self-esteem (Russell & McAuley, 1986; Weiner, 2010).

Gender differences exist in relation to school performance (Holmlund & Sund, 2005). Additionally, research findings indicate systematic gender differences in relation to testing anxiety. Girls are more likely to experience and report testing anxiety—a phenomenon that has been shown to affect academic achievement and life satisfaction negatively (DordiNejad et al., 2011; OECD, 2017; Segool et al., 2013; von der Embse et al., 2018). In light of evidence that gender affects school performance and that gender-related differences in responses to testing may not be captured or evident at baseline, we included gender as a critical confounder.

In each of the risk of bias assessments of outcome measures, we examined how the study authors had considered the two predefined critical confounding factors, either at the design stage or in the analysis.

Measures of treatment effect

For continuous data, we calculated the standardised mean difference (SMD) where possible, as our outcomes were measured and reported with a wide range of different scales. To correct for upward bias in small samples, we used the small sample bias-corrected Hedges' *g* in our analysis (Borenstein et al., 2009; Hedges, 1981; Lipsey & Wilson, 2001). Hedges' *g* is calculated as (Lipsey & Wilson, 2001, pp. 47‑49):

$$ES_g = \left[1 - \left(\frac{3}{4N} - 9\right)\right]\left(\frac{X_t - X_c}{S_p}\right), \quad (1)$$

with the variance given as

$$Var_g = \left(\frac{N}{n_t + n_c}\right) + \left(\frac{ES_g^2}{2N}\right), \quad (2)$$

where $N = n_t + n_c$ is the total sample size, X_t and X_c are the means in the treatment and the control group, respectively. We used the covariate-adjusted means or regression-adjusted estimates of the mean difference whenever available. S_p is the pooled standard deviation defined as:

$$S_p = \sqrt{\frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c - 2}}$$

Here, S_t and S_c denote the unadjusted standard deviations of the treatment and control groups. As recommended by Hedges et al. (2023), we included the small-sample adjustment when calculating effect sizes but not when calculating the standard errors/variances of the effect sizes.

We recorded both post- and pre-test standard deviations when available. For our main analysis, we calculated SMD's with post-test standard deviations, as these values are more likely to be reported and due to the possibility of floor effects at pre-test. Only when post-test standard deviations were not available, but pre-tests standard deviations were, we used these as replacements for possible missing values.

All outcomes were either continuous or reported as the mean proportion of correct answers in the treatment and control group. Effect sizes based on proportions might not be fully

comparable to SMDs based on continuous measures, and it is not clear what type of effect sizes are most comparable. As this issue was not mentioned in our protocol, we therefore tested four different effect size measures (reported in section Included studies). In the end, *g* turned out to work as well or better than other measures, meaning that we used *g* for both continuous outcome measures and those reported as proportions (see Figure 5).

Some studies reported only results for lower-level subgroups (e.g., based on pre-intervention achievement), which were not represented in other studies. Because effect sizes based on subgroups omit the between-subgroup variation, standard deviations will be smaller and effect sizes mechanically larger. We therefore used the methods recommended by Wilson (2017) to aggregate the subgroup results to a level comparable to our other included studies.

Our sample consisted of effect sizes from between-subject and within-subject designs. Because students are their own control group in within-subject designs, such designs keep a very large number of factors constant. Consequently, the standard deviation ought to be smaller and effect sizes mechanically larger in within-subject compared to between-subject designs. In our primary analyses, we therefore keep the two designs separate. However, to improve statistical power, we combined the designs in exploratory analyses. To make effect sizes more comparable, we followed the recommendations in Morris and DeShon (2002) and adjusted effect sizes from within-subject designs, as

$$ES_g^{adj} = ES_g\sqrt{2(1-\delta)}\,, \qquad (3)$$

where δ is the within-subject correlation on tests in different conditions. Because information about δ was lacking in most studies, we used three different values: 0.6, 0.75, and 0.9. We adjusted the variances using the formula (2) for Var_g, but substituted ES_g with ES_g^{adj}.

Units of analysis issues

Errors in statistical analysis can occur when the unit of allocation differs from the unit of analysis. In cluster-randomised trials, participants are randomly assigned to treatment and control

groups in clusters (e.g., classes or schools). QESs may also include clustered assignment of treatment. Effect sizes and standard errors from such studies may be biased if the unit of analysis is the individual and an appropriate cluster adjustment is not used (Higgins & Green, 2011). We adjusted effect sizes by study using the methods suggested by Hedges (2007), Hedges et al. (2023), and What Works Clearinghouse (2021) as well as information about intra-cluster correlation coefficient (ICC) and realised cluster sizes. As for individually assigned treatments, we included a small-sample adjustment in the calculation of effect sizes but not in the calculation of standard errors/variances (Hedges et al., 2023).

Whenever the study included information about the ICC and cluster sizes, we used this information to adjust the study. However, in many cases this information was not included and we therefore adjusted the effect sizes using estimates from the literature of the ICC in Hedges and Hedberg (2007), assuming equal cluster sizes. We used an ICC of 0.11, which approximately corresponds to the average of ICCs taken over grades, math and reading tests reported in Table 2 and 3 in Hedges and Hedberg (2007, pp. 68–69). We tested if our results were sensitive to this choice by using ICCs of 0 (the theoretical lower bound) and 0.32 (the empirical upper bound in the same two tables). To calculate an average cluster size, we divided the total sample size in a study by the number of clusters (typically the number of classrooms or schools).

Criteria for determination of independent findings

We did not include any studies implementing multiple interventions per individual or separate studies using the same samples. Many studies included multiple dependent effect sizes (e.g., because the same students were tested with several outcome tests). As described further in section Data synthesis, we used the correlated-hierarchical effects model with robust-variance estimation (CHE-RVE) developed by Pustejovsky and Tipton (2022) to account for dependent effect sizes.

Dealing with missing data

tudies had to permit calculation of a numeric effect size for the outcomes to be eligible for inclusion in the meta-analysis. Where studies had missing summary data, such as missing standard deviations or means, we derived these where possible from, for example, *F*-ratios, *t*-values, χ^2-values, and correlation coefficients using the methods suggested by Lipsey and Wilson (2001). We used the web-based tool WebPlotDigitizer (https://automeris.io) to extract data from figures in studies that did not report other effect estimates and uncertainty measures.

If these statistics were also missing, we requested information from the study authors. If we could not retrieve the necessary information, we reported the study results in as much details as possible, i.e., the study was included in the review, but we had to exclude it from the meta-analysis. Missing data and attrition rates for the individual studies were assessed with the risk of bias tools, where both ROB-2 and ROBINS-I have specific domains focusing on biases arising from missing data (Sterne et al., 2016, Sterne et al., 2019).

Assessment of heterogeneity

Heterogeneity can stem from either an expected variation in effects or from sampling errors in included studies. In this review, we assume that variation in effects will occur and we therefore used a random effects model in our main analysis (see section Data synthesis). We assessed the level of heterogeneity with the Q-statistic, the I^2 (Higgins et al., 2003), the within- (ω) and between-study (τ) heterogeneity, and total standard deviation ($\sigma = \sqrt{(\omega^2 + \tau^2)}$), as well as prediction intervals (defined below).

We report prediction intervals to show how effects are dispersed. Prediction intervals are based on the mean effect size and the standard deviation of effect sizes, instead of standard errors, which are used in the calculation of confidence intervals. We calculated prediction intervals

wherein effects will lie 95% of the time using the *predict*-function in *metafor* (see Viechtbauer, 2025), which takes both the within-study and the between-study heterogeneity into account.

Assessment of reporting bias

Reporting bias might refer to both publication bias and selective reporting of outcome data and results. We assessed bias from selective reporting of outcome data and results in both ROB-2 and ROBINS-I.

As the relative performance of methods aiming at accounting for publication bias depends on the severity of selection, we used different methods to assess the extent of publication bias. Chen and Pustejovsky (2025) and Pustejovsky, Citkowicz, et al. (2025) have shown that regression-based methods generally perform better than selection models when publication bias is weak, whereas selection models tend to outperform regression-based methods when selection is moderate to strong. We therefore aimed to employ a combination of publication bias tests that perform well under different levels of selective reporting. Moreover, we chose these models because all have been shown to perform reasonably well with heterogeneous and dependent effect sizes.

First, we constructed funnel plots and examined whether they were asymmetric (Higgins & Green, 2011). To formally test for asymmetry, we used a version of Egger's test (Egger et al., 1997) suggested by Rodgers and Pustejovsky (2020) and further developed by Chen and Pustejovsky (2025). As Rodgers and Pustejovsky (2020), we interpret the rejection of the null hypothesis of no asymmetry in a one-sided test with significance level 0.05 as an indication of asymmetry. This test is also described as the PET/PEESE test. Stanley and Doucouliagos (2014) recommend that "[w]hen the PET test is not rejected, meaning that the average effect is not sta-

tistically distinguishable from zero, then the PET intercept is used for estimating the adjusted average effect. However, if the PET test is rejected and the average effect is statistically distinct from zero, the PEESE is used for estimating the adjusted average effect" (Chen & Pustejovsky, 2025, p. 6). We followed this decision rule.

It is important to note that asymmetric funnel plots are not necessarily caused by publication bias (and publication bias does not necessarily cause asymmetry in a funnel plot). We tested how sensitive our results were to publication bias using the worst-case meta-analysis method developed by Mathur and VanderWeele (2020). As a further sensitivity analysis, we used selection models that may identify and correct for the presence of publication bias (e.g., Hedges, 1992; Hedges & Vevea, 2005, Pustejovsky, Citkowicz et al., 2025). Our protocol described a choice of selection model based on Andrews and Kasy (2019), Hedges and Vevea (2005), and Coburn & Vevea, (2019), none of which are adapted to data structures with dependent effect sizes. However, since the publication of our protocol, models that do take dependent effect sizes into account have been developed by Pustejovsky et al. (2025). As this model fits our data structure, we decided to use it alongside our pre-specified cutoffs.

Specifically, we applied two sets of selection models: a three-parameter selection model (3PSM) and a four-parameter selection model (4PSM). In the 3PSM, we used a single step of $\alpha_1 = .025$, which assumes that positive effects are statistically significant at the two-sided level of $p < .05$ have a different probability of selection than effects that are either not statistically significant or not in the anticipated direction. In contrast, the 4PSM includes two selection steps, $\alpha_1 = .025$ and $\alpha_2 = .5$, allowing for different probabilities of selection among effects that are positive but not statistically significant and those that are negative (i.e., in the opposite direction of the

intended effect; Pustejovsky, Joshi et al., 2025). Note that we calculated *p*-values using the adjusted standard errors of the effect sizes, which are not necessarily equal to the *p*-values reported by authors of the primary studies.

For all but the worst-case meta-analysis, we used a modified version of effect size variance. Specifically, we calculated $Var_g^{mod} = N/(n_t + n_c)$, as recommended by Pustejovsky and Rodgers (2019). By removing the second term of Var_g from Equation (2), we eliminated the artificial correlation between *g* and its corresponding variance, which arises because the effect size itself is used to construct the second term in Equation (2). For cluster-bias-corrected studies, we multiplied the design effect $1 - (m-1)ICC$ with Var_g^{mod}, where *m* is the average cluster size and *ICC* is the intraclass correlation.

Data synthesis

We conducted all statistical analyses in R. The data used in the analysis and the analytic code will be made at OSF upon publication.

The data synthesis was conducted in the following steps: First, we provide descriptive summaries of the contextual, methodological, and outcome characteristics for the studies included in the data synthesis. Second, we present our main effects analysis. Along with the main analysis, we present forest plots, prediction intervals, and heterogeneity statistics. Third, we conducted our proposed moderator and sensitivity analyses (described in the next sections).

The main effects analysis compared the high frequency test conditions with the low frequency test conditions of interventions. As mentioned in section Measures of treatment effect, we found both studies using between-subject and within-subject designs. Because effect sizes from the two designs are not comparable without further assumptions, we kept the two designs separate in our primary analyses.

We assumed a random-effects model. We used inverse-variance weighted mean effect sizes for all parts of the analysis and included effect sizes from both treatment-control and treatment-comparison studies (as they contrasted groups with different testing frequencies in comparable ways). To estimate the overall effect size and heterogeneity parameters, we used the RVE methods developed by Pustejovsky and Tipton (2022). Their method is implemented in three steps.

First, we identified an appropriate working model based on the features of our sample. There were two types of dependencies between effect sizes: (1) correlated effects, which arise when, for example, the same sample is tested on multiple outcomes or several treatment groups are compared to the same control group, and (2) hierarchical effects, which occur when different, independent samples are included within the same study. There were 30 studies with a correlated structure, 6 studies with a hierarchical structure, and 16 studies with both a correlated and a hierarchical structure (the remaining 7 studies reported one effect size that we could use). Thus, the CHE-RVE model suited our data structure. Furthermore, a baseline value for the correlation between pairs of effect sizes from the same study (ρ) has to be specified. We chose 0.6, as in Pustejovsky and Tipton (2022), but tested if our results were sensitive to lower (0.4) and higher (0.9) values. We chose the latter value because some of the results in Pustejovsky and Tipton (2022) were sensitive to using values of ρ higher than 0.8.

Second, based on the chosen working model, we estimated a meta-regression using the metafor package in R (Viechtbauer, 2010). We estimated the random effects variance components, including heterogeneity statistics, inverse-variance weight matrices, and the meta-regression coefficients using the restricted maximum likelihood (REML) procedure.

Third, we calculated confidence intervals based on the RVE standard errors. These standard errors are adjusted for small-sample bias as suggested by Tipton (2015) and Tipton and Pustejovsky (2015). We report 95% confidence intervals for all analyses.

Our primary outcome variable is effect sizes based on measures of academic achievement and our secondary outcome variable is effect sizes based on measures of testing anxiety. The latter could, as mentioned, not be meta-analysed. In the analysis of academic achievement, we included all types of tests and subjects in the main effects analysis. Corresponding to our first two research questions, the meta-regression model included indicators for how the testing frequency of the treatment group differed from the baseline provided by the control/comparison group. We included indicators for 0, 1, 2, 3, and 4 or more tests in the control group (the treatment group was always tested more frequently).

Due to relatively few studies in the latter four categories, the statistical power and degrees of freedom were limited. As pre-specified in our protocol, we coarsened the frequency categories when the adjusted Sattertwaithe degrees of freedom were below 4 for any coefficient. We chose this threshold because Tipton (2015) suggested that RVE standard errors are unreliable when the degrees of freedom are below 4. This resulted in a specification with two indicators, one for control testing frequencies of zero, and one with one or more tests in the control group. This specification was the starting point for the subgroup analysis and investigation of heterogeneity, which we describe next.

Subgroup analysis and investigation of heterogeneity

To answer our third research question, we conducted moderator analyses to identify the characteristics that are possibly associated with smaller and larger effects on the primary out-

comes. For the moderator analysis, we used a similar meta-regression method as for the main effects analysis. We included all pre-specified moderators in the meta-regressions to reduce the risk of misleading results due to correlated independent variables (i.e., that an included moderator captures the association of an omitted moderator). We again conducted separate analyses for between and within-subject designs. Our protocol pre-specified the following types of moderators:

1. Test subject (reading/math/etc.)
2. Grade level (kindergarten/1st-3rd grade/4th-6th grade/etc.)
3. Type of test (formative/summative and high/low stakes)
4. Duration of intervention (measured in weeks)
5. Gender

Because our data showed limited variation for some moderators, and because the number of studies was too small to estimate multiple variables with sufficient adjusted degrees of freedom (i.e., above 4), we were unable to conduct the pre-specified analyses exactly as planned. We defined the moderators as follows:

Test subject. We included an indicator equal to one for language arts tests, which was by far the largest subject category in our sample, and zero for all other subjects.

Grade level. We used the US education system that has kindergarten as the first year of primary school as the reference point. US kindergarten was coded as grade 0 and other countries were recoded so that the first year of primary school was coded as 0 as well. That is, our grade level variable measured the number of years in primary school at the time the intervention started. If information about grade was missing, we used the reported mean age and information about the country's school system to impute the mean grade level. We included grade level as a mean-centered continuous variable.

Type of test. We intended to contrast practice tests with a formative or summative purpose, and whether the tests were high or low stakes for the students. However, nearly all included

tests were formative and low stakes, making it impossible for us to conduct this planned analysis. This moderator could therefore not be included in the analyses.

Duration of intervention. We included a mean-centered variable measuring the duration of the intervention (in weeks). The starting point was the first practice test and the end point the last practice test.

Gender. We included a mean-centered variable measuring the proportion of girls in the treatment group (or in the full sample, if information for the treatment group was not reported).

Of the pre-specified moderators, only Gender had missing observations (77 effect sizes). We imputed the mean in these cases (separately in the between and within-subject sample).

In exploratory (i.e., non-pre-specified) analyses, we examined moderators including time between end of intervention and the outcome test, test identicality, whether the outcome test assessed skill or rote memorization, and the type of control condition. The latter were divided into different categories: 1) at least one practice test for the control group, 2) treatment as usual or a filler task, 3) restudy or review, 4) worked examples, and 5) other idiosyncratic, active conditions (e.g., summary or rainbow writing, concept mapping, discussions). We also examined whether the observed heterogeneity was reduced when we excluded studies with small effective sample sizes. We again imputed the mean for moderators with missing observations (see Table 1 for information about missing observations).

Sensitivity analysis

We conducted sensitivity analyses to examine the robustness of our main effects analysis. As mentioned, we examined if our results were sensitive to the choice of ρ and ICC. We also performed an examination of the distribution of effect sizes, study design, and risk of bias assessments – all analyses are described further below. In our protocol, we specified a potential sensi-

tivity analysis using the pre-test standard deviations to calculate the SMDs instead of post-test standard deviations. However, there were very few studies that reported pre-test standard deviations (missing in 145 between-subject designs, and in all within-subject designs). We therefore refrained from conducting this analysis.

Distribution of effect sizes. We examined the distributions of effect sizes for the presence of outliers and examined the sensitivity of the results by Winsorising the outliers to the nearest non-outlier value (Lipsey & Wilson, 2001). We considered effect sizes that were larger or smaller than three times the interquartile range to be outliers in this analysis (Pustejovsky, Zhang et al. 2025).

Study design and risk of bias assessments. To assess methodological quality, we conducted sensitivity analyses by restricting the analysis to RCTs only, and by restricting the sample – within each risk-of-bias domain – to outcomes rated better than high/serious risk of bias.

Differences between protocol and review

We implemented a much more comprehensive search strategy than the one specified in our protocol (Thomsen et al., 2022. See Supplementary Information Appendix B for more information.

The protocol stated that we would not include studies in which the practice tests used during the intervention were identical to the outcome test. The motivation for this exclusion criterion was that identical tests could confer an unfair advantage on the treatment group because they had more exposure to the learning material. In line with this motivation, we chose to include interventions and effect sizes based on identical practice and outcome tests if this feature did not confer an unfair disadvantage for either the treatment or control group.

We intended to contrast practice tests with formative and summative purposes, and whether the tests were high or low stakes for the students. However, nearly all included tests were formative and low stakes, making it impossible for us to conduct this planned analysis. Similarly, we could not conduct a meta-analysis of our secondary outcome, testing anxiety, due to only including one study reporting this outcome.

We used the *predict*-function in *metafor* to calculate prediction intervals instead of the formulas mentioned in our protocol. The reason was that the *predict*-function incorporates within-study heterogeneity and not only between-study heterogeneity, which was more suited to the rest of our analytical framework and our data.

We chose to use a different set of selection models to examine publication bias than those specified in our protocol. These new models, which were not available at the time we wrote the protocol, were better suited to our data structure with dependent effect sizes.

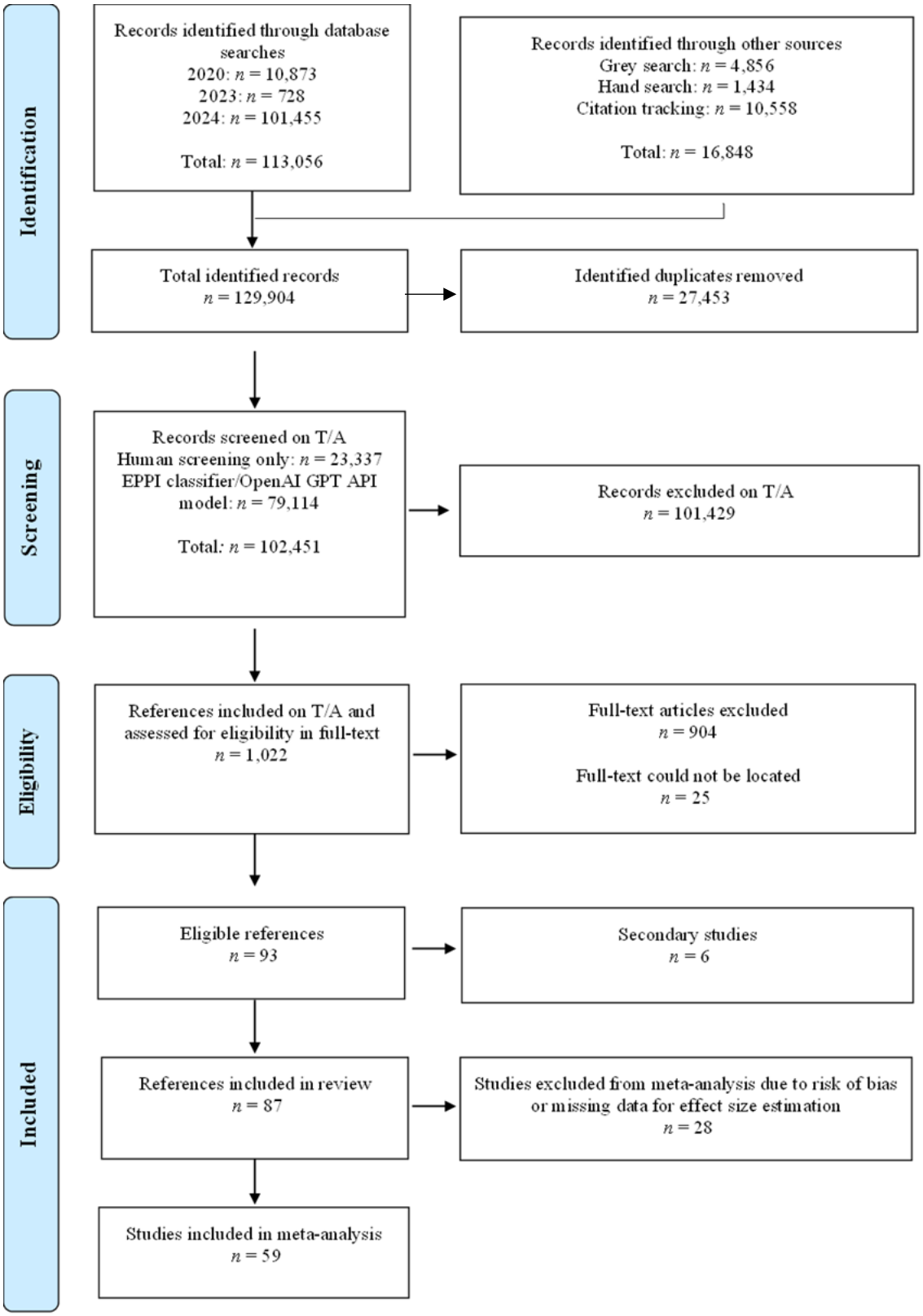
Results

Description of studies

Results of the search

Figure 1 reports the results of the search and screening process using a PRISMA-type flow chart (developed by Page et al., 2021). The search included in total 129,904 records (electronic databases: 113,056, other sources: 16,848). After excluding duplicates, 102,451 potentially eligible records remained.

FIGURE 1. *Flow chart of the search and screening process*



We excluded 67,500 records using a built-in classifier in EPPI trained on all studies from the initial (2020) and updated (2023) searches (which were all screened by two review team members). We excluded an additional 9,375 records using AIscreenR (Vembye & Olsen, 2025). The remaining 2,239 records were screened by two review team members. In total, combining the original and updated electronic database search, the search of other sources, and the extra broadened search, we screened 23,337 records using human screeners, and 79,114 records were excluded by ML algorithms or AIscreenR (see Supplementary Information Appendix B for a more detailed description of the broadened search and screening process). An extensive evaluation of all screening performance (i.e., human and AI recall and specificity measures) of this review can be found in Vembye et al. (2025).

There were 1,022 potentially eligible records after title and abstract screening. We could not locate 25 of these, and included 93 records after full text screening. Six of these records were secondary studies, that is, they examined the same intervention as another record, which we used for coding and risk of bias assessment. There were thus 87 studies included in the review.

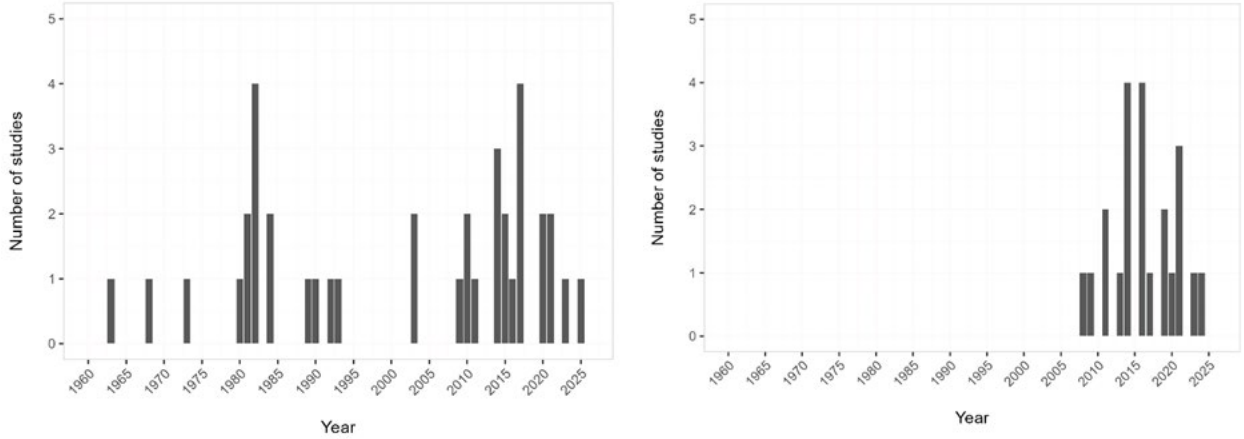
Included studies

The search resulted in a final selection of 87 studies, which met the inclusion criteria for this review. The full coding of these studies and the risk of bias assessments are available in Supplementary Information Appendix D and E, respectively. We excluded 28 studies from the meta-analysis. Twenty-two were excluded from the meta-analysis because they received a critical risk of bias rating. We excluded five studies because they lacked the necessary information to calculate an effect size, and we were not able to retrieve the information from the authors. One study (Lavy, 2024) investigated a type of testing frequency intervention–how tight exam schedules affect student achievement (as measured by the number of final exam tests taken during one day or

during a week)–that did not include practice tests in the same way as other studies in our sample did. It met our inclusion criteria, but we deemed it not comparable to the other included studies. We describe this study narratively along with the results from the 5 studies lacking information in the section Results of studies excluded from the meta-analysis due to lack of information or comparability.

The rest of this section describes the 59 studies that were included in the meta-analysis. Most of these were studies from the United States (US, 28 studies). Studies from the following countries were also included: Australia (3), Brazil (3), China (1), Germany (1), Saudi Arabia (1), Spain (1), Sweden (6), Taiwan (2), UK (1), the Netherlands (11), and the Philippines (1). Figure 2 shows the number of studies included in the meta-analysis by year of publication. As in the main effects and most of the moderator analyses, we present the data by study design (within- and between-subject). The 38 between-subject designs are spread out over a broad range of years. The oldest study was published in 1962 and the most present in 2025. The 22 within-subject designs are newer in comparison, the oldest study being from 2007 (note that one study contained both a between- and a within-subject design, which is why they do not sum to 59).

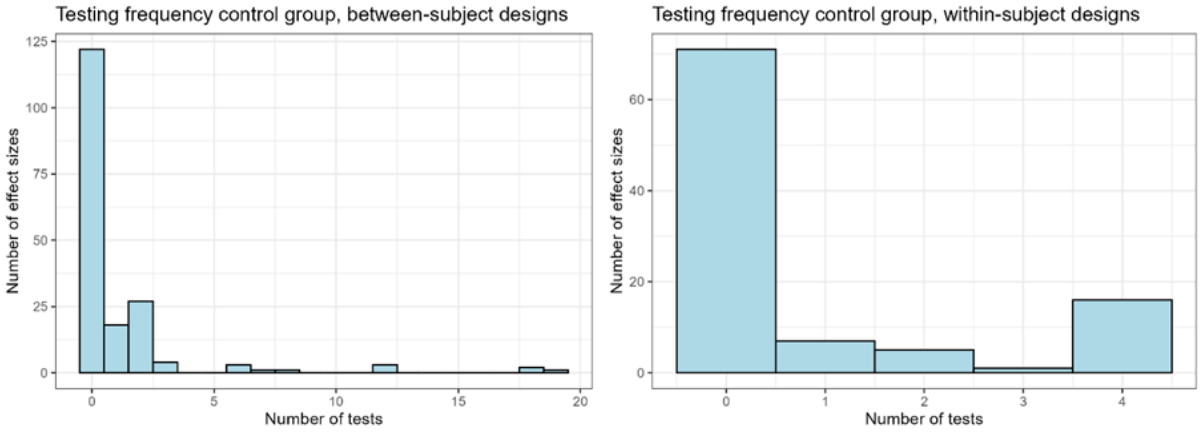
FIGURE 2. *Number of studies included in the meta-analysis by year of publication*



Note: The left panel shows between-subject designs and the right panel shows within-subject designs. The frequency on the y-axis counts studies.

Figure 3 displays the testing frequency in the control group, again separated by between- (left panel) and within-subject designs (right panel). As the frequency sometimes differs within studies, the figure counts the number of effect sizes (y-axis) in each frequency category (x-axis). For both study designs, the large majority of effect sizes compare a treatment group that is tested with a control group that receives no practice tests. In between-subject designs, the range of control group frequencies is relatively broad but there are few effect sizes with more than two tests in the control group. In within-subject designs, no effect size had a control group test-frequency higher than four, and there are relatively few in the categories one, two, and three tests.

FIGURE 3. *Testing frequency in the control group*



Note: The left panel shows between-subject designs and right panel shows within-subject designs. The frequency on the y-axis counts effect sizes.

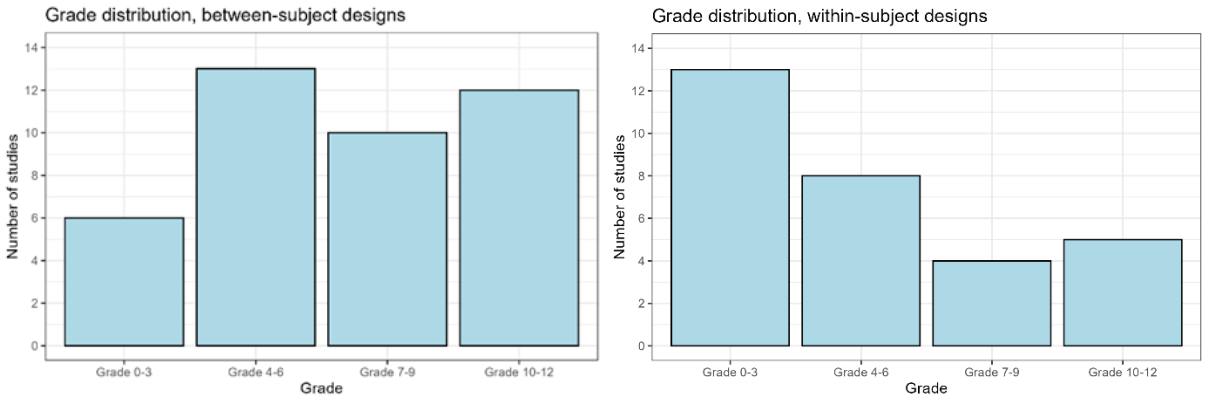
Table 1 contains additional descriptive statistics of the variables used in the analysis. Panel A describes between-subject designs (38 studies and 182 effect sizes) and Panel B within-subject designs (22 studies and 100 effect sizes). As some statistics differ within studies, we report statistics on the effect size level. That is, if the "Effect sizes" column reports less than 182 and 100, respectively, then information is missing from some studies. As mentioned, we could

not meta-analyse our secondary outcome, meaning that all effect sizes concern our primary outcome, academic achievement.

For both between and within-subject designs, the sample sizes were often small. The mean effective sample size was 56.7 (between) and 107.9 (within) and the minimum was as small as 10.0 (between) and 14.7 (within). Many between-subject designs therefore seem to be under-powered to find anything but quite large effects. Within-subject designs ought to have better power, both because they are on average larger and because the design keeps factors related to student characteristics constant between treatment and control conditions.

Both the mean age and grade indicated that studies covered a large share of primary and secondary school. Mean age is missing in some studies, while we have complete information about grades and we therefore focus on this variable henceforth. The grade level ranges from 3-12 (between) and 1-12 (within), with a mean of around seven in both designs. Figure 4 shows how studies were distributed across grades. From here, it can be seen that while means appeared to be rather identical across designs, Figure 4 shows that comparatively more within-subject designs examined testing frequency in earlier grades. The share of girls also varies considerably, with a mean slightly below 50% in both designs (but note that there is quite a lot of missing data).

FIGURE 4. *Grade distribution*



Note: The left panel shows between-subject designs and right panel shows within-subject designs. The frequency on the y-axis counts effect sizes.

Among the between-subject designs, most effect sizes come from RCTs (93%), whereas slightly less than half come from RCTs among the within-subject designs (45%). For the between-subject designs, we have complete information regarding the testing frequency of both the control and treatment group, whereas the treatment group testing frequency is missing in two cases for the within-subject designs. The treatment group testing frequency is wide-ranging in both designs. All effect sizes derive from low-stakes, formative practice tests in the within-subject designs, and nearly all in the between-subject designs. The practice tests used in the interventions were often short and relatively simple quizzes conducted during one or just a few sessions. These features reflect the literature's focus on the testing effect and retrieval practice-type of treatment conditions. Our sample is therefore not well suited to examine the consequences of frequent testing with high-stakes or summative tests.

The sample is more balanced in terms of whether the students received feedback on their practice tests, with 50% (between) and 61% (within) of effect sizes containing feedback. Note though, that we do not have complete information regarding this variable and that the indicator is equal to one in interventions where feedback was explicitly mentioned. It is possible that other

interventions also included feedback, but because it was not an explicit part of the intervention, it may not have been reported.

The outcome tests also differed. The most common subject was language arts for both designs. Other subjects represented were biology, drivers' license education, geography, history, math, physics, psychology, and science. Most tests were developed for the intervention in question, and there were very few standardized tests. Many outcomes were, as mentioned, reported as proportions, especially in the within-subject designs. We examine these effect sizes further below. A large majority of outcomes were measured at or close to the end of intervention. Mean follow-up length was less than a week after the last practice test in the between-subject designs, and 1.5 weeks in the within-subject designs. Relatively few outcome tests were identical in the between-subject designs and around half tested some form of skill (often of the problem-solving type) instead of the remembrance of some form of item, for example, history facts or word definitions. In the within-subject designs, a majority of outcome tests were identical to the practice tests. Few were tests of skills; instead, they tested whether students remembered simple items like word definitions or science facts.

In both designs, there were a relatively wide range of control conditions. We categorised these as treatment as usual (TAU)/filler tasks, conditions in which the control group was tested at least once, restudy and review conditions, worked examples, and other active control conditions (e.g., rainbow writing, summary writing, concept mapping, home work time and feedback, discussion). The first three conditions were most common in both designs.

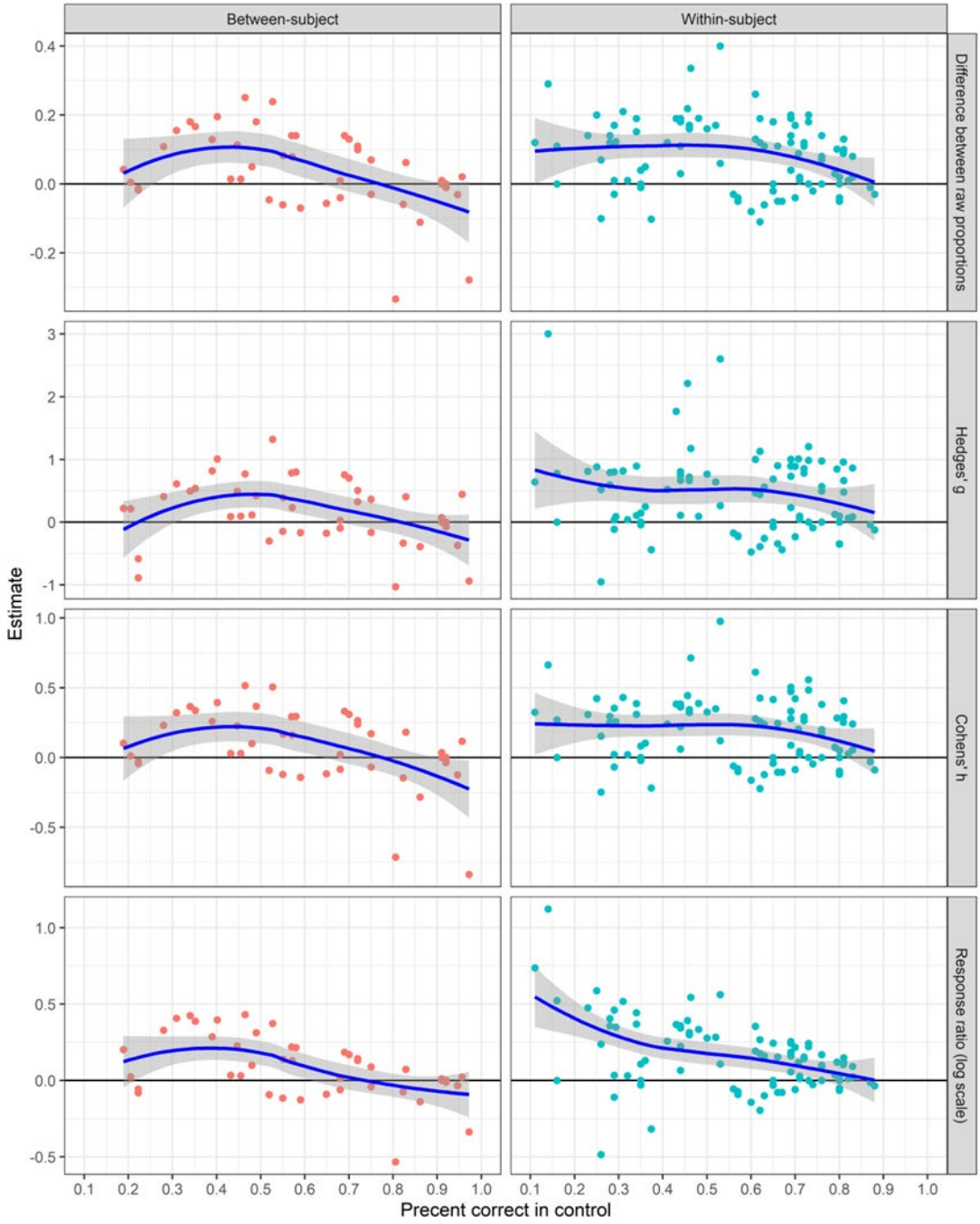
TABLE 1. *Descriptive statistics of effect sizes included in the meta-analysis*

Variable	Effect sizes	Min	Max	Median	Mean	SD
Panel A: Between-subject designs						
Participants						
Sample size	182	10.00	1951.00	44.00	84.50	192.03
Effective sample size	182	10.00	182.00	44.00	56.46	34.91
Mean age	95	8.50	18.00	11.14	12.06	2.77
Mean grade	182	3.00	12.00	7.00	7.47	3.00
Percent girls	105	0.00	90.00	52.00	46.04	17.46
Study and intervention design						
RCT	182	0.00	1.00	1.00	0.93	0.26
Number of tests in control group	182	0.00	19.00	0.00	1.14	2.96
Number of tests in treatment group	182	1.00	90.00	1.00	10.10	16.69
Formative	182	0.00	1.00	1.00	0.99	0.11
Low-stakes	182	1.00	1.00	1.00	1.00	0.00
Feedback	162	0.00	1.00	0.50	0.50	0.50
Intervention duration (weeks)	182	0.00	30.50	0.14	4.19	5.99
Properties of outcome tests						
Language arts	182	0.00	1.00	0.00	0.41	0.49
Outcome reported as proportion	182	0.00	1.00	0.00	0.24	0.43
Follow-up length (weeks)	169	0.00	9.00	0.43	0.79	1.06
Identical outcome and practice tests	182	0.00	1.00	0.00	0.04	0.21
Skill outcome test	182	0.00	1.00	1.00	0.53	0.50
Control conditions						
TAU/Filler task	182	0.00	1.00	0.00	0.21	0.41
Control group tested	182	0.00	1.00	0.00	0.33	0.47
Restudy	182	0.00	1.00	0.00	0.26	0.44
Worked examples	182	0.00	1.00	0.00	0.18	0.38
Other active control	182	0.00	1.00	0.00	0.03	0.16
Panel B: Within-subject designs						
Participants						
Sample size	100	20.00	648.00	108.00	109.30	86.09
Effective sample size	100	14.69	648.00	108.00	107.87	86.82
Mean age	89	7.93	18.03	12.08	12.34	3.04
Mean grade	100	1.00	12.00	6.00	6.72	3.16
Percent girls	86	0.00	77.00	48.90	47.74	13.28
Study and intervention design						
RCT	100	0.00	1.00	0.00	0.45	0.50
Number of tests in control group	100	0.00	4.00	0.00	0.84	1.50
Number of tests in treatment group	98	1.00	12.00	2.00	3.80	3.93
Formative	100	1.00	1.00	1.00	1.00	0.00
Low-stakes	100	1.00	1.00	1.00	1.00	0.00
Feedback	100	0.00	1.00	1.00	0.52	0.50
Intervention duration (weeks)	100	0.14	7.80	0.29	1.19	1.67
Properties of outcome tests						
Language arts	100	0.00	1.00	1.00	0.54	0.50
Outcome reported as proportion	100	0.00	1.00	1.00	0.89	0.31

Variable	Effect sizes	Min	Max	Median	Mean	SD
Follow-up length (weeks)	100	0.00	17.40	0.57	1.59	3.15
Identical outcome and practice tests	100	0.00	1.00	1.00	0.64	0.48
Skill outcome test	100	0.00	1.00	0.00	0.11	0.31
Control conditions						
TAU/Filler task	100	0.00	1.00	0.00	0.08	0.27
Control group tested	100	0.00	1.00	0.00	0.29	0.46
Restudy	100	0.00	1.00	0.50	0.50	0.50
Worked examples	100	0.00	1.00	0.00	0.06	0.24
Other active control	100	0.00	1.00	0.00	0.07	0.26

As mentioned in section Measures of treatment effect, effect sizes based on proportions are not necessarily comparable to those based on continuous outcome measures. In particular, there is a risk that effect sizes differ at the ends of the spectrum. Furthermore, it is unclear which type of effect size should be used. In Figure 5, we examine the distributions of effect sizes based on proportions (y-axis) by plotting them against the mean proportion in the control group (x-axis). We examine four effect size measures: the treatment–control difference in mean proportions; Hedges' *g*, Cohen's *h*, and the log of the response ratio (see figure note for details on how they are calculated). The patterns are generally similar regardless of the measure, but vary slightly across research designs.

FIGURE 5. *Effect sizes based on proportions and their relation to percent correct in the control group*



Note: The panels show the relation between the mean proportion of correct answers in the control group (x-axis) and the effect size (y-axis) for four alternative effect size measures. From the top: the treatment – control difference in mean proportions; Hedges' *g*, calculated as in section *Measures of treatment effect*; Cohen's *h*, calculated as $2\text{asin}(X_t^{0.5}) - 2\text{asin}(X_c^{0.5})$; and the log of the response ratio, calculated as $\log(X_t/X_c)$, where X_t and X_c are the post-intervention means in the treatment and control group.

For the within-subject designs, all measures but the response ratio exhibit an almost straight line across all measures of the percent correct answers in the control group. Yet, we see a minor decline when the percent correct answers in the control approaches 90% correct, but this decline is driven by a few effect sizes only. Therefore, it is hard to infer from the plot that the within-subject design is subject to ceiling effects.

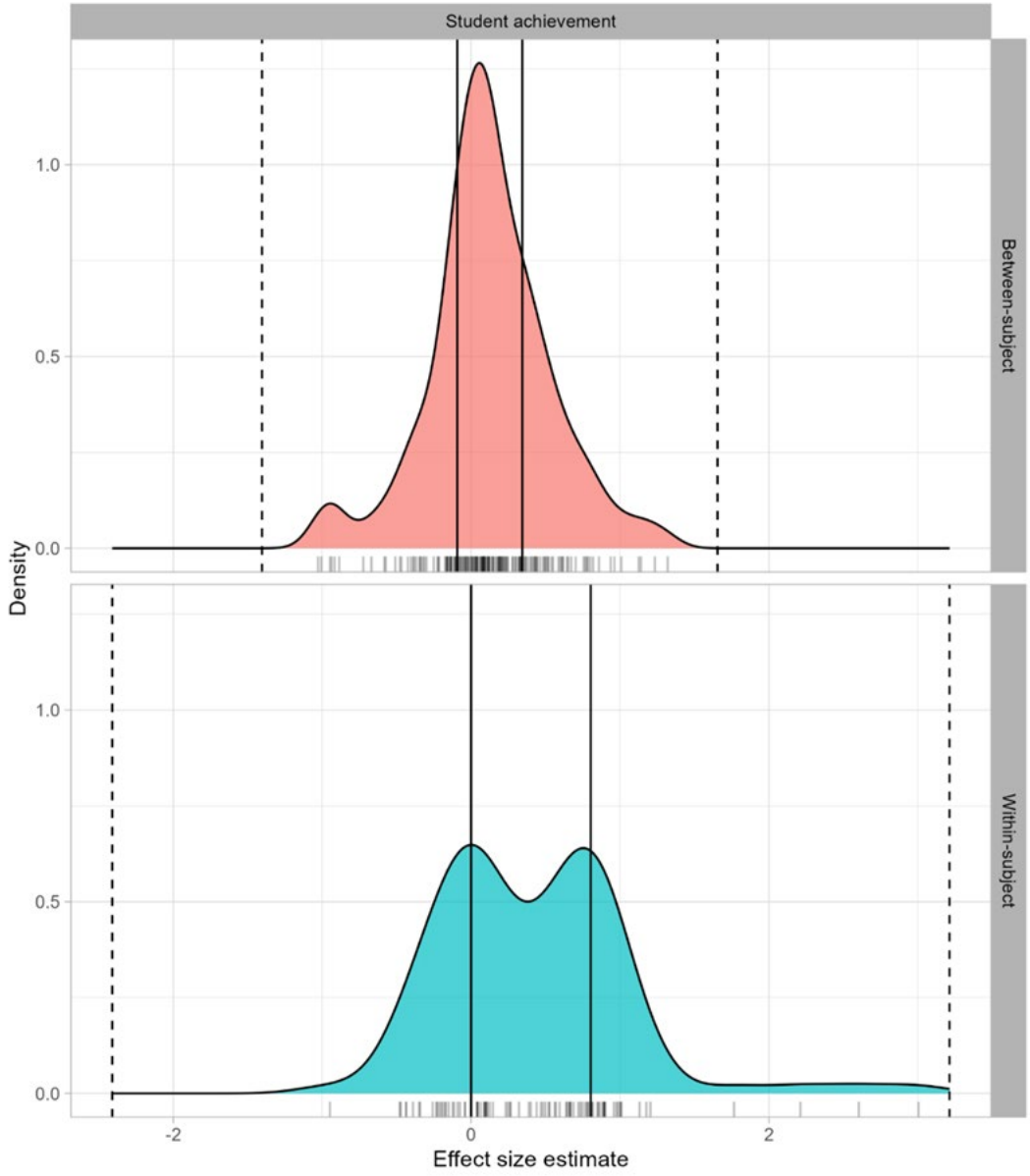
By contrast, we observe an inverted U-shaped pattern across all measures for the between-subject design effects, suggesting that these effects may be influenced by both floor and ceiling effects. On the one hand, the ceiling effect appears more pronounced in between-subject designs and emerges earlier (around 80% correct in the control group). Although this pattern is based on relatively few effect sizes, we cannot clearly rule out the presence of a ceiling effect. For this reason, we conducted exploratory sensitivity analyses to examine its potential impact on the main results.

On the other hand, we are cautious about drawing any conclusion regarding a floor effect, as this pattern is driven by only four effect sizes. What is clear, however, is that further research is warranted, and future meta-analyses using proportions of correct answers should be attentive to issues related to floor and ceiling effects.

To conclude, as the pattern does not depend on the chosen measure, we use g for effect sizes based on proportions as well.

In Figure 6, we show density plots of the full set of effect sizes in the between-subject (orange) and within-subject designs (green). There are more large and positive effect sizes (and some potential outliers) among the within-subject designs than among the between-subject designs, providing some justification for our decision to analyse the two designs separately.

FIGURE 6. *Effect size density plot*



Note: Empirical distribution of effect size estimates. Solid vertical lines indicate lower and upper quartiles. Dashed lines indicate the 1st quartile minus 3 times the inter-quartile range and the 3rd quartile plus 3 times the interquartile range. Effect sizes outside of the range of dashed lines would be considered outliers (Pustejovsky, Zhang, et al., 2025)

Excluded studies

Due to the large number of studies screened on full text, we were unable to describe all excluded studies. In this section, we exemplify how we applied the inclusion criteria by describing a selection of studies that met many, but not all, of our inclusion criteria and that readers may expect to be included.

The most common reason for excluding studies that at first glance looked relevant was lack of information about the testing frequency employed in the control group. If we could be certain that control students received no tests, we included the study and used it in contrasts of zero tests vs. one or more tests.

However, the control group testing frequency was often unlikely to be zero. For long intervention durations, e.g., an entire school year or a whole semester, it would not be plausible to assume that control students received no tests at all (given that testing/assessment in some shape or form during the school year is ubiquitous in most school systems). For example, we excluded studies by Supovitz (2016), where the intervention period was two years, and Yin et al. (2008), where control students were exposed to "regular teaching practices" (likely entailing some form of testing/monitoring) for a period of approximately six months. In other cases, studies only counted certain types of tests (e.g. standardized tests or central exit exams), with no information about other types of testing occurring during the intervention period (Im et al., 2020). Similarly, Wößmann (2003), used TIMMS data to compare countries with central exit exams to countries with no central exit exams, and Jürges et al. (2012) compared federal states with central exit exams to states without central exits exams. In both cases, we judged it as unlikely that no testing occurred in control countries/states, since school systems without central exit exams usually apply other types of testing/exam procedures to assess students.

Another common reason to exclude studies was the existence of co-interventions. When considering the issue of co-interventions, we were interested in whether there was a difference in the exposure to a given co-intervention between the treatment and the control group. If both groups received the exact same co-intervention, meaning that testing frequency was the only factor varying between groups, we would be able to isolate our treatment effect of interest. On the contrary, differences in the use of co-interventions between groups would inhibit us from establishing an unbiased treatment effect, as was the case with Fuchs et al. (1991) where the co-intervention favoured the treatment group. Another example was Lang et al. (2014), where the intervention was a web-based electronic performance support system, including both formative assessment tasks, rubrics, professional development modules, and other resources for mathematics teachers.

In a number of cases, studies were excluded because they turned out to revolve around other types of interventions that did not entail a variation in testing frequency between two or more groups. As an example, Ackermans et al. (2019) examined two forms of rubrics, Alitto (2008) investigated peer-mediated goal setting and feedback, and Beyers et al. (2013) focused on exploring the technical adequacy of a form of curriculum-based measurement (CBM) and the effects of a reading comprehension intervention. Other studies did look at the effects of tests or assessments of students, but without incorporating differences in the frequency of testing, as in the case of Gustafson et al. (2019) and Mintert (2019). Finally, some studies were excluded due to the tasks given to students falling outside of our definition of a test; as an example, two studies by Baars and co-authors (2014, 2018) gave tasks to students with the purpose of exploring students' judgments of learning (JOLs) and their self-explanations, without these tasks being presented as tests.

As specified in the protocol, studies employing only one unit of allocation in the treatment or control group (or both) were excluded since it would not be possible to separate potential treatment effects from unit effects. Examples of such studies were Naseem (2021) and Abu-Hamour & Mattar (2013), in which one class unit was compared to another class unit, as well as Hamilton (2013) and Karuza (2014), where there was only one treatment school.

An exclusion criterion that proved difficult to assess was the use of outcome tests identical to those applied during the intervention (i.e., practice tests). As previously mentioned, we judged this criterion by considering both whether practice and outcome tests were completely identical and whether the presence of identical tests constituted an unfair advantage for any group. As an example, we excluded the study by Jaeger et al. (2014) as the treatment group got an unfair advantage by being exposed to the same target facts twice. We chose to include a number of studies where control students had similar exposure to the facts tested through restudy/review since neither group had an advantage over the other. However, there were studies in which control students had restudy options, but where we still judged that one group had an unfair advantage over the other. An example was Mitchell (2009), where the practice and outcome tests contained the same items in a different order and where the control condition consisted of restudy. We decided to exclude this study because the learning material was a text, and the questions seen by the treatment group during practice pointed out what was important to take away from the text, giving them an unfair advantage over the control group.

We defined eligible study designs as those applying a treatment-control group design or a comparison design. We therefore excluded studies without a well-defined control or comparison group, an example being the study of January et al. (2018) where there was no control group. All students tried all three monitoring schedules one after another with no counterbalancing.

As noted in the protocol, the primary outcome of interest for this review was academic achievement. We therefore excluded studies in which the outcome of interest was not academic achievement, such as Capizzi & Fuchs (2005), where the outcome was teacher planning, as well as Rohrer et al. (2010) and Tempel & Sollich (2023) where the outcomes were pure memory tests. We also excluded studies where test responses were not recorded (on paper, computer, tablet, or otherwise), examples being the studies of Cavanaugh et al. (1996) and Reimer (2019).

Finally, we excluded interventions conducted outside of a school setting. Cohonner & Mayer (2018) conducted their intervention in an extra-curricular learning environment, the Experimental Biology Lab FLOX, and was excluded.

Risk of bias in included studies

We excluded 22 studies from the synthesis because all outcomes received a critical rating in at least one domain of ROBINS-I. Most of these studies were originally QES. We deemed the randomisation not valid in two RCTs, which were subsequently assessed with ROBINS-I and received critical risk of bias. Almost all outcomes received critical risk of bias in the confound-ing bias domain. The two exceptions received critical in the selection bias domain or the meas-urement bias domain. Common reasons for the critical rating in the confounding bias domain were that there was little or no adjustment for confounders; that students or teachers could self-select into the treatment; or that the assignment to treatment was not clearly described; and large imbalances at baseline. It is important to note that the ratings are assessments of the risk of bias, not actual bias, and that the ratings are not an assessment of a study's value. Studies can be very valuable, despite having too high risk of bias for our purposes.

In Figure 7 (RCT) and 8 (QES), we visualize the risk of bias of the 228 outcomes included in the meta-analysis. All plots represent unweighted risk of bias plots. Each figure has two panels, one for between-subject designs (the top panel in both figures) and one for within-subject designs. There were 169 effect sizes (93%) nested in 36 studies in the between-subject design sample, and 45 effect sizes (45%) nested in 12 studies in the within-subject design sample that were derived from an RCT.

Figure 7 displays the risk of bias ratings in RCTs. No outcome received low risk of bias in all domains. The main reason was the universal lack of pre-specified analysis plans, which is a necessary requirement for receiving low risk of bias in the selection of reported results domain. We did not systematically search for analysis plans, that is, if a plan was not mentioned in the study, we took this as an indication that there was none. We rated all outcomes as having some concerns in this domain, which meant we found no indications in the studies of selective reporting (e.g., a study mentioning an outcome in the methods section but not reporting it), although this was of course difficult to assess given the lack of pre-specified analysis plans.

In the between-subject designs, there were also relatively few studies that received a low risk of bias rating in any domain. In particular, almost all outcomes received some concerns in the measurement of the outcome domain, and some concerns or high risk of bias in the randomisation process domain. The deviations from intended interventions and the missing outcome data domains have slightly better ratings, but risk of bias was still relatively high. In comparison, the within-subject design RCTs received better ratings in all these domains.

Risk in the measurement domain arose mainly either from a lack of information about whether assessors were blind to treatment status, or from clear evidence that the assessors knew the treatment status, combined with the possibility that such knowledge could influence the out-

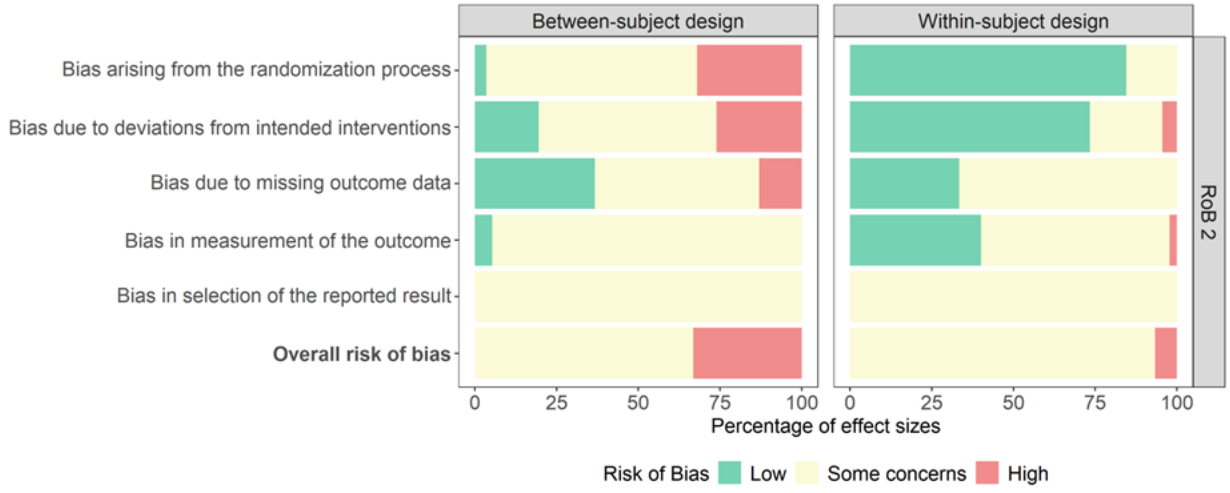
come assessment. Most studies further lacked a description of the method of random sequence generation. However, because there is no tradition of reporting this method in educational interventions, we opted for trusting the authors when they wrote that the study was randomized (if there were signs to the contrary, we used ROBINS-I). To receive a rating of some concerns, there also had to be indications of baseline imbalances, implying a risk that the randomisation had not produced balanced treatment and control groups. The risk of such imbalances is much larger in between-subject than within-subject designs.

The risk of bias in the deviations from intended interventions and the missing outcome data domain were relatively high in both designs. The reasons for ratings higher than low in the former category were mainly connected to problems with implementation fidelity or lack of information about fidelity, in combination with participants (teachers, students, caregivers) being aware of treatment status. In the latter domain, lack of information or substantial and differential attrition, especially when study authors reported no formal tests of differential attrition, yielded some concerns or high risk of bias ratings.

Figure 8 displays the risk of bias judgements in QESs. With the exception of the selection of the reported results domain, where all studies received a “moderate” rating for not having a pre-specified analysis plan, there is again an advantage for the within-subject designs. Nearly all outcomes in between-subject designs received serious risk of bias in the confounding bias domain, whereas relatively few of the within-subject designs did. This difference reflected the inherently easier task of balancing the treatment and control group on student characteristics in within-subject designs. Confounding is still a risk in these designs, for example, if the assigned facts to be learned in the intervention are not properly counterbalanced between conditions. Risk

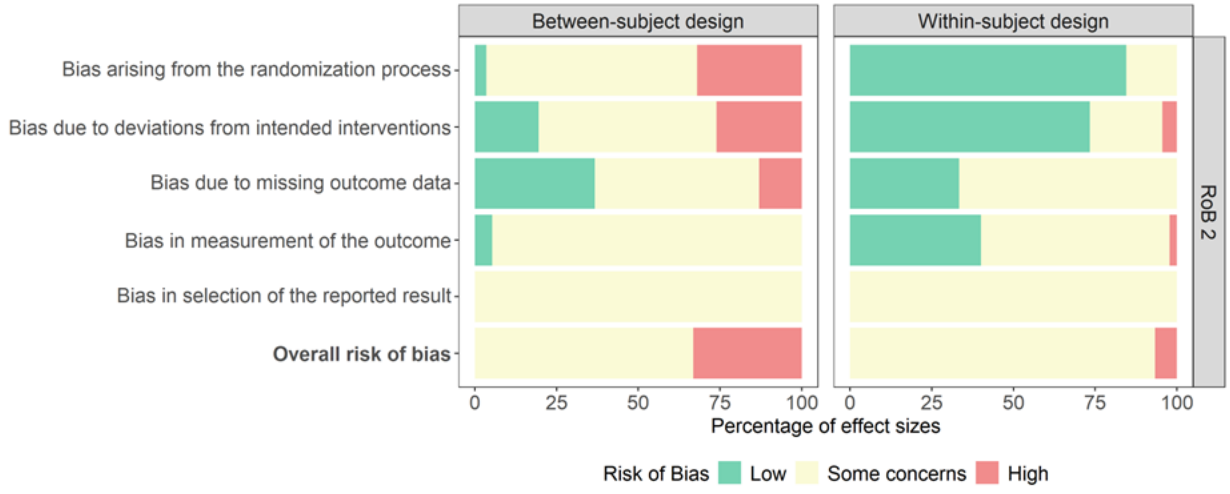
of bias due to selection of participants, in the classification of interventions, and due to deviations from intended interventions were rare, whereas most outcomes had a moderate risk of bias due to missing data and in the measurement of outcomes (for similar reasons as the RCTs).

FIGURE 7. *Risk of bias of outcomes included in the meta-analysis from RCTs*



Note: The plot is based on 169 effect sizes from 36 between-subject design studies and 45 effect sizes from 12 within-subject design studies.

FIGURE 8. *Risk of bias of outcomes included in the meta-analysis from QES*



Note: The plot is based on 13 effect sizes from three between-subject design studies and 55 effect sizes from 11 within-subject design studies.

Synthesis of results

Main effects

Table 2 displays the results of the analysis of main effects. Panel A contains the estimates from the between-subject designs (182 effect sizes nested in 38 studies) and Panel B contains the estimates from within-subject designs (100 effect sizes nested in 22 studies). We report the coefficient estimates, small-sample adjusted degrees of freedom, the 95% confidence interval, and the within-study, between-study, and total heterogeneity (in standard deviation units) from two specifications for each sample. The first specification contains an intercept representing the weighted average effect of interventions where the control group received zero tests, and indicators for control group testing frequencies of one, two, and three or more tests. As we have few observations for at least one indicator in both samples and correspondingly low degrees of freedom, we report a second specification containing the same intercept and an indicator equal to one if the control group was tested.

We found a positive and statistically significant intercept in both samples and specifications (henceforth, significant always refers to statistical significance). The estimate is more than twice as large in the within-subject design sample compared to the between-subject design sample (0.46-0.47 compared to 0.21-0.22). None of the indicators for control group testing frequencies of one, two, and three or more tests, or the indicator for one or more tests are significantly different from the intercept in any of the samples. In the between-subject design sample, coefficients are positive but relatively small, and not significantly different from the intercept.

TABLE 2. *Main effects*

Variable	(1) Coef	Df	(2) Coef	Df
Panel A: Between-subject designs				
Intercept = 0 tests in control	0.21	24.8	0.22	24.4
	[0.08, 0.34]		[0.09, 0.34]	
1 test in control	0.00	2.9		
	[-0.24, 0.23]			
2 tests in control	-0.01	6		
	[-0.32, 0.31]			
3 or more tests in control	0.07	7.7		
	[-0.13, 0.26]			
1 or more tests in control			0.02	10.1
			[-0.16, 0.20]	
Between-Study SD	0.21		0.21	
Within-study SD	0.26		0.26	
Total SD	0.34		0.33	
QE-statistic	555.0		556.2	
I-squared	67.65		68.5	
Prediction interval, intercept	[-0.49, 0.92]		[-0.49, 0.92]	
Panel B: Within-subject designs				
Intercept = 0 tests in control	0.47	18.9	0.46	19.4
	[0.3, 0.64]		[0.29, 0.62]	
1 test in control	0.00	2.3		
	[-0.29, 0.30]			
2 tests in control	-0.16	1.1		
	[-0.61, 0.28]			
3 or more tests in control	-0.54	1.1		
	[-2.02, 0.93]			
1 or more tests in control			-0.16	3.6
			[-0.49, 0.16]	
Between-study SD	0.25		0.27	
Within-study SD	0.34		0.34	
Total SD	0.42		0.43	
QE-statistic	578.7		591.6	
I-squared	81.9		82.6	
Prediction interval, intercept	[-0.43, 1.37]		[-0.46, 1.37]	

Note: The table displays the coefficient estimates (Coef-columns) Sattertwhaite small-sample adjusted degrees of freedom (Df-columns), and the 95% confidence intervals in brackets below the coefficient estimate. The sample sizes are: 182 effect sizes nested in 38 studies in the between-subject design sample and 100 effect sizes nested in 22 studies in the within-subject design sample.

The weighted average effect is only significant in interventions with three or more tests in the control group. In the within-subject design sample, the coefficients are zero (one test) or increasingly negative and reasonably large, when the control group also received tests. None are

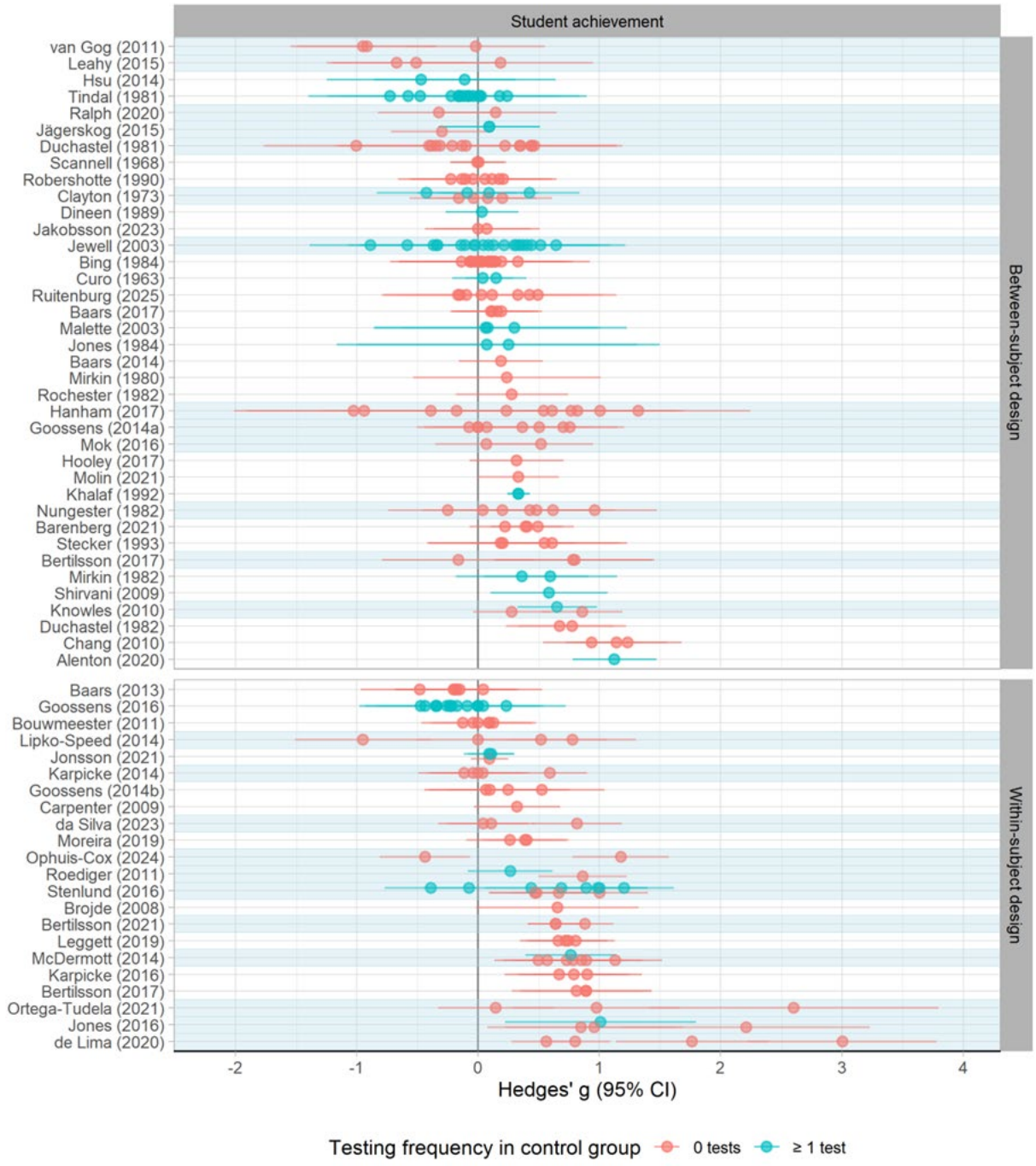
however significantly different from the intercept, and only interventions in which the control group is tested once have a significant weighted average effect.

Figure 9 shows how effect sizes are distributed across research designs and testing frequency. It also displays (in light blue) all the studies exhibiting more heterogeneity than is expected from sampling variation alone. The heterogeneity is substantial in both samples, both within and between studies. For example, the total within- and between-study heterogeneity in between-subject designs is larger than the weighted average effect size of interventions in which the control group is not tested. In within-subject designs, it is smaller than this effect size but very close. The Q-test rejects the null hypothesis of no residual heterogeneity in all specifications in Table 2 ($p < 0.0001$). The I^2 is 68-69% in the between-subject specifications and 82-83% in the within-subject specifications. The prediction interval for the intercepts is wide-ranging in all specifications and includes substantial negative effect sizes.

Results of the sensitivity analysis

Table 3 displays the results of the sensitivity analysis in which we examined different values of ρ and *ICC*s, winsorised outliers, and restricted the analysis to RCTs or to outcomes with better than high/serious risk of bias ratings. We used the specification with an intercept representing zero tests in the control group and only one indicator equal to one if the control group was tested to improve our statistical power. Columns 1 and 2 show the coefficients and confidence intervals from the between-subject designs, and columns 3 and 4 from the within-subject designs.

FIGURE 9. *Forest plot exhibiting effect sizes across study design and testing frequency*



Note: Studies exhibiting within-study heterogeneity beyond what is expected from sampling error alone are highlighted with light blue shades. Studies are ordered by the study mean effect size obtained from fitting the within-study effect sizes to a univariate random effects meta-analysis model, as suggested by Fernández-Castilla et al. (2020).

TABLE 3. *Results of the sensitivity analysis*

	(1) Between-subject		(2) Within-subject	
Analysis	**0 tests**	**1+ tests**	**0 tests**	**1+ tests**
rho = 0.4	0.21	0.03	0.46	-0.16
	[0.08, 0.33]	[-0.15, 0.33]	[0.29, 0.63]	[-0.48, 0.63]
rho = 0.9	0.22	0.01	0.45	-0.17
	[0.09, 0.35]	[-0.18, 0.35]	[0.28, 0.61]	[-0.51, 0.61]
ICC = 0	0.21	0.02	0.47	-0.17
	[0.09, 0.33]	[-0.14, 0.33]	[0.29, 0.64]	[-0.50, 0.64]
ICC = 0.32	0.20	0.02	0.48	-0.17
	[0.08, 0.33]	[-0.14, 0.32]	[0.30, 0.65]	[-0.50, 0.65]
Outliers winsorised	0.22	0.02	0.46	-0.16
	[0.09, 0.34]	[-0.16, 0.34]	[0.29, 0.62]	[-0.49, 0.62]
RCTs only	0.19	0.03	0.46	-0.15
	[0.05, 0.33]	[-0.18, 0.33]	[0.19, 0.73]	[-0.45, 0.73]
Overall risk bias less than high/serious	0.23	-0.01	0.50	-0.10
	[0.05, 0.41]	[-0.32, 0.41]	[0.31, 0.69]	[-0.46, 0.69]

Note: The table displays the coefficient estimates and the 95% confidence intervals in brackets below the coefficient estimate of the indicator for 0 tests in the control group (the intercept in the meta-regressions) and for 1 or more tests in the control group. The row headers refer to a sensitivity analysis. The sample sizes are: 182 effect sizes nested in 38 studies in the between-subject design sample and 100 effect sizes nested in 22 studies in the within-subject design sample for the analyses of rho, ICC, and outliers. The RCTs only samples are 169 effect sizes nested in 36 studies in the between-subject design sample, and 55 effect sizes nested in 12 studies in the within-subject design sample. The overall risk of bias less than high/serious are 113 effect sizes nested in 26 studies in the between-subject sample, and 74 effect sizes nested in 20 studies in the within-subject design sample.

All coefficients and confidence intervals are close to the primary analysis. For interventions where the control group was not tested, the largest difference in the between-subject design sample that the effect size is 0.03 lower in RCTs. In the within-subject design sample, we obtained the largest difference when we restricted the analysis to effect sizes with overall risk bias less than high/serious, which was 0.04 higher than in the primary analysis. For between- and

within-subject designs, this restriction also yielded that largest difference regarding interventions where the control group was tested (-0.03 and -0.06, respectively).

We also examined sensitivity for each domain in the risk of bias assessments by removing outcomes, as specified in our protocol, which received a rating of either 'high' or 'serious' (reported in text and not in Table 2). Because there were very few QESs among the between-subject designs (three studies), we did not conduct this analysis for them. Furthermore, no outcome received a high risk or serious risk of bias rating in several of the domains (see Figure 7 and 8), and we did not conduct this analysis for those domains.

The estimates were again mostly close to the main effects estimates. For the RCTs in the between-subject design sample, the intercept varied between 0.19 and 0.22 (always significant) and the indicator for one or more tests in the control group from 0.001 to 0.06 (never significant). For the RCTs in the within-subject design sample, the intercept varied between 0.46 and 0.47 (always significant). The indicator for one or more tests in the control group was -0.16 in all specifications. There was some more variation among the QESs in the within-subject design sample, as the intercept varied between 0.45 and 0.54 (always significant), and the indicator for one or more tests in the control group between -0.24 and -0.02 (never significant).

Lastly, we examined sensitivity to ceiling effects in the between-subject design sample (see discussion in section Included studies). We removed effect sizes based on proportions that had a control group mean above 0.9 and 0.8, and re-ran the primary analysis. The differences to our primary analysis results were very small. With the 0.9-restriction, the average effects in interventions where the control group was not tested was 0.23 (95% CI = [0.095, 0.357]) and the coefficient for the indicator of interventions where the control group was tested was 0.02 (95% CI

= [-0.158, 0.199]). With 0.8, the corresponding estimates were 0.23 (95% CI = [0.095, 0.356]) and 0.02 (95% CI = [-0.152, 0.203]).

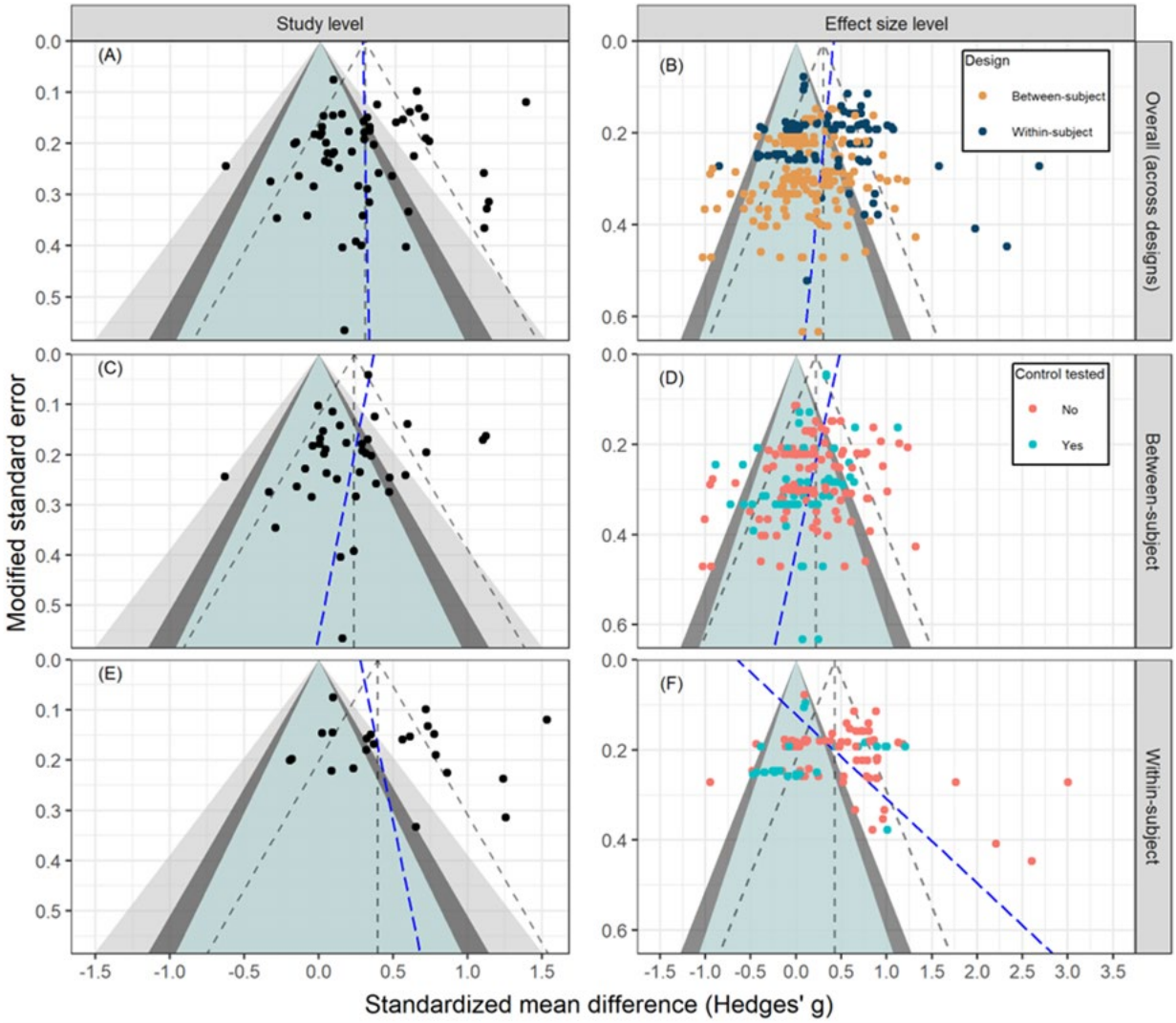
Summing up the sensitivity analysis, the main effects estimates seem robust. Next, we examine publication bias.

Publication bias

In Figure 10, we show funnel plots on the study level (left-side panels A, C, and E) and effect size levels (right-side panels, B, D, E). The top panels show between-subject designs and within-subject designs combined, the middle panel only between-subject designs, and the bottom panel only within-subject designs. There is very little asymmetry when we combine the designs, especially on the study-level. The middle and bottom panels indicate that this is a combination of two opposite asymmetries. In between-subject designs, larger studies have slightly larger effects on average, whereas larger studies have smaller effects on average in within-subject designs (potentially at least partly driven by a few large outliers).

Table 4 reports the results of the worst-case sensitivity analysis, the PEESE- or PET-versions of Egger's test (depending on relevance), and of a three- and a four-level selection model. In the worst-case analysis, the average weighted effect size remains positive in both between- and within-subject designs, but it is not significant. In line with the funnel plot, the PEESE test indicates reversed publication bias in between-subject designs and the PET test regular bias in within-subject designs. That is, that the primary analysis underestimates the weighted average effect sizes in between-subject designs and overestimates them in within-subject designs.

FIGURE 10. *Funnel plots*



Note: The figure shows funnel plots on the study level (left-side panels A, C, and E) and effect size level (right-side panels, B, D, E). The top panels show between-subject designs (blue dots in B) and within-subject designs (yellow dots in B) combined, the middle only between-subject designs, and the bottom only within-subject designs. In the latter, we separate effect sizes in which the control was tested (red dots) and not (green). The plot depicts the relationship between effect size estimates and their corresponding standard errors across the time from the end of the intervention. The colors of the points visualize study affiliations. Following Pustejovsky and Rodgers (2019), we used the adjusted standard errors based on sample sizes (i.e., same as the weights in our meta-regressions) to avoid biasing the plots because of the correlation between the effect size and its regular standard error. Effects in the green region have *p*-values above 0.1, effects in the dark gray area have *p*-values between 0.05-0.1, effects in the light gray area contain *p*-values ranging from 0.01 to 0.05, and finally effects in the white area have *p*-values less than 0.01. We calculated *p*-values using the adjusted standard errors of the effect sizes, which are not necessarily equal to the *p*-values reported by authors of the primary studies. The dashed grey lines mark the distribution about the mean effect size. Effects falling outside the dashed triangle are statistically different from the estimated mean effect size. The dashed blue line is the estimated slope from a meta-regression. For the study-level plot, we fitted classical Egger's regression models, whereas we fitted meta-regression models as suggested by Rodgers and Pustejovsky (2021) for the effect-size level plots.

TABLE 4. *Results of the publication bias analysis*

	Between-subject design				**Within-subject design**			
Model/ Coefficient	**Worst-case**	**PEESE**	**3PSM**	**4PSM**	**Worst-case**	**PET**	**3PSM**	**4PSM**
	Est [95% CI]	Est [95% CI]	Est [95% CI]	Est [95% CI]	Est [95% CI]	Est [95% CI]	Est [95% CI]	Est [95% CI]
Intercept	0.073 [-0.006, 0.151]	0.346 [0.125, 0.560]	0.194 [-0.002, 0.387]	0.133 [-0.080, 0.338]	0.041 [-0.076, 0.158]	-0.6112 [-1.606, 0.384]	0.615 [0.317, 0.854]	0.579 [0.195, 0.846]
Control group tested	-0.048 [-0.197, 0.100]	0.030 [-0.147, 0.230]	-0.104 [-0.283, 0.147]	-0.112 [-0.305, 0.161]	-0.042 [-0.333, 0.250]	-0.146 [-0.482, 0.191]	-0.338 [-0.729, 0.273]	-0.371 [-0.805, 0.300]
$\beta SE^2/\beta SE$	-	-2.106 [-4.687, 0.475]	-	-	-	5.267 [0.058, 10.476]	-	-
λ_1	-	-	1.249 [0.470, 3.458]	1.409 [0.508, 3.666]	-	-	1.884 [0.318, 5.729]	2.411 [0.576, 6.693]
λ_2	-	-	-	0.825 [0.302, 2.287]	-	-	-	1.318 [0.148, 5.441]
$\hat{\tau}$	0.000	0.220	0.388	0.405	0.000	0.4217	0.536	0.562
$\hat{\omega}$	0.204	0.266	-	-	0.144	0.343	-	-
Total SD	0.204	0.345	0.387	0.405	0.144	0.543	0.536	0.562
Total studies	33	38	38	38	15	22	22	22
Group sizes	(25, 10)	(27, 14)	(27, 14)	(27, 14)	(12, 4)	(21, 6)	(21, 6)	(21, 6)
Total Effects	154	182	182	182	51	100	100	100

The selection models indicate partly the opposite answer. Average effect sizes are slightly smaller in be-tween-subject designs for both interventions in which the control group was tested and in those where it was not. In within-subject designs, the average effect size in interventions in which the control group was tested is larger in the selection models, whereas the coefficients on interven-tions with some tests in the control group are more negative than in the primary analysis.

In sum, neither the funnel plots nor the formal tests show consistent evidence of publication bias. We have no theoretical reason to expect that publication bias would operate differently on the level of study design, so the inconsistencies between the designs in the funnel plot may be the results of outliers or, more generally, heterogeneity of effect sizes.

Confirmatory moderator analysis

Table 5 contains the results of the confirmatory moderator analysis using the specification with two indicators for control group testing frequency (the zero test (intercept) and one or more tests). Column 1 shows the between-subject designs and column 2 the within-subject designs. Including the pre-specified moderators did not result in any significant moderators and did not reduce the heterogeneity in any sample (we focused on the within, between, and total SD measures of heterogeneity, reported at the bottom of the table).

Two of the moderators in the within-subject design sample had a substantial, albeit insignificant, association with effect sizes. As in the main effects analysis, the effect sizes were smaller when the control group was tested, although the coefficient was smaller when more moderators were included (-0.06 compared to -0.16 in primary analysis). The coefficient on the inter-

vention duration was -0.09, meaning that longer interventions might have smaller effects. However, as indicated by the low degrees of freedom, there were only few longer interventions in the within-subject design sample.

TABLE 5. *Confirmatory moderator analyses*

	(1) Between-subject		(2) Within-subject	
Variable	**Coef**	**Df**	**Coef**	**Df**
Intercept = 0 tests in control	0.22	17.9	0.44	8.9
	[0.05, 0.38]		[0.04, 0.84]	
1 or more tests in control	0.01	6.9	-0.06	2.9
	[-0.17, 0.2]		[-0.27, 0.15]	
Language arts	0.02	17.5	-0.01	16.2
	[-0.22, 0.26]		[-0.46, 0.44]	
Grade	-0.01	12.5	0.02	8.7
	[-0.05, 0.03]		[-0.03, 0.06]	
Intervention duration	0.00	7.6	-0.09	3.8
	[-0.01, 0.02]		[-0.21, 0.04]	
Share of girls	0.00	4.7	0.01	4.7
	[-0.01, 0.01]		[-0.02, 0.04]	
Between-study SD	0.24		0.37	
Within-study SD	0.26		0.33	
Total SD	0.35		0.49	

Note: The table displays the coefficient estimates (Coef-columns), the Sattertwhaite small-sample adjusted degrees of freedom (Df-columns), and the 95% confidence intervals in brackets below the coefficient estimate. The sample sizes are: 182 effect sizes nested in 38 studies in the between-subject design sample and 100 effect sizes nested in 22 studies in the within-subject design sample.

Exploratory analyses

Thus far, the results indicate substantial heterogeneity of the effect sizes. In this section, we explore the heterogeneity further in two ways: First, we added groups of moderators related to features of the outcome tests, the intervention, and the control group condition to our primary

specification. We used three separate meta-regressions to improve statistical power and to improve power further, we also ran meta-regressions on the combined sample. Second, we examined how heterogeneity estimates changed when we excluded studies with small effective sample sizes from the analysis.

Exploratory moderator analyses. Table 6 (between-subject designs) and 7 (within-subject designs) contain the results of the exploratory moderator analysis. Although there are some moderators with substantial associations with effect sizes, the main result is that the heterogeneity estimates remain large in all specifications and in both samples (see bottom of the tables for estimates of the within, between, and total heterogeneity reported in standard deviation units). Below, we comment on the moderators with substantial associations with effect sizes.

In both samples and tables, column 1 indicates that effect sizes are substantially, but not significantly larger when based on outcome tests that are identical to the practice tests used in the intervention. Effect sizes based on skill outcome tests are also substantially but not significantly smaller than effect sizes based on rote memorization outcomes in the between-subject sample. In the within-subject sample, the coefficient is also very large and significant, although the degrees of freedom are so small that the standard errors might be unreliable. In column 2, the coefficient on the feedback-indicator is negative in the between-subject sample, and positive, large, and significant in the within-subject sample. When controlling for either groups of moderators related to the outcome tests or the intervention (i.e., in column 1 and 2), the coefficient on the indicator for one or more tests in the control group is negative and relatively large, although not significant in the within-subject sample.

Column 3 of Table 6 and 7 shows results from a specification with different control group conditions (without an intercept). In both samples, interventions that increase the testing frequency have positive and significant effects when compared to a control group that receives either TAU /filler task or a restudy condition. The effect is positive but small and not significant in the between-subject sample when compared to control conditions that involve worked examples. In the within-subject sample, the same comparison yields a negative and significant coefficient. The effect is positive for other active controls (which there are few of in both samples) but smaller than the effects for TAU/filler tasks and restudy conditions. Lastly, the effect estimate is positive and significant for interventions where the control group was tested one or more times in the between-subject sample, and positive but not significant in the within-subject sample.

In sum, there are indications that the effects of testing students at least once depend on features of the outcome test, intervention, and the control condition. However, the results are not always consistent across the between- and within-subject design samples. A potential reason for the inconsistency is that the number of effect sizes and studies is often small and we have too little variation in the data to obtain precise estimates. Another reason may be that moderators are correlated, and therefore capture partly the same underlying tendency. To investigate this further we adjusted effect sizes from within-subject designs by the methods described in section Measures of the treatment effect, combining the two samples, and including all exploratory moderators as well as an indicator for within-subject designs in one meta-regression.

TABLE 6. *Exploratory moderator analyses of between-subject designs*

	(1) Outcome		(2) Intervention		(3) Control	
Moderator	**Coef**	**Df**	**Coef**	**Df**	**Coef**	**Df**
Intercept = 0 test in control	0.24	15.2	0.24	14.9		
	[0.07, 0.4]		[-0.01, 0.5]			
1 or more tests in control	0.02	11.5	0.02	7.6	0.24	10.8
	[-0.16, 0.21]		[-0.19, 0.22]		[0.06, 0.42]	
Follow-up length	0.03	3.0				
	[-0.05, 0.11]					
Identical outcome and practice tests	0.20	3.7				
	[-0.10, 0.50]					
Skill outcome test	-0.07	11.8				
	[-0.25, 0.10]					
Testing frequency treatment group			0.00	3.2		
			[-0.01, 0.01]			
Feedback			-0.05	29.0		
			[-0.32, 0.22]			
TAU/Filler task					0.35	10.3
					[0.14, 0.55]	
Restudy					0.22	8.0
					[0.06, 0.38]	
Worked examples					0.06	5.4
					[-0.25, 0.37]	
Other active control					0.14	3.0
					[-0.16, 0.45]	
Between-study SD	0.22		0.22			0.19
Within-study SD	0.26		0.26			0.26
Total SD	0.34		0.34			0.32

Note: The table displays the coefficient estimates (Coef-columns), the Sattertwhaite small-sample adjusted degrees of freedom (Df-columns), and the 95% confidence intervals in brackets below the coefficient estimate. The sample size is 182 effect sizes nested in 38 studies. Note that the meta-regression reported in column 2 contains an indicator for the outcomes missing information about feedback, which we omitted from the table for brevity.

TABLE 7. *Exploratory moderator analyses of within-subject designs*

	(1) Outcome		(2) Intervention		(3) Control	
Moderator	**Coef**	**Df**	**Coef**	**Df**	**Coef**	**Df**
Intercept = 0 test in control	0.46	11.6	0.13	5.9		
	[0.26, 0.66]		[-0.21, 0.47]			
1 or more tests in control	-0.25	3.9	-0.19	3.6	0.30	4.1
	[-0.66, 0.17]		[-0.59, 0.21]		[-0.1, 0.7]	
Follow-up length	-0.01	2.3				
	[-0.05, 0.04]					
Identical outcome and practice tests	0.16	10.8				
	[-0.08, 0.4]					
Skill outcome test	-0.77	2.5				
	[-1.47, -0.08]					
Testing frequency treatment group			0.00	3.0		
			[-0.11, 0.11]			
Feedback			0.56	12.8		
			[0.20, 0.92]			
TAU/Filler task					0.61	1.6
					[0.38, 0.84]	
Restudy					0.5	16.0
					[0.33, 0.67]	
Worked examples					-0.19	1.0
					[-0.27, -0.11]	
Other active control					0.26	1.1
					[-4.74, 5.27]	
Between-study SD	0.24		0.24		0.23	
Within-study SD	0.31		0.30		0.34	
Total SD	0.40		0.39		0.41	

Note: The table displays the coefficient estimates (Coef-columns), the Sattertwhaite small-sample adjusted degrees of freedom (Df-columns), and the 95% confidence intervals in brackets below the coefficient estimate. The sample size is 100 effect sizes nested in 22 studies. Note that the meta-regression reported in column 2 contains an indicator for the outcomes missing information about feedback, which we omitted from the table for brevity.

Table 8 presents the results from the exploratory analysis of the combined sample. In column 1, we adjusted effect sizes and variances from within-subject designs assuming a within-subject correlation of 0.6, in column 2 we assumed 0.75, and in column 3 a correlation of 0.9. With the exception of the indicators for skill outcome tests and feedback, the coefficient estimates are reasonably stable across the columns. The indicator for within-subject designs is not significant in any specification, but comes close to being significant in column 3, where it is relatively large as well. The coefficient is closest to zero in column 2, which may be an indication that a correlation of 0.75 comes closest to equalizing the differences between effect sizes from the two designs. Because we adjust within-subject designs to be comparable to between-subject designs, these estimates are more directly comparable to the earlier between-subject design estimates.

The differences between the heterogeneity estimates are more pronounced: heterogeneity is lower in column 2 and, especially, in column 3 than in column 1, and also lower than the heterogeneity observed in the main effects and earlier moderator analyses. It is especially the between-study heterogeneity that is reduced quite a lot in column 3.

TABLE 8. *Exploratory moderator analyses of the combined sample*

	(1) δ = 0.6		(2) δ = 0.75		(3) δ = 0.9	
Moderator	**Coef**	**Df**	**Coef**	**Df**	**Coef**	**Df**
Intercept = 0 test in control and TAU/filler tasks	0.36	15.8	0.38	15.8	0.39	15.7
	[0.12, 0.61]		[0.16, 0.60]		[0.20, 0.59]	
1 or more tests in control	-0.20	13.9	-0.19	13.6	-0.17	13.1
	[-0.37, -0.03]		[-0.34, -0.04]		[-0.3, -0.04]	
Follow-up length	-0.01	3.3	-0.01	3.3	-0.00	3.3
	[-0.03, 0.02]		[-0.03, 0.02]		[-0.02, 0.01]	
Identical outcome practice tests	0.12	14.6	0.11	14.8	0.10	15.1
	[-0.06, 0.30]		[-0.03, 0.26]		[-0.00, 0.21]	
Skill outcome test	-0.19	10.3	-0.16	10.2	-0.12	10.1
	[-0.43, 0.05]		[-0.36, 0.04]		[-0.27, 0.03]	
Testing frequency treatment group	0.00	3.0	0.00	3.1	0.00	3.2
	[-0.01, 0.01]		[-0.01, 0.01]		[-0.01, 0.01]	
Feedback	0.17	33.7	0.11	33.4	0.03	31.9
	[-0.06, 0.39]		[-0.08, 0.3]		[-0.11, 0.17]	
Restudy	-0.14	7.7	-0.14	7.8	-0.13	8.1
	[-0.28, -0.01]		[-0.26, -0.01]		[-0.24, -0.02]	
Worked examples	-0.25	8.7	-0.25	8.6	-0.24	8.5
	[-0.54, 0.04]		[-0.52, 0.02]		[-0.49, 0.00]	
Other active control	-0.31	6.8	-0.29	6.7	-0.25	6.5
	[-0.66, 0.04]		[-0.58, 0.00]		[-0.46, -0.03]	
Within-subject design	0.05	20.2	-0.02	20.5	-0.12	20.2
	[-0.15, 0.26]		[-0.18, 0.15]		[-0.24, 0]	
Between-study SD	0.18		0.14		0.05	
Within-study SD	0.26		0.22		0.18	
Total SD	0.31		0.26		0.18	

Note: The table displays the coefficient estimates (Coef-columns), the Sattertwhaite small-sample adjusted degrees of freedom (Df-columns), and the 95% confidence intervals in brackets below the coefficient estimate. The sample size is 282 effect sizes nested in 59 studies. TAU/filler tasks is the omitted reference category for the control conditions and the intercept is therefore the effect of interventions in which a tested treatment group is con-trasted with a control group that is not tested and that receives TAU or filler tasks. Note that the meta-regressions contain an indicator for the outcomes missing information about feedback, which we omitted from the table for brevity.

The intercept in Table 8 represents the effect of interventions in which the control group was not tested and received TAU or a filler task, with none of the binary moderator features present and with the continuous moderators set at their mean values. The effect of more frequent testing is robustly positive and significant across the specifications. The coefficient of the indicator for one or more tests in the control group is similarly robust across specifications, but negative and significantly different from the intercept. It suggests that when we adjust for other moderators, the advantage of testing students more frequently is about half as large when the control group is also tested.

Other moderators with stable coefficients of noteworthy size across specifications include identical outcome and practice tests, for which effects are significantly larger than the intercept, and skill outcome test, restudy, worked examples, and other active control, all of which are associated with smaller effect sizes. Of these, only the coefficients on restudy and other active controls is significant in at least one specification. The total marginal effect (intercept plus coefficient) is still positive for all the negatively associated moderators.

Lastly, it is also noteworthy that the coefficient for testing frequency in the treatment group is close to zero in all specifications in Table 8, as well as in Table 6 and 7. Meaning that, once the testing frequency of the control group is accounted for, there is almost no association between the number of tests in the treatment group and effect sizes.

Excluding studies with small effective sample sizes. All our previous analyses indicated substantial heterogeneity. This heterogeneity might be a sign that there are important moderators missing from the analyses, that the functional form of the relation between effect sizes and moderators is more complex than the linear form we used, or that there are unmeasured me-

diators (e.g., implementation fidelity).Another explanation might be that there is additional random noise that the heterogeneity parameters are capturing. The within and between-study heterogeneity measures take sampling error into account (i.e., they measure heterogeneity over and above sampling error). However, sampling error does not include sources of uncertainty such as measurement error in the out-come (Hedges, 1981), or "design-based" uncertainty caused by the lack of information about all potential outcomes of the treatment and control group (Abadie et al., 2020). In relation to de-sign-based uncertainty, there may be baseline differences between the treatment and control group (Hedges, 1983). Even properly conducted RCTs are only balanced in expectation, and there may be chance bias–random baseline differences between treatment and control groups (Roberts & Torgerson, 1999).

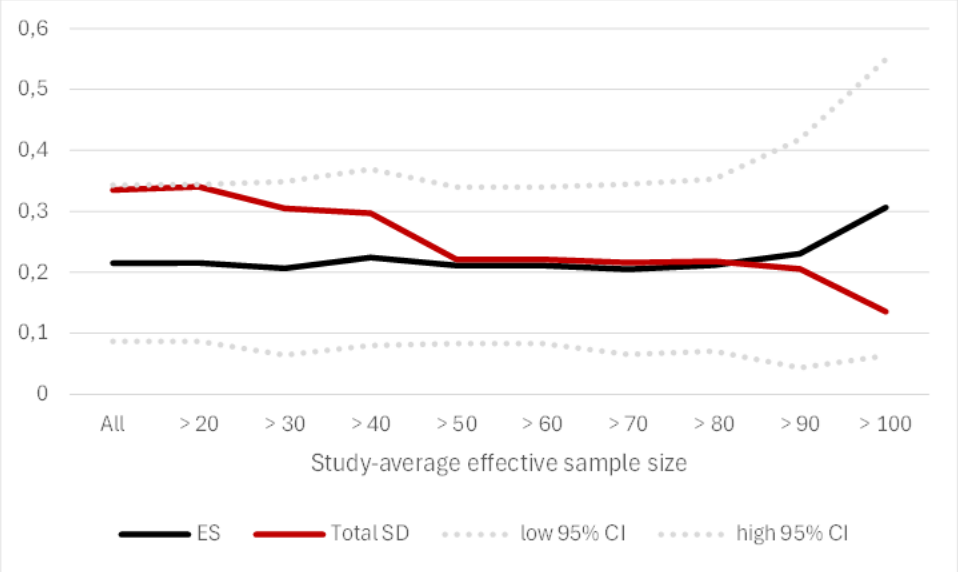
Because these sources of uncertainty are more likely to appear in small sample sizes, examining the relation between the total heterogeneity and the effective sample size (i.e., sample size adjusted for clustering) may be informative. That is, if there is less heterogeneity in large studies, it may indicate that part of the heterogeneity is due to sources that are random and therefore not explainable. The included studies contained many studies with very small sample sizes, so this issue may be pertinent.

In Figure 11 and 12, we plot the estimate of the effect size in interventions where the control group was not tested (black line) from our primary specification. The meta-regressions progressively exclude studies with small effective samples (the x-axis shows the cutoff used). On the y-axis in each figure, we also show the 95% confidence interval of the effect estimate (grey dotted line) and the total within and between-study heterogeneity (red line). We kept the between (Figure 11) and within-subject designs (Figure 12) separated because in designs that control for

more baseline differences, we expect heterogeneity to decrease less when we exclude small studies. Furthermore, if outcome measurement error is caused by random differences in testing conditions (e.g., weather, temperature), then within-subject designs should also be less affected. The reason is that the testing conditions are more likely to be balanced across treatment and control conditions when subjects in different conditions are tested at exactly the same time. We thus expect less of a decline of the heterogeneity in within-subject designs.

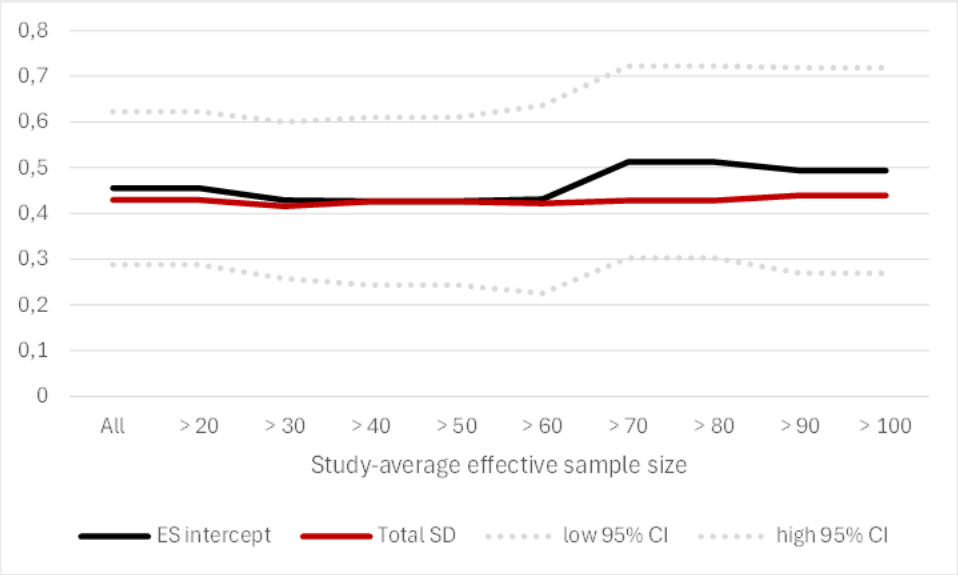
This is also the pattern we found. The effect estimates are very stable in both Figures 11 and 12. It is only when we have excluded all studies with effective sample sizes over 100 that we see an increase of the effect in both figures and then we are down to relatively few studies (5 in the between-subject design sample, 12 in the within-subject design sample). In Figure 11, we see a relatively large decline in heterogeneity: it is almost 40% lower when we exclude effective sample sizes of 40 to 90, and 60% lower in the last regression. By contrast, the heterogeneity estimate is almost completely flat in Figure 12. That is, the patterns are consistent with the idea that there are additional random sources of uncertainty in the between-subject designs. A substantial proportion of the within and between-study heterogeneity may thus not be possible to explain.

FIGURE 11. *Heterogeneity and effective sample size in between-subject designs*



Note: The y-axis shows the estimated weighted effect size (black line), the total heterogeneity (red line), and the 95% confidence interval of the estimated weighted effect size in standard deviation units (dotted grey lines). The x-axis shows the cutoff for the effective sample size used in the meta-regression corresponding to the cutoff.

FIGURE 12. *Heterogeneity and effective sample size in within-subject designs*



Note: The y-axis shows the estimated weighted effect size (black line), the total heterogeneity (red line), and the 95% confidence interval of the estimated weighted effect size in standard deviation units (dotted grey lines). The x-axis shows the cutoff for the effective sample size used in the meta-regression corresponding to the cutoff.

Results of studies excluded from the meta-analysis due to lack of information or comparability

In this section, we briefly discuss the studies that were excluded from the meta-analysis due to either lack of information needed to calculate an effect size or because we deemed the intervention not comparable to other included interventions.

We included five studies in the review that we were unable to include in the meta-analysis due to lack of information. Two studies had control groups exposed to at least one test and the three remaining studies had control groups who were exposed to some type of restudy.

Schreiner (1973) and Mayer and Rojas (1982) investigated varying testing frequencies in 7th and 8th grade students, respectively. Both were RCTs. Schreiner assigned daily or one-time practice tests combined with varying combinations of like/unlike materials and reinforcement/no reinforcement. Mayer and Rojas assigned participants to three groups: one tested every day, one tested approximately every fourth day, and one tested approximately every tenth day. None of the studies found significant differences between the groups tested more and less frequently.

Three studies–two RCTs (Hayes-Roth, 1976; Paas, 1992) and one QES (Sweller & Cooper, 1985, experiment 2)–had interventions that increased the testing frequency in the treatment group(s) whereas the control group was not tested. The studies were set in upper secondary school/high school. In Hayes-Roth (1976), participants read sixteen ten-sentence passages in either a fill-in-the-blank testing condition or a restudy condition. The testing condition produced better performance on both free and cued recall outcome tests. Paas (1992) found the opposite result, the problem-solving test condition showed a lower score than both conditions without tests: studying the problem solution or completing a partly worked-out problem solution (the statistical tests for the specific comparisons are not reported). Sweller & Cooper (1985) found

mixed effects on algebra tests of problem-solving test conditions compared to control conditions involving worked examples.

To sum up, the two studies where the control groups received at least one test could not detect any effect of increased testing. The three studies where the control groups were exposed to some kind of restudy or worked example condition were more varying in their conclusions, pointing to zero, positive, and negative effects of testing compared to restudy or worked examples.

Lavy (2024) examined high-stakes exit exams at the end of high school in Israel. He used a natural experiment providing exogeneous within-subject variation in the number of weekly tests and in the number of tests per day. This intervention was different from all others we included: there was no intention to improve student achievement, and there was no test of whether students learned anything from the test taking. The results show that increasing the number of tests per week or per day reduces test performance in a nonlinear pattern. Two tests in a week on different days negatively affect test scores relative to one weekly exam. The negative effect was doubled when both tests were on the same day, although the difference was not significant. Adding more weekly exams (more than three exams a week was rare), each on a different day, was three times worse than moving from one to two weekly exams on separate days.

Discussion

Summary of main results

We searched for studies in primary and secondary school of the effects of changing the testing frequency on student achievement and testing anxiety. We found few examples of the latter outcome and could not conduct meta-analysis related to our second research question. For our

primary research question, we included 59 studies in the meta-analysis. The studies examined interventions, which, almost universally, used practice tests that had a formative purpose and were low-stakes for the students. Most interventions were short, with few practice tests for the treatment group and no or a small number of tests in the control group. The control conditions were otherwise quite heterogeneous, and included, for example, TAU, filler tasks, restudy, and worked examples. The interventions covered a broad range of subjects, but around half were categorized as language arts. Students were tested with outcome tests conducted shortly after the end of intervention. The outcome tests were, for the most part, researcher developed and designed to explore the effects of testing on recall of, for example, word definitions, and science or history facts. There were also tests of more complex skills, typically problem-solving tasks.

We found positive (i.e., beneficial for the students), robust, and significant effects of interventions that compared treatment groups to control groups that were not tested. We separated our sample by study design, and found an average weighted effect size of 0.21-0.22 in the sample of between-subject designs and 0.46-0.47 in the within-subject design sample. These effect sizes may seem relatively large compared to for example the criteria in Kraft (2020), where effect sizes above 0.2 are considered large. However, Kraft's criteria are based on standardized tests, which are not fully comparable to the tests we included (e.g., Wolf et al., 2022). Using practice tests is likely to be a low-cost method. The effect sizes therefore seem large in relation to expected costs, and indicate that low-stakes, formative testing can be highly cost-effective.

In interventions where the control group was also tested, our results were more uncertain. While the estimates were always positive, they were not always significant, and, in some analyses, diminished with the control group testing frequency. The main reason for the uncertainty is that there were comparatively few studies in which the control group was tested at all.

We found substantial heterogeneity in our main effects analyses. We examined pre-specified moderators of the effects, and found no significant ones and no reduction of the heterogeneity. We therefore also examined heterogeneity in a range of exploratory analyses. Some moderators related to features of the outcome tests and the control condition appeared to be influential: effect sizes were substantially larger when outcome tests were identical to the practice tests used in the intervention, and smaller when outcome tests measured skills such as problem-solving. The effects were largest when the control group received TAU or a filler task, smaller in restudy conditions and when the control group was tested as well, and smallest in control conditions that involved worked examples or other active conditions. It is also noteworthy that when we included the treatment group testing frequency alongside the control group testing frequency in the meta-regressions, the former association with effect sizes was very close to zero. That is, adding information about treatment group frequency explained hardly any variation over and above the control group frequency.

We want to emphasise that the results of the exploratory analysis should be interpreted cautiously. As the moderators were not consistently significant across specifications (or significantly different from one another), not pre-specified, and in the combined analysis, relied on the assumptions of the within-subject correlations used to adjust effect sizes, we see these results primarily as suggestions for hypotheses to be tested in further research.

Moreover, while the within and between-study heterogeneity was reduced in some of the exploratory analyses, it remained clearly above zero. One reason may be that there were unmeasured moderators or mediators left out of our analysis. We discuss potential examples in section Implications for research below. However, in the between-subject design sample, we also found evidence suggesting that a substantial share of the between- and within-study heterogeneity

stems from sources (e.g., measurement error, chance bias) that are unlikely to be explainable. When we excluded studies with small effective sample sizes, the effect of testing compared to control groups that were not tested was highly stable whereas the heterogeneity parameters declined. Because nearly all studies were small (effective sample size below 100), we could not fully examine how far this decline continued as studies got bigger.

Overall completeness and applicability of evidence

Although nearly half of our studies were from the United States, 11 other countries contributed studies to the meta-analysis. These countries were diverse in terms of geography. We included studies from Western Europe, East and West Asia, Oceania, as well as North and South America. The countries were mainly high or middle-income countries according to the World Bank (2025) classification. Low-income countries were thus underrepresented, indicating a need for further research. Furthermore, we only included studies published in English, German, Danish, Swedish, and Norwegian due to language restrictions in the review team. Studies from countries where these languages are not used may therefore be underrepresented in our review, although we included studies from countries with other first languages (Brazil, China, Saudi Arabia, Spain, Taiwan, the Netherlands, and the Philippines).

The included studies covered almost the full range of grades in primary and secondary school. Only kindergarten class, which is sometimes a part of primary school, was not included. We also included studies from different time periods, our oldest study was published in 1962 and the newest in 2025. Most of the studies were relatively new though and published in the last 10-15 years.

The included studies imposed several important restrictions on what we could analyse. While the range of testing frequencies was relatively broad, most interventions contained only

one practice test. There were comparatively few studies in which the control group was tested, and very few studies compared high testing frequencies in both the treatment and control group. We found few studies examining how frequent high-stakes and summative tests affect students, and few studies measured testing anxiety. We could therefore not conduct the planned analyses of such interventions and outcomes. Outcomes were typically measured at or shortly after the end of intervention. Thus, we do not know whether the short-run positive effects persist, or whether they, like in most educational interventions, fade out over time (e.g., Bailey et al., 2017, 2020; Dietrichson et al., 2026; Hart et al., 2024).

Quality of the evidence

Our meta-analyses were based on a relatively large sample of 59 studies. However, the studies used both between-subject and within-subject designs, which were unlikely to be fully comparable without additional assumptions that were difficult to verify. We therefore split the analyses by design, which reduced the statistical power. Our main effects-estimates were nevertheless precisely estimated and robust across all sensitivity analyses. We found inconsistent results in the analysis of publication bias. The inconsistencies were found both between the between-subject and the within-subject samples, and between different tests within these samples. Visual and formal tests of asymmetric funnel plots based on the combined sample found no evidence of asymmetries. Thus, although publication bias cannot be ruled out, there was no robust evidence of such bias in our data.

The between-subject design sample contained a large proportion of RCTs (93%). The within-subject design sample contained 45% RCTs. We found similar effects in RCTs as in QES. Thus, our results were not driven by the inclusion of QES.

There were 22 QESs where we assessed that all outcomes had critical risk of bias in at least one domain of ROBINS-I. They were therefore not included in the synthesis (as per the guidelines for ROBINS-I, Sterne et al., 2016). Almost all outcomes received critical risk of bias in the confounding bias domain. The overall risk of bias ratings of the outcomes included in the meta-analysis were relatively high, and no outcome received a low risk of bias rating in all domains. The main reason was the universal lack of pre-specified analysis plans. Furthermore, the risk of bias in the deviations from intended interventions and the missing outcome data domain were relatively high in both designs, but few outcomes in the within-subject design samples received a high risk of bias-rating. We found no evidence of effect sizes being associated with risk of bias ratings.

Most studies in our sample were small. Especially in the between-subject design sample, many studies did not have adequate statistical power to find effects similar in magnitude to the estimated average effect size. As expected with small sample sizes across studies, we observed considerable variation of effect sizes, some of which could be explained by sampling error. However, the within and between-study parts of the heterogeneity were also substantial and, for the most part, remained unexplained in our moderator analyses. We presented evidence that parts of this heterogeneity in the between-subject sample might be coming from sources (e.g., measurement error, chance bias) that are difficult to capture in moderator analyses based on data like ours.

Potential biases in the review process

We conducted a substantial search and screening effort. The title and abstract screening comprised 102,451 records, and the full text screening 1,022 records. The large number of records in the title and abstract screening was the result of a concern that our initial search strategy

was too narrow. To handle the large number of records, we used ML algorithms to prioritize records and LLMs to screen on title and abstract (see Supplementary Information Appendix B for details). These methods worked well according to our own, in-review, tests, and have worked well when tested by others (e.g., Borge et al., 2023; Vembye et al., 2025). However, we cannot rule out that relevant records were missed in this process. We conducted a thorough citation tracking effort of both related reviews and primary studies, which should provide additional safeguards against missing relevant studies.

There were 25 potentially relevant full-texts that we could not locate. These records were disproportionally older studies, which were less likely to be included in the meta-analysis. The studies included in the meta-analysis was 5.8% of the total number of records screened in full text (59/1,022). Under the conservative assumption that the studies we could not locate were equally likely to be included as other studies screened in full text, the expected number of missing studies is 1.4. Because 1-2 additional studies would be unlikely to move our primary results much, we believe that the risk of bias to our results from not finding all full texts is low.

The screening, coding, and data extraction included many review team members. This increases the risk of inconsistencies. To mitigate these risks, we held regular and frequent meetings where we discussed issues related to screening and coding.

Our protocol specified that we would search ProQuest Dissertations & Theses Global. However, at the time of the updated and extra searches, we no longer had access to this database. We searched EBSCO Open Dissertations, which also specifically covers dissertations. Furthermore, dissertations are also included in other databases used in our search, as well as in the searches of other sources.

There were five studies that we could not include in the meta-analysis due to lack of information needed to calculate effect sizes. However, our narrative analysis of these studies did not indicate that they would have greatly changed the results, had we been able to include them.

Agreements and disagreements with other studies or reviews

We found positive and significant effects of interventions that compared treatment groups to control groups that were not tested. Earlier reviews have also found positive and significant effects examining similar contrasts (Adesope et al., 2017; Phelps, 2012; Yang et al., 2021). The magnitudes were more difficult to compare because both the inclusion criteria and analytic strategies differed across reviews. With those caveats in mind, our average effect sizes seem smaller than those found in earlier reviews. We separated our sample by study design, and found an average weighted effect size of 0.21 in the sample of between-subject designs and 0.46 in the within-subject design sample. Adesope et al. (2017) overall found larger effect sizes than we did (0.64 in primary school and 0.83 in secondary school), but no significant difference between within and between-subject designs. One explanation may be that they included laboratory experiments, which we did not. Yang et al. (2021) found larger effects in within-subject designs than in between-subject designs (0.67 versus 0.42). These estimates included students in tertiary education, which may explain why they are larger. Yang et al. (2021) found effects of 0.33 in elementary school, 0.60 in middle school, and 0.66 in high school in analyses that included both types of design.

Phelps (2012) found an effect size of 0.85 in studies where the treatment group was tested more frequently than the control group, which we believe was the closest estimate to ours. This estimate included a broader set of study designs and participants than ours, which may explain the difference. Adesope et al. (2017) and Yang et al. (2021) did not include studies in which the

control group was also tested. That is, there is no direct counterpart to our estimate of effects in interventions where the control group received one or more practice tests.

All three reviews conducted subgroup analyses that corresponded, at least partly, to our moderator analyses. However, these analyses included one subgroup at a time and were not confined to participant groups that corresponded to primary and secondary school students. The magnitudes of the associations are therefore not directly comparable to ours. We therefore comment on the direction here. Phelps (2012) included studies in which the type of tests differed, and found an advantage for the treatment group also when tested with higher stakes, something we were unable to examine. Yang et al. (2021) also found a positive effect of testing with high-stakes tests.

Adesope et al. (2017) and Yang et al. (2021) both identified smaller effects in restudy conditions versus filler/no treatment conditions, which we also did in one analysis. Yang et al. (2021) observed a relatively small, positive but insignificant effect in interventions using elaborative strategies as the control condition. Their definition of elaborative strategies (p. 414) appeared similar to our moderators indicating “other active” and, possibly, “worked example” control groups, which were associated with smaller effects in our analyses. Both reviews also found larger effects when the outcome and practice test format were matched and similar advantages for material that were covered over untested material. These results were similar to our result that identical outcome and practice tests were associated with larger effects.

In Yang et al.’s analyses, corrective feedback and higher testing frequency in the treatment group were associated with larger effects. Adesope et al. (2017) found similar effects with and without feedback on practice tests, and smaller effects when the treatment group was tested two or more times compared to only one time. Our estimates for feedback were not robust across

specifications, and we found no association between the treatment group testing frequency and effect sizes, once the control group's frequency was included in the meta-regression.

Lastly, like us, Adesope et al. (2017) found significant heterogeneity in almost all of their analyses (measured with the Q and I2 statistics). The overall analysis in Yang et al. (2021) also indicated significant heterogeneity using the Q-statistic. Neither review reported estimates of the between and within-study heterogeneity.

Implications for practice and policy

Our results indicated that formative and low-stakes practice tests can be an effective pedagogical method. In line with earlier reviews, we found positive, robust, and significant effects of interventions in which the treatment group was tested and the control group was not. Practice tests seem especially effective when the learning material is less complex, like remembering word definitions or history and science facts. Furthermore, practice tests seem like a comparatively inexpensive intervention, and thus has the potential to be cost-effective as well.

There are three important caveats to this conclusion: First, the heterogeneity was substantial, and there were also interventions that showed negative effects. Our exploratory analyses suggested some potentially important moderators, which together underlined that the effects of practice tests are likely smaller when the learning material is more complex, and when the alternative pedagogical method involves more active forms of learning (such as restudy or exposure to worked examples). However, our results also indicated that some of the heterogeneity in the between-subject design sample might be due to small studies being more heterogeneous for random reasons. If the heterogeneity estimates capture random factors, then the risk of obtaining real negative effects would be lower. But more research is needed to learn why between-subject

designs with small sample sizes seem to be more heterogeneous than those with larger sample sizes.

Second, the risk of bias was relatively high in the included studies. We found no study that reported having a pre-registered analysis plan, which made it difficult to assess selective reporting in studies. There was also relatively high risk of bias in domains that assessed the measurement of outcomes, the randomisation process, deviations from intended interventions, missing outcome data, and confounding. We rated the within-subject designs as having lower risk of bias than the between-subject designs. As the within-study designs also had better statistical power, the results from within-study designs seem more secure.

Third, we could not meta-analyse the effects on testing anxiety. The positive effects on student achievement could potentially be accompanied by negative effects on anxiety. However, Yang et al. (2023) examined the effects of testing versus not testing on testing anxiety and found the opposite result. That is, testing reduced anxiety. The one included study that measured effects on testing anxiety in our review also found that testing reduced anxiety, but it is unclear to what extent those results hold for primary and secondary school students in general, or when stakes are high.

Our results indicated that never testing students is unlikely to be an optimal strategy. As our results were more uncertain when the control group was also tested, how frequently students should be tested is still an open question. The answer likely depends on the learning material and the alternative pedagogical methods available to the school and its teachers. The optimal mix of pedagogical methods for learning, for example, problem-solving skills, may include both practice tests and other active forms of learning, like worked examples.

Implications for research

Although the literature on the effects of testing students is large, our review pointed to several areas where improvements in study designs are possible and more research is needed. Starting with study designs, many studies in our sample seemed underpowered and would have benefitted from larger sample sizes. This was especially problematic for the between-subject design sample, where many studies were so small that there is little reason to expect even a proper randomisation to produce balanced treatment and control groups (see Goldberg, 2019 for simulations of chance differences). Although combining small studies to gain power and even out chance differences are key advantages of meta-analysis, it is possible that small studies also increase heterogeneity. Substantial within and between-study heterogeneity, as we found, may indicate that some interventions have adverse effects, which should caution against strong conclusions. For these reasons and despite the possibility that small-study heterogeneity may also be caused by random factors (as we found indications of), the literature would clearly benefit from larger sample sizes.

Using within-subject designs partly circumvents both the power and the imbalance problems. In line with this notion, we found smaller heterogeneity estimates in relation to the average effects and no decline of the heterogeneity in within-subject designs when we excluded small studies. For some research questions, a within-subject design is therefore a better choice than a between-subject design. However, other research questions are difficult to study in within-subject designs. Any intervention with "carry-over" effects between conditions – when what you learn in one condition affect the outcome in the others – would yield biased estimates in a within-subject design. These risks were often relatively small in our setting because a) intervention content was not connected in the sense that learning, for example, one word definition or one fact,

does not imply that learning another word or fact is easier, and b) interventions targeting skills that might transfer, such as problem solving, were typically short and skill transfer usually takes some time.

Another way to improve power and reduce chance bias is to adjust for pre-intervention variables, which comparatively few studies in our sample did. Lack of adjustment for pre-intervention variables was also a major reason why our risk of bias assessments yielded relatively high ratings for QES. We rated all outcomes in 22 studies as having critical risk of bias and they were excluded from the synthesis. Almost all received the critical rating in the confounding bias domain and a common reason for the rating was that there was little or no adjustment for confounders. No outcome included in the meta-analyses received a low risk of bias rating in all domains. The main reason was the universal lack of pre-specified analysis plans, and the literature would be improved if studies pre-registered their outcomes and hypotheses.

Our review points to areas where more research is needed. There were comparatively few studies in which the control group was also tested, and especially few in which the control group received more than a handful of tests. A related issue was that most interventions were short and outcomes were measured shortly after the end of intervention. Studies examining effects of testing students more frequently over an extended period of time and conducting longer-run follow-ups would be very valuable. Because carry-over effects are more likely in longer interventions, the difficulty of using within-subject design may be an explanation why relatively few included interventions were longer than a few weeks.

We could not conduct our planned analyses of our secondary outcome, testing anxiety. We could also not analyse the effects of testing students more frequently using summative and high-stakes tests. As the debate on the effects of testing has often revolved around the potentially

harmful effects of summative and high-stakes tests on non-achievement measures like testing anxiety, there is a pressing need for such studies. We believe there are at least two, related, reasons for the lack of studies in our sample. Controlled experiments of potentially harmful interventions are for obvious reasons problematic. We found studies exploiting natural experiments where some students for exogenous reasons are tested with summative tests more often than others (e.g., Green et al., 2025). However, these studies lacked information about the control group testing frequency, which is clearly more problematic to keep track of in retrospective studies.

We found substantial heterogeneity in our main effects analyses. One reason may be that there were unmeasured moderators and mediators. An example of the former might be that effects on motivation could be heterogeneous across achievement, as high-achieving students get positively reinforced, in contrast to students with difficulties. To study this issue, within-study variation in the effects across the achievement spectrum would be preferable. An example of a potential mediator is the quality of implementation. We extracted data on implementation problems, but few studies reported such problems. As many interventions were very short, it is plausible that there were few or no problems. However, more information about implementation would still be valuable. As mentioned, a second reason for the substantial heterogeneity, at least in the between-subject design sample, may be the small effective sample sizes. That small samples may be more heterogeneous has been suggested in other areas too (e.g., IntHout et al., 2015; Turner et al., 2013; Viechtbauer & Lopez-Lopez, 2022; Williams et al., 2021), but more research on the generality and reasons for this phenomenon is needed.

References

Achord, R. L. K. (2015). *The effect of frequent quizzing on student populations with differing preparation and motivation in the high school biology classroom* [Master's thesis, Louisiana State University and Agricultural and Mechanical College]. LSU Scholarly Repository. https://doi.org/10.31390/gradschool_theses.871

Alenton, J. C. (2020). Daily practice test: effects on mathematics performance in solving the fundamental operations on fractions. *International Journal of Educational Research and Innovation (IJERI)*, *18*, 47–61. https://doi.org/10.46661/ijeri.4519.

Alexander, B., Owen, S., & Thames, C. B. (2020). Exploring differences and relationships between online formative and summative assessments in Mississippi career and technical education. *Asian Association of Open Universities Journal*, *15*(3), 335–349. https://doi.org/10.1108/AAOUJ-06-2020-0037.

Allen, J. (2019). *Does adoption of ACT Aspire Periodic Assessments support student growth?* ACT, Inc. Research Report 2019-1. https://eric.ed.gov/?id=ED596133

Augustin, M. A. (2015). *Effect of progress monitoring on reading achievement for students in a middle school setting* [Doctoral Dissertation, Missouri Baptist University]. ProQuest Dissertations & Theses, 2015. 3687670.

Barenberg, J., Berse, T., Reimann, L., & Dutke, S. (2021). Testing and transfer: Retrieval practice effects across test formats in English vocabulary learning in school. *Applied Cognitive Psychology*, *35*(3), 700–710. https://doi.org/10.1002/acp.3796.

Bertilsson F. (2023). *Retrieval practice and individual differences: exploring factors relevant to the benefit and use of retrieval practice* [Doctoral Dissertation, Umeå University]. https://www.diva-portal.org/smash/get/diva2:1801541/FULLTEXT01.pdf

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist C., & Jonsson, B. (2021). Retrieval practice: Beneficial for all students or moderated by individual differences? *Psychology Learning & Teaching*, *20*(1), 21–39. https://doi.org/10.1177/1475725720973494

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *Journal of Cognitive Education and Psychology*, *16*(3), 241–259. https://doi.org/10.1891/1945-8959.16.3.241.

Bing, S. B. (1984). Effects of testing versus review on rote and conceptual learning from prose. *Instructional Science, 13*, 193–198. https://doi.org/10.1007/BF00052385

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). The effect of instruction method and relearning on Dutch spelling performance of third through fifth-graders. *European Journal of Psychology of Education*, *26*, 61–74. http://dx.doi.org/10.1007/s10212-010-0036-3

Brojde, C. L., & Wise, B. W. (2008). An evaluation of the testing effect with third grade students. *Proceedings of the 3-Th Annual Meeting of the Cognitive Science Society*,1362–1367. http://citeseerx.ist.psu.edu/viewdoc/summary?doi 10.1.1.547.2878

Brown, B. R. (1999). *A study on the use of frequent quizzing as a teaching strategy: Does it effect achievement in mathematics?* [Doctoral Dissertation, George Mason University]. ProQuest Dissertations & Theses, 1999. 9919829.

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology*, *28*, 382–391. https://doi.org/10.1002/acp.3008

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology 37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Baars, M., Visser, S., van Gog, T., de Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology*, *38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*, 760–771. http://dx.doi.org/10.1002/acp.1507

Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, *32*, 265–282. http://dx.doi.org/10.1080/09500690802650055

Clayton, M. (1973). *The differential effects of three types of structured reviews on the learning and retention of mathematics* [Doctoral Dissertation, North Carolina State University]. ProQuest Dissertations & Theses, 1973. 7418980.

Curo, D. M. (1963). *An investigation of the influence of daily pre-class testing on achievement in high school American History classes* [Doctoral Dissertation, Purdue University]. ProQuest Dissertations & Theses, 1963. 6404574.

da Silva, F. V., Ekuni, R., & Jaeger, A. (2023). Retrieval practice benefits for spelling performance in fifth-grade children. *Memory*, *31*(9), 1197–1204. https://doi.org/10.1080/09658211.2023.2248420.

Damayanti, A. F., Rahmat, A., & Suwandi, T. (2024). Retrieval practice: Strategy for reducing cognitive anxiety through students' concept mastery and cognitive ability. *Jurnal Inovasi Pendidikan IPA*, 10(2), 120–134. https://doi.org/10.21831/jipi.v10i2.71416

de Lima, N. K., & Jaeger, A. (2020). The effects of prequestions versus postquestions on memory retention in children. *Journal of Applied Research in Memory and Cognition*, *9*(4), 555–563. https://doi.org/10.1016/j.jarmac.2020.08.005.

Deboer, G. E. (1980). Can repeated testing of en route objectives improve end-of-course achievement in high school chemistry? *Science Education*, *64*(2), 141–147.

Dineen, P., Taylor, J., & Stephens, L. (1989). The effect of testing frequency upon the achievement of students in high school mathematics courses. *School Science and Mathematics, 89*(3), 197–200.

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, *6*, 217–226. http://dx.doi.org/10.1016/0361-476X(81)90002-3.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research*, *75*(5), 309–313. https://doi.org/10.1080/00220671.1982.10885400

Evans, D. D. (2013). *Quizzing and retention in the high school science class* [Master's thesis, Louisiana State University], 2013. Lousiana Scholarly Repository, https://digitalcommons.lsu.edu/gradschool_theses/3780

Gates, A. I., (1931). An experimental comparison of the study-test and test-study methods in spelling. *Journal of Educational Psychology*, *22*(1), 1–19. https://doi.org/10.1037/h0075394

Gates, A. I., & Bennett, C. C. (1933). Two tests versus three tests weekly in teaching spelling. *Elementary School Journal*, *34*, 44–50. https://doi.org/10.1086/456976

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., & Tabbers, H. K. (2014). The effect of retrieval practice in primary school vocabulary learning. *Applied Cognitive Psychology*, *28*(1), 135–142. https://doi.org/10.1002/acp.2956

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multiclassroom study. *Applied Cognitive Psychology*, *30*, 700–712. http://dx.doi.org/10.1002/acp.3245

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory & Cognition*, *3*, 177–182. http://dx.doi.org/10.1016/j.jarmac.2014.05.003.

Güven, U. (2017). *The relationship between testing frequency and student achievement in eighth-grade mathematics: An international comparative study based on TIMSS 2011* [Doctoral Dissertation, Duquesne University]. ProQuest Dissertations & Theses, 2017.10263961.

Güven U. (2021). Testing frequency and student's mathematics achievement relationship. *Third Sector Social Economic Review*, *56*(3), 1508–1521. https://doi.org/10.15659/3.sektor-sosyal-ekonomi.21.08.1577.

Hanham J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology*, *31*(3), 265–280. https://doi.org/10.1002/acp.3324

Hayes-Roth, B. (1976). *Effects of repetitions and questions at varying lags during self-paced learning from text*. Rand Corp.

Hirschman B. (2017). *The effects of daily quizzes on student achievement in a chemistry class* [Unpublished Doctoral Dissertation]. Montana State University.

Hooley, D., & Thorpe, J. (2017). The effects of formative reading assessments closely linked to classroom texts on high school reading comprehension. *Educational Technology Research & Development*, *65*(5), 1215–1238. https://doi.org/10.1007/s11423-017-9514-5

Hsu C. (2014). *The effects of principle-based information on the sequence of pairing worked examples and problems in physics learning* [Unpublished Doctoral Dissertation]. School of Education, Faculty of Arts and Social Sciences, University of New South Wales. https://doi.org/10.26190/unsworks/16951.

Jakobsson, A., Loberg, J., & Kjörk, M. (2024). Retrieval-based learning versus discussion; which review practice will better enhance primary school students' knowledge of scientific content? *International Journal of Science Education*, *46*(12), 1216–1238. https://doi.org/10.1080/09500693.2023.2283906

Jewell, J. (2003). *The utility of curriculum-based measurement writing indices for progress monitoring and intervention* [Doctoral Dissertation, Northern Illinois University]. ProQuest Dissertations & Theses, 2003. 3102754.

Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., & Rickard, T. C. (2016). Beyond the rainbow: Retrieval practice leads to better spelling than does rainbow writing. *Educational Psychology Review*, *28*, 385–400. https://doi.org/10.1007/s10648-015-9330-6

Jones, E. M. (1984). *An experimental comparison of the test-study-test and three-test methods for teaching spelling in the fifth grade* [Doctoral Dissertation, The University of Wisconsin-Milwaukee]. ProQuest Dissertations & Theses, 1984. 8509257.

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology*, *113*(5), 972–985. https://doi.org/10.1037/edu0000627

Jägerskog, A. S. (2015). *Pictures and a thousand words: Learning psychology through visual illustrations and testing* [Licentiate thesis, Stockholm University]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A883530&dswid=8179.

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*, *7*, Article 350. https://doi.org/10.3389/fpsyg.2016.00350

Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory & Cognition*, *3*, 198–206. https://doi.org/10.1016/j.jarmac.2014.07.008

Khalaf, A. S. S. (1989). *The effects of classroom testing frequency on student achievement in tenth-grade biology in Saudi Arabia* [Doctoral Dissertation, Kansas State University]. ProQuest Dissertations & Theses, 1989. 9005063.

Khalaf, A. S. S, & Hanna, G. S. (1993). The impact of classroom testing frequency on high school students' achievement. *Contemporary Educational Psychology*, *17*(1), 71–77. https://doi.org/10.1016/0361-476X(92)90047-3

Knowles, N. P. (2011). *The relationship between timed drill practice and the increase of automaticity of basic multiplication facts for regular education sixth graders* [Doctoral Dissertation, Walden University]. ProQuest Dissertations & Theses, 2010. 3427303.

Lavy, V. (2024). *The effect of multitasking on educational outcomes and academic dishonesty*. NBER Working Paper no. w31699, National Bureau of Economic Research, Cambridge, MA.

Leahy, W., Hanham, J., & Sweller, J (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review 27*(2), 291–304. https://doi.org/10.1007/s10648-015-9296-4

Leggett, J. M. I., Burt, J. S., & Carroll, A. (2019). Retrieval practice can improve classroom review despite low practice test performance. *Applied Cognitive Psychology*, *33*, 759–770. https://doi.org/10.1002/acp.3517

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory & Cognition*, *3*, 171–176. https://doi.org/10.1016/j.jarmac.2014.04.002

Mallette, G. S., (2003) *The effects of cognitive load expectancy on fernald-assisted spelling instruction using high-frequency words* [Doctoral Dissertation Mississippi State University]. ProQuest Dissertations & Theses, 2003.3080201.

Maloney, E. L., & Ruch, G. M. (1929). The use of objective tests in teaching as illustrated by grammar. *The School Review*, *3*, 62–66.

Martinez, E. Jr. (2016). *The effects of classroom assessment frequency on common core state standards-aligned benchmarks for second-grade English learners* [Doctoral Dissertation, Keiser University]. ProQuest Dissertations & Theses, 2016.10248080.

Mayer, V. J., & Rojas, C. A. (1982). The effect of frequency of testing upon the measurement of achievement in an intensive time-series design. *Journal of Research in Science Teaching*, *19*(7), 543–551. https://doi.org/10.1002/tea.3660190703

McDaniel, M. A., McDermott, K. B., Agarwal, P. K., & Roediger, H. L. (2008). *Test-enhanced learning in the classroom: The Columbia Middle School project*. Poster presented at the meeting of the Institute of Education Sciences Research, Washington, DC.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L. III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*, 3–21. https://doi.org/10.1037/xap0000004

Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. (1982). Frequency of measurement and data utilization as factors in standardized behavioral assessment of academic skill. *Journal of Behavioral Assessment*, *4*(4), 361–70. https://doi.org/10.1007/BF01341230.

Mirkin, P. K., & Deno, S. L. (1979). *Formative evaluation in the classroom: An approach to improving instruction*. https://eric.ed.gov/?id=ED185754.

Mirkin, P. K., & Deno, S. L. (1980). *The effects of selected variations in the components of formative evaluation to improved academic performance*. https://eric.ed.gov/?id=ED186489.

Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science*, *44*, 567–581. https://doi.org/10.1007/s11251-016-9393-x

Molin, F., Cabus, S., Haelermans, C., & Groot, W. (2021). Toward reducing anxiety and increasing performance in physics education: Evidence from a randomized experiment. *Research in Science Education*, *51*, 233–249. https://doi.org/10.1007/s11165-019-9845-9

Moreira, B. F.T., Pinto, T. S. S., Justi, F. R. R., & Jaeger, A. (2019). Retrieval practice improves learning in children with diverse visual word recognition skills. *Memory*, *27*(10), 1423–1437. https://doi.org/10.1080/09658211.2019.1668017

Nejati R. (2016). The durability of the effect of the frequent quizzes on Iranian high school students' vocabulary learning. *International Journal of the Humanities*, *23*, 29–42.

Norton, C. B. (2013). *The effect of frequent quizzing on student learning in a high school physical science classroom* [Master's thesis, Louisiana State University]. ProQuest Dissertations & Theses, 2013. 29124133.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

Ojugo, A. A., Ugboh, E., Onochie, C. C., Eboka, A. O., Yerokun, M. O., & Iyawa, I. J. B. (2013). Effects of formative test and attitudinal types on students' achievement in mathematics in Nigeria. *African Educational Research Journal*, *1*(2), 113–117.

Ophuis-Cox, F. H. A., Rozendal, L., Catrysse, L., Joosten-ten Brinke, D., & Camp, G. (2024). The effects of summarization and factual retrieval practice on text comprehension and text retention in elementary education. *Journal of Experimental Psychology: Applied*, *30*(2), 258–267. https://doi.org/10.1037/xap0000507

Ortega-Tudela, J. M., Lechuga, M-T., Bermúdez-Sierra, M., & Gómez-Ariza, C. J. (2021). Testing the effectiveness of retrieval-based learning in naturalistic school settings. *SAGE Open*, *11*(4). https://doi.org/10.1177/21582440211061569.

Paas F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Ralph, S. M. (2020). *Retrieval practice as a way to enhance learning and transfer in a high school biology classroom* [Unpublished Doctoral Dissertation]. Kansas State University, https://krex.k-state.edu/server/api/core/bitstreams/34daff9e-0f81-4f96-b58d-849fd7a677be/content.

Requa, L. O. (1988). *Interim assessment testing and reading achievement of second grade students in DoDDS - Germany* [Doctoral Dissertation, University of Southern California]. ProQuest Dissertations & Theses, 1988.0562974.

Robershotte, L. A. (1990). *Factual and inferential learning under varying levels of information processing* [Unpublished manuscript]. Department of Educational Technology, Arizona State University.

Robertson, W. L. (2010). *The impact of various quizzing patterns on the test performance of high school economics students* [Doctoral Dissertation, Walden University]. ProQuest Dissertations & Theses, 2010.3398381.

Rochester, M. (1982). *The effect of formative assessment and correctives on learning achievement* [Doctoral Dissertation, University of South Carolina]. ProQuest Dissertations & Theses, 1982. 8311357.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Rojas, C. A. (1980). *A study of the effect on frequency of testing upon the measurement of student attitudes toward science class and measurement of achievement on a crustal evolution unit using a time-series design* [Unpublished Doctoral Dissertation]. Ohio State University, http://rave.ohiolink.edu/etdc/view?acc_num=osu1487088320404442

Ruitenburg, S. K., Ackermans, K., Kirschner, P. A., Jarodzka, H., & Camp, G. (2025). After initial acquisition, problem-solving leads to better long-term performance than example study, even for complex tasks. *Learning and Instruction*, *95*, 102027. https://doi.org/10.1016/j.learninstruc.2024.102027.

Rumanová, L., Vallo, D., & Záhorská, J. (2020). The impact of formative assessment on results of secondary school pupils in mathematics: One case of schools in Slovakia. *TEM Journal*, *9*(3), 1200–1207. https://doi.org/10.18421/TEM93-47

Scannell, D. P., & Haugh, O. M. (1968). *Teaching composition skills with weekly multiple choice tests in lieu of theme writing*. Office of Education (OHEW), Washington, D.C. Bureau of Research, BR-6-813.

Schreiner, R. L. (1973). *Verbal coding as an instructional strategy in improving pupil performance on standardized measures of reading comprehension* [Unpublished manuscript]. University of Minnesota.

Shirvani, H. (2009). Examining an assessment strategy on high school mathematics achievement: Daily quizzes vs. weekly tests. *American Secondary Education*, 34–45.

Stecker, P. M. (1993). *Effects of instructional modifications with and without curriculum-based measurement on the mathematics achievement of students with mild disabilities* [Doctoral Dissertation, Vanderbilt University]. ProQuest Dissertations & Theses, 1993.9416516.

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710–1727. http://dx.doi.org/10.1080/01443410.2014.953037.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3

Tindal, G. (1981). *The relationship between student achievement and teacher assessment of short- or long-term goals*. Office of Special Education and Rehabilitative Services (ED), Washington, DC. IRLD-RR-61.

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. http://dx.doi.org/10.1016/j.cedpsych.2010.10.004.

Wiklund-Hörnqvist, C., Stillesjö, S., Andersson, M., Jonsson, B., & Nyberg, L. (2022). Retrieval practice is effective regardless of self-reported need for cognition – Behavioral and brain imaging evidence. *Frontiers in Psychology*, *12*, Article 797395. https://doi.org/doi: 10.3389/fpsyg.2021.797395.

Zgraggen, F. D. *The effects of frequent testing in the mathematics classroom* [Unpublished Master's Thesis]. The Graduate School, University of Wisconsin-Stout. https://minds.wisconsin.edu/bitstream/handle/1793/43403/2009zgraggenf.pdf?sequence=1

References to excluded studies

Ackermans, K., Rusman, E., Nadolski, R., Specht, M., & Brand-Gruwel, S. (2019). Video-or text-based rubrics: What is most effective for mental model growth of complex skills within formative assessment in secondary schools? *Computers in Human Behavior*, *101*, 248-258. https://doi.org/10.1016/j.chb.2019.07.011

Alitto, J. M. (2008). *The effects of peer-mediated goal setting and performance feedback on curriculum-based measurement indices of written expression* (Doctoral dissertation). Northern Illinois University.

Abu-Hamour, B., & Mattar, J. (2013). The applicability of curriculum-based-measurement in math computation in Jordan. *International Journal of Special Education*, *28*(1), 111-119.

Baars, M., Leopold, C., & Paas, F. (2018). Self-explaining steps in problem-solving tasks to improve self-regulation in secondary education. *Journal of Educational Psychology*, *110*(4), 578-595. https://doi.org/10.1037/edu0000223

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, *33*, 92-107. https://doi.org/10.1016/j.learninstruc.2014.04.004

Beyers, S. J., Lembke, E. S., & Curs, B. (2013). Social studies progress monitoring and intervention for middle school students. *Assessment for Effective Intervention*, *38*(4), 224-235.

Capizzi, A. M., & Fuchs, L. S. (2005). Effects of curriculum-based measurement with and without diagnostic feedback on teacher planning. *Remedial and Special Education*, *26*(3), 159-174. https://doi.org/10.1177/07419325050260030401

Cavanaugh, R. A., Heward, W. L., & Donelson, F. (1996). Effects of response cards during lesson closure on the academic performance of secondary students in an earth science course. *Journal of Applied Behavior Analysis*, *29*(3), 403-406. https://doi.org/10.1901/jaba.1996.29-403

Cohonner, A., & Mayer, J. (2018). Retrieval-based learning in the context of inquiry-based learning. In Gericke, N. & Grace, M. (eds.) *Challenges in Biology Education Research - A selection of papers presented at the XIth conference of European Researchers in Didactics of Biology (ERIDOB)*, pp. 273-287.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991). The contribution of skills analysis to curriculum-based measurement in spelling. *Exceptional Children*, *57*(5), 443-452. https://doi.org/10.1177/001440299105700507

Gustafson, S., Nordström, T., Andersson, U. B., Fälth, L., & Martin, I. (2019). Effects of a formative assessment system on early reading development. *Education*, *40*(1), 17-27.

Hamilton, M. D. (2013). *Using formative reading assessments and data utilization to improve ELL Spanish speaking students' achievement test scores* (Doctoral dissertation). Wilmington University (Delaware).

Im, H., Kwon, K. A., Jeon, H. J., & McGuire, P. (2020). The school-level standardized testing policy and math achievement in primary grades: The mediational role of math instructional approach. *Studies in Educational Evaluation*, *66*, 100877. https://doi.org/10.1016/j.stueduc.2020.100877

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology*, *35*(4), 513-521. https://doi.org/10.1080/01443410.2014.963030

January, S. A. A., Van Norman, E. R., Christ, T. J., Ardoin, S. P., Eckert, T. L., & White, M. J. (2018). Progress monitoring in reading: Comparison of weekly, bimonthly, and monthly assessments for students at risk for reading difficulties in grades 2–4. *School Psychology Review*, *47*(1), 83-94. https://doi.org/10.17105/SPR-2017-0009.V47-1

Jürges, H., Schneider, K., Senkbeil, M., & Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, *31*(1), 56-65. https://doi.org/10.1016/j.econedurev.2011.08.007

Karuza, H. (2014). *A math intervention model for middle school: How the combination of formative assessment, feedback, academic vocabulary, and word problems affect student achievement in mathematics* (Doctoral dissertation). The Claremont Graduate University.

Lang, L. B., Schoen, R. R., LaVenia, M., & Oberlin, M. (2014). Mathematics formative assessment system – common core state standards: A randomized field trial in kindergarten and first grade. Conference paper, *Society for Research on Educational Effectiveness*.

Mintert, A. L. (2019). *The Effects of Formative Assessment on Student Motivation for Learning and Achievement in Standards-Based Grading* (Doctoral dissertation). Evangel University.

Mitchell, D. S. (2009). *The impact of communalism and testing on recall among African-American elementary school students* (Master's thesis). Howard University

Naseem, A. (2021). Effect of Quizzes on Anxiety and Performance in Mathematics at Middle Level. *Bulletin of Education and Research*, *43*(1), 59-75.

Reimer, C. K. (2019). *The effect of retrieval practice on vocabulary learning for children who are deaf or hard of hearing* (Doctoral dissertation). Washington University in St. Louis.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233 –239. https://doi.org/10.1037/a0017678

Supovitz, J. (2016). *Overview of the OGAP Formative Assessment Project and CPRE's large-scale experimental study of implementation and impacts*. Conference paper, Society for Research on Educational Effectiveness.

Tempel, T., & Sollich, S. (2023). Retrieval-based learning in special education. *Journal of Research in Special Educational Needs*, *23*(3), 244-250. https://doi.org/10.1111/1471-3802.12594

Wößmann, L. (2002). How central Exams affect educational achievement: International evidence from TIMSS and TIMSS-Repeat. Paper presented at the "Taking Account of Accountability: Assessing Politics and Policy" Conference (Cambridge, MA, June 10-11, 2002).

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., ... & Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education*, *21*(4), 335-359. https://doi.org/10.1080/08957340802347845

Additional references

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), 265–296. https://doi.org/10.3982/ECTA12675

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Aloe, A. M., Dewidar, O., Hennessy, E. A., Pigott, T., Stewart, G., Welch, V., Wilson, D. B., & Group, C. M. W. (2024). Campbell Standards: Modernizing Campbell's Methodologic Expectations for Campbell Collaboration Intervention Reviews (MECCIR). *Campbell Systematic Reviews*, *20*(4), e1445. https://doi.org/10.1002/cl2.1445

Andrews, I., & Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, *109*(8), 2766–2794. https://doi.org/10.1257/aer.20180310

Azin, M., & Resendez, M. G. (2008). Measuring student progress: Changes and challenges under No Child Left Behind. *New Directions for Evaluation¸ 117*, 71–84. https://doi.org/10.1002/ev.253

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, *94*(7–8), 435–452. https://doi.org/10.1016/j.jpubeco.2010.04.001

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, *21*(2), 55–97. https://doi.org/10.1177/1529100620915848

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*(2), 85–99. https://doi.org/10.1080/00220671.1991.10702818

Bergbauer, A. B., Hanushek, E. A., & Woessmann, L. (2018). *Testing*. NBER Working Paper no. w24836. National Bureau of Economic Research, Cambridge, MA.

Bernatzky, M., Cabrera, J. M., & Cid, A. (2017). Frequency of testing. Lessons from a field experiment in higher education. *Journal of Economics and Economic Education Research*, *19*(1), 1–11.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons Ltd.

Borge, T. C., Ames, H., Jardim, P. J., Meneses-Echavez, J. F., Himmels, J., Rose, C., Hestevik, C., & Muller, A. E. (2023). *Implementation of machine learning in evidence syntheses in the Cluster for Reviews and Health Technology Assessments: Final report 2021-2022.* Oslo: Norwegian Institute of Public Health.

Bruhn, J., Gilraine, M., Ludwig, J., & Mullainathan, S. (2025). *Do test scores misrepresent test results? An item-by-item analysis*. NBER Working Paper no. w34484*,* National Bureau of Economic Research, Cambridge, MA.

Buck, S., Ritter Gary, W., Jensen Nathan, C., & Rose Caleb, P. (2010). Teachers say the most interesting things—An alternative view of testing. *Phi Delta Kappan*, *91*(6), 50–54. https://doi.org/10.1177/003172171009100613

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Cheek, J. R., Bradley, L. J., Reynolds, J., & Coy, D. (2002). An intervention for helping elementary students reduce test anxiety. *Professional School Counseling*, *6*(2), 162–164. https://www.jstor.org/stable/42732406

Chen, M., & Pustejovsky, J. E. (2025). Adapting methods for correcting selective reporting bias in meta-analysis of dependent effect sizes. *Psychological Methods*, Advance online publication. https://doi.org/10.1037/met0000773

Coburn, K. M., & Vevea, J. L. (2019). weightr – Estimating weight-function models for publication bias in R. R package version 2.0.2. https://CRAN.R-project.org/package=weightr

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Crooks, T. J. (1988). The impact of evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481. https://doi.org/10.3102/00346543058004438

Dempster, F. N. (1996). Chapter 9—Distributing and managing the conditions of encoding and practice. In E. L. Bjork, & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Elsevier Inc.

DePaepe, J., Matthews, D., Mathias, K., Harris, S., Davison, S., Lee, E., & Braskamp, D. (2015). *CWU teacher time study: How Washington public school teachers spend their work days*. Technical Report, Central Washington University.

de Diego-Lázaro, B. (2024). Retrieval practice and word learning in children who are hard of hearing, *International Journal of Speech-Language Pathology*, *27*(5), 735–748. https://doi.org/10.1080/17549507.2024.2381465

Department for Education (2018). *Childcare and early years survey of parents in England*. Report, Department for Education, UK. https://www.gov.uk/government/statistics/childcare-and-early-years-survey-of-parents-2018

Dietrichson, J., Bhatnagar, R., Filges, T., & Vembye, M. H. (2026). On the mechanisms of intervention effect fade-out: A meta-analytic review of interventions targeting at-risk students' achievement. *Psychological Bulletin*, accepted for publication. https://osf.io/preprints/psyarxiv/2eg67_v1

DordiNejad, F. G., Hakimi, H., Ashouri, M., Dehghani, M., Zeinali, Z., Daghighi, M. S., & Bahrami, N. (2011). On the relationship between test anxiety and academic performance. *Procedia—Social and Behavioral Sciences*, *15*, 3774–3778. https://doi.org/10.1016/j.sbspro.2011.04.372

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(Suppl. 1), 4–58. https://doi.org/10.1177/1529100612453266

Egger, M., Smith, G. D. S., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, N., Onghena, P., & Van den Noortgate, W. (2020). Visual representations of meta-analyses of multiple outcomes: Extensions to forest plots, funnel plots, and caterpillar plots. *Methodology*, *16*(4), 299–315. https://doi.org/10.5964/meth.4013

Fuchs, L. S., & Fuchs, D. (2001). *What is scientifically-based research on progress monitoring?* National Center on Student Progress Monitoring. https://files.eric.ed.gov/fulltext/ED502460.pdf

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*(40). Retrieved from https://archive.org/details/recitationasfact00gaterich/page/n11/mode/2up.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Goldberg, M. H. (2019). How often does random assignment fail? Estimates and recommendations. *Journal of Environmental Psychology*, *66*, 101351. https://doi.org/10.1016/j.jenvp.2019.101351

Green, C. P., Nyhus, O. H., & Salvanes, K. V. (2025). Does testing young children influence educational attainment and wellbeing? *Journal of Population Economics*, *38*(1), 20. https://doi.org/10.1007/s00148-025-01060-z

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2024). Fadeout and persistence of intervention impacts on social–emotional and cognitive skills in children and adolescents: A meta-analytic review of randomized controlled trials. *Psychological Bulletin*, *150*(10), 1207. https://doi.org/10.1037/bul0000450

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255. https://doi.org/10.1214/ss/1177011364

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–70. https://doi.org/10.3102/1076998606298043

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 259-282. https://doi.org/10.1111/bmsp.12296

Hedges, L. V., & Vevea, J. (2005). Selection method approaches. In R. Rothstein Hannah, J. Sutton Alexander, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 145–174). Chichester, England: John Wiley & Sons, Ltd.

Higgins, J. P. T. & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0. The Cochrane Collaboration.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher?. *Labour Economics*, *15*(1), 37–53. https://doi.org/10.1016/j.labeco.2006.12.002

Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, *7*, 24–51. https://doi.org/10.1093/jleo/7.special_issue.24

IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: a meta-meta-analysis. *Journal of Clinical Epidemiology*, *68*(8), 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Jones, H. E. (1923). *Experimental studies of college teaching: The effect of examination on permanence of learning* [Unpublished Doctoral Dissertation]. Columbia University.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401–418. https://doi.org/10.1007/s10648-012-9202-2

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology*, *25*(6), 427–436.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kulp, D. H. (1933). Weekly tests for graduate students. *School and Society*, *38*(970), 157–159.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

Ma, X., Li, T., Duzi, K., Li, Z. Y., Ma, X., Li, Y., & Zhou, A. B. (2020). Retrieval practice promotes pictorial learning in children aged six to seven years. *Psychological Reports*, *123*(6), 2085-2100. https://doi.org/10.1177/0033294119856553

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, *14*(2), 200–206. https://doi.org/10.3758/BF03194052

McShane, B. B., Bäckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/174569161666224

National Research Council. 2011. *Incentives and test-based accountability in education*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12521

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., ... & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885-958. https://doi.org/10.1037/bul0000153

Organization for Economic Co-operation and Development. (2017). *PISA 2015 results (Volume III): Students' well-being*. OECD Publishing.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910-2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330–342. https://doi.org/10.1002/jrsm.1354

Pustejovsky, J. E., Citkowicz, M., & Joshi, M. (2025). *Estimation and inference for step-function selection models in meta-analysis with dependent effects* [Unpublished manuscript]. MetaArXiv, https://doi.org/10.31222/osf.io/qg5x6_v1

Pustejovsky, J. E., Joshi, M., & Citkowicz, M. (2025). *metaselection: Meta-analytic selection models with cluster-robust and cluster-bootstrap standard errors for dependent effect size estimates* (Version 0.1.5) [R]. https://github.com/jepusto/metaselection

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, *10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*, 425–438. https://doi.org/10.1007/s11121-021-01246-3

Pustejovsky, J. E., Zhang, J., & Tipton, E. (2025). *A preliminary data analysis workflow for meta-analysis of dependent effect sizes*. MetaArXiv, https://doi.org/10.31222/osf.io/vfsqx_v1

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76*(1), 85–97. https://doi.org/10.1037/0022-0663.76.1.85

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PLoS One*, *8*(11), e78976. https://doi.org/10.1371/journal.pone.0078976

Roberts, C., & Torgerson, D. J. (1999). Baseline imbalance in randomised controlled trials. *BMJ*, *319*(7203), 185. https://doi.org/10.1136/bmj.319.7203.185

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

Roediger, H. L. III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L. III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 181–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rowley, T., & McCrudden, M. T. (2020). Retrieval practice and retention of course content in a middle school science classroom. *Applied Cognitive Psychology*, *34*(6), 1510–1515. https://doi.org/10.1002/acp.3710

Russell, D., & McAuley, E. (1986). Causal Attributions, causal dimensions, and affective reactions to success and failure. *Journal of Personality and Social Psychology*, *50*(6), 1174–1185. https://doi.org/10.1037/0022-3514.50.6.1174

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. *Journal of Educational Change*, *11*(1), 45–61. https://doi.org/10.1007/s10833-008-9098-2

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

Segool, N. K., Carlson, J. S., Goforth, A. N., Embse, N. V. D., & Barerian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, *50*(5), 489–499. https://doi.org/10.1002/pits.21689

Smith, W. C. (2016). *The global testing culture*. Symposium Books.

Snedden, D. (1931). Practice effect. *Journal of Educational Research*, *24*(5), 376–380. https://doi.org/10.1080/00220671.1931.10880225

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. https://doi.org/10.1037/h0063404

Standlee, L. S., & Popham, W. J. (1960). Quizzes' contribution to learning. *Journal of Educational Psychology*, *51*(6), 322–325. https://doi.org/10.1037/h0048442

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

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.… Higgins, J. P. T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J, Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Stoll, C. R. T., Izadi, S., Fowler, S., Green, P., Suls, J., & Colditz Graham, A. (2019). The value of a second reviewer for study selection in systematic reviews. *Research Synthesis Methods*, *10*(4), 1–7. https://doi.org/10.1002/jrsm.1369

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

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Turner, R. M., Bird, S. M., & Higgins, J. P. (2013). The impact of study size on meta-analyses: Examination of underpowered studies in Cochrane reviews. *PloS One*, *8*(3), e59202. https://doi.org/10.1371/journal.pone.0059202

Turney, A. H. (1931). The effect of frequent short objective tests upon the achievement of college students in educational psychology. *School and Society*, *33*(858), 760–762.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(1), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & López-López, J. A. (2022). Location-scale models for meta-analysis. *Research Synthesis Methods*, *13*(6), 697-715. https://doi.org/10.1002/jrsm.1562

von der Embse, N., & Hasson, R. (2012). Test anxiety and high-stakes test performance between school settings: Implications for educators. *Preventing School Failure*, *56*(3), 180–187. https://doi.org/10.1080/1045988X.2011.633285

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist*, *45*(1), 28–36. https://doi.org/10.1080/00461520903433596

Williams, D. R., Rodriguez, J. E., & Bürkner, P. C. (2021). *Putting variation into variance: modeling between-study heterogeneity in meta-analysis* [Unpublished manuscript]. https://files.osf.io/v1/resources/9vkqy_v1/providers/osfstorage/60a2793cd137cb00f5f6545d?action=download&direct&version=5

Wilson, D. B. (2017). *Formulas used by the "Practical Meta-Analysis Effect Size Calculator"* [Unpublished manuscript]. https://mason.gmu.edu/~dwilsonb/downloads/esformulas.pdf

Wolf, B., & Harbatkin, E. (2022). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134–161. https://doi.org/10.1080/19345747.2022.2071364

Wolters, C. A., & Pintrich, P. R. (1998). Contextual differences in student motivation and self-regulated learning in mathematics, English, and social studies classrooms. *Instructional Science*, *26*, 27–47. https://doi.org/10.1023/A:1003035929216

World Bank (2017). *World development report 2018: Learning to realize education's promise*. World Bank, https://www.worldbank.org/en/publication/wdr2018.

World Bank (2025). *World Bank country classifications by income level for 2024-2025*. https://blogs.worldbank.org/en/opendata/world-bank-country-classifications-by-income-level-for-2024-2025.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Li, J., Zhao, W., Luo, L., & Shanks, D. R. (2023). Do practice tests (quizzes) reduce or provoke test anxiety? A meta-analytic review. *Educational Psychology Review*, *35*(3), 87. https://doi.org/10.1007/s10648-023-09801-w

Zeidner, M. (1998). *Test anxiety: The state of the art*. Springer Nature.

Online supplemental material

Supplementary Information Appendix A: Search Documentation

Supplementary Information Appendix B: Extra Search

Supplementary Information Appendix A: Search documentation

This appendix provides documentation of our searches. First, we provide the full searches and results per electronic database in the initial search in 2020, the updated search in 2023, and the extra search in 2024. Second, we provide documentation for the searches of other sources: hand searches of relevant journals; searches for dissertations, searches for grey literature, and searches for systematic reviews; citation tracking; and contacts to experts in the field.

Initial search

SocIndex. EBSCO-host. 17/11/2020. No limiters.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	144
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	824,950
S10	S7 OR S8 OR S9	713,042
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	530,998
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	316,719
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	93,091
S6	S4 OR S5	64,458
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	307
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	64,193
S3	S1 OR S2	714
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	59
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	656

ERIC. 17/11/2020. EBSCO. NO LIMITERS.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	1,692
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	1,014,635
S10	S7 OR S8 OR S9	1,215,812
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	998,561
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	709,977
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	261,886
S6	S4 OR S5	66,758
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,065
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	64,170
S3	S1 OR S2	3,078
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	847
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	2,261

ACADEMIC SEARCH. 17/11/2020. EBSCO. NO LIMITERS.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	2,230
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	13,251,377
S10	S7 OR S8 OR S9	4,291,122
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	2,578,571
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	2,172,651
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	469,499
S6	S4 OR S5	1,086,497
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	2,648
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	1,084,422
S3	S1 OR S2	15,801
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	732
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	15,115

PSYCHINFO. 17/11/20. EBSCO. NO LIMITERS.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	1,863
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	2,814,723
S10	S7 OR S8 OR S9	1,705,303
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	1,426,359
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	669,388
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	195,758
S6	S4 OR S5	340,004
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,218
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	337,481
S3	S1 OR S2	5,554
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	1,001
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	4,587

TEACHER REFERENCE CENTER. 17/11/2020. EBSCO. NO LIMITERS.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	434
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	270,689
S10	S7 OR S8 OR S9	485,972
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	335,338
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	283,563
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	90,369
S6	S4 OR S5	9,418
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	727
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	8,786
S3	S1 OR S2	792
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	252
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	549

ECONLIT. 17/11/20. EBSCO. NO LIMITERS.

#	Query	Results
S12	S3 AND S6 AND S10 AND S11	31
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	611,408
S10	S7 OR S8 OR S9	140,225
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	65,034
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	83,889
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	22,579
S6	S4 OR S5	34,944
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	42
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	34,909
S3	S1 OR S2	616
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	21
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	600

INDEXES=SCI-EXPANDED TIMESPAN=ALL YEARS 17/11/20. WEB OF KNOWLEDGE.

Set	Results	Save History / Create AlertOpen Saved History
# 12	**1,991**	#11 AND #10 AND #6 AND #3
# 11	**19,355,610**	TI=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR at tain* OR success* OR “test* anxiety” OR wellbeing) OR AB=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR “test* anxiety” OR wellbeing)
# 10	**3,818,645**	#9 OR #8 OR #7
# 9	**1,803,213**	TI=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*)
# 8	**2,113,026**	AB=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB=(class* OR school* OR kindergarten* OR “primary education” OR “secondary e ducation”)
# 7	**374,780**	TI=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI=(class* OR school* OR kindergarten* OR “primary education” OR “secondary ed ucation”)
# 6	**1,790,251**	#5 OR #4
# 5	**1,248**	AB=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")
# 4	**1,789,371**	AB=(test* OR assess* OR measur* OR exam* OR quiz*) AND AB=(frequen* OR repeat* OR i nterim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)
# 3	**27,745**	#2 OR #1
# 2	**370**	TI=(“test enhanced learning” OR “testing effect*” OR “testing phenomenon” OR “frequency o f testing” OR “cumulative testing” OR “progress monitoring” OR “curriculum-based”)
# 1	**27,395**	TI=(test* OR assess* OR measur* OR exam* OR quiz*) AND TI=(frequen* OR repeat* OR in terim* OR formative* OR dail* OR weekl* OR monthl* OR annual*) *Indexes=SCI-EXPANDED Timespan=All years*

INDEXES=SSCI-EXPANDED TIMESPAN=ALL YEARS 17/11/20. WEB OF KNOWLEDGE.

Set	Results	Save History / Create AlertOpen Saved History
# 12	**1,303**	#11 AND #10 AND #6 AND #3 *Indexes=SSCI Timespan=All years*
# 11	**2,691,431**	TI=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) *Indexes=SSCI Timespan=All years*
# 10	**1,252,896**	#9 OR #8 OR #7 *Indexes=SSCI Timespan=All years*
# 9	**974,902**	TI=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) *Indexes=SSCI Timespan=All years*
# 8	**466,400**	AB=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB=(class* OR school* OR kindergarten* OR "primary education" OR "secondary education") *Indexes=SSCI Timespan=All years*
# 7	**163,272**	TI=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI=(class* OR school* OR kindergarten* OR "primary education" OR "secondary education") *Indexes=SSCI Timespan=All years*
# 6	**298,937**	#5 OR #4 *Indexes=SSCI Timespan=All years*
# 5	**1,724**	AB=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based") *Indexes=SSCI Timespan=All years*
# 4	**297,592**	AB=(test* OR assess* OR measur* OR exam* OR quiz*) AND AB=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*) *Indexes=SSCI Timespan=All years*
# 3	**5,307**	#2 OR #1 *Indexes=SSCI Timespan=All years*
# 2	**613**	TI=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based") *Indexes=SSCI Timespan=All years*
# 1	**4,710**	TI=(test* OR assess* OR measur* OR exam* OR quiz*) AND TI=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*) *Indexes=SSCI Timespan=All years*

SOCIOLOGICAL ABSTRACTS. 17/11/20. PROQUEST. NO LIMITERS.

Set	Search	
S12	S3 AND S6 AND S10 AND S11	**78**
S11	ti(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR ab(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	**742,031**
S10	S7 OR S8 OR S9	**562,701**
S9	ti(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR ab(student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	**424,247**
S8	ab((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12))) OR ab(class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	**223,892**
S7	ti((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12))) OR ti(class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	**75,148**
S6	S4 OR S5	**52,534**
S5	ab("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	**172**
S4	ab(test* OR assess* OR measur* OR exam* OR quiz*) AND ab(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	**52,383**
S3	S1 OR S2	**382**
S2	ti("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	**19**
S1	ti(test* OR assess* OR measur* OR exam* OR quiz*) AND ti(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	**363**

PROQUEST DISSERTATIONS & THESES GLOBAL. 17/11/2020. PROQUEST. NO LIMITERS.

Set	Search	Results
S12	S3 AND S6 AND S10 AND S11	**885**
S11	ti((achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)) OR ab((achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing))	**2,650,794**
S10	S7 OR S8 OR S9	**1,063,932**
S9	ti((student* OR pupil* OR child* OR adolescen* OR youth* OR young*)) OR ab((student* OR pupil* OR child* OR adolescen* OR youth* OR young*))	**659,122**
S8	ab(((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)))) OR ab((class* OR school* OR kindergarten* OR "primary education" OR "secondary education"))	**623,976**
S7	ti(((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)))) OR ti((class* OR school* OR kindergarten* OR "primary education" OR "secondary education"))	**210,868**
S6	S4 OR S5	**308,008**
S5	ab(("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based"))	**1,954**
S4	ab(test* OR assess* OR measur* OR exam* OR quiz*) AND ab(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	**306,508**
S3	S1 OR S2	**2,637**
S2	ti(("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based"))	**443**
S1	ti((test* OR assess* OR measur* OR exam* OR quiz*)) AND ti((frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*))	**2,209**

EBSCO Open Dissertations. EBSCO. 19/11/2020. No limiters.

Search	Terms	
S13	S10 AND S11 AND S12	(222)
S12	S3 AND S6	(476)
S11	S7 OR S8 OR S9	(232,092)
S10	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	(559,264)
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	(138,226)
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	(125,992)
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	(47,619)
S6	S4 OR S5	(69,879)
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	(346)
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	(69,626)
S3	S1 OR S2	(819)
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	(92)
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	(734)

Updated search

EBSCO Research Databases: SocINDEX with Full Text
Limiters - Published Date: 20201101-20231231 Search date - 20230829

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	19
S12	S3 AND S6 AND S10 AND S11	151
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	864,591
S10	S7 OR S8 OR S9	730,137
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	543,432
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	326,898
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	95,260
S6	S4 OR S5	65,736
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	313
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	65,467
S3	S1 OR S2	725
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	63
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	664

EBSCO Research Databases: ERIC
Limiters - Published Date: 20201101-20231231 Search date - 20230829

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	271
S12	S3 AND S6 AND S10 AND S11	2,030
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	1,122,736
S10	S7 OR S8 OR S9	1,324,574
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	1,093,338
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	775,061
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	286,870
S6	S4 OR S5	74,994
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,410
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	72,116
S3	S1 OR S2	3,526
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	934
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	2,623

EBSCO Research Databases: Academic Search Premier
Limiters - Published Date: 20201101-20231231 Search date - 20230830

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S9 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	156
S12	S3 AND S6 AND S9 AND S10 AND S11	765
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	16,390,661
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	2,984,953
S9	S7 OR S8	2,738,837
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	2,606,015
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	549,247
S6	S4 OR S5	1,361,530
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,247
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	1,359,015
S3	S1 OR S2	19,442
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	865
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	18,629

EBSCO Research Databases: APA PsycINFO
Limiters - Published Date: 20201101-20231231 Search date - 20230830

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S9 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	145
S12	S3 AND S6 AND S9 AND S10 AND S11	1,024
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	3,221,270
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	1,603,144
S9	S7 OR S8	794,263
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	754,347
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	219,653
S6	S4 OR S5	393,042
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,730
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	390,115
S3	S1 OR S2	6,407
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	1,124
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	5,321

EBSCO Research Databases: Teacher Reference Center
Limiters - Published Date: 20201101-20231231 Search date - 20230830

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S9 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	32
S12	S3 AND S6 AND S9 AND S10 AND S11	235
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	291,267
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	355,497
S9	S7 OR S8	318,752
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	297,571
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	95,831
S6	S4 OR S5	10,900
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	796
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	10,210
S3	S1 OR S2	884
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	268
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	625

EBSCO Research Databases: EconLit
Limiters - Published Date: 20201101-20231231 Search date - 20230830

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S3 AND S6 AND S9 AND S10 AND S11 Limiters - Published Date: 20201101-20231231	1
S12	S3 AND S6 AND S9 AND S10 AND S11	7
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	735,132
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	78,147
S9	S7 OR S8	109,435
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	98,973
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	25,318
S6	S4 OR S5	43,820
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	46
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	43,782
S3	S1 OR S2	712
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	24
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	693

Sociological Abstracts searched through the ProQuest interface
Limiters - Published Date: 20201101-20231231 Search date - 20230830

Search History

#	Query	Results
S12	[S3] AND [S6] AND [S9] AND [S10] AND [S11]	5
S11	title(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR abstract(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) AND pd(20201101-20231231)	82,200
S10	title((student* OR pupil* OR child* OR adolescen* OR youth* OR young*)) OR abstract((student* OR pupil* OR child* OR adolescen* OR youth* OR young*)) AND pd(20201101-20231231)	40,707
S9	[S7] OR [S8]	19,642
S8	abstract(((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)))) OR abstract((class* OR school* OR kindergarten* OR "primary education" OR "secondary education")) AND pd(20201101-20231231)	18,884
S7	title(((grade* NEAR/1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)))) OR title((class* OR school* OR kindergarten* OR "primary education" OR "secondary education")) AND pd(20201101-20231231)	5,622
S6	[S4] OR [S5]	7,049
S5	abstract("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based") AND pd(20201101-20231231)	25
S4	abstract(test* OR assess* OR measur* OR exam* OR quiz*) AND abstract(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*) AND pd(20201101-20231231)	7,031
S3	[S1] OR [S2]	65
S2	title("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based") AND pd(20201101-20231231)	3
S1	(title(test* OR assess* OR measur* OR exam* OR quiz*) AND title(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)) AND pd(20201101-20231231)	62

Social Science Citation Index & Science Citation Index Expanded searched through the Web of Science platform
Limiters - Publication Date: 20201101-20231231 Search date - 20230831

Search History

#	Query	Results
#13	#3 AND #6 AND #9 AND #10 AND #11 AND Publication Date: 20201101-20231231	176*
#12	#3 AND #6 AND #9 AND #10 AND #11	678
#11	TI=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	82,200
#10	TI=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	2,910,314
#9	#7 OR #8	3,206,529
#8	AB=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB=(class* OR school* OR kindergarten* OR “primary education” OR “secondary education”)	3,001,805
#7	TI=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI=(class* OR school* OR kindergarten* OR “primary education” OR “secondary education”)	603,119
#6	#4 OR #5	2,330,979
#5	AB=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,239
#4	AB=(test* OR assess* OR measur* OR exam* OR quiz*) AND AB=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	2,328,528
#3	#1 OR #2	19,111
#2	TI=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	1,082
#1	TI=(test* OR assess* OR measur* OR exam* OR quiz*) AND TI=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	18,045

* The RIS-file generated from this database contained only 50 hits, even though the number listed in #13 is 176. We are unable to determine the cause of this discrepancy, but have counted only 50 references from this database in our flowchart, since these are the references that were possible to import and screen in EPPI-Reviewer at the time of the search.

ProQuest Dissertations & Theses Citation Index searched through the Web of Science platform
Limiters - Publication Date: 20201101-20231231 Search date - 20230831

Search History

#	**Query**	**Results**
#13	#3 AND #6 AND #9 AND #10 AND #11 AND Publication Date: 20201101-20231231	46
#12	#3 AND #6 AND #9 AND #10 AND #11	647
#11	TI=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB=(achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	3,063,199
#10	TI=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB=(student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	758,162
#9	#7 OR #8	854,922
#8	AB=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB=(class* OR school* OR kindergarten* OR “primary education” OR “secondary education”)	774,108
#7	TI=(grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI=(class* OR school* OR kindergarten* OR “primary education” OR “secondary education”)	229,870
#6	#4 OR #5	374,864
#5	AB=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	2,243
#4	AB=(test* OR assess* OR measur* OR exam* OR quiz*) AND AB=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	373,135
#3	#1 OR #2	3,291
#2	TI=("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	485
#1	TI=(test* OR assess* OR measur* OR exam* OR quiz*) AND TI=(frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	2,822

EBSCOhost Research Databases - OpenDissertations
Limiters - Published Date: 20201101-20231231 Search date - 20230829

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S14	S10 AND S11 AND S12 Limiters - Published Date: 20201101-20231231	3
S13	S10 AND S11 AND S12	230
S12	S3 AND S6	607
S11	S7 OR S8 OR S9	290,415
S10	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	719,596
S9	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	173,021
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	159,350
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	53,530
S6	S4 OR S5	87,723
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	401
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	87,426
S3	S1 OR S2	1,072
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	108
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	971

Extra search

EBSCO Research Databases: ERIC
Search date - 20240206

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S9 AND S10 AND S11 AND S12	28,315
S12	S3 OR S6	77,540
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	1,144,791
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	1,112,674
S9	S7 OR S8	824,664
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	788,518
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	292,087
S6	S4 OR S5	76,666
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,477
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	73,734
S3	S1 OR S2	3,588
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	944
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	2,675

EBSCO Research Databases: APA PsycINFO
Search date - 20240206

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S9 AND S10 AND S11 AND S12	35,317
S12	S3 OR S6	403,213
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	3,289,756
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	1,632,442
S9	S7 OR S8	808,633
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	768,486
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	223,789
S6	S4 OR S5	401,895
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,806
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	398,911
S3	S1 OR S2	6,534
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	1,140
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	5,433

EBSCO Research Databases: Academic Search Premier
Search date - 20240209

Expanders - Apply equivalent subjects
Search modes - Boolean/Phrase

Search History

#	Query	Results
S13	S9 AND S10 AND S11 AND S12	37,823
S12	S3 OR S6	1,408,249
S11	TI (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing) OR AB (achiev* OR learn* OR perform* OR improv* OR progress* OR result* OR impact* OR attain* OR success* OR "test* anxiety" OR wellbeing)	16,905,459
S10	TI (student* OR pupil* OR child* OR adolescen* OR youth* OR young*) OR AB (student* OR pupil* OR child* OR adolescen* OR youth* OR young*)	3,055,281
S9	S7 OR S8	2,811,810
S8	AB (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR AB (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	2,675,246
S7	TI (grade* N1 (1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12)) OR TI (class* OR school* OR kindergarten* OR "primary education" OR "secondary education")	563,826
S6	S4 OR S5	1,403,111
S5	AB ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	3,350
S4	AB (test* OR assess* OR measur* OR exam* OR quiz*) AND AB (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	1,400,514
S3	S1 OR S2	19,948
S2	TI ("test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" OR "cumulative testing" OR "progress monitoring" OR "curriculum-based")	895
S1	TI (test* OR assess* OR measur* OR exam* OR quiz*) AND TI (frequen* OR repeat* OR interim* OR formative* OR dail* OR weekl* OR monthl* OR annual*)	19,106

Search of other sources

Hand searches of relevant journals

We searched the following journals and volumes published between 2018-2023/2024:

- *Assessment in Education: Principles, Policies and Practice* (volumes 25(1) – 30(5-6)): 230 records screened on title/abstract, 0 included.
- *Journal of Educational Research* (volumes 111(1) – 116(5)): 295 records screened on title/abstract, 0 included.
- *School Psychology Review* (issues 47(1) – 53(1)): 257 records screened on title/abstract, 0 included.
- *Journal of Educational Psychology* (volumes 110(1) – 116(2)): 542 records screened on title/abstract, 0 included.
- *Educational Assessment* (volumes 23(1) – 28(4)): 110 records screened on title/abstract, 0 included.

Searches for dissertations

Searches for dissertations are included under the headings of Initial search and Updated search above.

Grey literature search

See Table A.1 below.

Citation tracking

We backward and forward citation tracked all included records and the following reviews: Adesope et al. (2017), Bangert-Drowns et. al. (1991), Phelps (2012) and Yang et al. (2021, 2023). We backward citation tracked by screening the reference lists, and forward citation tracked by examining which records that had cited an included record using Google Scholar (and in some cases, Web of Science). In total, we screened 10,558 references on title and sometimes abstract.

Contacts to experts in the field

We contacted three international experts to identify unpublished and ongoing studies, and pro-vided them with the inclusion criteria for the review along with the list of included studies, ask-ing for any other published, unpublished, or ongoing studies relevant to the review. We did not receive any reply.

TABLE A.1*: Results of the search for grey literature*

Ressource	Method	Search terms	Results	Included	Date
e.g. Social Science Research Network	**e.g. free text, advanced search, standard search, title search, abstract search etc. Note if filters were used**	**intervention keywords or combinations thereof**	**please list results individually for each search term/combination of terms, not only by overall resource**	**Number of hits included**	**Date of search**
Social Care Online - https://www.scie-socialcareonline.org.uk/	Basic Search	Testing frequency and student achievement	53 hits	0	25.03-2024
Social Care Online - https://www.scie-socialcareonline.org.uk/	Basic search	Testing frequency and anxiety	7 hits	0	25-03-2024
Social Care Online - https://www.scie-socialcareonline.org.uk/	Basic search	test enhanced learning	58 hits	0	25-03-2024
Social Science Research Network - https://www.ssrn.com/index.cfm/en/	basic search	test enhanced learning	187 hits	0	25-03-2024
Social Science Research Network - https://www.ssrn.com/index.cfm/en/	Basic search	Test and anxiety	425 hits (first 100 hits screened)	0	25-03-2024
Social Science Research Network - https://www.ssrn.com/index.cfm/en/	Basic search	frequency of tests	788 hits (first 100 screened)	0	25-03-2024
Mathematica - https://mathematica.org/	Basic search	testing frequency and student achievement	5 hits	0	25-03-2024
Mathemactica - https://mathematica.org/	Basic search	testing frequency and anxiety	0 hits	0	25-03-2024
Mathematica - https://mathematica.org/	Basic search	test enhanced learning	167 hits	0	25-03-2024

-					
Google Schoolar - https://scholar.google.dk/	Basic search	what are the effects of different testing frequencies on student achievement? Search from 2000-2024	274.000 hits (first 100 hits screened)	0	25-03-2024
Google Scholar - https://scholar.google.dk/	Basic search	"test enhanced learning" OR "test* effect*" OR "test* phenomenon" OR "frequency of test*" OR "cumulative test*" OR "progress monitoring" OR "curriculum-based"	17.900 hits (first 100 hits screened)	0	07-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	test enhanced learning	7.460.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	test* effect	8.260.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	test* frequency	6.600.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	Cumulative test*	5.440.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	Progress monitoring*	6.940.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	Test* anxiety	3.959.000 (first 100 hits screened)	0	27-06-2024
Google Scholar - https://scholar.google.dk/	Basic search	test OR assess OR measure OR exam OR quiz AND frequent OR repeat OR interim OR formativ OR summativ OR dail OR week OR month OR year OR annual	6.700.000 (first 100 hits screened)	0	27-06-2024
MDRC - https://www.mdrc.org/	Specific Search (by topic)	Topic: Early childhood. Document type: Reports, working papers, issue focus	104 hits	0	07-05-2024

MDRC - https://www.mdrc.org/	Specific search (by topic)	Topic: K-12 Education. Document type: reports, working papers, issue focus	211 hits	0	07-05-2024
MDRC - https://www.mdrc.org/	Specific search (by topic)	Topic: Post-secondary education. Document type: Reports, working papers, issue focus.	178 hits	0	07-05-2024
MDRC - https://www.mdrc.org/	Specific search (by topic)	Topic: Young People. Document type: Reports, working papers, issue focus	183 hits	0	07-05-2024
Abt Associates - https://www.abtglobal.com/	Specific search (by topic)	Topic: Education, youth and families	87 hits	0	07-05-2024
American Institutes for Research - https://www.air.org/	Specific search (by topic)	Topic: Education measurement & assessment	15 hits	0	07-05-2024
American Institutes for Research - https://www.air.org/	Specific search (by topic)	Topic: Multi-tiered systems of supports	21 hits	0	07-05-2024
WestED - https://www.wested.org/	Basic search	test* OR assess* OR measur* OR exam* OR quiz* AND frequen* OR repeat* OR interim* OR formative* OR summativ* OR dail* OR week* OR month* OR year* OR annual*	3 hits	0	07-05-2024
WestED - https://www.wested.org/	Basic Search	"test enhanced learning" OR "testing effect*" OR "testing phenomenon" OR "frequency of testing" and "anxiety"	0 hits	0	07-05-2024
WestED - https://www.wested.org/	Basic Search	"Testing Frequency & student achivement"	0 hits	0	07-05-2024
WestED - https://www.wested.org/	Basich Search	Testing frequency and anxiety	0 hits	0	07-05-2024
WestEd - https://www.wested.org/	Basic Search	test enhanced learning	83 hits	0	07-05-2024

Westat - https://www.westat.com	Specific search (by topic)	Topic: Education	35 hits	0	07-05-2024
Westat - https://www.westat.com	Specific search (by topic)	Topic: Educational Assessment	12 hits	0	07-05-2024
Westat - https://www.westat.com	Basic search	Testing frequency and student achievement	0 hits	0	07-05-2024
SRI - https://www.sri.com/	Basic Search	Testing frequency and student achievement	0 hits	0	07-05-2024
SRI - https://www.sri.com/	Specific search	Student Behaviour publications	29 hits	0	07-05-2024
SRI - https://www.sri.com/	Specific search	Early childhood learning and development publications	117 hits	0	07-05-2024
Google searches - google.com	Basic search	test OR assess OR measur OR exam OR quiz AND frequen OR repeat OR interim OR formative OR summativ OR dail OR week OR month OR year OR annual	13.000.000.000 (first 100 hits)	0	19-08-2024
Google searches - google.com	Basic search	"test enhanced learning" OR "test* effect" OR "test* phenomenon" OR frequency of test*	1.880.000.000 (first 100 hits)	0	19-08-2024
Google searches - google.com	Basic search	Testing frequency and student achievement	71.400.000 (first 100 hits)	0	19-08-2024
Google searches - google.com	Basic search	Progress monitoring*	1.010.000.000 (first 100 hits)	0	19-08-2024
For Systematic reviews:					
Cochrane Library - https://www.cochranelibrary.com/	Basic search	test* OR assess* OR measur* OR exam OR quiz AND frequen* OR repeat* OR interim OR formative* OR summativ* OR dail* OR week* OR month* OR year* OR annual*	8711 (first 100 hits)	0	07-06-2024

Cochrane Library - https://www.cochranelibrary.com/	Basic search	"test enhanced learning" OR "test* effect" OR "test* phenomenon" OR frequency of test*	243 (first 100 hits)	0	07-06-2024
-					
Centre for Reviews and Dissemination Databases - https://www.crd.york.ac.uk/CRDWeb/	Basic search	test* effect OR test enhanced learning OR test* phenomenon OR frequency of test* OR cumulative test* OR progress monitoring OR curriculum-based.	47 hits	0	07-06-2024
-					
Systematic reviews, evidence synthesis - The Campbell Collaboration - https://onlinelibrary.wiley.com/journal/18911803	Campbell Collaboration, systematic reviews journal	We have gone through all published Campbell Systematic reviews	229 hits	0	20-09-2024
-					
EPPI Centre - https://eppi.ioe.ac.uk/cms/Databases/tabid/185/Default.aspx	Advanced search under Education	We have gone through all reviews via "Publications" under Education	248 hits	0	24-06-2024

Supplementary Information Appendix B: Description of the extra search

The initial search strategy for this review did not sufficiently cover relevant results. We therefore decided to run an additional, broader search. The search terms were kept but the the search string with field search of intervention search terms was changed from *title AND abstract* to *title OR abstract*.

The result of the extra and broadened search strategy was a new collection of records from three different databases: 28,315 records identified from ERIC; 35,317 records identified from APA PsycINFO; and 37,823 records identified from Academic Search Premier (see *Supplementary Information Appendix A*). These three databases were chosen based on the results from the initial and updated search of electronic databases. Included records from the initial and updated search after screening on title and abstract and on full text with lower than critical risk of bias rating were exclusively from ERIC, APA PsycINFO, Academic Search Premier and ProQuest Dissertations & Theses. Together, these bibliographical databases contribute to a widespread search of relevant literature for this review. While ERIC and APA PsycINFO, to a larger degree, are more subject-specific for educational and psychology research, respectively, Academic Search Premier and ProQuest contribute with a more multidisciplinary focus. Unfortunately, we no longer had access to ProQuest, which meant that the final search was made for ERIC, APA PsycINFO, and Academic Search Premier.

In total, the extra database search resulted in 101,455 records. Using both a semi-automated duplicate check in EPPI Reviewer with a threshold value of 0.85 and a further manual duplicate check, we found 22,341 duplicates and a total number of unique and new records from the broader search was 79,114.

Using EPPI Reviewer, we built a classifier model that we trained on the 6,572 records from the initial and updated search. The classifier model was trained to include and exclude records based on the title and abstract screening, which was independently screened by two humans. We ran the classifier model on the 79,114 unique records, which yielded the distribution of probabilities of relevance shown in Table B.1.

TABLE B.1. *Classifier-generated probability distribution of relevance*

Note: Distribution of records, results from the classifier model. Number of records on the Y-axis and probability of relevance on the X-axis. Red bars indicates probability categories of records below the cutoff chosen to exclude automatically. Green bars indicate categories above this cutoff, which were screened by *AIscreenR*.

Different studies research the most optimal way of defining the cutoff between irrelevant and relevant records using stopping rules (Campos et al., 2024; Boetje & Schoot, forthcoming). In general, there are, however, no consensus on how to define a threshold of relevance. We used the following procedure: We started with a 40% cutoff. Two review authors then screened a random sample of fifty records with 39% probability of relevance on title and abstract. None of the fifty records with 39% probability were deemed relevant by either of the screeners. We thus kept the cutoff at 40% probability and excluded records with lower probability. The remaining subset of records consisted of 11,614 records.

To ease the screening of the 11,614 additional records, we used the newly developed R-package *AIscreenR* (Vembye & Olsen, 2024), which enables reviewers to conduct title and abstract screening based on OpenAI's GPT API models (for the first evaluations of this technique, see Guo et al., 2024; Syriani et al., 2023). For our screening, we drew on the gpt-4-0613 API model from OpenAI reached from the v1/chat/completions endpoint. All codes for this screening are available at https://osf.io/6b4ah.[1] We used six different prompts to conduct a hierarchical screening, meaning that only the included records after running the first prompt were screened using the second prompt, and so on. Each prompt was developed using the published protocol of this review, the screening

[1] To reproduce the code one has to get an unique API KEY from OpenAI. This can be found at https://platform.openai.com/api-keys. If one wants to reproduce the screening, be aware that there is a cost.

guidelines. and by looking at the abstract elements of the included records from the initial and updated search. Several tests were made using *AIscreenR* on multiple subsets of records (respectively 154, 235, 135 etc.) consisting of the included and excluded records from the initial and updated search to identify the most important criteria and the optimal order of the prompts, and to distinguish the relevant records from the irrelevant records. We tested the tool until we reached a recall of a minimum of 85% (i.e., the GPT model had an agreement rate with us regarding relevant studies on at least 85%) and a specificity of 95% (i.e., the GPT model had at least 95% percent agreement with us regarding irrelevant studies) and when using the hierarchical screening on the subset consisting of 135 records, recall was 89% and specificity 97%. The applied thresholds were set based on the average recall and specificity we typically find between two human screeners in our Campbell Review. Furthermore, these thresholds are in line with Khraisha et al. (2024) recommendation that agreement rates between 80% to 95% can be considered to be on par with common agreement rates between human screeners. Each of the six used prompts was initiated with the following statement: "*We are conducting a systematic review, which examines the potential effects of different frequencies or intensities in testing on the academic achievement or testing-related anxiety of primary and secondary school students*." The following phrases are specific for each developed prompt. An overview of the elements of the prompts is given in Table B.2:

TABLE B.2. *Overview of prompts*

Prompt number	**Specific phrases for each prompt**	**Specific question**
1	We want to include studies with quantitative measures. For each study, we would like you to assess:	1)Does the study report quantitative measures?
2	Only investigations performed in a school setting on children or students (ages 4-18 years old) are relevant for this review. This means that experiments performed in laboratories must be excluded, because we are only interested in real school settings and educational systems. For each study, we would like you to assess:	1) Does the intervention take place within a school setting?
3	We only want to include studies that investigate children or students attending either primary or secondary school, this means from kindergarten until grade 12. In other words, we are looking for studies where the participants are students 4-18 years old. For each study, we would like you to assess:	1) Are the participants in the study children or students attending either primary or secondary school, this means from kindergarten until grade 12.
4	The study must entail testing students or children. The testing can be standardized and non-standardized tests as well as formative assessments and summative tests, and high-stakes and low-stakes exams.	1) Does the study report on tests or testing of students or children?

	This also include repeated testing, interim assessment testing, class quizzes, multiple choice testing, progress monitoring assessments or measures, curriculum-based measurement or assessments, retrieval practice measures or assessments, etc. For each study, we would like you to assess:	
5	We like to include randomized controlled trials (RCT), fields experiments, quasi-experimental studies (QES), or observational studies, which use a control/comparison research design to examine effects. This means that the study must compare at least two groups of students or children. Such studies can have many labels and the different designs can have different notations. The most common sub-categories of randomised controlled trials and quasi-experimental studies are: individual randomised assignment, cluster randomised assignment, stratified/blocked random assignment, pseudo-randomisation, matching cohort studies, difference-in-differences, regression-discontinuity designs, instrumental variable designs, propensity score matching, case-control studies, etc. Studies employing a within-subject design are also eligible for inclusion. For each study, we would like you to assess:	1) Is the study a randomized controlled trial (RCT), a field experiment, a quasi-experimental study, an observational study, or a study employing a within-subject design?
6	In the review, we would like to include studies that measures students' academic achievement. In this review, we do not restrict measures of academic achievement to specific subjects. For each study, we would like you to assess:	1) Does the study report on measures of academic achievement or academic skills?

AIscreenR included 2,086 records based on the six prompts. This set represents all records that were included on all prompts. The included records were, thereafter, screened on title and abstract by review team members to either confirm the inclusion or decide on exclusion. The 153 records that either had no abstract or had the terms: “meta”, “meta-analysis”, “review” or “systematic review” in their title, were also screened to confirm the exclusion. Out of the 2,239 records, only 321 were included for the full text screening.

Concerns regarding the reproducibility of our GPT screening: Technical details

A clear disadvantage of using OpenAI GPT API models is that older models deprecate over time, meaning that they eventually will be removed from OpenAI’s servers and cannot be reached by any users. This is relevant for the model that we used via the *AIscreenR* package. For now, there has not yet been set a date for when this gpt-4-0613 model deprecate but we expect it to happen over the next year, as has happened with other older models such as the 0301 models. It is also important to notice that one cannot reproduce our screening with newer gpt-4 models unless one codes these screening functions oneself. The reason is that the newer gpt-4 models, such as gpt-4-turbo, do not

handle function calls the same way as the gpt-4-0613 model. The main difference is that function calls in newer GPT API models should be added to the *tool* argument in the request body whereas the gpt-4-0613 model receives function calls via the the deprecated arguments *function_call* and *functions* in the request body. The issue here is that the current version of the *AIscreenR* package draws on the latter. Function calling (read JSON function) is used to make clear instructions about how the GPT model should respond to a given request to ensure a systematic and standardized response output. This makes the responses more accurate, reliable, and efficient relative to entering this information in the main prompt. Therefore, function calling was all-important for the performance of the models we used. Thus, function calling must be used if one wants to make a reliable attempt to reproduce our screening results.

Literature:

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res*, *26*, e48996. https://doi.org/10.2196/48996

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, *n/a*(n/a). https://doi.org/https://doi.org/10.1002/jrsm.1715

Syriani, E., David, I., & Kumar, G. (2023). Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews. *ArXiv Preprint ArXiv:2307.06464*.

Vembye M, & Olsen T (2024). *AIscreenR: AI screening tools for systematic reviews*. R package version 0.0.0.9016, https://mikkelvembye.github.io/AIscreenR/.