# A Meta-Analysis Investigating the Association Between Metacognition and Math Performance in Adolescence

Gemma Muncer[1]*, Philip A. Higham[2], Corentin J. Gosling[3, 4], Samuele Cortese[5],

Henry Wood-Downie[6,] & Julie A. Hadwin[7]

[1]BCP Educational Psychology Service, Dolphin Centre, Poole, BH15 1SA
[2]Centre for Perception and Cognition, School of Psychology, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[3]DysCo Lab, Department of psychology, Paris Nanterre University, F-92100 Nanterre, France
[4]Université de Paris, Laboratoire de Psychopathologie et Processus de Santé, F-92100 Boulogne Billancourt, France
[5]Centre for Innovation in Mental Health-Developmental Lab, School of Psychology, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[6]Centre for Research in Inclusion (CRI), Southampton Education School, University of Southampton, Highfield, Southampton SO17 1BJ, UK
[7]Childhood Research Forum and Centre for Education and Policy Analysis, School of Education, Liverpool Hope University, Hope Park, Liverpool, L16 9JD

*Corresponding author: gcs1n17@southamptonalumni.ac.uk

## Abstract

Poor math and numeracy skills are associated with a range of adverse outcomes, including reduced employability and poorer physical and mental health. Research has increasingly focused on understanding factors associated with the improvement of math skills in school. This systematic literature review and meta-analysis investigated the association between metacognition and math performance in adolescence (11-16-year-olds). A systematic search of electronic databases and grey literature (to 04.01.2020) highlighted 31 studies. The quantitative synthesis of 74 effect sizes from 29 of these studies (30 independent populations) indicated a significantly positive correlation between metacognition and math performance in adolescence ($r = .37$, 95% CI = [.29, .44], $p < .001$). There was significant heterogeneity between studies. Consideration of online (versus offline) measures of metacognition and more complex (versus simple) measures of math performance, and their combination, were associated with larger effect sizes; however heterogeneity remained high for all analyses.

*Keywords*: adolescence, metacognition, math

**A Meta-Analysis Investigating the Association Between Metacognition and Math**

**Performance in Adolescence**

Math and numeracy skills (the ability to use numbers and solve mathematical problems in everyday life; National Numeracy, 2020) are often used in daily tasks, including managing money and finances, using travel timetables or following a recipe (Price & Ansari, 2013). Studies have highlighted the societal implications of numeracy skills. For example, Martin et al. (2014) estimated the cost of poor numeracy to the UK economy to be £20.2 billion per year in 2012 (approximately 1.3% of Gross Domestic Product). Further studies have found that low numerical ability in childhood and adolescence was associated with outcomes in adulthood that negatively impacted employability and prospective earnings (Crawford & Cribb, 2013; Wolf, 2011), and was linked to increased youth offending and criminality (Meltzer et al., 1984; Parsons, 2002).

Several initiatives have aimed to develop an international profile of math achievement. For example, the Trends in International Mathematics and Science Study (TIMSS 2019; Mullis et al., 2020) reported that across 64 countries there was a general increase in math achievement in 9-10 and 13-14-year-old pupils since this USA initiative was started in 1995. Further reports have highlighted that most countries require pupils to study math to the age of 16-17 years, where the qualification achieved at this point in education typically represents a critical gateway to further study or training (Hodgen & Pepper, 2010). To meet the requirements for progression in the UK context, for example, all pupils are expected to achieve a pass grade in GCSE math at 15-16 years.  While a pass is required to access most UK higher education courses and employment, around one-quarter of adolescents do not achieve this level (Ofqual, 2019).

Recognising the impact of math on developmental outcome in adolescence and across the lifespan has increasingly focused research agendas on understanding factors associated

with achievement in this subject. Mullis et al., (2020) highlighted a complex profile of factors linked to a positive outcome in math, including gender (with males outperforming females in > 90% of countries), the home context (e.g., books in the home, parent occupation), and the school context (e.g., more school resources, more time spent studying math, a school focus on achievement, fewer pupil behavioural problems). In addition, the report outlined several pupil self-reported factors associated with math achievement, including more enjoyment and value of the subject, as well as increased confidence and metacognitive skill (i.e., pupils' reported awareness of their own ability to solve complex mathematical problems).

**Defining Metacognition**

Metacognition (MC) refers to an individual's self-regulation of their own learning, including an awareness of their own strengths and weaknesses, as well as a recognition of the strategies that may be useful to progress in specific tasks (e.g., how well the individual monitors progress during the completion of a task, and the extent to which they recognise what behavioural change is needed to reach an outcome; Credé & Phillips, 2011; Hacker et al., 1998). Metacognition is typically divided into two or three parts in the research literature. The dyadic model of MC (Nelson & Narens, 1990) includes two components linked to (1) an individual's awareness of their strategic knowledge associated with memory and learning, and (2) their ability to monitor (e.g., "How well am I doing?") and flexibly control ("What do I need to do?") cognitive processing as they complete a task. The ternary model divides metacognition into three components (e.g., Efklides, 2008). The first two components fit with those proposed in the dyadic model, including knowledge (the extent to which a person is aware of what they know) and cognitive skill (strategy use to monitor and regulate cognition and effort to meet task goals). The third component of MC in the ternary model is proposed to reflect feelings that emerge (e.g., satisfaction and confidence) when an individual engages in a task. This component has been termed "experience" (Flavell, 1979, p. 906), and it is

linked to the implicit use of cues associated with a student's knowledge and skill as they progress through a task (Dent & Koenka, 2016; review by Schneider, 2008).

Researchers have proposed that different aspects of MC are closely related. If an individual is aware of their own knowledge, then this awareness can increase attentional focus on what is still to be learnt (Metcalfe & Finn, 2008), and effectively guide self-directed learning (Garrett et al., 2006; Ohtani & Hisasaka, 2018). In addition, monitoring is a necessary pre-requisite to regulate cognitive activity and behavioural response (Baker, 1989). In support, several studies have found a positive correlation between MC knowledge and the use of self-regulation MC strategies in learning (e.g., Schraw et al., 2012; Schraw & Dennison, 1994).

**Measuring Metacognition**

A systematic review with 4-16-year-olds identified 84 MC measures across 149 papers (Gascoine et al., 2017). Measures of MC are typically categorized as online or offline. Offline measures are questionnaires that aim to capture an individual's self-reported perception of their own MC ability based on previous learning experiences (see Saraç and Karakelle, 2012). For example, the MC self-regulation sub-scale in the motivated strategies for learning questionnaire (MSLQ, Pintrich, 1991) asks respondents to read 12 statements (e.g., "I ask myself questions to make sure I know the material I have been studying") and to indicate how true each statement is for them. The metacognitive awareness inventory (MAI, Schraw & Dennison, 1994) similarly asks individuals to indicate whether each of 52 statements related to learning is true (1 point) or false (0 points) for them (e.g., "I try to use strategies that have worked in the past"). The junior metacognitive awareness inventory (Jr. MAI, Sperling et al., 2002) asks children (8-11-years; 12 items) and adolescents (12-15-years; 18 items) about the frequency of use of metacognitive strategies during learning (e.g., "I think of several ways to solve a problem and then choose the best one"). In contrast, online

measures capture an individual's MC via ongoing behaviour and performance as they complete a task (Saraç & Karakelle, 2012; Veenman & van Cleef, 2019). These include think-aloud protocols, for example, where individuals verbalise their thoughts while engaging in a task. Verbalisations are recorded and later coded for the quality and/or quantity of MC activity (e.g., Veenman et al., 2005).

Recent discussion has focused on the distinction between online and offline measures, and particularly whether self-reported MC measures that are relevant to processes in ongoing tasks should be classed as online or offline. These measures include individuals' confidence judgments, accuracy measures (i.e., the difference between an individual's predicted score and actual score; also known as calibration accuracy) and judgment of learning (JOL) scores. JOLs typically ask individuals how confident they are from 0% to 100% that they would recall learnt information in a later test (e.g., Myers et al., 2020). Saraç and Karakelle (2012) classified confidence judgments and JOLs as online, as they pertain to a specific task at hand. In contrast, Veenman and van Cleef (2019) considered them as offline judgements, as they typically follow (i.e., but sit outside) the completion of a task. More recently, Craig et al. (2020) categorized confidence judgments during a task (where individuals reported their confidence in their answers to specific questions immediately after each item and before completing further items) as online, and those made following the completion of an entire task as offline.

Some research has reported poor correspondence between offline and online MC measurements. Sperling et al. (2004) found a significant correlation between undergraduate responses to two offline self-reported questionnaires: the MAI (Schraw & Dennison, 1994) and the MC self-regulation scale of the MSLQ (Pintrich, 1991). In addition, both offline questionnaires were not linked to the accuracy of students' confidence against their predicted test scores - categorized in this paper as online. An earlier study similarly found no

association between scores on the MAI (an offline questionnaire) with 14-17-year-old student

judgements about whether they could solve a math question (Tobias et al., 1999). The lack of

correspondence between online and offline measures across studies suggests that they may be

measuring different facets of MC. Sperling et al. (2002) suggested, for example, that the Jr.

MAI is a broad measure of MC, as compared to some existing measures that focus more

specifically on MC self-regulation. Saraç and Karakelle (2012) further proposed that online

measures may capture implicit experience-based judgements, whereas offline measures may

reflect more explicit knowledge-based judgements of MC.

**Metacognition and Academic Performance**

Several systematic reviews and meta-analyses have explored associations between

MC and academic achievement across different subjects in adult populations, highlighting

small but significant associations between offline MC measures and achievement. Credé and

Phillips (2011) found that student scores on the MC self-regulation scale of the MSLQ were

moderately significantly correlated with Grade Point Average (GPA; 98 correlations from 24

independent samples, $N = 9,696$, $r = .22$, 90% CI = [.03, .47]) and current course grade (431

correlations from 53 samples, $N = 15,321$, $r = .23$, 90% CI = [.02, .45]). Richardson et al.

(2012) considered 50 constructs associated with achievement and identified a small but

significant correlation ($r = .18$, 95% CI = [.10, .26]) between student MC and GPA ($N =$

6,205 across 9 studies).

Ohtani and Hisasaka (2018) extended these analyses to synthesise 149 effect sizes

from 118 independent samples, that included children and adults. The authors similarly

identified small and moderate correlations between MC with academic performance ($r = .28$,

95% CI = [.24, .31]) and intelligence ($r = .33$, 95% CI = [.26, .39]). In an earlier study, Dent

and Koenka (2016) also showed a small but significant correlation between MC and

academic achievement ($r = .20$, 95% CI = [.16, .24]) across 61 studies carried out in North

America and Canada with school-aged children and adolescents aged 6-19 years. Both meta-analyses found that online MC (vs. offline) measurements were most clearly associated with achievement. Respective effects sizes for online and offline associations were $r$ = .53 (90% CI = [.45, .61]) and $r$ = .23 (90% CI = [.20, .26]; Ohtani & Hisasaka, 2018), and $r$ = .39 (95% CI = [.34, .43]) and $r$ = .15 (95% CI = [.12, .18]; Dent & Koenka, 2016).

Previous research has found that while MC thinking is evident in young children, its use in learning contexts to efficiently plan and control effort and attention to focus on what needs to be learned increases across childhood (Paulus et al., 2014; review by Schneider, 2008). In a review of interventions on achievement in primary and secondary school children aged 5-16 years, Dignath and Büttner (2008) reported a large effect size for the impact of self-regulated programmes (including those based on MC strategy) on mathematics and academic achievement more broadly. The review further indicated that secondary (vs. primary) school-aged children and adolescents were most able to benefit from interventions that included the promotion of MC strategies in school.

Meta-analyses have also considered whether associations between MC and achievement were moderated by age. Ohtani and Hisasaka (2018) reported an increased effect size for children relative to adults between MC and achievement across subjects, however, overall age (separated into three broad categories including elementary (6-12 years), secondary (13-18), and adults (18 years and above)) did not moderate this association. Dent & Koenka (2016) similarly tested the hypothesis that the association between MC and achievement would increase with age. Comparisons between age were also not significant, and the results showed a small positive effect size across elementary ($r$ = .24, 95% CI = [.15, .32]), secondary ($r$ = .21, 95% CI = [.16, .25]) and high school students ($r$ = .18, 95% CI = [.10, .25]).  Contrary to expectations, however, a further age analysis indicated a stronger effect size for 5-6-year-olds ($r$ = .42, 95% CI = [.36, .48]) compared with 8-11-year-olds ($r$ =

.11, 95% CI = [.02, .20]). The analysis for the younger age group was, however, based on a small number of studies ($n$ = 4) and all had used online MC measures, indicating a conflation of age and measurement in this analysis.

**Metacognition and Math Performance**

Several studies have reported an association between MC and math achievement in children and adolescents (e.g., Özsoy, 2011; van der Walt et al., 2008). Consistently, Dent and Koenka (2016) carried out a further analysis in their review that focused on subject specific associations, and reported a small correlation between MC and math achievement across children and adolescents ($n$ = 39 studies; $r$ =.21, 95% CI = [.03, .27]). The effect size of the association was similar to that reported for English and science, but significantly smaller compared with social studies ($r$ =.34, 95% CI = [.27, .40]).

Further research has, however, demonstrated non-significant associations between MC and math achievement (e.g., Maras et al., 2019; Young & Worrell, 2018). Some researchers have suggested that the disparity in findings across MC studies may be due to differences in how MC is conceptualised and assessed (Desoete & Roeyers, 2006; Veenman et al., 2006). The measurement of MC goes some way to explain differences in findings; with online (vs. offline) measures being most clearly linked to academic achievement (reviews by Dent & Koenka, 2016; Ohtani & Hisasaka, 2018). The disparity in results between studies may also be a function of how math performance is measured.

Campbell (2005) proposed that mathematical ability is made up of two key elements: numerical ability (basic number representation and simple arithmetic and operations) and mathematical problem-solving (the generation of solutions from abstract representations of mathematical relations in context-rich problems). Other researchers have divided mathematical challenges into routine (i.e., questions that test student knowledge of what was recently covered) and non-routine problems (i.e., those that cannot be solved immediately

and often require complex multi-step problem-solving; Mayer, 1998). Non-routine problems typically go beyond existing knowledge and skills, requiring the solver to plan, monitor and review their solution (Mayer, 1998; Verschaffel et al., 2010), and these are increasingly integrated into the math curriculum across development (e.g., UK Department for Education (DfE), 2014). For example, Mokos and Kafoussi (2013) asked 10-year-olds to think aloud when completing open-ended, every-day (authentic) and complex mathematical problems. The results showed that MC control and monitoring were most evident when children were asked to complete complex math problems. In a review of think-aloud methods, Jordano and Touron (2018) also reported that children's use of MC strategy increased with more complex and open-ended mathematical tasks.

**Aims of the Systematic Review**

Based on previous studies and reviews of existing research there is an emerging consensus that MC plays a small but consistent role in understanding individual differences in achievement across childhood and adolescence. Studies have also demonstrated evidence for a specific association between MC and math performance. These findings are, however, more mixed and may reflect differences in the way in which researchers have measured MC and achievement in math. In the current paper, we extend existing research to consider the strength of association between MC and math in adolescence. Adolescents are recognised to utilise MC more efficiently (Dermitzaki, 2005; Veenman et al., 2006), are faced with increasingly more complex mathematical problems to solve (DfE, 2014) and are working towards key examinations (Hodgen & Pepper, 2010). It therefore represents a stage of education that is critical for identifying factors that education stakeholders and practitioners can utilise to promote optimal achievement in school for the best outcome of pupils.

Following previous reviews (e.g., Dent & Koenka, 2016; Ohtani & Hisasaka, 2018), we anticipated that the association with MC and math performance in adolescents would be

most evident when MC is measured using online (vs. offline) measures. We extended

previous analyses to test the possibility that, across studies, the association between MC and

math performance would be stronger for complex (vs. simple) math tasks. We additionally

carried out exploratory analyses to consider the combination of MC measure and math

assessment and anticipated that associations would be most evident for studies using online

measures and more complex math assessments. Furthermore, we provided a comprehensive

quality assessment of existing research, and broadened the scope and focus of previous

systematic reviews and meta-analyses by placing no limit on literature searches with respect

to year or language of publication, and via the inclusion of a comprehensive quality

assessment of existing research.

## Method

This review was carried out following the best practice guidelines for conducting a

systematic review published by Siddaway et al. (2019) and the Preferred Reporting Items for

Systematic Review and Meta-Analysis guidelines (PRISMA; Moher et al., 2015). The

protocol was determined before starting the review and a title registration was pre-registered

with the Campbell Collaboration (review number 19-009).

### Search Strategy

We used variations of the terms *metacognition*, *math* and *performance* (see Table 1)

to search the titles, abstracts and keywords of records in four databases: Education Resources

Information Centre (ERIC; 1966-2019; $n = 542$), Web of Science Core Collection (1990-

2019; $n = 880$), and PsycINFO and PsycARTICLES via EBSCO (1887-2019; $n = 628$).

Searches were initially conducted up to 15.07.2019 and were repeated on 04.01.2020,

following data extraction, to identify papers that had become available since initial searches

($n = 28$). No limiters were imposed on publications (e.g., relating to publication date or

language). The syntax was adapted to meet the requirements of each database (see

Supplementary Material A for an example search). To include unpublished research, we additionally searched ProQuest Dissertations and Theses Global (using the terms in Table 1; $n = 327$) and OpenGrey ($n = 11$). Due to input restrictions, the keywords metacogniti* AND math* were used to search OpenGrey. The reference lists of papers included in the final sample were also manually screened for additional potentially relevant studies ($n = 93$). Two researchers independently carried out all database searches and yielded identical results (i.e., 100% agreement). Pilot searches included three additional terms for MC (resolution, calibration, and self-regulation) which were subsequently removed due to producing a high number of irrelevant papers.

**Inclusion and Exclusion Criteria**

The titles and abstracts of all records retrieved via the systematic search ($n = 1,985$ after duplicates were removed) were screened against the pre-determined inclusion criteria. Studies were included if: (i) the research reported the strength of association between MC and math performance (e.g., by reporting the Pearson correlation coefficient). Where studies investigated the impact of a MC intervention, these were only included if the statistical relationship between MC and math performance was reported before participants took part in the intervention (at baseline) or in a control group, (ii) participants were aged 11-16 ($\pm$ two years if $\geq 80\%$ of the sample were aged 11-16), (iii) the study included an objective measure of math performance (e.g., school assessment or standardised score), and (iv) the study included a measure of MC. Studies were excluded if: (i) they did not include primary data, (ii) the only measure of math performance was self-reported, (iii) participants were reported to have a complex neurodevelopmental disorder such as autism spectrum condition (ASC), or (iv) the only measure of MC was a broad measure of self-regulation as defined by Zimmerman (1989; i.e., it included other variables such as motivation and effort).

**Study Selection**

Searches yielded 2,509 records. These were exported into EndNote Desktop, and 524 duplicates were removed. Two researchers independently screened the titles, abstracts, and keywords of the remaining 1,985 records for relevance by applying the inclusion criteria stated above. This process was carried out using the web application, Rayyan (Ouzzani et al., 2016). Cohen's Kappa indicated substantial agreement between the two researchers regarding the inclusion or exclusion of records ($\kappa = .77$). Conflicts were resolved using the consensus model with reference to the inclusion criteria. Following this process, the full texts of 115 papers were retained for secondary screening.

Where the full text of a study was unavailable ($n = 16$), we contacted the corresponding author to request the paper. The authors of four studies were contacted, and two replied by sending the relevant paper. Where a contact address was not available, or the author did not reply, the paper was requested via the University of Southampton inter-library loan service ($n = 14$ requested, $n = 10$ received). Six of the retrieved papers were not in English. Two of these papers were translated for screening using the online translation programme, Google Translate, and four papers were read and screened by native speakers.

Two researchers, who were blind to the decision of the other, read the full texts of the 115 records to further consider eligibility to the current review. Cohen's Kappa indicated substantial agreement between the two researchers regarding the inclusion or exclusion of studies at this stage ($\kappa = .61$). Disagreements were resolved using the consensus model, and on two occasions, discussions took place with a third researcher to further consider inclusion. To avoid duplication of samples, where data was reported from the same participants in more than one study, the paper that reported the largest number of participants was included. If the number of participants was equal, the earliest study was included. Where the author(s) had measured MC and math performance but had not reported the association between the two ($n$

= 6), we contacted the author(s) to request this information. Two authors responded, one author provided the required data, and one reported that this information was not available. In total, 84 papers were excluded during secondary screening. Supplementary Materials: Table B shows the reason for exclusion for papers. The procedure of how the final sample of studies was reached in the qualitative synthesis ($n$ = 31) and quantitative synthesis ($n$ = 29) is shown in Figure 1.

**Data Extraction**

We extracted key data for the quality assessment of each paper and for the meta-analysis (see Table 2). The data of included papers were extracted by the first researcher. A second researcher checked the extracted data from 35% of studies (11 of 31) and agreed that this was accurate. Where only some participants within a paper fitted the inclusion criteria (e.g., a typically developing control group in a study primarily focused on individuals with ASC), data were extracted for typically developing participants only. For longitudinal studies ($n$ = 3), time 1 data was extracted.

**Quality Assessment**

Two researchers independently assessed the quality of each study using the Critical Appraisal Skills Programme (CASP) Cohort Study Checklist (2018). The checklist includes 12 questions and two sub-questions (14 items in total). Two questions are open-ended. Question 6 (a, b) was removed because it relates to longitudinal data and was not relevant to the current review. One item asked how precise the results were. We used the remaining nine items to generate a scoring system whereby a *yes* response scored 1 and *can't tell* or *no* both scored 0. Higher scores therefore reflected greater methodological quality. A table of the adapted checklist items is included in Supplementary Materials Table C.

**Analytic Strategy**

All analyses were conducted in R environment.

*Main analysis*

Pearson's *r* coefficients reported by primary studies were converted into Fisher's Z (Borenstein et al., 2009). These effect sizes were then categorized according to the math measure (simple vs. complex vs. unclear) and to reflect the MC measure (online vs. offline) based on the distinction proposed by Veenman and van Cleef (2019). Several studies reported more than one effect size as a result of multiple MC or math measures being used to quantify this association. Moreover, several research groups have conducted multiple studies across different publications. To account for non-independence, a two-stage random-effects multivariate meta-analysis was performed (using the 'metafor' and 'clubSandwich' packages in R; Pustejowsky, 2021; Viechtbauer, 2010). In addition to the multivariate structure, the data had some forms of hierarchical structure (i.e., one study included two independent groups, both having completed two outcomes). For this study, we assumed that the groups were entirely independent so that the number of independent studies in the model was 30.

Following the approach recommended by Pustejowsky and Tipton (2021), we started analyzing data by conducting a random-effects multivariate meta-analysis known as *subgroup correlated effects*. In this model, we included random effects for each outcome within each study and each research group. We used a diagonal variance structure and a restricted maximum likelihood estimation. To implement this model, we had to impute the covariance matrix for all primary studies. This was performed using the 'clubSandwich' package. We used the subgroup option proposed by this package to consider the categorization of the effect sizes according to the math and MC measures. This analysis assumes a mean correlation of *r* = 0.8 between effect sizes coming from the same study and category. We computed cluster-robust standard errors. We clustered the standard errors by research group to account for the possibility of dependence across studies conducted by the same group. Even when this model converged, the inspection of the profile likelihood plot suggested some overparameterization.

Therefore, we simplified the model by deleting the random effects for the research group. Even if research group was no longer included in our working model, we still maintained the clustering of standard errors by this factor to address the potential dependency. Throughout the manuscript, the model described here is referred to as the 'primary model'.

We then reassessed the pooled effect size of the association between MC and math performance using different statistical approaches. Here, the overall association of math with MC was reassessed by (i) refitting the primary model, but assuming different correlations between effect sizes of a same study and category (four other values were assessed: 0.05, 0.2, 0.5 and 0.95), (ii) refitting the primary model, but without classifying effect sizes according to math performance and MC measure when imputing the covariance matrix, and without including the factor in the random effects, (iii) using a classic robust variance estimation approach to handle the dependence of effect sizes within studies (Tipton, 2015) and (iv) using the aggregation approach (see Borenstein et al., 2009) to handle the dependence of effect sizes within studies. These approaches did not affect the statistical significance of this analysis.

### *Sensitivity analyses*

A total of five additional analyses were conducted. First, we performed a leave-one-out analysis (i.e., we re-ran our primary model, but leaving out each study sequentially) to assess the impact of each study on the pooled effect size. Models converged for all exclusions except one. Second, we re-ran our primary model but excluded effect sizes with standardized residuals superior to 2 (results of this analysis is not reported because no effect size was associated with a standardized residual superior to 2), with hat values superior to twice the mean of hat values (five effect sizes were excluded), or with Cook's distance superior to twice the mean of Cook's distance (six effect sizes were excluded). Third, we re-ran our primary model, but excluding studies with less than 80% power to detect the effect size of the

study with the lowest variance (eight studies and 12 effect sizes were excluded). Fourth,

using the results of our quality assessment we identified three questions in the CASP (2018)

relating to critical biases (see Supplementary Material C). Six studies (28 effect sizes) had at

least one unclear/high risk of bias score (indicating by an *unclear* or *no* response to the item),

and we re-ran our primary excluding studies these studies. Finally, we re-ran our primary

model but excluding the two very large PISA studies (*n* = 2).

## *Moderation analysis*

Because we anticipated that effect sizes would differ depending on the MC measure

and assessment of math performance, we ran a moderation analysis assessing the influence of

these factors. This analysis investigated whether there were significant differences in the

pooled effect sizes of studies that used online versus offline measures of MC, and studies that

used simple versus complex math assessments. To consider the different distinctions in the

literature between online and offline measures of MC, this subgroup analysis was carried out

four times. In all four analyses self-reported questionnaires were classed as offline and think-

aloud protocols and behaviour observations were classed as online. However JOLs,

confidence judgments and calibration scores were classed differently between analyses. The

four analyses carried out were: (i) online defined as not self-reported (JOLs, confidence

judgments and calibration scores classed as offline in-line with the definition proposed by

Veenman & van Cleef, 2019), (ii) online defined as pertaining to a specific task at hand (i.e.,

JOLs, confidence judgments and calibration scores classed as online, in line with the

definition proposed by Saraç & Karakelle, 2012), (iii) online defined as taking place during a

task (JOLs, confidence judgments and calibration scores classed as online where they were

carried out during a task, with those carried out before/after a task classed as offline, as in

Craig et al., 2020) and (iv) JOLs, confidence judgments and calibration scores (i.e., student-

reported MC score relevant to a specific task at hand) categorized separately from other online and offline measures.

Last, we considered the combination of MC measures and math assessment. We combined the math performance (according to its complexity: simple vs. complex vs. unclear) and the MC measure (according to Veenman & van Cleef's categorization) into a single measure. This combination yielded a moderator with six potential modalities (e.g., *complex* math task conducted *online*; *simple* math task conducted *offline,* etc.). We found that five combinations were explored in primary studies, and that only four were represented by at least two studies/research groups. We conducted a meta-regression with this factor as the moderator and without including a model intercept. This analysis generated a pooled effect size for each of the four combination of math performance and MC measure assessed by at least two studies/research groups. Raw and adjusted (with a Bonferroni correction) *p* values are also reported.

## Results

Of 1,985 papers (115 full texts) screened for eligibility, 31 studies met the inclusion criteria. Details of the included studies, including quality assessment ratings, are displayed in Table 2. Studies included in this review are peer-reviewed journal articles ($n = 25$), unpublished doctoral theses ($n = 4$) and conference research papers ($n = 2$). Included studies were published/made available between 1991 and 2019. One study included data from students in 34 countries using the PISA 2000 database and another used data from students in 63 countries using the PISA 2009 database. Other studies were conducted in 11 countries including the Netherlands ($n = 6$), Turkey ($n = 6$), Israel ($n = 1$), the United States of America (USA; $n = 7$), Malaysia ($n = 1$), the United Kingdom (UK; $n = 2$), Japan ($n = 1$), Spain ($n = 1$), Singapore ($n = 1$), China ($n = 2$) and South Africa ($n = 1$). Collectively, the 31 papers included 572,559 participants. Apart from the two studies which used PISA data and

involved high numbers of participants (88,590 and 475,460), the number of participants ranged from 18 to 1,019. All participants were aged 11-16 years. There was inconsistent reporting of age; some studies reported mean age and/or age range, and some studies did not report age ($n = 9$). In these instances, age was derived from the reported stage of schooling (see Table 2). The lowest reported mean age was 11.3 years, and the highest was 16.39 years. Twenty-six (/31) studies reported the sex/gender split of participants; from these studies, the total participant sample was 483,145, and of these, 243,061 (50.3%) were female.

**Qualitative Results**

Nineteen of the 31 papers reported a statistically significant positive association(s) between MC and math performance ($p$s <. 05). Eight studies reported positive association(s) that were not statistically significant. Four studies reported mixed findings (i.e., more than one correlation was reported due to measuring MC and/or math performance using more than one measure/scale, and at least one correlation was significant, and one correlation was not significant).

**Quality Assessment**

We utilised the scoring system whereby increased *yes* responses in the adapted CASP questionnaire (2018) are indicative of higher quality research. In seven studies, we rated all nine items as *yes*, six studies were given eight *yes* responses, ten were given seven *yes* responses, five were given six, two were given five, and one study was awarded two *yes* responses. (See Supplementary Materials E for responses for each item).

All studies addressed a focused issue that was evidenced by clear research aims, and all were considered to be sufficiently precise. Correlations/associations were reported with at least 95% confidence ($p < .05$), and in most cases, with greater confidence (e.g., 99%, $p < .01$). Generally, participants were recruited in a way that meant they were likely to be

representative of their cohort. However, *n* = 7 studies did not report how participants were recruited.

Most studies (*n* = 18) used pre-published and validated questionnaires. One study used a measure designed by the researcher. In this case, the measure had high inter-rater reliability, but there was no reference to validity testing. Most studies (*n* = 25) used acceptable measures of math performance. In two studies, the measure of math performance was unclear. One study did not report how math performance was measured, and the second provided unclear information. Additionally, one study reported coefficient alphas that fell within a questionable range ($\alpha$s < .70), one study used a measure on a single topic within mathematics that was reported to be particularly difficult for teachers to teach and in one study, the selection of question items was made on the basis that males had previously out-performed females on chosen questions.

The impact of confounding factors on the relationship between MC and math performance was not consistently considered. Three studies reported semi-partial correlations between MC and math performance to control for the contribution of general intelligence. Most studies included participants who were relevant to the population of interest in the current review. It was unclear whether some studies that selected participants based on specific characteristics (e.g., learning disability, having made below-expected progress in math, having below or above average anxiety, or being academically talented) were representative of 11-16-year-olds. Furthermore, one study included only female participants.

**Meta-Analysis Results**

Correlations between MC and math performance were available for 29 (/31) studies, and these were included in the meta-analysis (participant *N* = 570,575, *n* = 30 independent populations, *k* = 74).

*Small-Study Effects*

We assessed small-study effects using a modified version of the Egger's test (Egger et al., 1997). More precisely, we re-ran our primary model but including the inverse of the sample size (or its square root) as the moderator. These analyses showed some evidence of small study effects ($p = .052$). A very similar result was obtained when using the standard error as moderator ($p = 0.045$). However, when adjusting our primary model by the standard error, the inverse of the sample size, or its square root, the pooled effect size remained systematically statistically significant (all $r$ coefficients > .24, all $p$ values < .01).

### *Primary Analysis*

The primary analysis revealed a positive and significant association between MC and math performance ($r = .37$, 95% CI = [.29, .44], $p < .001$).

### *Heterogeneity*

Heterogeneity was significant and high in the primary analysis ($Q(73) = 16646.6539$, $p < .001$), indicating substantial variation in effect sizes across included studies (Higgins et al., 2003). To quantify inconsistency, we refitted our primary model but estimating an unique variance component corresponding to all outcomes. This new model allowed the computation of one overall $I^2$ statistic across all outcomes (see Jackson et al., 2012). The results of this analysis revealed an $I^2$ statistic of 99.92%, meaning that almost all the variability in effect estimates can be attributed to heterogeneity rather than sampling error. Several sensitivity analyses were then conducted to investigate the source of this heterogeneity and assess the robustness of our primary result (see Table 3). In all of these analyses, the heterogeneity and the pooled effect size of the association of math with MC remained statistically significant. While these analyses suggest that the association between math and MC is robust, they failed to identify the source of heterogeneity. Several moderation analyses were then conducted for this purpose.

### *Moderator Analysis*

Subgroup analysis explored whether measure of MC (online vs. offline) moderated the relationship between MC and math performance. To consider the different distinctions in the literature between online and offline measures, this analysis was carried out four times. First, we considered JOLs, confidence judgments and calibration scores classed as offline so that we compared think-aloud protocols and behaviour observations versus self-report questionnaires, JOLs, confidence judgments and calibration scores). In the subsequent reporting of results *n* indicates the number of independent samples and *k* indicates the number of effect sizes. In this analysis the pooled effect size was significantly larger when online MC measures were employed ($n = 9$, $k = 14$, $r = .54$, $I^2 = 79.12\%$) than when offline measures were employed ($n = 23$, $k = 60$, $r = .30$, $I^2 = 99.94\%$; $p = .034$). Second, when JOLs, confidence judgments and calibration scores were classed as online and so that we compared JOLs, confidence judgments, calibration scores, think-aloud protocols and observations versus offline questionnaires, the pooled effect size was significantly higher for online studies ($n = 16$, $k = 29$, $r = .46$, $I^2 = 89.66\%$) than for offline studies ($n = 16$, $k = 45$, $r = .28$, $I^2 = 99.9\%$; $p = .034$). Third, when JOLs, confidence judgments and calibration scores that were completed before or after each individual item during a math task were classed as online (alongside think-aloud protocols and behaviour observations), and those that were completed before or after an entire math task were classed as offline (alongside self-report questionnaires), the pooled effect size was significantly higher for studies that measured MC during a math task ($n = 14$, $k = 26$, $r = .48$, $I^2 = 91.23\%$) than not during a math task ($n = 18$, $k = 48$, $r = .28$, $I^2 = 99.95\%$; $p = .027$). Finally, when JOLs, confidence judgments and calibration scores were coded separately to other online and offline measures, an omnibus test showed that the effect estimates for the three categories (online, offline, JOL/accuracy) were not equivalent ($p = .0080$). The pooled effect sizes and $I^2$ statistics for online ($n = 7$, $k = 12$ correlations), offline ($n = 16$, $k = 45$) and JOL/accuracy scores ($n = 9$, $k = 17$) were $r = .59$, $I^2$

= 30.21, $r$ = .28, $I^2$ = 99.96% and $r$ = .38, $I^2$ = 92% respectively. Collectively, the results of this analysis suggest that consideration of offline versus online measures were not sufficiently precise to achieve homogeneity.

Similar difficulties to obtain homogeneous categories was observed when performing a moderation analysis considering math measure as the moderator (comparing effect sizes for studies that used simplex and complex math measurement and those where measurement was unclear). This analysis showed a significant omnibus test ($p$ < .01), indicating differences in the pooled effect sizes between effect estimates produced by complex math measurement ($n$ = 12, $k$ = 21, $r$ = 0.48, $I^2$ = 86.8%), simple math measurement ($n$ = 4, $k$ = 4, $r$ = 0.10, $I^2$ = 86.75%) and unclear measurement ($n$ = 15, $k$ = 49, $r$ = 0.33, $I^2$ = 99%; $p$ = .03). However, The number of studies using simple math tasks was small, and inconsistency remained high in each category.

The observed heterogeneity motivated a more precise exploratory moderation analysis combining both math and MC measures. An omnibus test revealed that the effect estimates marginally differed between the categories ($p$ = 0.0574). More precisely, a very large effect size was observed when an online measure of MC and a complex math task were employed ($n$ = 13, $r$ = 0.54, $I^2$ = 79%, $p$ = .058). Moderate effect sizes were observed when an offline measure of MC and a complex ($n$ = 8, $r$ = 0.35, $p$ = .001, $I^2$ = 86%), or unclear math tasks ($n$ = 48, $r$ = 0.32, $I^2$ = 100%, $p$ < .001) were employed. A low effect size was observed when an offline measure of MC and a simple math task were employed ($n$ = 4, $r$ = 0.10, $I^2$ = 87% $p$ > .99).

**Discussion**

The current paper investigated the association between MC and math performance in adolescents aged 11-16 years via a systematic review and quality assessment of existing research. In addition, it included a meta-analysis to investigate the strength of the association

between MC and math performance in adolescents across studies. The meta-analysis also considered whether measurements of MC (online vs. offline) and math performance (simple vs. complex) and their combination were important in understanding links between MC and performance in math tests. The systematic search yielded 31 studies. The synthesis of 74 effect sizes from 29 of these studies (N = 570,575, 30 independent populations) indicated a significantly positive, medium-sized correlation between MC and achievement in math ($r$ = .37, CI = [.29, .44], $p$ < .001). This relationship indicates that in a key stage of education where students are working towards exams critical for progression in further education or career pathways, individuals who showed or reported increased MC skill also performed better in math tasks. While the calculation of a pooled effect size generates an important understanding of the data, the association between MC and math performance indicated significant heterogeneity between studies. Moreover, efforts to reduce heterogeneity (via e.g., a sequential leave-one-out analysis, the removal of outliers, the exclusion of studies with low statistical power or possible risk of bias, exclusion of the two very large PISA studies) were unsuccessful in identifying its source.

Subgroup analyses explored moderators that may be potentially important in understanding whether theoretical and empirical differences between studies underpinned heterogeneity across studies. The current paper replicated the findings from previous reviews which have found that online (vs. offline) MC measures were most associated with performance in math assessments (Dent & Koenka, 2016; Ohtani & Hisasaka, 2018). We extended these analyses to more closely utilise researcher definitions of online versus offline MC across four separate analyses (see Saraç & Karakelle, 2012; Veenman & van Cleef, 2019). These considered whether MC processes were used by adolescents as they completed the math problem (e.g., using think-aloud methods), or occurred immediately before or after its completion (e.g., using JOL or calibration scores), or were measured outside of the math

task (i.e., using questionnaires). The results showed that the use of online measures were consistently associated with the largest effect size across analyses. Moreover, while heterogeneity remained high in most analyses, we found that when effect sizes were separated by MC measure (online vs. offline vs. JOL/calibration score), the 12 effect size estimates for the online measure were relatively consistent ($I^2 = 30\%$) and strong (11 out of 12 effect size estimates were stronger than $r = .40$). This finding supports the proposition that the active use of MC during math problems is associated with increased performance. Because only two research groups produced these 12 effect sizes, future studies conducted by different research groups will be important to assess the robustness of this finding.

Building on previous research (e.g., Mokos & Kafoussi, 2013; Verschaffel et al., 2010; review by Jordano & Touron, 2018), we further investigated in a subset of studies ($n = 16$) studies whether increased complexity of the math task was most linked with the use of MC processes, and could potentially explain heterogeneity across studies. In support of the hypothesis, the results showed that the effect size was largest when MC measures were linked to more complex math tasks, however heterogeneity for all analyses remained high and the number of papers that included simple maths task was small. Further exploratory moderation analyses showed that when combining our two moderators, the combination of an online MC measure and a complex math task produced the largest effect size ($r = .54$). Conversely, the combination of an MC offline measure and simple math task was associated with the smallest effect size ($r = .10$). Though heterogeneity in these analyses also remained high, the findings provide indicative evidence that the association between MC and performance may be stronger when adolescents are completing tasks that demand some awareness of strategic knowledge and ability to monitor performance and control cognitive processing as they move through the math problem (see e.g., Credé & Phillips, 2011; Nelson & Narens, 1990).

Our moderation analyses demonstrated that the heterogeneity in the overall association between MC and performance in math tasks could not be only attributed to the MC or math measures used in primary studies. Previous meta-analyses using similar moderation approaches to explain heterogeneity between studies considering MC and achievement have also reported high heterogeneity (e.g., Dent & Koenka, 2016; Ohatani & Hisasaka, 2018; Richardson et al., 2012). Consistent with the findings in this paper, for example, Ohatani & Hisasaka (2018) reported reductions in heterogeneity for moderation analysis including comparisons of online (versus offline) tasks, though it still remained moderately high.

Given the range of MC measures used in this evidence base (Gascoine et al., 2017), the varied contexts in which measures are taken (e.g., during learning, retrospectively during testing, or outside of the learning context altogether and via MC questionnaire scales), and the types of information the measures are intended to generate, variability between studies is not surprising. For example, offline measures reflect more stable trait-like characteristics indicating student awareness and potential use of MC strategies in their approach to solving math problems; while online measures capture thought processes associated with working through specific material (review by Jordano & Touron, 2018). One important goal for optimising the potential for math achievement in school should focus on identifying MC strategies that will help students to identify areas of the curriculum that require additional attention or study (e.g., Son & Metcalfe, 2000). Gascoine et al., (2017) reported that 60% of research exploring MC with children and adolescents use offline questionnaire measures. Questions about MC that are not specifically related to curriculum content currently being learned do not, however, help students specify which areas of e.g., math knowledge need more attention. For example, the response to the question, "I try to use strategies that have worked in the past" (Schraw & Dennison, 1994) does not inform students that while they

have a good understanding of geometry, they need to work more on algebra. Conversely, the generation of online item-by-item JOLs while studying math material may facilitate greater student awareness in making this type of discrimination.

This argument suggests that the process of making MC judgments during learning is neither static nor neutral (i.e., it can modify study habits that in turn can affect learning). In other words, MC judgements made during learning reflect both reactive (to guide students to specific material) and reciprocal (to understand what has been learned) processes. For example, several studies that have focused on JOLs suggest that the very act of making a MC judgement alters what is remembered (e.g., Fiacconi et al., 2019; Janes et al., 2018; Myers et al., 2020; Tekin & Rodiger, 2020; see Double et al., 2018 for a meta-analysis). This research suggests that students who focus on the quality of their learning by providing MC responses impacts what is learnt. Although JOL reactivity clearly indicates a link between MC judgements and learning outcomes, the mechanisms underpinning this relationship are unclear. Furthermore, specifying how JOL reactivity might affect the relationship between MC judgements and achievement is complicated by the fact that making (versus not making) JOLs sometimes improves learning, sometimes causes learning to deteriorate, and sometimes has no effect. Furthermore, to our knowledge, no study has investigated whether the magnitude of making a JOL (or another MC judgement) about learning produces differential reactivity. For example, if a student reports low levels of confidence that they would be able to recall learnt information in a later test (e.g., Myers et al., 2020), that judgement may lead to a selective enhancement of learning, and this behaviour may result in no (or even an inverse relationship) between MC and achievement. In contrast, if low levels of confidence do not change learning behaviour, then the correlation between the JOL and learning will be higher. The consideration of the dynamic and reciprocal interaction between MC judgements with subsequent strategies for learning and achievement are less relevant to offline measures,

or online measures made at test (e.g., retrospective confidence ratings), because these judgements do not have the potential to causally affect learning in the same way.

Another factor that can affect the relationship between MC measures and achievement is student ability. Students who perform poorly on a task often show poor MC insight into their own limitations, causing them to overestimate their abilities (e.g., Kruger & Dunning, 1999) or give MC ratings that are poorly related to performance (e.g., Higham & Arnold, 2007). With respect to lower-performing students, this effect is sometimes referred to as the "double curse" (Kruger & Dunning, 1999) or the "unskilled-and-unaware" phenomenon (Hartwig & Dunlosky, 2014). This phenomenon may be dependent on whether the MC measure being used is clearly integrated with the material being learned. Vuorre and Metcalfe (2021) noted that if academic tests are used that permit guessing (e.g., multiple-choice tests), for example, then "metacognitive misses" (i.e., correct guesses assigned low MC rating) can undermine the relationship between the MC measure and performance. Such misses occur more often with low-performing students, thereby lowering MC accuracy specifically for those students. In tasks that don't allow MC misses (e.g., recall of learnt material), the relationship between MC and performance may remain relatively intact. While the current paper focused on specific combinations of MC measures (online, offline) and math performance (simple, complex), these studies suggest that further aspects of the tasks employed to measure MC and task performance can affect the MC rating/performance link.

More generally, to understand the relationship between MC measures and academic performance, we suggest it is critical to examine the specifics of both the MC measure and the measure of performance under scrutiny. This principle is true not just of MC measures, but of other academic measures as well. For example, Muryama et al., (2013) found that student self-reported intrinsic motivation and deep learning strategies were unrelated to math achievement at 11 years of age. On the surface, this finding might seem like a

counterintuitive, however, the authors noted that students with high intrinsic motivation might have little concern about performing well on an upcoming test. Also, deep learning strategies may be slower and more effortful than more superficial learning strategies, which may be costly in tests written in the short term (but more effective over the long term). Collectively, these studies indicate that associations between student self-reported approaches to learning and achievement cannot be studied in a vacuum; the particulars of the measurement instruments, both those designed to measure MC and those designed to measure math achievement, are as fundamental to this relationship as the overarching concepts themselves.

The results of the current review and meta-analysis have gone someway to highlight that measurement of MC, math tasks and their combination are important in understanding associations between these variables. The methodological quality of studies in the review was acceptable or good, though there were difficulties accessing some papers. At face value, the results indicate that the use of MC strategies in learning math are best understood when this cognitive process is situated within the learning activity (via online tasks) and utilised when students are engaged with complex (versus simple) math tasks. They further indicate that existing conceptual differences between online, offline and JOLs, confidence judgments and calibration scores may be too simplistic. Nevertheless, the study highlighted significant heterogeneity between studies in all analyses. The discussion has focused on the measurement and dynamic interplay between MC and math achievement to start to understand this heterogeneity. It suggests that future research should focus more closely on how students utilise MC processes to change their own learning behaviour, and to understand how any adjustments are reflected in learning outcomes. In addition, further studies have highlighted other factors that could potentially moderate the association between MC and math achievement, and that have not been considered in existing research, including anxiety

(Moran 2016) and executive functioning (Steinmayr et al., 2010), and their interaction. Moreover, the TIMSS 2019 report (Mullis et al., 2020) highlighted a complex picture with respect to identifying factors beyond the influence of student self-reported MC on math achievement that future research could explore. This focus could include, for example, whether males and females utilise MC strategy more or less effectively, or whether access to learning resources in the home and school learning environments influences the development of MC and its use in the classroom.

**Declarations**

*Conflicts of Interest* The authors have no relevant financial or non-financial interests to disclose.

References

Ahmed, W., van der Werf, G., Kuyper, H., & Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: A growth curve analysis. *Journal of Educational Psychology*, *105*(1), 150–161. https://doi.org/10.1037/a0030160

Aschbacher, P. R., Koency, G., & Schacter, J. (1995). *Los Angeles Learning Center alternative assessments guidebook (resource paper no. 12*). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Aşık, G., & Erktin, E. (2019). Metacognitive experiences: Mediating the relationship between metacognitive knowledge and problem solving. *Egitim ve Bilim*, *44*(197), 85–103. https://doi.org/10.15390/EB.2019.7199

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*(1), 3–38. https://doi.org/10.1007/BF01326548

Bishara, S., & Kaplan, S. (2018). The relationship of locus of control and metacognitive knowledge of math with math achievements. *International Journal of Disability, Development and Education*, *65*(6), 631–648. https://doi.org/10.1080/1034912X.2018.1432033

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

Callan, G. L., & Cleary, T. J. (2019). Examining cyclical phase relations and predictive influences of self-regulated learning processes on mathematics task performance. *Metacognition and Learning*, *14*(1), 43–63. https://doi.org/10.1007/s11409-019-09191-x

Callan, G. L., Marchant, G. J., Finch, W. H., & German, R. L. (2016). Metacognition, strategies, achievement, and demographics: Relationships across countries. *Kuram ve*

*Uygulamada Egitim Bilimleri*, *16*(5), 1485–1502.

https://doi.org/10.12738/estp.2016.5.0137

Campbell, J. I. (2005). *Handbook of mathematical cognition*. Psychology Press.

Çetinkaya, P., & Erktin, E. (2002). Assessment of metacognition and its relationship with

reading comprehension, achievement, and aptitude. *Boğaziçi Üniversitesi Eğitim*

*Dergisi*, *19*(1), 1-11.

Chiu, M. M., Chow, B. W. Y., & Mcbride-Chang, C. (2007). Universals and specifics in

learning strategies: Explaining adolescent mathematics, science, and reading

achievement across 34 countries. *Learning and Individual Differences*, *17*(4), 344–365.

https://doi.org/10.1016/j.lindif.2007.03.007

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-

reports: systematic reviews of the value of self-report in metacognitive

research. *Metacognition and Learning*, *15*(2), 155-213.

Crawford, C., & Cribb, J. (2013). *Reading and math skills at age 10 and earnings in later*

*life: a brief analysis using the British cohort study.* Institute of Education, University of

London.

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the motivated strategies for

learning questionnaire. *Learning and Individual Differences*, *21*(4), 337–346.

https://doi.org/10.1016/j.lindif.2011.03.002

Critical Appraisal Skills Programme (2018). *CASP cohort study checklist*. CASP Checklists.

https://casp-uk.net/wp-content/uploads/2018/01/CASP-Cohort-Study-

Checklist_2018.pdf

Dennison, R. S., Krawchuk, C. M., Howard, B. C., & Hill, L. (1996, April 8-11). *The*

*development of a children's self-report measure of metacognition.* Annual meeting of

the American Educational Research Association, New York.

Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, *28*(3), 425–474. https://doi.org/10.1007/s10648-015-9320-8

Department for Education (2014). *Mathematics programmes of study: key stage 4 National curriculum in England*. https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study

Dermitzaki, I. (2005). Preliminary investigation of relations between young students' self-regulatory strategies and their metacognitive experiences. *Psychological Reports*, *97*(3), 759-768. https://doi.org/10.2466%2Fpr0.97.3.759-768

Desoete, A., & Roeyers, H. (2006). Metacognitive macroevaluations in mathematical problem solving. *Learning and Instruction*, *16*(1), 12–25. https://doi.org/10.1016/j.learninstruc.2005.12.003

Desoete, A., Roeyers, H., & Buysse, A. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities*, *34*(5), 435-447. https://doi.org/10.1177/002221940103400505

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*(3), 231–264. https://doi.org/10.1007/s11409-008-9029-x

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. Memory, 26(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Efklides, A. (2006). Metacognitive experiences: The missing link in the self-regulated learning process. *Educational Psychology Review*, *18*(3), 287–291. https://doi.org/10.1007/s10648-006-9021-4

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277-287. https://doi.org/10.1027/10169040.13.4.277.

Efklides, A., & Vlachopoulos, S. P. (2012). Measurement of metacognitive knowledge of self, task, and strategies in mathematics. *European Journal of Psychological Assessment, 28*(3), 227–239. https://doi.org/10.1027/1015-5759/a000145

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

Erktin, E. (2004). Teaching thinking for mathematics through the enhancement of metacognitive skills. *Research in the Schools, 11*(1), 3–13.

Fadlelmula, F. K., Cakiroglu, E., & Sungur, S. (2015). Developing a structural model on the relationship among motivational beliefs, self-regulated learning strategies, and achievement in mathematics. *International Journal of Science and Mathematics Education*, *13*(6), 1355-1375. https://doi.org/10.1007/s10763-013-9499-4

Fiacconi, C. M., Mitton, E. E., Laursen, S. J., & Skinner, J. (2019). Isolating the contribution of perceptual fluency to judgments of learning (JOLs): Evidence for reactivity in measuring the influence of fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 926. https://doi.org/10.1037/xlm0000766

Fitzpatrick, C. (1994). *Adolescent mathematical problem solving: The role of metacognition, strategies and beliefs* (ED374969). ERIC. https://eric.ed.gov/?id=ED374969

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Fusco, D. R. (1995). *The role of strategy, effort and unknown attributions in a metacognitive model of mathematical problem solving* [Doctoral dissertation, The City University of New York]. ProQuest Digital Collections. https://search.proquest.com/pqdtglobal/docview/304181779/19A59985BBA0439CPQ/1?accountid=13963

Garrett, A. J., Mazzocco, M. M. M., & Baker, L. (2006). Development of the metacognitive skills of prediction and evaluation in children with or without math disability. *Learning Disabilities Research & Practice, 21*(2), 77-88. https://doi.org/10.1111/j.1540-5826.2006.00208.x

Gascoine, L., Higgins, S., & Wall, K. (2017). The assessment of metacognition in children aged 4-16 years: A systematic review. *Review of Education*, *5*(1), 3–57. https://doi.org/10.1002/rev3.3077

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (1998). *Metacognition in educational theory and practice*. Routledge.

Haddad Center. (2012). *Battery math tests*. Bar Ilan University.

Harris, M. M. (2015). *The role of metacognition in a montessori environment and the effects on academic achievement* [Doctoral dissertation, Union University]. ProQuest Digital Collections. https://search.proquest.com/docview/1733667442?accountid=13963

Hartwig, M. K., & Dunlosky, J. (2014). The contribution of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over- (and under-) confidence. *Memory & Cognition, 42(1*), 164–173. https://doi.org/10.3758/s13421-013-0351-4

Hassan, N. M., & Rahman, S. (2017). Problem solving skills, metacognitive awareness, and mathematics achievement: A mediation model. *The New Educational Review*, *49*(3), 201-212.

Henfi, J. (1990). *Redactiesommen.* Ajodakt.

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

Higham, P. A., & Arnold, M. M. (2007). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology, 19,* 718–742. https://doi.org/10.1080/09541440701326121

Hodgen, J. & Pepper, D. (2010). *An international comparison of upper secondary mathematics education*. Nuffield Foundation 2010. https://wwwnuffieldfoundation.org.

Hong, E., & Peng, Y. (2004). *Test-taking strategies questionnaire*. Unpublished document, University of Nevada.

Ichihara, M., & Arai, K. (2006). Moderator effects of meta-cognition: A test in math of a motivational model. *Japanese Journal of Educational Psychology*, *54*(2), 199-210.

Jackson, D., White, I. R., & Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine, 31*(29), 3805–3820. https://doi.org/10.1002/sim.5453

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review*, 25(6), 2356–2364. https://doi.org/10.3758/s13423-018-1463-4

Jordano, M. L., & Touron, D. R. (2018). How often are thoughts metacognitive? Findings from research on self-regulated learning, think-aloud protocols, and mind-wandering. *Psychonomic Bulletin & Review*, *25*(4), 1269-1286. https://doi.org/10.3758/s13423-018-1490-1

Kramarski, B., Rich, I., Mevarech, Z., & Libereman, A. (2005). *The effect of metacognitive processes on achievement motivation and mathematical thinking among students in middle schools*. Bar Ilan University.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lucangeli, D., & Cornoldi, D. (1997). Mathematics and metacognition: What is the nature of the relationship? *Mathematical Cognition*, *3*(2), 121-139. https://doi.org/10.1080/135467997387443

Maras, K., Gamble, T., & Brosnan, M. (2019). Supporting metacognitive monitoring in mathematics learning for young people with autism spectrum disorder: A classroom-based study. *Autism*, *23*(1), 60–70. https://doi.org/10.1177/1362361317722028

Martín, E., Martínez-Arias, R., Marchesi, A., & Pérez, E. M. (2008). Variables that predict academic achievement in the Spanish compulsory secondary educational system: A longitudinal, multi-level analysis. *The Spanish Journal of Psychology*, *11*(2), 400–413. https://doi.org/10.1017/S113874160000442X

Martin, R., Hodgson, H., Maloney, A. & Rayner, I. (2014). *Pro bono economics report for national numeracy: Cost of outcomes associated with low levels of adult numeracy in the UK*. National Numeracy. https://www.probonoeconomics.com/sites/default/files/files/PBE%20National%20Numeracy%20costs%20report%2011Mar.pdf

Mathematics Diagnostic Testing Project (2006). *The MDTP assessment system*. The California State University/University of California Mathematics Diagnostic Testing Project. https://mdtp.ucsd.edu/assessments/index.html

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1-2), 49-63. https://doi.org/10.1023/A:1003088013286

Meltzer, L. J., Levine, M. D., Karniski, W., Palfrey, J. S., & Clarke, S. (1984). An analysis of the learning styles of adolescent delinquents. *Journal of Learning Disabilities*, *17*(10), 600-608. https://doi.org/10.1177/002221948401701006

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review*, *15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Missouri Department of Elementary and Secondary Education (1990). *Mastery and achievement tests: Guide to score, interpretation and Use*. Jefferson City.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … Whitlock, E. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Revista Espanola de Nutricion Humana y Dietetica*, *20*(2), 148–160. https://doi.org/10.1186/2046-4053-4-1

Mokos, E., & Kafoussi, S. (2013). Elementary students' spontaneous metacognitive functions in different types of mathematical problems. *Journal of Research in Mathematics Education*, *2*(2), 242-267. http://dx.doi.org/10.4471/redimat.2013.29

Montague, M., & Bos, C. S. (1990). Cognitive and metacognitive characteristics of eighth grade students' mathematical problem solving. *Learning and Individual Differences*, *2*(3), 371–388. https://doi.org/10.1016/1041-6080(90)90012-6

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*, 831–864. http://dx.doi.org/10.1037/bul0000051

Moreno, A. (2002). The assessment of metacognitive skills. In A. Marchesi & E. Martín (Eds.), *Secondary education assessment* (pp. 119-136).  Editorial SM.

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., & Fishbein, B. (2020). *TIMMS 2019 International results in maths and science*. International Association for the Evaluation of Educational Achievement. Boston College, USA. https://timssandpirls.bc.edu/timss2019/

Murayama, K., R. Pekrun, S. Lichtenfeld, and R. Vom Hofe. 2013. Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development 84 (4):* 1475–1490. doi:10.1111/cdev.12036

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, *48*, 1-14. https://doi.org/10.3758/s13421-020-01025-5

National Numeracy. (2020). *What is numeracy?* National Numeracy. https://www.nationalnumeracy.org.uk/what-numeracy

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 125–141). Academic. https://doi.org/10.1016/S0079-7421(08)60053-5

Ning, H. K. (2016). Examining heterogeneity in student metacognition: A factor mixture analysis. *Learning and Individual Differences*, *49*, 373–377. https://doi.org/10.1016/j.lindif.2016.06.004

Ofqual (2019). *An infographic: GCSEs in 2019.* Key Stage 3 and 4 Exam Marking, Qualifications and Results. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826795/GCSE_infographic_17__1_.pdf

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance.

*Metacognition and Learning*, *13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

O'Neil, H. F., & Abedi, J. (1996). Reliability and validity of a state metacognitive inventory: Potential for alternative assessment. *The Journal of Educational Research*, *89*(4), 234-245. https://doi.org/10.1080/00220671.1996.9941208

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Özcan, Z. Ç. (2016). The relationship between mathematical problem-solving skills and self-regulated learning through homework behaviours, motivation, and metacognition. *International Journal of Mathematical Education in Science and Technology*, *47*(3), 408–420. https://doi.org/10.1080/0020739X.2015.1080313

Özcan, Z. Ç., & Eren Gümüş, A. (2019). A modeling study to explain mathematical problem-solving performance through metacognition, self-efficacy, motivation, and anxiety. *Australian Journal of Education*, *63*(1), 116–134. https://doi.org/10.1177/0004944119840073

Özsoy, G. (2005). The relationship between problem solving skills and mathematical achievement. *Gazi University Journal of Education*, *25*(3), 179-190.

Özsoy, G. (2011). An investigation of the relationship between metacognition and mathematics achievement. *Asia Pacific Education Review*, *12*(2), 227–235. https://doi.org/10.1007/s12564-010-9129-6

Panaoura, A., & Philippou, G. (2003). *The construct validity of an inventory for the measurement of young pupils' metacognitive abilities in mathematics* (ED501054). ERIC. https://eric.ed.gov/?id=ED501054

Parsons, S (2002). *Basic skills and crime.* Basic Skills Agency.

https://discovery.ucl.ac.uk/id/eprint/1566250/

Paulus, N. Tsalas, J. Proust, B., & Sodian, B. (2014). Metacognitive monitoring of oneself

and others: Developmental changes during childhood and adolescence. *Journal of*

*Experimental Child Psychology, 122,* 153-165.

https://doi.org/10.1016/j.jecp.2013.12.011

Pearson Education. (2008). *AIMSweb® benchmark and progress monitoring system for*

*Grades K–8*. Bloomington.

Peng, Y., Hong, E., & Mason, E. (2014). Motivational and cognitive test-taking strategies

and their influence on test performance in mathematics. *Educational Research and*

*Evaluation*, *20*(5), 366–385. https://doi.org/10.1080/13803611.2014.966115

Perry, J., Lundie, D., & Golder, G. (2019). Metacognition in schools: What does the literature

suggest about the effectiveness of teaching metacognition in schools? *Educational*

*Review*, *7*(4), 483-500.  https://doi.org/10.1080/00131911.2018.1441127

Pintrich, P. R. (1991). *A manual for the use of the motivated strategies for learning*

*questionnaire (MSLQ).* (ED338122). ERIC. https://eric.ed.gov/?id=ED338122

Price, G., & Ansari, D. (2013). Dyscalculia: Characteristics, causes, and treatments.

*Numeracy*, *6*(1), 1-16. https://doi.org/10.5038/1936-4660.6.1.2

Pustejovsky, JE, & Tipton, E. (2021). Meta-Analysis with Robust Variance Estimation:

Expanding the Range of Working Models. OSF, https://osf.io/mq9hj/

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university

students' academic performance: A systematic review and meta-analysis. *Psychological*

*Bulletin*, *138*(2), 353.

Saraç, S., & Karakelle, S. (2012). On-line and off-line assessment of metacognition.

*International Electronic Journal of Elementary Education*, *4*(2), 301-315.

Sato, J., & Arai, K. (1998). The relation between the use of learning strategies, learning goals and causal attributions. *Tsukuba Psychological Research*, *20*, 115-124.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460-475. https://doi.org/10.1006/ceps.1994.1033

Schraw G., Olafson L., Weibel M. & Sewing, D. (2012). Metacognitive knowledge and field-based science learning in an outdoor environmental education program. In A. Zohar & Y. Dori (Eds.), *Metacognition in science education: Trends in current research* (pp. 57-77). Springer. https://doi.org/10.1007/978-94-007-2132-6_4

Schneider, W. (2008). The Development of Metacognitive Knowledge in Children and Adolescents: Major Trends and Implications for Education. *Mind, Brain and Education*, *2*, 114-121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418

Sink, C. A., Barnett, J. E., & Hixon, J. E. (1991, April 3-7). *Self-regulated learning and academic performance in middle school children.* Annual meeting of the American Educational Research Association, Chicago.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. Journal of Experimental Psychology: *Learning, Memory, and Cognition, 26(1)*, 204–221. https://doi.org/10.1037/0278-7393.26.1.204

Sperling, R. A., Howard, B. C., Miller, L. A., & Murphy, C. (2002). Measures of children's knowledge and regulation of cognition. *Contemporary Educational Psychology*, *27*(1), 51–79. https://doi.org/10.1006/ceps.2001.1091

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *International Journal of Phytoremediation*, *21*(1), 117–139. https://doi.org/10.1076/edre.10.2.117.27905

Steinmayr, R., Ziegler, M., & Träuble, B. (2010). Do intelligence and sustained attention interact in predicting academic achievement? *Learning and Individual Differences*, *20*(1), 14–18. https://doi.org/10.1016/j.lindif.2009.10.009

Tekin, E., & Rodiger, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift Für Psychologie, 228(4),* 278–290. https://doi.org/10.1027/2151-2604/a000425

Tian, Y., Fang, Y., & Li, J. (2018). The effect of metacognitive knowledge on mathematics performance in self-regulated learning framework-multiple mediation of self-efficacy and motivation. *Frontiers in Psychology*, *9*, 2518. https://doi.org/10.3389/fpsyg.2018.02518

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

Tobias, S., Everson, H. T., & Laitusis, V. (1999). *Towards a performance-based measure of metacognitive knowledge monitoring: Relationships with self-reports and behaviour ratings*. (ED432590). ERIC. https://eric.ed.gov/?id=ED432590

van der Stel, M., & Veenman, M. V. J. (2014). Metacognitive skills and intellectual ability of young adolescents: A longitudinal study from a developmental perspective. *European Journal of Psychology of Education*, *29*(1), 117–137. https://doi.org/10.1007/s10212-013-0190-5

van der Stel, M., Veenman, M. V. J., Deelen, K., & Haenen, J. (2010). The increasing role of metacognitive skills in math: A cross-sectional study from a developmental perspective.

*International Journal on Mathematics Education*, *42*(2), 219–229.

https://doi.org/10.1007/s11858-009-0224-2

van der Walt, M. S., Maree, J. G., & Ellis, S. M. (2008). Metacognition in the learning of mathematics in the senior phase :Some implications for the curriculum. *International Journal of Adolescence and Youth*, *14*(3), 205–235.

https://doi.org/10.1080/02673843.2008.9748004

Veenman, M. V. J., Kerseboom, L., & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress and Coping*, *13*(4), 391–412. https://doi.org/10.1080/10615800008248343

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211.

https://doi.org/10.1007/s11251-004-2274-8

Veenman, M. V. J., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: Students' self-reports versus on-line assessment methods. *Mathematics Education*, *51*(4), 691–701. https://doi.org/10.1007/s11858-018-1006-5

Veenman, M. V. J., van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3-14. https://doi.org/10.1007/s11409-006-6893-0

Veenman, M. V., & Spaans, M. A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*(2), 159-176.

https://doi.org/10.1016/j.lindif.2004.12.001

Verschaffel, L., van Dooren, W., Greer, B., & Mukhopadhyay, S. (2010). Reconceptualising word problems as exercises in mathematical modelling. *Journal Fur Mathematik-Didaktik*, *31*(1), 9–29. https://doi.org/10.1007/s13138-010-0007-x

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48. v36i03.pdf

Vuorre, M., & Metcalfe, J. (2021). Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacognition and Learning*. Advance online publication. https://doi.org/10.1007/s11409-020-09257-1-1

Walker, E. (2013). *Understanding the role of metacognition and working memory in math achievement* [Doctoral dissertation, University of Southampton]. University of Southampton Digital Collections. https://eprints.soton.ac.uk/358501/

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test (WRAT4)*. Psychological Assessment Resources.

Wolf, A. (2011). *Review of vocational education- The Wolf report*. Department for Education. http://www.educationengland.org.uk/documents/pdfs/2011-wolf-report-vocational.pdf

Wolters, C. A., Pintrich, P. R., & Karabenick, S. A. (2006). Assessing academic self-regulated learning. In K. A. Moore & L. H. Lippman (Eds.), *What do children need to flourish?* (pp. 251–270). Springer. https://doi.org/10.1007/0-387-23823-9_16

Yap, E. G. (1993). *A structural model of self-regulated learning in math achievement* [Doctoral dissertation, University of Southern California]. ProQuest Digital Collections. https://search.proquest.com/docview/1627936537?accountid=13963

Young, A. E., & Worrell, F. C. (2018). Comparing metacognition assessments of mathematics in academically talented students. *Gifted Child Quarterly*, *62*(3), 259–275. https://doi.org/10.1177/0016986218755915

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329–339. https://doi.org/10.1037/0022-0663.81.3.329

**Table 1**

*The Search Terms Inputted into Databases to Identify Relevant Studies*

| Metacognition | Maths | Performance |
|---|---|---|
| Metacogniti* | Math* | Performance |
| "Meta-cogniti*" | Arithmetic | Attainment |
| "Judgment* learn*" | Numeracy | Achievement |
| Metamemor* | Statistics | Grade |
| "Meta-memor*" | | Score |
| Metacomprehen* | | Mark |
| "Meta-comprehen*" | | |
| Metaknowledge | | |
| "Meta-knowledge" | | |
| "Metacognitive monitoring" | | |
| "Meta-cognitive monitoring" | | |
| Overconfiden* | | |
| "Over-confiden*" | | |
| "Under-confiden*" | | |
| "Self-assessment" | | |

*Note*. The Boolean operator "OR" was applied to the words within each column and the operator "AND" was applied to combine the three columns of words.

**Table 2**

*Details of the Studies That met the Inclusion Criteria Including Publication Information, Participant Characteristics, Design and Setting, Math Performance Measures (Characterised as Simple and Complex), Metacognitive Measures (Characterised as Online and Offline), Key Findings, and Quality Assessment Rating (QA)*

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 1 | Ahmed et al., (2013) | Journal article | The Netherlands | 495 (252) *M* = 12.8 years | N/A | Longitudinal/ Two secondary schools in two middle- | School assessment (Graded 1-10) *Unclear* | MSLQ (Wolters et al., 2006) *Offline* | *r* = .36 (*p* < .01) | 9 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | income suburban communities (21 classes) | | | | |
| 2 | Aşık & Erktin (2019) | Journal article | Turkey | 406 (195) *M* = 14 years | N/A | Correlational/ Three public and two private inner-city schools | Three word math problems (algebra/ arithmetical operations). Scored from 0 (entirely incorrect) – 4 (entirely | MSI (Çetinkaya & Erktin, 2002); MES (Efklides, 2006) - before the three math problems *Offline; online* | MSI: *r* = .40 (*p* < .01); MES: *r* = .53 (*p* < .01) | 6 |

| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | QA (/9) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | *Publication Information* | *Participant Characteristics* | *Design & Setting* | *Measures* |

| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance | Metacognition (MC) | MC and Math Performance Association (Sig.) | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | correct) based on the Holistic Scoring Rubric (Aschbacher et al., 1995) *Complex* | | | |
| 3 | Bishara & Kaplan (2018) | Journal article | Israel | 60 (26) Age not reported but assumed to be 13-14 years | 30 adolescents with learning disabilities enrolled in mixed classes in a | Correlational (group design)/ One public middle school | Math Aptitude Test (Haddad Center, 2012) - 10 questions set by the Ministry of Education's | MC questionnaire re knowledge of math (Kramarski et al., 2005; | Before problem-solving: $r = .69$ (sig., $p < .001$); During and after problem-solving: $rs = .83$ and $.69$ ($ps < .001$); | 6 |

| Study | | Publication Information | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | mainstream school | | curriculum for 7th-grade (score 0 to 10). *Unclear* | Montague & Bos, 1990) *Offline* | Beliefs about solving math problems: $r = .84$ ($p < .001$) | |
| 4 | Callan & Cleary (2019) | Journal article | The USA | 96 (54) Age not reported but assumed to be 13-14 years | 90.7% met the criteria for a free or reduced lunch | Correlational / One urban school | Three multi-step algebra word problems from a NAEP past assessment *Complex* | Pupils responded on a 1-7 scale to, "How sure are you that you solved this problem correctly?", and | $r = .37$ ($p < .001$) | 8 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | | this judgement was compared with actual performance to calculate an accuracy score *Offline* | | |
| 5 | Callan et al., (2016) | Journal article | Used PISA data (2009) from 63 countries | 475,460 (239,156) 15 years | N/A | Correlational / Schools in 63 countries | The PISA multiple-choice international achievement test | The PISA metacognitive indexes (understanding, remembering, summarising) - | $r = .46$ ($p < .001$) | 7 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | *Unclear* | students were asked how useful they thought various reading strategies were to solve a reading text *Offline* | | |
| 6 | Chiu et al., (2007) | Journal article | Uses PISA data (2000) | 88,590 (gender not reported) | N/A | Correlatio nal/ | PISA multiple-choice achievement test | PISA self-reported metacognitive | $r = .04$ (not sig.) | 9 |

| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | QA (/9) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | from 34 countries | but sample included males and females) 15 years | | Schools in 34 countries | *Unclear* | strategy use questionnaire *Offline* | | |
| 7 | Erktin (2004) | Journal article | Turkey | 100 (not reported but approximately 50% of the total | N/A | Correlational/ One inner-city private high school | A researcher-designed multiple-choice test on probability *Unclear* | MSI (Çetinkaya & Erktin, 2002) *Offline* | $r = .42$ ($p < .05$) | 7 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | sample were reported to be female) Age not reported but assumed to be 15-16 years | | | | | | |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 8 | Fadlelmul a et al., (2015) | Journal article | Turkey | 1019 (48) Assumed to be 13 years | N/A | Correlatio nal/ 11 inner-city public schools (34 classroom s) | A 10-item multiple-choice test consisting of items linked to studied topics (numbers, geometry, algebra) *Unclear* | Items (planning, monitoring, regulating) from the MSLQ (Pintrich, 1991)- Participants were asked to think about math when answering items *Offline* | $\beta = -1.41$ (not sig.) | 6 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 9 | Fitzpatrick (1994) | Conference paper | The USA | 100 (50) Age not reported but assumed to be 14-17 years | Students were selected based on their Preliminary Scholastic Aptitude Test (MPSAT) score: 50 participant' scores fell on or below the 40th centile, and 50 scores fell on or | Correlational (group design)/ Two urban public and three private high schools | Six multiple-choice questions from the 1987 SAT math test Three questions were word problems *Unclear* | The researcher-designed metacognition awareness assessment (MAA) *Offline* | $r = .28$ (not sig.) | 5 |

56

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | above the 60th centile | | | | | |
| 10 | Fusco (1995) | Doctoral thesis | The USA | 30 (30) 13-16 years, most were 14-15 years | Students selected based on how they attributed math problem-solving performance: (n=10 students each attributed strategy, effort, | Correlational (group design)/ One urban Catholic high school | One non-routine word problem. Scored from 1 to 5 (5 = completely correct answer) *Complex* | Think-aloud protocols and behaviour observations during problem-solving *Online* | $r = .68$ (sig., $p < .001$) | 7 |

57

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | unknown causes) | | | | | |
| 11 | Harris (2015) | Doctoral thesis | The USA | 27 (not reported but sample included males and females) 11-14 years | N=6 students with learning disabilities. (Students with language impairment, autism and intellectual giftedness were excluded) | Correlational/ One public Montessori school | A standardised grade-level skills assessment in mathematics (AIMS web; Pearson Education, 2008) *Simple* | Jr. MAI (Dennison et al., 1996) *Offline* | $\beta$ = -.30 (not sig.) | 7 |

| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | QA (/9) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | **Design & Setting** | **Measures** | | **Key Findings** | |
| 12 | Hassan & Rahman (2017) | Journal article | Malaysia | 333 (not reported) Age not reported but assumed to be 15-16 years | N/A | Correlational/ Ten secondary schools | Not reported *Unclear/missing* | Offline Questionnaire based on the MAI (Schraw & Dennison, 1994) *Offline* | β = .48 (*p* < .001) | 2 |
| 13 | Ichihara & Arai (2006) | Journal article | Japan | 543 (264) Age not reported but | N/A | Correlational/ One public | End of term school assessment (score = 0 - 100) | MQ (Sato & Arai, 1998) *Offline* | *r* = .31 (sig.) | 7 |

| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
|---|---|---|---|---|---|---|---|---|---|---|
| | (*Japanese*) | | | assumed to be 12-14 years | | junior high school | *Unclear* | | | |
| 14 | Maras et al., (2019) | Journal article | The UK | 49 (18) *M* = 13.4 years (11-15 years) | Participants were working at age-related expectations in math | Intervention experiment *One secondary school* | Three mental math questions (block one) selected from past papers of national UK examinations or revision | Four questions relating to awareness of own performance, confidence, and strategies used | $r_s$ = -.12 (not sig.) | 8 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | workbooks. Questions were scored as correct or incorrect *Simple* | during the math task *Offline* | | |
| 15 | Martín et al., (2008) | Journal article | Spain | 965 (435) Most were 12-13 years old | N/A | Longitudinal/ 17 private and ten public inner-city | A multiple-choice researcher-designed test *Simple* | Four scales: meta-comprehension (accuracy of predicted score), verification of one's results, the | $\beta = 3.24$ ($p < .001$) | 9 |

|  | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
|  |  |  |  |  |  | secondary schools |  | consciousness of the strategies one uses and consciousness of one's own comprehension (Moreno, 2002) *Offline* |  |  |
| 16 | Ning (2016) | Journal article | Singapore | 873 (441) *M* = 15.36 years | N/A | Correlational/ 10 schools | A multiple-choice standardised test *Unclear* | Jr. MAI (Dennison et al., 1996) *Offline* | Knowledge of cognition scale: *r* = .00 (not sig.); | 6 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 17 | Özcan (2016) | Journal article | Turkey | 268 (not reported after attrition but 145 of 323 of the original sample, 45%, | N/A | Correlatio nal/ Two inner-city public schools | Six math problems related to problems in the students' course textbooks. Each scored from 0 to 4 with 4 | YPMAiM (Panaoura & Philippou, 2003); MES (Efklides, 2006) *Offline; online* | Regulation of cognition scale *r* = .07 (not sig.) YPMAiM: *r* = .17 (*p* < .05); MES: *r* = .33 (*p* < .01) | 7 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | were female) Age not reported but assumed to be 11-14 years | | | indicating a wholly correct and clear answer *Complex* | | | |
| 18 | Özcan & Eren Gümüş (2019) | Journal article | Turkey | 517 (265) Age not reported but assumed | N/A | Correlational/ Two inner-city public | Four multi-step word problems on linear equations | MES (Efklides, 2006) *Offline* | $r = .5$ ($p < .01$) | 8 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | to be 12-13 years | | middle schools | taken from the seventh-grade coursebook. Responses were scored using the Holistic Scoring Rubric (0-4 where 4 is a completely correct answer) *Complex* | | | |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 19 | Özsoy (2011) | Journal article | Turkey | 242 (134) *M = 11.3 years* | N/A | Correlational/ Six urban public schools | Researcher designed math test designed (Özsoy, 2005) *Unclear* | MSA-TR (Desoete et al., 2001) *Offline* | $r = .65$ ($p < .01$) | 6 |
| 20 | Peng et al., (2014) | Journal article | China | 438 (256) 15-16 years | N/A | Correlational/ One inner-city high school | Final test score (not reported but assumed to be an end of year school-administered test) | Items from the TTSQ (Hong & Peng, 2004) *Offline* | Planning scale: $r = .11$ ($p < .05$); Self-checking scale: $r = .05$ (not sig.); Strategy | 6 |

| Study | | Publication Information | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | *Unclear* | | selection scale: $r$ = .04 (not sig.) | |
| 21 | Sink et al., (1991) | Conference paper | The USA | 62 (34) $M$ = 11.6 years (11-13 years) | N/A | Correlational/ One middle school in a small town | School assessment (teacher-designed math test); MMAT (Missouri Department for Elementary and Secondary | Accuracy of predicted test score (to actual achieved score) *Offline* | School assessment: $r$ = .29 ($p$ < .05); MMAT: $r$ = .43 (sig., $p$ < .01) | 8 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | Education, 1990) *Unclear* | | | |
| 22 | Tian et al., (2018) | Journal article | China | 569 (324) *M* = 16.39 years | N/A | Correlational/ One high school | Three successive mathematics examinations *Unclear* | MKMQ (Efklides & Vlachopoulos, 2012) *Offline* | Separate correlations reported for each of the three math exams. MK of self (easiness/fluency): $r$s ≥ .19, $p$s < .05; MK of self (difficulty/lack of | 8 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | | | fluency): $r$s $\geq$ .14, $p$s < .05; MK of tasks (easy/low demands): $r$s -.05 to -.06, $p$s > .05; MK of tasks (difficult/high demands): $r$s $\leq$ -.17, $p$s < .05; MK of strategies (cognitive/metacognitive strategies): $r$s $\geq$ .026, $p$s < .05; | |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | | | MK of strategies (competence-enhancing strategies): $r$s $\geq .16$, $p$s $< .05$; MK of strategies (avoidance strategies): $r$s $\leq -.15$, $p$s $< .05$ | |
| 23 | van der Stel & Veenman (2014) | Journal article | The Netherlands | 25 (not reported but sample included | Students with known learning or conduct | Longitudinal/ One urban | Five word problems adapted from an age-appropriate | Think-aloud protocols were analysed for use of metacognitive skills, according | (Semi-partial correlations account for intellectual ability) | 9 |

|  | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
|  |  |  |  | both males and females) 13 years | disorders were excluded | secondary school | math textbook – one point for using the correct procedure and one point for a correct answer *Complex* | to the quantity (frequency) and quality of utterances *Online* | Quality of utterances: $r = .70$ (semi-partial $= .30$) (sig., $p < .01$); Quantity of utterances: $r = .73$ (semi-partial $= .30$) (sig., $p < .01$) |  |
| 24 | van der Stel et al., (2010) | Journal article | The Netherlands | 59 (36) 13-15 years | N/A | Correlational (group design)/ | Five (2nd years) or six (3rd years) word problems | Think-aloud protocols were analysed for MC skills, according | *Second-year participants* Quality of utterances: $r = .53$ | 9 |

| Study | | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | Two suburban schools | adapted from a commonly used math textbook (maximum of 10 points per question) *Complex* | to the quantity (frequency) and quality of utterances *Online* | (sig., $p < .01$) ; Quantity of utterances: $r$ = .29) (not sig.) *Third-year participants* Quality of utterances: $r$ = .78 (sig., $p < .01$) ; Quantity of utterances: r = .40 (sig., $p < .01$) | |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 25 | van der Walt et al., (2008) | Journal article | South Africa | 339 (199, 58.7%) $M$ = 13.64 years (12-17 years) | N/A | Correlational/ Six urban schools | School assessment (exam score); A geometry problem (calculate the surface area of a parallelogram within a rectangle) *Unclear; complex* | The Lucangeli-Cornoldi instrument (Lucangeli & Cornoldi, 1997) was used while pupils were solving the geometry problem *Offline* | School assessment: Prediction of success: $r_s$ = .26 ($p < .05$); Degree to which learner could monitor steps in the solution: $r_s$ = .21 ($p < .05$); Evaluation of success: $r_s$ = .30 ($p < .05$); | 7 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | | | Reflection on solution: $r_s$ = .11 (< .05) Geometry problem: Prediction of success: $r_s$ = .37 ($p < .05$); Degree to which learner could monitor steps in the solution: $r_s$ = .33 ($p < .05$); | |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | | | Evaluation of success: $r_s = .39$ ($p < .05$); Reflection on solution: $r_s = .04$ ( | |
| 26 | Veenman et al., (2000) | Journal article | The Netherlands | 30 (not reported) 12-13 years | Students were selected based on anxiety questionnaire scores: 20 reported high anxiety and 10 reported low anxiety | Intervention experiment / One secondary school | Three mathematical word problems, adapted from Henfi (1990). Scored as correct (1 point) or | Systematic behaviour observations and analysis of think-aloud protocols for the quality of MC skilfulness | Systematic observation: $r = .41$ (.38 corrected for extreme anxiety groups) ($p < .05$); Think-aloud protocols: $r = .52$ | 5 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | incorrect (0 points) *Complex* | during problem-solving *Online* | (.50 corrected) (*p* < .01) | |
| 27 | Veenman et al., (2005) | Journal article | The Netherlands | 41 (~ 50%) 12-13 years | N/A | Intervention experiment/ Two urban secondary schools | Three math word problems, adapted from Henfi (1990). Scored as correct (1 point) or incorrect (0 | Systematic behaviour observations and analysis of think-aloud protocols *Online* | Semi-partial correlations account for intellectual ability Word problems: *r* = .48 (semi-partial = .47) (*p* < .01); Math | 9 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | | | points); GPA for math at the end of the previous school year *Complex; unclear* | | GPA: $r = .40$ (semi-partial= .30) ($p < .01$) | |
| 28 | Veenman & Spaans (2005) | Journal article | The Netherlands | 31 (18) *M* = 13.9 years | Pupils selected based on intelligence test score: 16 pupils scored as low in | Correlational (group design)/ One urban | Six math problems adapted from Henfi (1990). Scored as correct (1 | Systematic behaviour observations and analysis of think-aloud protocols | (Semi-partial correlations account for intelligence) $r = .75$ (semi-partial = .35, *p* | 7 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | intelligence and 15 scored highly | secondary school | point) or incorrect (0 points *Complex* | *Online* | < .01) (corrected for extreme intelligence groups, *r* = .66, semi-partial= .45) | |
| 29 | Walker (2013) | Doctoral thesis | The UK | 18 (11) *M* = 13.15 years (13-14 years) | All had made age-appropriate progress in math by the end of primary school but had not made expected | Intervention experiment/ One secondary school | The oral math scale and computation scale from the WRAT4 (Wilkinson & Robertson, 2006) | Jr. MAI; (Sperling et al., 2002) *Offline* | *r* = -.14 (not sig.) | 7 |

| Study | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | | progress at secondary school | | *Simple* | | | |
| 30 | Yap (1993) | Doctoral thesis | The USA | 591 (285) Age not reported but assumed to be 13-14 years | N/A | Correlational/ 18 schools | NAEP math tests (standardised)- 41 multiple-choice items *Simple* | The self-checking subscale from the state self-regulatory inventory (O'Neil & Abedi, 1996) *Offline* | $r = .21$ ($p < .01$) | 9 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| 31 | Young & Worrell (2018) | Journal article | The USA | 179 (math grade and GPA available) / 183 (MDT and summer course data available) (97) | Attending a university summer programme for academically talented youth | Correlational/ One university summer programme | Most recent math school grade (MG); GPA for math; Mathematics diagnostic test (MDT; Mathematics Diagnostic Testing Project, 2006); Final course grade in a math course at | Jr. MAI (Sperling et al., 2002) *Offline* | MG: $r$ = .05 (not sig.); GPA: $r$ = .00 (not sig.); MDT: $r$ = -.12 (not sig.); SCG: $r$ = .01 (not sig.) | 7 |

| | Publication Information | | | Participant Characteristics | | Design & Setting | Measures | | Key Findings | QA |
|---|---|---|---|---|---|---|---|---|---|---|
| Study | Author(s) (Year) | Publication Type | Country | N (Females) Age | Defining Characteristics | Design/ Setting | Math Performance *(Simple vs. Complex vs. Unclear)* | Metacognition (MC) *(Online vs. Offline[a])* | MC and Math Performance Association (Sig.) | (/9) |
| | | | | *M* = 13.29 years (11-17 years) | | | the end of the summer program (SCG) *Unclear* | | | |

*Note*. GPA = grade point average; Jr. MAI = Junior metacognitive awareness inventory; MC = metacognition; MES = Metacognitive experiences scale; MI = Metacognitive inventory; MKMQ = Metacognitive knowledge in mathematics questionnaire; MMAT = Missouri mastery and achievement test; MQ = Metacognitive questionnaire; MSA-TR = Metacognitive knowledge and skills assessment; MSI = Metacognitive skills inventory; MSLQ = Motivated strategies for learning questionnaire; N = number of participants; NAEP = National association of education programme; PISA = Programme for international student assessment; QA = quality assessment; SAT = Scholastic assessment test; sig = statistical significance/statistically significant; T1 = time 1 (pre-intervention); T2 = time 2 (post-intervention); TTSQ = Test-taking strategies questionnaire; WRAT4 = Wide-ranging achievement test, 4th Edition; YPMAiM = The young pupils' metacognitive abilities in mathematics; β = beta coefficient; $p$ = .05 = 95% confidence in significance; $p < .01$ = 99% confidence in significance, $r$ = Pearson correlation coefficient

ᵃCategorisation according to the distinction proposed by Veenman & van Cleef (2019). See Supplementary Materials Table D for categorisations according to alternative distinctions between online and offline metacognition.

**Fig 1**

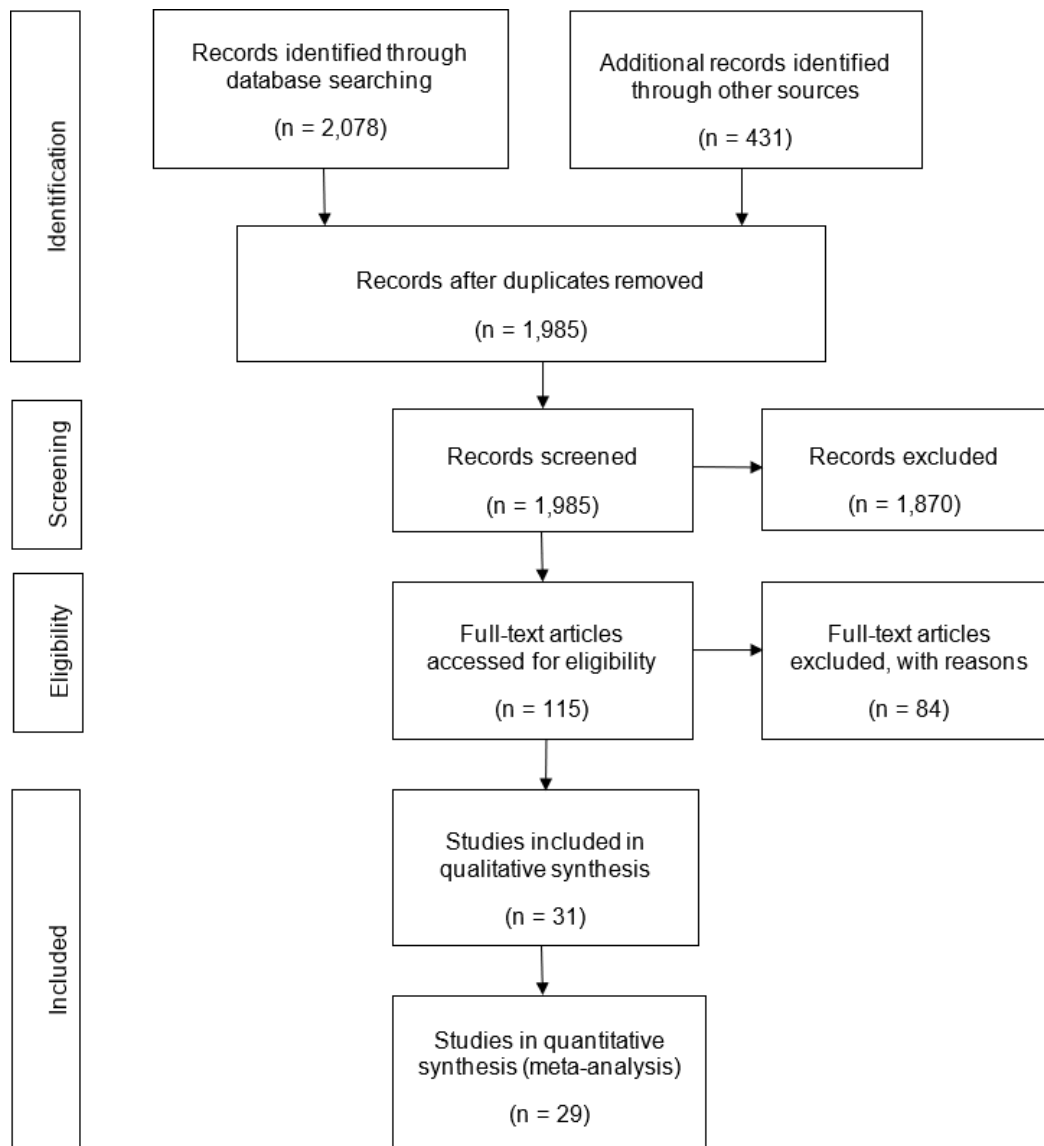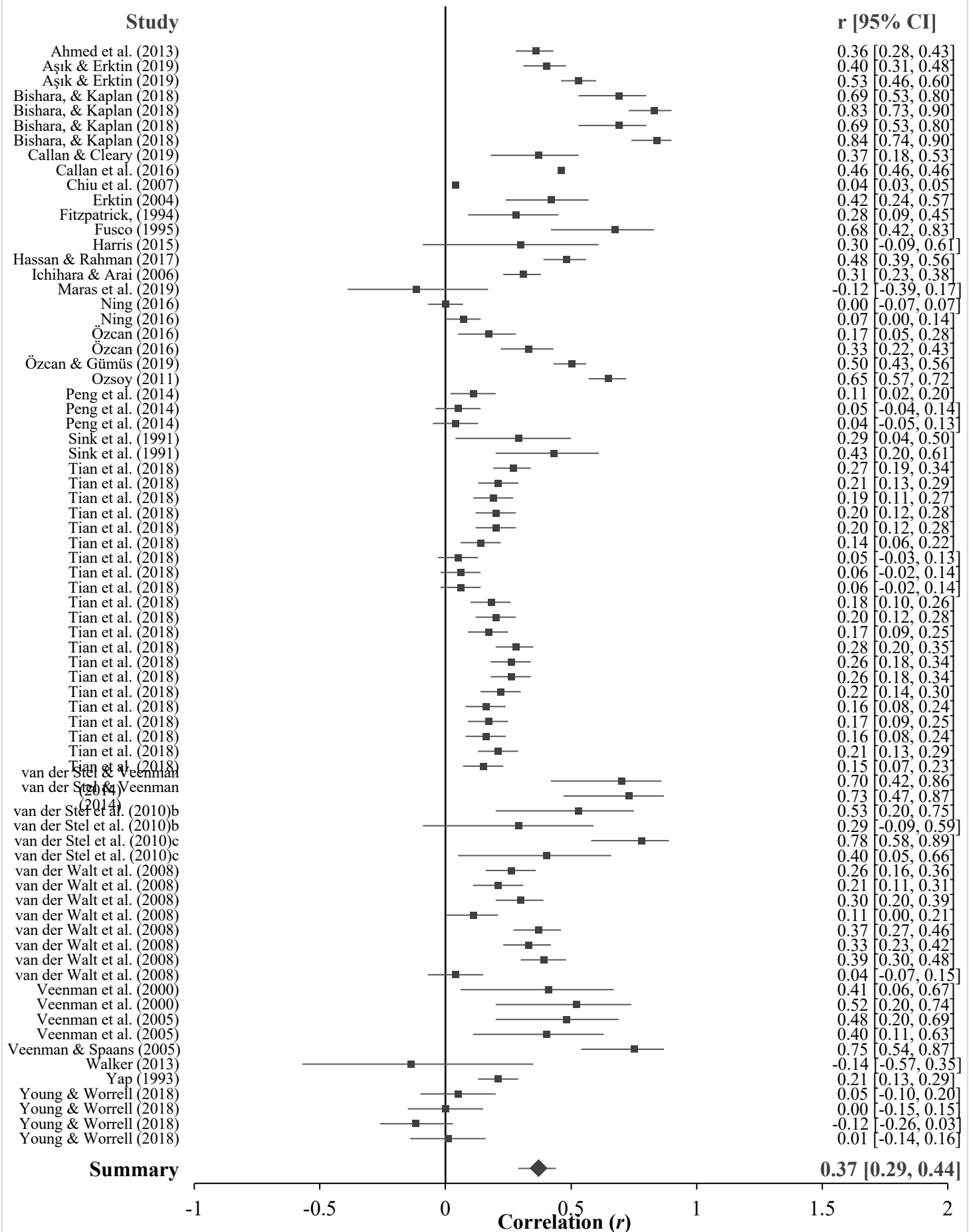*The Process by Which the Final Sample of Studies was Reached in the Current Systematic Review and Meta-Analysis*

**Fig 2.** *A Forest Plot of the Effect Sizes for Each Study and the Overall Effect Size*

| Study | r [95% CI] |
|---|---|
| Ahmed et al. (2013) | 0.36 [0.28, 0.43] |
| Aşık & Erktin (2019) | 0.40 [0.31, 0.48] |
| Aşık & Erktin (2019) | 0.53 [0.46, 0.60] |
| Bishara, & Kaplan (2018) | 0.69 [0.53, 0.80] |
| Bishara, & Kaplan (2018) | 0.83 [0.73, 0.90] |
| Bishara, & Kaplan (2018) | 0.69 [0.53, 0.80] |
| Bishara, & Kaplan (2018) | 0.84 [0.74, 0.90] |
| Callan & Cleary (2019) | 0.37 [0.18, 0.53] |
| Callan et al. (2016) | 0.46 [0.46, 0.46] |
| Chiu et al. (2007) | 0.04 [0.03, 0.05] |
| Erktin (2004) | 0.42 [0.24, 0.57] |
| Fitzpatrick, (1994) | 0.28 [0.09, 0.45] |
| Fusco (1995) | 0.68 [0.42, 0.83] |
| Harris (2015) | 0.30 [-0.09, 0.61] |
| Hassan & Rahman (2017) | 0.48 [0.39, 0.56] |
| Ichihara & Arai (2006) | 0.31 [0.23, 0.38] |
| Maras et al. (2019) | -0.12 [-0.39, 0.17] |
| Ning (2016) | 0.00 [-0.07, 0.07] |
| Ning (2016) | 0.07 [0.00, 0.14] |
| Özcan (2016) | 0.17 [0.05, 0.28] |
| Özcan (2016) | 0.33 [0.22, 0.43] |
| Özcan & Gümüs (2019) | 0.50 [0.43, 0.56] |
| Ozsoy (2011) | 0.65 [0.57, 0.72] |
| Peng et al. (2014) | 0.11 [0.02, 0.20] |
| Peng et al. (2014) | 0.05 [-0.04, 0.14] |
| Peng et al. (2014) | 0.04 [-0.05, 0.13] |
| Sink et al. (1991) | 0.29 [0.04, 0.50] |
| Sink et al. (1991) | 0.43 [0.20, 0.61] |
| Tian et al. (2018) | 0.27 [0.19, 0.34] |
| Tian et al. (2018) | 0.21 [0.13, 0.29] |
| Tian et al. (2018) | 0.19 [0.11, 0.27] |
| Tian et al. (2018) | 0.20 [0.12, 0.28] |
| Tian et al. (2018) | 0.20 [0.12, 0.28] |
| Tian et al. (2018) | 0.14 [0.06, 0.22] |
| Tian et al. (2018) | 0.05 [-0.03, 0.13] |
| Tian et al. (2018) | 0.06 [-0.02, 0.14] |
| Tian et al. (2018) | 0.06 [-0.02, 0.14] |
| Tian et al. (2018) | 0.18 [0.10, 0.26] |
| Tian et al. (2018) | 0.20 [0.12, 0.28] |
| Tian et al. (2018) | 0.17 [0.09, 0.25] |
| Tian et al. (2018) | 0.28 [0.20, 0.35] |
| Tian et al. (2018) | 0.26 [0.18, 0.34] |
| Tian et al. (2018) | 0.26 [0.18, 0.34] |
| Tian et al. (2018) | 0.22 [0.14, 0.30] |
| Tian et al. (2018) | 0.16 [0.08, 0.24] |
| Tian et al. (2018) | 0.17 [0.09, 0.25] |
| Tian et al. (2018) | 0.16 [0.08, 0.24] |
| Tian et al. (2018) | 0.21 [0.13, 0.29] |
| Tian et al. (2018) | 0.15 [0.07, 0.23] |
| van der Stel & Veenman (2014) | 0.70 [0.42, 0.86] |
| van der Stel & Veenman (2014) | 0.73 [0.47, 0.87] |
| van der Stel et al. (2010)b | 0.53 [0.20, 0.75] |
| van der Stel et al. (2010)b | 0.29 [-0.09, 0.59] |
| van der Stel et al. (2010)c | 0.78 [0.58, 0.89] |
| van der Stel et al. (2010)c | 0.40 [0.05, 0.66] |
| van der Walt et al. (2008) | 0.26 [0.16, 0.36] |
| van der Walt et al. (2008) | 0.21 [0.11, 0.31] |
| van der Walt et al. (2008) | 0.30 [0.20, 0.39] |
| van der Walt et al. (2008) | 0.11 [0.00, 0.21] |
| van der Walt et al. (2008) | 0.37 [0.27, 0.46] |
| van der Walt et al. (2008) | 0.33 [0.23, 0.42] |
| van der Walt et al. (2008) | 0.39 [0.30, 0.48] |
| van der Walt et al. (2008) | 0.04 [-0.07, 0.15] |
| Veenman et al. (2000) | 0.41 [0.06, 0.67] |
| Veenman et al. (2000) | 0.52 [0.20, 0.74] |
| Veenman et al. (2005) | 0.48 [0.20, 0.69] |
| Veenman et al. (2005) | 0.40 [0.11, 0.63] |
| Veenman & Spaans (2005) | 0.75 [0.54, 0.87] |
| Walker (2013) | -0.14 [-0.57, 0.35] |
| Yap (1993) | 0.21 [0.13, 0.29] |
| Young & Worrell (2018) | 0.05 [-0.10, 0.20] |
| Young & Worrell (2018) | 0.00 [-0.15, 0.15] |
| Young & Worrell (2018) | -0.12 [-0.26, 0.03] |
| Young & Worrell (2018) | 0.01 [-0.14, 0.16] |
| **Summary** | **0.37 [0.29, 0.44]** |

*The Boxes Represent the Effect Size (r) for Each Study, the Lines Represent 95% Confidence Intervals, and the Diamond Represents the Synthesised Effect Size.*

*b Second-year students c Third-year students*