Metacognition and Learning (2020) 15:155–213
https://doi.org/10.1007/s11409-020-09222-y

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

Kym Craig[1] · Daniel Hale[1] · Catherine Grainger[2] · Mary E. Stewart[1]

Received: 4 June 2019 / Accepted: 2 March 2020 / Published online: 9 May 2020

Abstract
Metacognitive skills have been shown to be strongly associated with academic achievement and serve as the basis of many therapeutic treatments for mental health conditions. Thus, it is likely that training metacognitive skills can lead to improved academic skills and health and well-being. Because metacognition is an awareness of one's own thoughts, and as such is not directly observable, it is often measured by self-report. This study reviews and critiques the use of self-report in evaluating metacognition by conducting systematic reviews and a meta-analysis of studies assessing metacognitive skills. Keyword searches were performed in EbscoHost, ERIC, PsycINFO, PsycArticles, Scopus, Web of Science, and WorldWideScience.org to locate all articles evaluating metacognition through self-report. 24,396 articles from 1982 through 2018 were screened for inclusion in the study. Firstly, a systematic review of twenty-two articles was conducted to review the ability of self-report measures to evaluate a proposed taxonomy of metacognition. Secondly, a systematic review and meta-analyses of 37 studies summarizes the ability of self-report to relate to metacognitive behavior and the possible effects of differences in research methods. Results suggest that self-reports provide a useful overview of two factors – metacognitive knowledge and metacognitive regulation. However, metacognitive processes as measured by self-report subscales are unclear. Conversely, the two factors of metacognition do not adequately relate to metacognitive behavior, but subscales strongly correlate across self-reports and metacognitive tasks. Future research should carefully consider the role of self-reports when designing research evaluating metacognition.

Keywords Metacognition · Cognitive ability · Self-report · Factor structure · Psychological theories · Student characteristics

✉ Kym Craig
kc71@hw.ac.uk

Extended author information available on the last page of the article

Importance.

Flavell (1979) was the first to utilize the term metacognition. He defined it as "thinking about thinking" and described metacognition as one's awareness of and understanding of their own and other's thoughts. Since then a variety of interpretations and adjustments of Flavell's original definition have been made. Currently, most researchers subscribe to the notion that metacognition involves processes that monitor and increase the efficiency of cognitive procedures (Akturk and Sahin 2011; Bonner, 1998; Van Zile-Tamsen 1996). In other words, metacognition encapsulates an awareness of one's own learning and comprehension, the capacity to evaluate the demands of a task and subsequently choose the appropriate strategy for task completion, the ability to monitor one's progress towards a goal and adjust strategy usage, the ability to reflect on one's decision making process, and the ability to discern the mental states of others (Beran 2012; Flavell 1979; Lai 2011). Metacognition, then, is essential for learning, and training metacognitive skills has been repeatedly shown to increase academic achievement (e.g. Brown 1978; Bryce et al. 2015; Flavell 1979; Perry et al. 2018; van der Stel and Veenman 2010; van der Stel and Veenman 2014; Veenman and Elshout 1994; Veenman and Spaans 2005; Wang et al. 1993). Furthermore, therapies grounded in metacognition have been successful in treating those with mental health conditions (Wells 2011).

Because metacognition is defined as an awareness of one's own thought processes and as such is not easily observed, it is difficult to measure. The most cost effective and efficient way to evaluate metacognitive skills is through a self-report questionnaire. Currently, there is not a self-report questionnaire that is considered the industry standard. Instead there is a wide range of questionnaires that measure a variety of components of metacognition (see Table 1 for a complete list of the evaluated self-reports). Employing a wide range of self-report assessments that evaluate a variety of metacognitive components results in an inconsistent understanding of the concept of metacognition and may affect how lay personnel, such as teachers and therapists, work directly with the metacognitive skills of those in their care. Therefore, the aim of this work is to critique the value of self-reports in metacognitive research by summarizing their ability to measure metacognition in two inter-related but distinct reviews:

1) a systematic review of the entire body of metacognitive literature that evaluates whether self-report can adequately measure the distinct components of metacognition being assessed by the researcher's purported taxonomy
2) a separate systematic review and meta-analysis that analyzes the ability of self-report to adequately measure all aspects of purported taxonomies and the ability of self-report scales to relate metacognitive components to metacognitive behavior.

To our knowledge this is the first systematic review and meta-analysis to comprehensively investigate the use of self-report measures and their utility as a valid measure of distinct metacognitive components.

This review and meta-analysis were conducted and reported in accordance with the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement (Moher et al. 2009). Because both the systematic review and meta-analyses were not medical in nature, and do not investigate interventions, published scales for assessing risk of bias were not applicable. Consequently, bias was assessed following The Cochrane Collaboration's (2011) recommendation of a domain-based evaluation.

Table 1 Studies Evaluating the Factor Structure of Metacognition

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
Akin et al, 2007	Metacognitive Awareness Inventory translated to Turkish (MAI) – 52 item self-report questionnaire (Schraw & Dennison, 1994)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	607 University students Mean age = 20 310 men	Validity: Correlated with English MAI Internal Consistency: Cronbach's alpha excellent Structure: Multiple EFAs run to find Schraw & Dennison's 8 factors	r = .93 α = .95 Loadings ranged from .32 to .83	CFA was not run
Aydin & Ubuz, 2010	Junior Metacognitive Awareness Inventory B in Turkish (JrMAI) – 18 item Self-report questionnaire (Sperling et al, 2002)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	EFA – 314 10th grade students aged 17-18 142 boys CFA – 589 10th grade students aged 17-18 286 boys	Internal Consistency: Cronbach's alphas acceptable Structure: EFA found 4 factors EFA run again as 2 factors, one item failed to load and was removed. CFA run on one factor CFA run on two-factor model Two-factor model is better fit	KOC α = .75 ROC α = .79 49.3% of the variance 37.17% of the variance One-factor RMR = .06 GFI = .87 AGFI = .84 CFI = .79 RMSEA = .09 Two factor RMR = .05 GFI = .94 AGFI = .92 CFI = .91 RMSEA = .05	
Favieri, 2013	Metacognitive Awareness Inventory (MAI) translated to Spanish and reduced to 33 items to form the General Metacognitive Strategy Inventory (GMSI)	• KOC - Declarative - Procedural - Conditional • ROC - Planning	278 university students Mean age = 20	Internal consistency: Cronbach's alphas range from poor to acceptable Structure:	KOC α = .69 ROC α = .76 8 factors = 42% of the variance	CFA was not run

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
	(Schraw & Dennison, 1994)	- Information Management - Monitoring - Debugging - Evaluation		GMSI – 8 factors and second order model with 2 factors and 8 subcomponents	Second order model = 52%	
Harrison & Vallin, 2017	Metacognitive Awareness Inventory (MAI) – 52 item self-report questionnaire (Schraw & Dennison, 1994)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	622 university students 168 men	Structure: Study 1 – examined 4 models - unidimensional, Schraw & Dennison's theorized 2-factor model, Schraw & Dennison's resultant 2-factor model, and an 8-factor model based on Schraw & Dennison's theory No model was a good fit according to statistical standards Study 2 – eliminated items until good fit achieved with 2-factor theory from study 1 which resulted in 19 items	Unidmsl: CFI = .832 TLI = .825 RMSEA= .055 2-factor theory: CFI = .851 TLI = .845 RMSEA= .051 2-factor realized: CFI = .847 TLI = .841 RMSEA= .052 8-factor: no convergence CFI = .959 TLI = .954 RMSEA= .046	2/3 of participants were women
Kim et al, 2017	Junior Metacognitive Awareness Inventory B (JrMAI) – 18 item Self-report questionnaire (Sperling et al, 2002)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	1783 students in 6^{th}-12^{th} grades 893 boys	Structure: 2 factors as predicted	42% of the variance CFI = .91 RMSEA = .05 TLI = .89	There was some reassignment of items and 3 components loaded on both factors CFA showed reassignment to be a better fit
Magno, 2010	Metacognitive Awareness Inventory (MAI) – 52 item	• KOC - Declarative	240 university students	Structure:	2 factors: RMSEA = .13	

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
	self-report questionnaire (Schraw & Dennison, 1994)	- Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	Mean age = 16.45	2-factor model was not a good fit, although some of the results approach a good fit (.91). 8-factor model was also not an overall good fit, but did reach an acceptable fit with RMSEA and the Population Gamma.	McDonald Noncentrality = .89 Population Gamma = .91 Adj Pop Gamma = .81 8 factors: RMSEA = .05 McDonald Noncentrality = .84 Population Gamma = .95 Adj Pop Gamma = .93	Models of metacognition were looked at in relation to how they effected critical thinking Harrison & Vallin, 2017: Magno (2010) "… reported Akaike and Bayesian information criteria were smaller with the two-factor model, which suggests the opposite finding; additionally, because these were structural models with many other variables, rather than measurement (CFA) models, the evidence provides little information for other researchers and practitioners."
Ning, 2016	Junior Metacognitive Awareness Inventory B (JrMAI) – 18 item Self-report questionnaire (Sperling et al, 2002)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	873 secondary students Mean age = 15.36 (.32) 432 boys	Structure: A model of metacognition was looked for by how it fit with its respondents. The best fit showed 2 latent classes and 2 factors aligning with KOC and ROC. Other: There was a significant difference between the 2 classes of students – one had higher scores of metacognition, while the other had lower scores. Further investigation found that the group of students with lower scores fit a unidimensional model, while the group of students with higher scores fit a two-factor model.	Fit Indices AIC = 34,479 Adjusted BIC = 34,749 Entropy = .803 Differences between participant groups $t = 6.12$ $p < .001$ $d = 0.42$	
Ning, 2017	Junior Metacognitive Awareness Inventory A (JrMAI) – 12 item Self-report questionnaire (Sperling et al, 2002)	• KOC - Declarative - Procedural - Conditional • ROC	892 primary students Mean age = 11.35 (.29) 448 boys	Internal consistency: Composite reliability estimates good (KOC) and poor (ROC) Structure:	KOC ρ = .918 ROC ρ = .214 Unidmsl: CFI = .900 RMSEA= .062	

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
		- Planning - Information Management - Monitoring - Debugging - Evaluation		Looked at 4 models – unidimensional, 2 factors, second order factor, and bifactor The bifactor model showed the best fit across all statistics	SRMR = .042 AIC = 27,507 Adj. BIC = 27,565 Bifactor: CFI = .966 RMSEA= .035 SRMR = .028 AIC = 27,401 Adj. BIC =27,479 2nd Order: CFI = .914 RMSEA= .059 SRMR = .040 AIC = 27,484 Adj. BIC = 27,545 2-Factor: CFI = .914 RMSEA= .058 SRMR = .040 AIC = 27,482 Adj. BIC = 27,542	
Pour & Ghanizadeh, 2017	Metacognitive Awareness Inventory translated to Persian (MAI) – 52 item self-report questionnaire (Schraw & Dennison, 1994)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging	107 adults aged 18-43 35 males	Internal Consistency: Cronbach's alphas ranged from acceptable to good Structure: 8 factors as defined by Schraw & Dennison	Factor alphas range from .72 to .81 CFI = .91 GFI = .89 NFI = .90 RMSEA = .061	CFA on 8 factors, no other factor structure was assessed 67% female

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
		- Evaluation				
Schraw & Dennison, 1994	Metacognitive Awareness Inventory (MAI) – 52 item self-report questionnaire (created by authors)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	Study 1: 197 university students 85 males Study 2: 110 university students 41 males	Internal Consistency: Conbach's alphas excellent for measure and good for factors Structure: Study 1 – loaded on 6 factors Forced 2 factor loading saw items load properly on both KOC and ROC Study 2 – confirmed 2 factors	Study 1: α = .95 Study 2: KOC α = .88 ROC α = .88 All α = .93 6-factor: 78% of the variance 2-factor: 65% of the variance 2-factor: 58% of the variance	CFA was not run 59% female In both studies, there were items that failed to load on either factor – 3 for the first and 2 of the original 3 for the second. These items were not discarded.
Sperling et al, 2002	Junior Metacognitive Awareness Inventory versions A and B (JrMAI) – 12 and 18 item Self-report questionnaires (created by authors)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	Study 1: 344 3rd-9th grade students Study 2: 416 3rd-8th grade students	Structure: A – EFA found 5 factors 2 factors were forced and all items but 2 loaded on the 2 factors in study 1. All items loaded in study 2. B – EFA found 5 factors When limited to 2 factors the items didn't load as expected, and 6 of the 9 KOC items also loaded on ROC	Study 1: 5-factors – 60.4% of the variance 2-factors – 31% Study 2: 5-factors – 61.8% of the variance 2-factors – 46% Study 1: 5-factors – 55% of the variance 2-factors – not reported Study 2:	CFA was run Measurements were correlated with Problem solving skills and a reading inventory. Version B did not significantly correlate with either.

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
					5-factors – 52% of the variance 2-factors – 36%	
Teo & Lee, 2012	Metacognitive Awareness Inventory translated to Chinese (MAI) – 52 item self-report questionnaire (Schraw & Dennison, 1994)	• KOC - Declarative - Procedural - Conditional • ROC - Planning - Information Management - Monitoring - Debugging - Evaluation	245 university students majoring in education Mean age = 22.69 (4.3) 72 males	Structure: Initial EFA found 12 factors, scree plot suggested 4 factors With the 4-factor model, 7 items did not meet the loading cut off Measure reduced to 3-factors and 45 items CFA on 3-factor model was not a good fit Items reduced to 21 and CFA run as 3-factor, 1-factor, and second order None of the models meet the cut off criteria for good fit	12-factor: 67% of the variance 4-factor: 43.6% 3-factor: 44.63% 3-factor (52): TLI = .756 CFI = .768 RMSEA= .076 SRMR = .068 3-factor (21): TLI = .903 CFI = .914 RMSEA= .063 SRMR = .048 1-factor: TLI = .821 CFI = .839 RMSEA= .86 SRMR = .064 2nd Order: TLI = .903 CFI = .914 RMSEA= .063 SRMR = .048	Did not compare 3-factor model to either 2-factor or 8-factor models despite that being the aim of the study. Gave no theoretical explanation for choice of 3 factors
Allen & Armour–Thomas, 1993	Metacognition in Multiple Contexts Inventory (MMCI) – Problem solving inventory of	• Define problem • Select options	126 9th-11th grade students	Internal Consistency: Cronbach's alpha questionable Structure:	α = .67 62% of the variance	CFA was not run 126 students, 2/3 girls

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
	24 items (created by authors)	• Select strategy • Select representation • Allocate resources • Solution monitoring	51 boys	9 factors – confirmed Sternberg's idea of metacomponents		Hypothesized components interdependent and loaded on several factors
Altindag & Senemoglu, 2013	Metacognitive Skills Scale (MSS) – 30 item self-report questionnaire (created by authors)	• Use of learning strategies • Knowledge of own learning • Planning & Monitoring	239 university students	Internal Consistency: Cronbach's alpha excellent Structure: EFA run, 25 of the original 55 items eliminated based on poor factor loads, then EFA run again – one factor found	$\alpha = .94$ 35.74% of the variance	CFA was not run
Cetinkaya & Erktin, 2002	Metacognitive Inventory in Turkish (created by authors)	• Evaluation • Self-checking • Awareness • Cognitive strategies	111 6th grade students Mean age = 12 60 boys	Internal Consistency: Cronbach alpha good Structure: EFA performed and found 4 factors CFA performed	$\alpha = .87$ No indices were reported for factor analysis	Participants were gifted and 56% were male All factors would align with regulation Measure correlated with achievement – there were no significant results
Immekus & Imbrie, 2008	State measure of metacognition – 20 item self-report questionnaire (O'Neil & Abedi, 1996)	• Awareness • Cognitive strategy • Planning • Self-checking	3023 university students Mean age = 18.56 (.61) 2437 males	Structure: Tested both bifactor and unidimensional models. Models had similar fit scores. The bifactor had better chi square scores. However, items failed to significantly load under the bifactor model. Thus, the unidimensional model was a better fit.	Bifactor $\chi^2(1,389) = 25{,}520.58$ $p < .001$ Undmsl $\chi^2(1,409) = 26{,}396.72$ $p < .001$	81% male participants No participant in cohort 1 and only 6 from cohort 2 chose "strongly disagree" to 4 of the items, so "strongly disagree" and "disagree" were collapsed Very large samples (1000+) making chi square not the best measure
Meijer et al, 2013	Awareness of Independent Learning Inventory (AILI) – 63 item self-report	• Metacognitive Knowledge (MK)	1058 university students	Internal Consistency: Cronbach's alphas acceptable (MK/MRs) and good (MR)	MK $\alpha = .79$ MR $\alpha = .84$ MRs $\alpha = .77$	CFA was not run

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
	questionnaire there is also a 45 item version (created by authors)	o People o Strategies o Tasks • Metacognitive Regulation (MR) o Orientation o Monitoring o Evaluation • Metacognitive Responsiveness (MRs) o Sensitivity to experiences o Sensitivity to external feedback o Curiosity	(education majors)	Structure: Generalisability score obtained indicating findings can be generalised to a broader range of metacognitive components Validity: The AILI correlated significantly with the metacognitive section of the MSLQ	G = .79 MK r = .69 MR r = .73 MRs r = .67	AILI correlated significantly with 5 of the 6 scales of the MSLQ – all except test anxiety. 3 of the scales that correlated with the AILI were motivational scales and 2 were metacognitive scales.
O'Neil & Abedi, 1996	State measure of metacognition – 20 item self-report questionnaire (created by authors)	• Awareness • Cognitive strategy • Planning • Self-checking	219 university students 210 high school students	Internal Consistency: Cronbach's alphas acceptable Structure: 5 items per factor all loading on only one factor	Alphas ranged from .73 to .78 for each factor % of variance not reported Final version for 12th grade students only	CFA was not run Researchers ran several EFA and adjusted the measure with each one until the final form was reached. In all studies, participants were paid per item. Final form factors all align with regulation.
Pedone et al, 2017	Metacognition Self-Assessment Scale (MSAS) – 18 item self-report questionnaire (created by authors)	• Self o Monitoring o Integrating • Other o Differentiating o Decentring	6659 adults Mean age = 38.61 (13.97) 3049 males	Internal Consistency: Cronbach's alphas ranged from acceptable to good Structure: Four factors were found, but items didn't load as expected. New model: • Self-reflectivity	α's ranged from .72 to .87 57% of the variance NFI = 0.92 NNFI = 0.92 CFI = 0.92	Huge participant recruitment of over 6000 people Model focuses on theory of mind and knowledge of persons. Examples of items appear to fall predominantly under the knowledge factor.

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
				- Monitoring - Integrating of self • Critical Distance - Differentiating - Decentring • Mastery • Understanding other's mind monitoring CFA of new model was not a good fit	RMSEA= .065	
Porumb & Manasia, 2015	COMEGAM-ro – 36 item self-report questionnaire translated from French to Romanian (Richer et al, 2004)	• Metacognitive Knowledge (MK) - Persons - Strategies - Tasks • Metacognitive Management (MM) - Planning - Monitoring and Control - Regulation	575 secondary students aged 14-18 247 boys	Internal Consistency: Cronbach's alphas acceptable for scales and excellent for full measure. Structure: All 6 factors loaded as predicted and all indices show a good fit	Factor alphas range from .71 to .75 COMEGAM-ro α = .90 GFI = .963 SRMR = .05 IFI = .97	57% of participants female. Only published data on this measure. Measure published in French, but no data is reported.
Semerari et al, 2012	Metacognition Assessment Interview (MAI) – a way of interviewing with questions that cover 16 facets of metacognition (created by authors)	• Self - Monitoring - Integrating • Other - Differentiating - Decentring	175 adults Mean age = 30.69 (13.51) 60 males	Internal Consistency: Cronbach's alpha excellent overall and good for scales Structure: Initial EFA found 3 factors, but third was statistically weak Forced 2-factor did not load as expected. New solution named 1. Other oriented 2. Self-oriented	Self α = .90 Other α = .85 All α = .91 2-factor: 54% of the variance 1-factor: GFI = .70 CFI = .78 NNFI = .74 RMSEA = .16	66% women Model focuses on theory of mind and knowledge of persons. Examples of questions appear to fall entirely under the knowledge factor.

Table 1 (continued)

Article	Measure	Hypothesized Model	Participants	Results	Statistics	Notes
				CFA was done with 1-factor, 2-factor, and 2nd order models. The second order model proved to be the best fit: • Metacognition - Other oriented - Self-oriented	2-factor: GFI = .87 CFI = .92 NNFI = .91 RMSEA = .07 2nd Order: GFI = .91 CFI = .97 NNFI = .96 RMSEA = .05	
Yildiz et al, 2009	Metacognition Scale (MS) – 40 item self-report questionnaire (created by authors)	• Declarative knowledge • Procedural knowledge • Conditional knowledge • Planning • Self-control • Cognitive strategies •Self-assessment •Self-monitoring	426 students in 6th-8th grade 205 boys	Internal consistency: Cronbach's alpha excellent Structure: Initial EFA loaded on 6 factors, but researchers couldn't name the factors. After items eliminated, MS loaded on 8 factors.	α = .96 71.36% of the variance GFI = .85 NFI = .87 RMSEA = .04 AGFI = .81 RMR = .05	Structure aligns with structures that have 2 factors (knowledge and regulation) and subcomponents. Not all indices confirm a good fit.

Study 1: Systematic review: Can self-report assess distinct components of metacognition?

Introduction

Flavell's original theory and definition

Metacognition is widely used as an "umbrella term" to refer to a range of different cognitive processes, all of which crucially involve forming a representation about one's own mental states and/or cognitive processes. Whilst Flavell (1979) originally proposed a taxonomy of metacognition (Fig. 1), a range of other taxonomies are used within the field (e.g Brown 1978; Pedone et al. 2017; Schraw and Dennison 1994). As such, this has resulted in a wide variety of self-report questionnaires being used within the field, many of which are based on different taxonomies of metacognition. Flavell's 1979 (Fig. 1) original theory divides metacognition into four areas: metacognitive knowledge, metacognitive experiences, goals, and actions. Metacognitive knowledge refers to the knowledge one has gained regarding cognitive processes, both in oneself and in others. Metacognitive experiences describes the actual usage of strategies to monitor, control, and evaluate cognitive processes. For example, knowing study strategies would be metacognitive knowledge, using a strategy while studying would exemplify a metacognitive experience. Flavell (1979) also subdivides metacognitive knowledge into three areas of knowledge – person, task, and strategy. Knowledge of person is the understanding of one's own learning style and methods of processing information, as well as a general understanding of humans' cognitive processes. The understanding of a task as well as its requirements and demands is designated as knowledge of task. Lastly, knowledge of strategy includes the understanding of strategies and the manner in which each strategy can be employed (Livingston 1997). The remaining two factors of Flavell's description of metacognition are goals – one's intentions when completing a cognitive task, and actions – the behaviors or cognitive functions engaged in fulfilling a goal. Because actions are generally cognitive tasks, it is an area rarely addressed in more recent metacognitive theories as it blurs the necessary divide between cognitive and metacognitive activities.

Modifications to Flavell's taxonomy

From Flavell's pioneering work, many other theories of metacognition have been posited. Brown (1978) divided metacognition into knowledge of cognition (KOC) and regulation of

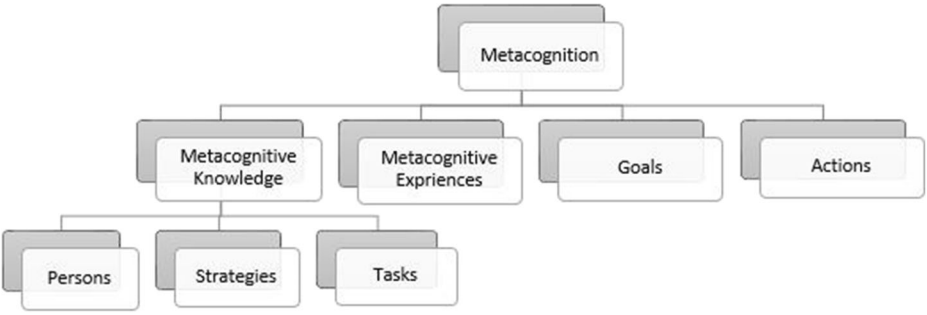

Fig. 1 Flavell's (1979) proposed taxonomy of metacognition

cognition (ROC) and referred to subcomponents of regulation such as planning, monitoring, and evaluating, or reflecting. Much like Flavell's theory, Brown's (1978) two factors comprise an understanding of one's ability to learn and remember (KOC) and one's ability to regulate their learning and memory (ROC). Paris and colleagues (1984) took Brown's model and divided knowledge of cognition into declarative, procedural, and conditional knowledge. Again, similar to Flavell, these subcomponents refer to one's knowledge of their own processing abilities (declarative), ability to solve problems (procedural), and knowledge of when and how to use specific strategies (conditional). Schraw and Dennison (1994; Fig. 2) further defined metacognition by adding information management and debugging to join planning, monitoring and evaluation as subcomponents of regulation of cognition.

Additional taxonomies

In contrast, some researchers look at metacognition as self versus other skills (Pedone et al. 2017; Semerari et al. 2012). In other words, they separate metacognitive awareness and understanding of one's own thoughts and actions from the awareness and understanding of other's thoughts and actions. Thus, subcomponents of self include monitoring and integrating, and subcomponents of others are defined as differentiating and decentring. Some researchers posit a third factor of metacognitive beliefs or attributions (Desoete et al. 2001) in addition to KOC and ROC. This factor encompasses individuals' attribution of their failures and successes, for example citing poor instructions as a reason for failure. However, there is a debate regarding whether attribution can be considered a true metacognitive process, and some researchers define it as an aspect of motivation, and not metacognition. Still other taxonomies build on those mentioned above by making slightly different distinctions, identifying more subcomponents, eliminating some subcomponents, and/or modifying the factors (see Pena-Ayala and Cardenas 2015 for a full comparison of all models of metacognition). Clearly there is a lack of consensus regarding a theoretical organization of metacognition, and available self-report questionnaires reflect this lack of consensus. A review of statistical representations of the structure of metacognitive self-reports may bring some clarity to this theoretical debate.

Methods

Searches and reviews were conducted in June and July of 2018 using EbscoHost, ERIC, PsycArticles, PsycINFO, Scopus, Web of Science, WorldWideScience.org, and bibliography reviews. The PRISMA chart in Fig. 3 details the searches as well as the inclusion and exclusion of papers. An initial search of all years of publication for the terms model, factor analysis and the various forms of metacognition (metacognition, metacognitive, meta-

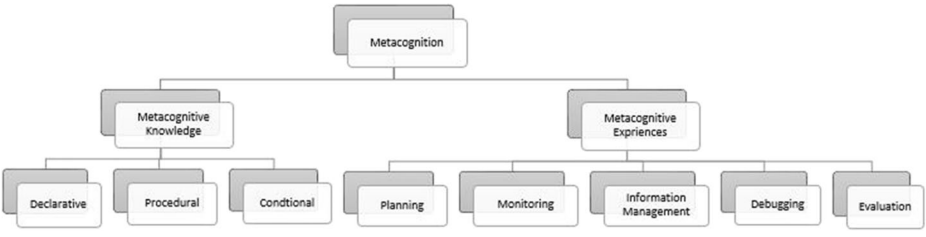

Fig. 2 Schraw and Dennison's (1994) proposed structure of metacognition

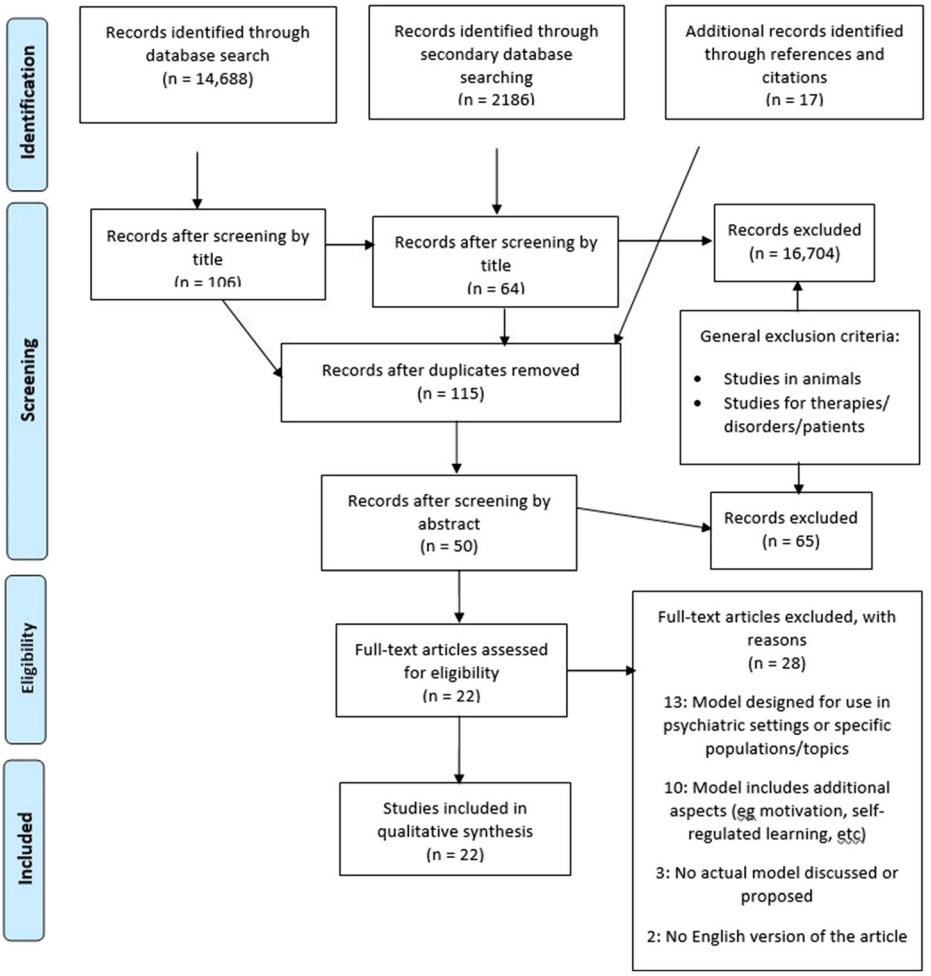

Fig. 3 PRISMA Flow chart of article searches from June and July 2018

cognition) was conducted. To evaluate a generalizable structure of metacognition, participants must represent the general population. Therefore, articles were included only if:

- they were from a peer reviewed journal or a chapter in a published book of articles
- they statistically evaluated metacognition in the general population
- the questionnaire used was widely applicable and not for a specific subset (thus research conducted in a mathematics class was included if the measures of metacognition were widely applicable and not specific to numeracy)

Articles were excluded if:

- participants had a condition or disability (e.g. schizophrenia, Parkinson's disease, learning disability)

- the questionnaire used was built for a specific subset of the population (e.g. patients, firefighters, chemistry students)
- the questionnaire used went beyond the scope of metacognition (e.g. included motivation or memory as part of the scales)
- and if the article could not be obtained in English.

If an article was in another language or could not be located, the authors of the research were contacted and a copy of the article in English was requested. Thanks to response from authors, only two articles were eliminated due to language barriers.

Thus, after a title search, 170 articles were further reviewed. Fifty-five articles were excluded as duplicates, and another 65 based on analysis of the article abstracts using the inclusion and exclusion criteria. Fifty full articles were read and 28 more excluded (see Fig. 3 for an itemized exclusion record with justification). A table was created to encapsulate the following data from each article; authors and year, evaluated structure as measured by questionnaire scales or confirmatory factor analysis, measures employed, narrative results, statistical analysis and any items of note (See Table 1). Thus, each of the 22 articles were reviewed for statistical analysis of internal consistency, validity, and fit indices. Measures were reviewed to ensure they were evaluating only metacognition. Finally, participants were reviewed to ensure compliance with inclusion and exclusion criteria and to note possible drawbacks with participant pools.

Results

Two-factor structure

In total, 22 articles spanning 25 years (1993–2018) of research were included (Table 1). All 22 articles evaluated the structure of metacognition using a self-report questionnaire, self-report through an interview, or task that included self-report questions. Twelve of the articles employed either confirmatory factor analysis (CFA) or exploratory factor analysis (EFA) on the same measure; the Metacognitive Awareness Inventory (MAI; Schraw and Dennison 1994). The remaining ten examined the factor structure proposed by the Metacognition in Multiple Contexts Inventory (MMCI), Metacognitive Skills Scale (MSS), Awareness of Independent Learning Inventory (AILI), the state form of a measure of metacognition as state and trait, Metacognition Self-Assessment Scale (MSAS), COMEGAM-ro, Metacognition Assessment Interview (MAI), Metacognition Scale (MS), and the Turkish Metacognitive Inventory. Of the 22 studies, 10 confirmed, either through factor analysis or theoretical reasoning, the existence of two overarching factors – a measure of metacognitive knowledge (Knowledge of Cognition or Metacognitive Knowledge; henceforth KOC) and a measure of metacognitive regulation (Regulation of Cognition or Metacognitive Experiences; henceforth ROC; see Table 1 and Figs. 1 and 2). The MS questionnaire (Yildiz et al. 2009) first loaded on 6 factors, but researchers failed to adequately name the factors based on item loadings. Therefore, the items were adjusted and finally loaded on the 8 sub factors defined by Schraw and Dennison (1994)), Fig. 2). The Turkish and Persian versions of the MAI (Akin et al. 2007; Pour and Ghanizadeh 2017) loaded onto the Schraw and Dennison 8 subcomponents. Schraw and Dennison's taxonomy defines metacognition as a two-factor structure of KOC and ROC with 8 subcomponents. Furthermore, Schraw and Dennison's MAI loads consistently on KOC

and ROC as factors. Thus, it is likely that all three of these studies would also load on KOC and ROC. In total, then 13 studies confirmed a 2-factor structure of metacognition separating knowledge from regulation.

Three-factor structure

In contrast, the AILI (Meijer et al. 2013) measure found three factors that were widely applicable using the generalizability coefficient *G* and validating it by correlating it to the Motivated Strategies for Learning Questionnaire (MSLQ). No factor analysis was run. The three factors – defined as knowledge, regulation, and metacognitive responsiveness – significantly correlated (all $rs > .34$) with all the subscales of the MSLQ except Test Anxiety. It should be noted that the MSLQ measures motivation as well as metacognition. In fact, the subscales of the AILI significantly correlated with the value scale ($rs > .57$), a motivational scale of the MSLQ. Additionally, the AILI included statements like "I think it's important that there are also personal aims linked to assignments". Therefore, motivation may help explain the third factor. Teo and Lee (2012) also confirmed a three-factor solution using a Chinese version of the MAI. However, as Harrison and Vallin (2018) aptly point out, no theoretical explanation for three factors was provided, and they utilized only 21 of the original 52 items. Additionally, there was no comparison of their structure with Schraw and Dennison's (1994) two factor findings for the MAI. Teo and Lee did report some fit indices on a two-factor structure (see Table 1), which ranged from statistically acceptable to scores just below the cutoff for acceptability. Thus, Teo and Lee's research can also be interpreted as lending some support for the two-factor structure.

Other structures

The MMCI (Allen & Amour-Thomas 1993) loaded on 6 factors, and both the state metacognitive measure (O'Neil and Abedi 1996) and the MI (Çetinkaya & Erktin, 2002) loaded on 4 factors (see Table 1). In all three cases, all of the resultant factors would align with only one of the overarching factors, suggesting the factors are all subcomponents of ROC. Similarly, the MSAS (Pedrone et al., 2017) and MAI (Semerari et al. 2012) loaded on 4 and 2 factors respectively. Again, all of the resultant factors would align with only one of the overarching factors defined in the two-factor structure, but in this case, it is KOC. Thus, these 5 studies also support the existence of a two-factor structure that distinguishes between knowledge and regulation, suggesting that the MMCI is best considered a self-report measure of metacognitive regulation, whilst the MSAS and MAI can be best considered self-report measures of metacognitive knowledge. None of the three self-reports provide suitable measures of knowledge *and* regulation.

Unidimensional

There were two studies that did not support the two factors of knowledge and regulation, but instead found a unidimensional structure (Altindağ & Senemoğlu, 2013; Immekus and Imbrie 2008). However, the single factor was reported after large adjustments to the original measures, that included eliminating almost half of the original items in one study and collapsing scores on one end of the Likert scale in the other study. Additionally, neither study reported fit indices other than chi square. Statistics that were reported were not ideal, for instance a

unidimensional model representing 35.74% of the variance (Altindağ & Senemoğlu, 2013) and a unidimensional model reporting $\chi^2(1409) = 26{,}396.72$, $p < .001$ (Immekus and Imbrie 2008).

Ability based structure

In addition to the 2017 study reported above that suggested a two-factor structure for the JrMAI, Ning (2016) completed a second study with the JrMAI. In this second study Ning chose to look at the structure of metacognition based on respondents. Participants were given the JrMAI and then divided into two groups – those with high scores, and those with low scores. A factor analysis of participants who self-reported weaker metacognitive skills by scoring lower on the questionnaire revealed a unidimensional structure of metacognition. Analysis of those with higher metacognitive scores found a two-factor structure that aligned with Schraw and Dennison's (1994) KOC and ROC. Ning's research suggests that level of metacognitive abilities may play a role in the factor structure of metacognition, lending credibility to both a two-factor and unidimensional structure of metacognition. As the JrMAI is for adolescents, Ning's research may also suggest that age could have an effect on factor structure as younger individuals have less sophisticated metacognitive skills (Dermitzaki 2005), however there is no discernable pattern of factor results based on age among the studies in this review. No other study attempted to divide participants by self-reported metacognitive abilities.

Subcomponent analysis

In sharp contrast to the strong support of a two-factor structure, the subcomponents of the factors are much more debatable. Component analysis varied widely both across the measures as well as on repeated assessments of the same measure. Structures with two, three, four, five, six, eight, and nine components were found (see Table 1). Just in the MAI, four, five, six, and eight subcomponents were found. Like the factor analysis, the number of components varied widely across ages and showed no discernable pattern of age influencing the number of subcomponents found.

Discussion

The papers systematically reviewed, despite the variance in results, lend strong support for the ability of various self-report measures to evaluate a two-factor structure. However, due to the wide range of results, no conclusion can be made regarding whether distinct subcomponents of these factors can be accurately assessed using a self-report measure. Of particular note, is that both the JrMAI and the MAI were unable to produce the same factor structure across studies. Ning's structural equation modelling of metacognition according to participant skill level gives a possible explanation for the diverse results. Participants in the studies ranged widely in age from primary school to university. The extent of abilities across this large spread in age coupled with the range of results reported in this paper lends support to Ning's supposition that reduced metacognitive skill operates with a less complex structure of metacognition. More research is required to determine whether varying metacognitive abilities effect the underlying structure of metacognition and are thus responsible for the wide variety of results. Regardless,

when taking all findings into consideration, it can be deduced that when participants self-report on their own metacognitive abilities they provide an overview of their knowledge and their experiences or ability to regulate cognition, but self-reports do not seem to be able to reliably reveal the more complex relationships found in the metacognitive process when evaluating subscales.

Based on fit indices, the most statistically noteworthy self-report analyses include the bifactor structure from the JrMAI (see Fig. 5; Ning 2017) and the two-factor structure with 6 subcomponents from the COMEGAM-ro (see Fig. 6; Porumb and Manasia 2015). Both had multiple indices (see Table 1) that declared the models to be a good fit for the corresponding questionnaire, as well as strong theoretical support. Ning's structure was evaluated on the JrMAI version A, which has had varying results. This study was the first attempt to compare several different theoretical structures alongside a bifactor structure. Results showed a bi-factor structure of general metacognition along with KOC and ROC to be the best fit (Fig. 4). However, upon looking at the reported Akaike and Bayesian analysis, it is questionable whether the bifactor structure is actually a better fit than the two-factor structure. In contrast the COMEGAM-ro model has strong statistical support in all areas (Porumb and Manasia 2015; Table 1). The results for the COMEGAM-ro revealed a two-factor structure of KOC and ROC with 6 subcomponents (Fig. 5). However, Porumb and Manasia's article is the only published analysis of the factor structure of the COMEGAM-ro, thus the structure has not been replicated.

Based on the systematic review, there is not a single self-report that can be recommended as the industry standard (i.e. reliable and replicable). However, results suggest that using self-report, in particular the COMEGAM-ro, are best suited to evaluate two distinctive metacognitive factors. Alternatively, Ning's (2016) novel approach of dividing participants by skill level may be a better method of evaluating self-reported metacognition. As both Ning's and Porumb and Manasia's results are each based on only one study, it is clear that more research is needed to determine the best method for using self-reports. Furthermore, based on the wide variety of subcomponent results, using a self-report to delineate the complexities of each factor may not be feasible. Thus, further research is also needed to

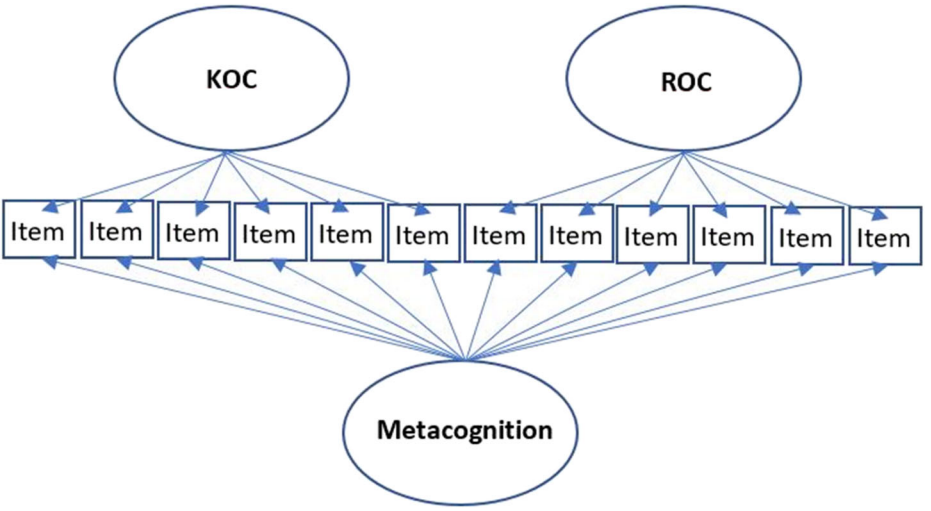

Fig. 4 Ning's (2017) bifactor structure of metacognition

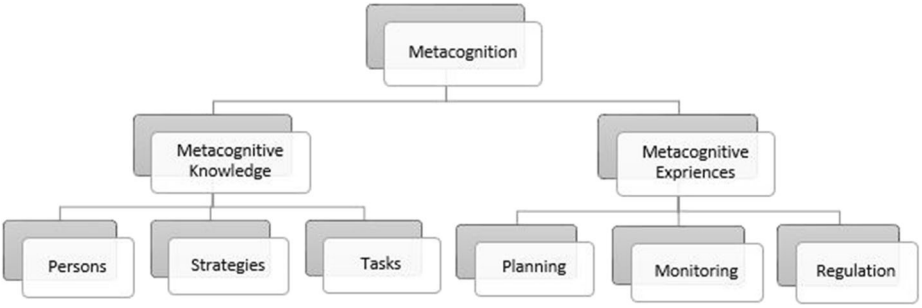

Fig. 5 Porumb and Manasia's (2015) metacognitive structure

explore the efficacy of measuring subcomponents with self-reports. Regardless, results of the review suggest that if a self-report analysis is included as part of a study, it can be used to evaluate general skills of two factors distinguishing knowledge from regulation but cannot adequately measure distinct subcomponents within the two factors.

If, as the systematic review suggests, knowledge and regulation can be adequately measured as distinct factors by self-reports, the subsequent question is whether those factors relate to participant behavior on experimental measures of knowledge and regulation.

Study 2: Systematic review and meta-analysis: Can self-report assess distinct components of metacognition and do those components relate to metacognitive behavior?

Introduction

Study 1 indicated that self-reports mostly measure two main factors of metacognition – knowledge and regulation. To date, the relationship between knowledge and regulation is not clear, in other words, knowledge of metacognitive skills may not relate to metacognitive behaviors. Much of the data seems to indicate that knowledge and regulation do not significantly correlate with each other, particularly when comparing knowledge to experimental measures of regulation (Jacobse and Harskamp 2012; Veenman 2005; Veenman 2013). Van Hout-Wolters & Schellings (2009) report *r*'s ranging from −.07 to .22 for self-report questionnaires and think aloud protocols, a method of measuring metacognition which asks participants to "think aloud" their thought processes as they complete a task. Correlations between retrospective task specific questionnaires and think aloud protocols fare a little better in that the *r*'s range from .10 to .42 (Van Hout-Wolters & Schellings 2009).

In contrast, correlations of subcomponents within each factor reveal larger effect sizes, albeit still with a range of results. Correlations of varying metacognitive behaviors (e.g. planning or monitoring) range from .64 to .98, and correlations of components of metacognitive knowledge (e.g. task or strategy knowledge) range from .02 to .80 (Schellings 2011; Van Hout-Wolters & Schellings 2009). The strength of the top end of these correlations within factors appears to verify the existence of two factors, but the low to moderate strength of the correlations between the factors questions the relationship between knowledge and behavior. The apparent contradictions of the results are often attributed to a variety of

methodological choices, including the type of instrument used, timing of the instruments, participant ages, and analysis that compares full scale scores instead of corresponding subscale scores.

Type of instrument

Because metacognition is not directly observable, measurement tends to involve either a mechanism for self-report or performance on a task (e.g. Akturk & Sahin, 2001; Georghiades 2004; Schraw and Moshman 1995; Veenman et al. 2005; Veenman et al. 2006). The measurements typically employed can be divided into two types – on-line and off-line. On-line measurement occurs during the performance of a task or during learning, for example evaluating one's judgement of learning or having a participant speak their strategies aloud as they complete a task. Off-line measurement occurs either before or after a task or learning has finished, such as interviewing a participant about the strategies they employed on the task they just completed or surveying participants about the general strategies they use to prepare for an exam. Due to its nature, knowledge is most often measured by self-report questionnaires or prospective interviews (off-line). Regulation is often measured with a task (on-line). Because, in general, on-line measures only weakly correlate with off-line measures (Veenman 2005), one interpretation of varied effect sizes is that the type of instrument (questionnaire versus task) may impact the results. Researchers agree that to truly understand the relationships between components of metacognition a multi-method approach using both on-line and off-line tasks is required (e.g. Desoete 2008; Schellings et al. 2013; Van Hout-Wolters & Schellings, 2009; Veenman 2005; Veenman et al. 2014). It is important to determine what off-line data (self-report) adds to understanding metacognition and metacognitive behaviors.

Timing

A similar interpretation for the variety of correlational analysis is the choice in timing of the measurement. Metacognition can be measured prior to performing a task (prospectively), during a task (concurrently), or following the completion of a task (retrospectively). It has been hypothesized that assessing metacognitive knowledge prospectively allows for too much bias as participants may be comparing themselves to others, what the teacher or supervisor thinks, or succumbing to social desirability (Schellings et al. 2013; Veenman 2005). A retrospective questionnaire allows participants to rely more heavily on actual behaviors just performed when evaluating the statements. Concurrent measures, like on-line measures, tend to obtain stronger correlations because they are evaluated during a task. However, not all skills are easily measured concurrently. For example, evaluating one's performance, by its nature, must be measured retrospectively. Thus, some researchers suggest employing concurrent and retrospective task specific measures to ensure more reliable measurement (Schellings et al. 2013; Van Hout-Wolters & Schellings, 2009).

Age and full score versus scale scores

The age of the participants and manner of statistical analysis may also impact effect sizes. Dermitzaki (2005) reports, it is likely that students in primary school have not fully developed

their metacognitive skills and may; therefore, not know how to apply their knowledge to a task or be fully aware of their own strategy use. Therefore, the variation in correlation coefficients could be due to lack of experience associated with chronological age. It has also been suggested that when comparing multiple measures of metacognition, they may be evaluating different subcomponents of the factors (e.g. planning and monitoring correlated to evaluation and reflection), resulting in poorer effect sizes. Thus, it has been suggested that correlational analysis be carried out by the corresponding subscales instead of the overall scores (Van Hout-Wolters & Schellings, 2009).

Meta-analysis

That we know of, there has never been a meta-analysis of the various relationships between and within factors of metacognition as assessed by self-reports and experimental procedures. Thus, based on the results of Study 1, this systematic review and meta-analysis will evaluate two factors of metacognition by summarizing the relationships between knowledge and regulation to first, determine the ability of self-report to measure proposed taxonomies and second, determine whether self-report relates to metacognitive behavior. Subcomponent correlations will be evaluated not only to determine relationships between self-report and behavior, but also to look again at whether self-report can capture more than a general overview of two factors. Furthermore, due to the current wide range of results, it is likely that meta-analysis results will be high in heterogeneity. Heterogeneity indicates that the pooled effect size estimate cannot be interpreted because another factor is moderating the results. Therefore, this analysis will also examine possible effects of moderators. When elevated heterogeneity is found, timing and type of instruments as well as age will be evaluated for their impact.

Methods

Searches and reviews were conducted in July and August of 2018 using EbscoHost, ERIC, PsycArticles, PsycINFO, Scopus, Web of Science, WorldWideScience.org, and bibliography reviews. The PRISMA chart in Fig. 6 details the searches and inclusion and exclusion criteria.

The aim of Study 2 is to determine the relationship between the varying components of metacognition, and whether measures of metacognitive knowledge relate to measures of metacognitive behavior (regulation). Consequently, several searches of all years of publication were performed. Since on-line tasks generally measure knowledge, and off-line tasks generally measure regulation, a search for these terms as well as the term multi-method was performed. The various forms of metacognition (metacognition, metacognitive, meta-cognition) were paired individually and with combinations of the terms online, on-line, offline, off-line, and multi-method (see the appendix for the specifics of the search).

Articles were included only if they compared at least two measures of pure metacognition. Thus, a comparison of the total scores of the Motivated Strategies for Learning Questionnaire (MSLQ) and a think aloud protocol would be excluded due to the generally accepted assumption that total scores on the MSLQ measure both participants' metacognitive abilities *and* motivation profile. However, a comparison of the metacognitive subscale of the MSLQ and a think aloud protocol would be included. Unlike the first search looking for an overall structure of metacognition, one of the aims of this search was to understand the extent to which self-report scales correlate to behavioral measures of metacognition. Thus, task specific

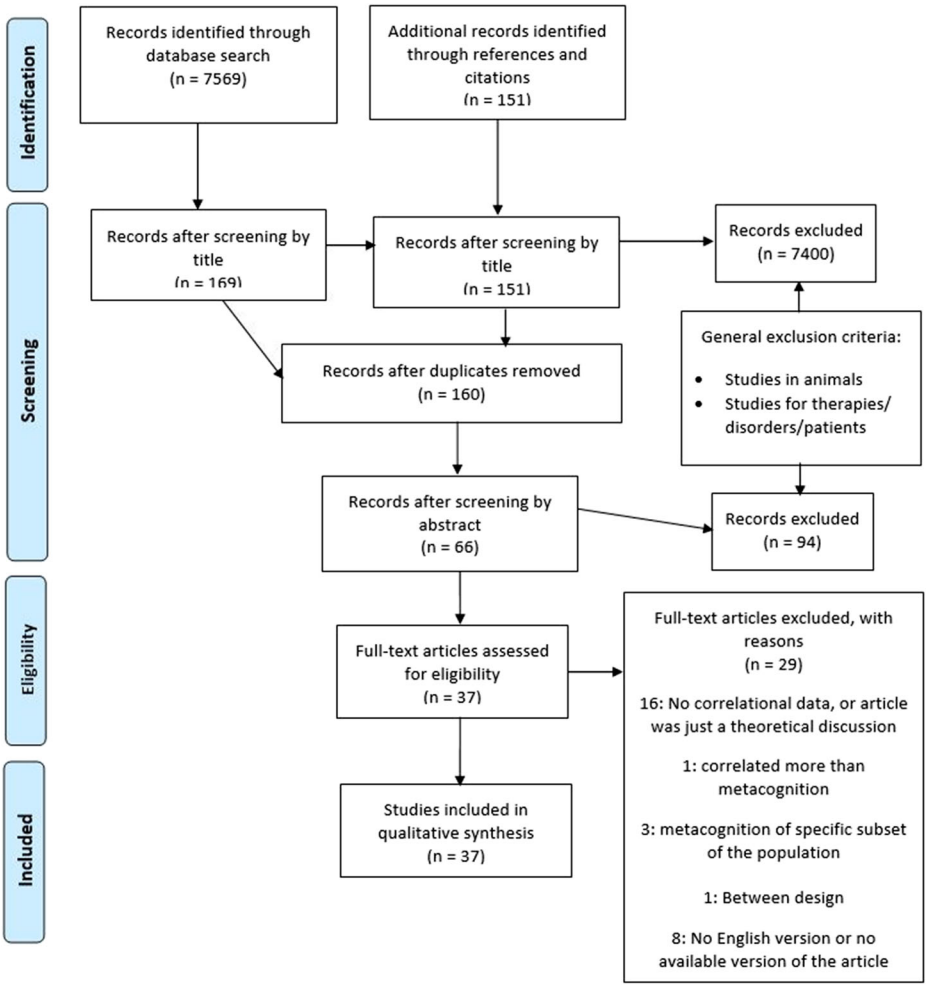

Fig. 6 PRISMA flow chart of article searches from July and August 2018

correlations were not excluded. Additionally, one task could be a measure of two components, provided the scales were listed separately and statistically compared. Therefore, articles were included if:

- they statistically compared components of metacognition using a within design method
- correlational effect sizes (e.g. Pearson's r, Kendall's tau) were provided
- the measures of metacognition employed did not include other skills (e.g. motivation)

Articles were excluded if:

- participants had a condition or disability (e.g. schizophrenia, Parkinson's disease, learning disability)
- there was no statistical data comparing components of metacognition (e.g. means and standard deviations listed, but no actual correlations run)

- the correlational data was between participants instead of within (i.e. comparing abilities of distinct groups of participants instead of components of an underlying structure)
- and the article could not be obtained in English.

Like the first systematic review, if an article was in another language or could not be located, the authors of the research were contacted and a copy of the article in English was requested. Thanks to the authors of the requested research, excluded studies based on lack of access were limited to 8 articles.

Ultimately, 320 articles were reviewed following a title search. One hundred sixty were excluded as duplicates. Another 94 articles were excluded after reviewing the article abstracts for relevance. Sixty-six full articles were read and 29 excluded based on the inclusion and exclusion criteria (see Fig. 6 for an itemized exclusion record with justification). A total of 37 articles spanning 33 years of research (1982–2015) were analyzed. A table was created summarizing authors and year, measures employed, components evaluated, age of participants, narrative results, statistical analysis and any items of note (see Table 2). In addition to this information, the type (on-line, off-line) and timing (prospective, concurrent, retrospective) of each instrument were noted. Thus, each of the 37 articles were reviewed for statistical relationships, and to ensure participant pools and metacognitive measures complied with inclusion and exclusion criteria. Any possible drawbacks to the study were also noted.

Statistical analysis

As recommended by researchers, most of the 37 articles used a multi-method approach to examine relationships or analyzed results by correlating corresponding subscales of measures. Thus, one article could feasibly contribute several pieces of data to the meta-analysis. In total, the 37 articles reported 328 correlations between factors and/or subcomponents of metacognition. Because only one statistic per population could be included in the meta-analysis, specific criteria for choosing the statistic was necessary. Correlations were chosen using the following hierarchy:

- from online measures – online measures such as think aloud protocols are less subject to bias and misinterpretation than offline measures (Schellings et al. 2013),
- correlations between two different measures as opposed to within one measure (e.g. correlations between subscales of a questionnaire) provide a more robust picture of relationships between metacognitive skills,
- from measures that, based on the systematic review, found a model closest to Porumb and Manasia's (2015) model (see Fig. 6 above) thus lessening possible interference of other factors, such as motivation,
- the better Cronbach's alpha scores for a more reliable measure,
- the median piece of data – if an even number of statistics was reported, then the range of each half of the data was calculated and the statistic chosen according to the larger range (e.g. correlation set {.27, .27, .28, .38} .28 was chosen; {.40, .45, .55, .63, .68, .72} .55 was selected).

All correlations were reported with either Pearson's r or Kendall's tau. Pearson's r and Kendall's tau cannot be directly compared. Thus, all Kendall's tau statistics were first converted to r using Kendall's formula $\sin(0.5*\pi*\tau)$ (Walker 2003). Data was then read into

Table 2 Studies Evaluating Relationships between Factors and Subcomponents of Metacognition

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
*Bannert & Mengelkamp, 2008	Off-line (LIST) and On-line (TAP) Prospective: LIST Concurrent: TAP Retrospective: LIST	Metacognitive knowledge (declarative, procedural, and conditional): LIST Metacognitive regulation (Orientation, Monitoring, regulation, organization, elaboration): TAP	70 University students, 84% of which were female. Mean age = 24.2	Think aloud (TAP; n=24) Prompted reflection throughout a task (n=24) Control group that just completed the task (n=22) Learning through reading, questionnaire (LIST) – given 1 week before and right after the task (modified by eliminating items inappropriate to the hypothetical learning situation)	Correlated retrospective LIST with performance on the tasks. No scale of the questionnaire correlated with performance on the task for any groups, except the think aloud group and elaboration scale. However, elaboration is a cognitive scale. No results reported for insignificant correlations and metacognitive scales. Prospective LIST not correlated.	TAP/LIST elaboration r = .54 TAP/ Metacognitive Scales r = no significant correlations
Bong, 1997	Off-line (MSLQ, Judgment) Prospective: Judgement Concurrent: NA Retrospective: MSLQ	Metacognitive Knowledge (procedural, declarative, conditional): MSLQ, judgments	588 high school students from 4 high schools in Los Angeles	Self-efficacy scale of the Motivated Strategies for Learning Questionnaire (MSLQ) Judgments on ability to solve actual problems (problems never completed)	Judgments of problem difficulty and general judgments of academic ability by class (MSLQ) significantly correlated in every subject.	English r = .45 Spanish r = .72 History r = .40 Algebra r = .63 Geometry r = .68 Chemistry r = .55
Chen, 2003	On-line (all judgments) Prospective: NA Concurrent: all judgments Retrospective: NA	Metacognitive regulation (planning, evaluation, reflection): all judgments	107 7th grade students in parochial school. 42 boys and 65 girls. 98% Caucasian	Pre and post judgements of ability/performance – measure of confidence to solve each problem (PJ), confidence in solution	All the judgment measures significantly correlated with each other.	PJ/CJ r = .77 PJ/EJ r = .49 CJ/EJ r = .47

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
				(CJ), evaluation of effort (EJ)		
Cooper et al, 2008	Off-line (MCA-I) and On-line (IMMEX) Prospective: MCA-I Concurrent: IMMEX Retrospective: NA	Metacognitive regulation (planning, monitoring, control, evaluating): MCA-I and IMMEX	209 Chemistry I students at a southeastern US research university	Metacognitive Activities Inventory (MCA-I) – 27 item self-report questionnaire Interactive Multi-media Exercises (IMMEX) – determines strategy use as participants solve online problems, classifies use from low to high metacognition	Correlated results of metacognition from self-report questionnaire and computer logs of metacognitive behavior combined with accuracy. The results correlated significantly, but with small effect size.	r = .20
*Cromley & Azevedo, 2006	Off-line (MARSI) and On-line (Think Aloud) Prospective: MARSI Concurrent: Think Aloud Retrospective: NA	Metacognitive knowledge (strategy use): MARSI Metacognitive regulation (planning & monitoring): Think Aloud	30 9th grade students in social studies classes, 17 girls & 13 boys, Mean age = 14.03, diverse ethnically	Metacognitive Awareness of Reading Strategies Inventory (MARSI) – a self-report questionnaire of strategy use Think aloud while reading American History text	The self-report measure did not correlate with any other measures. Most importantly, strategy use on the self-report did not significantly correlate with strategy use during the Think Aloud.	MARSI/Think Aloud r = -.02
Dermitzaki, 2005	On-line (Observation, Reflection) Prospective: NA Concurrent: Observation Retrospective: Reflections	Metacognitive regulation (judgments of confidence, estimate of task difficulty, reflection, planning, monitoring, strategy use): Observation, Reflections	25 2nd grade Greek students 13 boys, 12 girls Mean age = 7.6	Observations of completing a task (constructing a wooden toy). Observations were coded and rated using an instrument created and validated in a previous study by the author. Reflections on	The following aspects of metacognition were measured. There were only 2 significant correlations. Feeling of satisfaction (FS) Estimate of correctness (EC)	FS/EC r = .12 FS/EE r = .18 FS/EM r = .33 FS/P r = .29 FS/M r = .43 FS/E r = -.16 FS/AE r = .34 FS/LE r = .39 EC/EE r = -.18

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
				their performance of confidence, effort, and satisfaction.	Estimate of effort (EE) Effective use of model (EM) Planning (P) Monitoring (M) Evaluating (E) Awareness of errors (AE) Learning from errors (LE) Except for feeling of satisfaction and monitoring and estimate of correctness and learning from errors, no other section of the self-report correlated significantly with observed metacognitive behaviors.	EC/EM r = .19 EC/P r = .21 EC/M r = .24 EC/E r = .11 EC/AE r = .22 EC/LE r = .44 EE/EM r = .08 EE/P r = .30 EE/M r = .29 EE/E r = .19 EE/AE r = .18 EE/LE r = .18
Desoete, 2007	Off-line (PAC, RAC) and On-line (TAP, EPA2000) Prospective: PAC Concurrent: TAP, EPA2000 Retrospective: RAC	Metacognitive regulation (prediction, planning, monitoring, evaluation): PAC, RAC, TAP, and EPA2000	33 students tested in 3rd grade, and then again in 4th grade	Prospective Assessment of Children (PAC) and Retrospective Assessment of Children (RAC) – both self-report questionnaires of metacognitive regulation. The RAC is task specific as it is given after a task and students are asked to evaluate the recent performance. PAC is given before a task.	Only the PAC and RAC significantly correlated. Between years 3 and 4, a significant correlation was found between test/retest for the EPA2000. Test/retest correlations for Think Aloud, PAC, and RAC were not significant.	Prediction Skills TAP/PAC r = .06 TAP/RAC r = .02 TAP/EPA2000 r = .24 PAC/RAC r = .68 PAC/EPA2000 r = -.24 RAC/EPA2000 r = -.01

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
				EPA2000 – computer program measuring metacognitive regulation Think Aloud (TAP) – while solving 3 word problems		EPA2000 3rd/4th r = .40 Evaluation Skills TAP/PAC r = -.13 TAP/RAC r = -.27 TAP/EPA2000 r = .04 PAC/RAC r = .40 PAC/EPA2000 r = .12 RAC/EPA2000 r = .14 EPA2000 3rd/4th r = .39 Planning Skills TAP/PAC r = -.23 TAP/RAC r = -.25 PAC/RAC r = .57 Monitoring Skills TAP/PAC r = -.03 TAP/RAC r = -.03 PAC/RAC r = 1

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
Desoete, 2008	Off-line (PAC, RAC) and On-line (TAP, EPA2000) Prospective: PAC, EPA2000 Concurrent: TAP Retrospective: RAC, EPA2000	Metacognitive regulation (prediction, planning, monitoring, evaluation): PAC, RAC, TAP, and EPA2000	20 Third graders, 13 girls and 7 boys from one class in Flanders – the sample was ethnically diverse, though all were fluent in Dutch	Prospective Assessment of Children (PAC) and Retrospective Assessment of Children (RAC) – both self-report questionnaires measuring metacognitive regulation. The RAC is task specific as it is given after a task and students are asked to evaluate the recent performance. PAC is given before a task. EPA2000 – measures maths, predicting, and evaluating Think Aloud Protocols (TAP) – on 3 word problem solving tasks	All measures were broken down into subscales and correlated. For the most part, the self-reports did not significantly correlate with the tasks. However, there were two significant subtest correlations – the PAC and evaluation statements during Think Alouds, and PAC and evaluation questions from EPA 2000.	PAC & RAC r's ranged from .44 to .78 PAC & TAP r's ranged from -.10 to .55 PAC & EPA2000 Evaluation r's = -.02 and .42 RAC & TAP r's ranged from -.24 to .08 RAC & EPA2000 Evaluation r's = -.33 and -.24 TAP & EPA2000 r's = .14 and .42 PAC subscales r's ranged from -.29 to .58 RAC subscales r's ranged from .34 to .69 TAP subscales r's ranged from .02 to .84 EPA2000 subscales r = .89
Desoete, 2009	Off-line (CA) and On-line (EPA2000, CDR, TAP)	Metacognitive Regulation (Predicting, Evaluating): CA, CDR, EPA2000, TAP	66 Dutch students who were tested in 3rd and again in 4th grades	EPA2000 – measures maths, predicting, and evaluating	Significant correlations occurred between the CA and the EPA2000 as well as the CA and	CA/CDR r = .25 CDR/TAP r = .25 Prediction

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
	Prospective: CA Concurrent: CDR, EPA2000, TAP Retrospective: NA			Cognitive Developmental aRithmatics (CDR) – measures maths, predicting, and evaluating Think Aloud Protocol (TAP) Child Assessment (CA) – 12 item self-report of metacognitive skills	the Think Aloud. The CDR and the EPA2000, evaluation scales of the Think Aloud and the EPA2000 were also significant. There were no other significant correlations.	TAP/TAP Eval r = .35 TAP/ EPA2000 r = .14 EPA2000/CA r = -.02 Evaluation TAP/ EPA2000 r = .42 TAP/CA r = .55 EPA2000/CDR r = .92 EPA2000/CA r = .42
Desoete et al, 2001	Off-line (MAA) and On-line (MSA) Prospective: MAA, MSA Concurrent: NA Retrospective: NA	Metacognitive knowledge (declarative, procedural, and conditional): MSA Metacognitive regulation (prediction, planning, monitoring, and evaluation): MSA Attributions (internal stable/nonstable and external stable/ nonstable): MAA	80 Third grade Dutch students, 31 boys and 49 girls	Metacognitive Attribution Assessment (MAA) – 13-item self-report questionnaire Metacognitive Skills and Knowledge Assessment (MSA) – 75 items designed to test procedural, declarative, and conditional knowledge, as well as predicting, planning, monitoring and evaluation through a variety of tasks, such as evaluating item difficulty	There were no significant correlations between the online (MSA) and offline (MAA) measures. There were significant correlations among most sections of the online measure: Procedural knowledge (PK) Declarative knowledge (DK) Conditional knowledge (CK) Predicting (P) Planning (Pl) Monitoring (M)	MAA/MSA rs ranged from -.04 to .24 PK/DK r = .39 PK/CK r = .52 DK/CK r = .42 PK/P r = .10 PK/Pl r = .48 PK/M r = .24 PK/E r = .50 DK/P r = .16 DK/Pl r = .32 DK/M r = .34 DK/E r = .43 CK/P r = .18 CK/Pl r = .31 CK/M r = .28 CK/E r = .42

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
					Evaluation (E) The insignificant correlations: procedural knowledge and predicting, procedural knowledge and monitoring, declarative knowledge and predicting, conditional knowledge and predicting, predicting and evaluating, and monitoring and evaluating.	P/Pl r = .29 P/M r = .39 P/E r = .17 Pl/M r = .33 Pl/E r = .39 M/E r = -.04
Hadwin et al, 2001	Off-line (self-report questionnaire) Prospective: questionnaire Concurrent: NA Retrospective: NA	Metacognitive regulation (planning and monitoring): questionnaire	86 (planning) and 92 (monitoring) University students in Canada Mean age 21.9	Author created self-report questionnaire. It was given to rate metacognitive behaviors on learning text, writing a paper, and studying for an exam	Many of the measures did not correlate across contexts. However, monitoring and planning were consistent when reading/exam and writing a paper/exam were correlated. Reading/writing a paper did show some variance and a lower effect size.	Planning Reading/Writing a Paper r = .66 Reading/Studying for and Exam r = .80 Paper/Exam r = .81 Monitoring Reading/Writing a Paper r = .49 Reading/Studying for and Exam r = .56 Paper/Exam r = .67

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
Jacobse & Harskamp, 2012	Off-line (MSLQ) and On-line (VisA, TAP) Prospective: NA Concurrent: VisA, TAP Retrospective: MSLQ	Metacognitive regulation (monitoring, regulation, confidence judgments): VisA, TAP, and MSLQ	39 students from 5 grade 5 classes in the Netherlands. Mean age 10.91 SD = 0.28 24 boys, 18 girls 3 students didn't finish all the assessments	Think Aloud Protocol (TAP) on 2-word problems VisA metacognitive task using word problems Motivated Strategies for Learning Questionnaire (MSLQ) – only the metacognitive self-regulation scale used (12-items). General wording was replaced with wording specific to maths.	The MSLQ did not correlate with any measure. Think Aloud and the VisA significantly correlated. Reported from Veenman & Van Hout-Wolters, 2002 that on-line measures did not significantly correlate with off-line measures	MSLQ/TAP r = 0.16 MSLQ/VisA r = -0.20 TAP/VisA r = .29 Mean correlation did not exceed r = 0.17
Li et al, 2015	Off-line (SRMP) and On-line (Sokoban, TOL) Prospective: NA Concurrent: Sokoban, TOL Retrospective: SRMP	Metacognitive regulation (planning): SRMP, Sokoban, TOL	Beijing – 440 students from 4 grades (81 in 5th, 113 in 7th, 127 in 10th, and 119 in college: M = 11.6, 12.7, 15.9, 20.7 respectively) Boys and girls fairly even except in college (m-21, f-98)	Tower of London (TOL) and Sokoban – measures of metacognitive planning Correlation of time ratio (amount of time per move/total amount of time) Reduced version of the MAI called Self-Report on Metacognitive Planning (SRMP)	Behaviors during tasks did significantly correlate with what participants reported on questionnaire. This was true for both Tower of London and Sokoban. It was also true for the overall measure of metacognitive planning (MP).	TOL/SRMP r = 0.308 Sokoban/SRMP r = 0.180 MP/SRMP r = 0.179 TOL/Sokoban r = .616 MP/TOL r = .562
Merchie & Van Keer, 2014	Off-line (TLSI) and On-line (TAP) Prospective: NA Concurrent: TAP	Metacognitive regulation (planning, monitoring, evaluation): TLSI, TAP	20 5th and 6th grade students, 13 girls and 7 boys, Mean age = 11.64 SD = .62	Think Aloud Protocol (TAP) while studying a 300-word text Text Learning Strategies Inventory (TLSI) – 37 item self-report	Significant correlations were found between the self-report and specific behaviors, such as highlighting. However, correlations between	Planning τ = -.255 Monitoring τ = .238

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
	Retrospective: TLSI			requiring participants to rate their behaviors during a task they had just completed. In this case it was the 300-word text	metacognition during the think aloud were not significant with metacognition reported on the inventory. For these correlations, tau (τ) was used to calculate the relationship.	
*Minnaert & Janssen, 1997	Off-line (LERQ, ILS) Prospective: LERQ, ILS Concurrent: NA Retrospective: NA	Metacognitive regulation (monitoring, regulating): LERQ, ILS	517 freshman college students in a variety of classes	LERQ (Leuven Executive Regulation Questionnaire) – measuring monitoring and regulation ILS (Inventory of Learning Styles) – measures regulation	Both significant and non-significant correlations were found between the corresponding subscales of the questionnaires.	LERQ/ILS rs ranged from 0.13 to 0.80
Muis et al, 2007	Offline (LASSI, MSLQ, MAI) Prospective: LASSI, MSLQ, MAI Concurrent: NA Retrospective: NA	Metacognitive Regulation (organization, elaboration, self-regulation and evaluation): LASSI, MSLQ, MAI	318 students from various undergraduate courses 255 women, 61 men, 2 other Mean age = 23.08	Subscales of 3 self-report questionnaires. Subscales were chosen for having similar metacognitive items across all three scales. Scales included the Learning And Study Strategies Inventory (LASSI), Motivated Strategies for Learning Questionnaire (MSLQ) and Metacognitive Awareness Inventory (MAI)	Correlations across and within scales ranged from small to moderate.	Within MAI r's ranged from .51 to .70 Organization across all r's ranged from .29 to .37 Elaboration across all r's ranged from .54 to .60 Self-regulation across all r's ranged from .27 to .55

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
						Evaluation across all r's ranged from .41 to .50
Ofodu & Adedipe, 2011	Off-line (SAASRC) Prospective: SAASRC Concurrent: NA Retrospective: NA	Metacognitive Knowledge (questions pertaining to strategy awareness): SAASRC Metacognitive Regulation (questions pertaining to application of strategies): SAASRC	120 secondary schools students in Nigeria	Student Awareness and Application of some Strategies to Reading Comprehension (SAASRC) 20 item self-report questionnaire asking students about their knowledge of various strategies (15 items) and the usage of strategies (5 items)	Correlations revealed that students who are aware of metacognitive strategies do apply them	Metacognitive Knowledge/ Application of strategies r = .28
*Peterson et al, 1982	Off-line (SRI, CPQ) Prospective: NA Concurrent: NA Retrospective: CPQ, Interview	Metacognitive Knowledge (strategy use, self-efficacy judgement): CPQ, Interview Metacognitive Regulation (planning, monitoring, evaluation, regulation): CPQ, observation	72 5th and 6th grade students in Wisconsin	Stimulated Recall Interview (SRI) Cognitive Process Questionnaire (CPQ) – 23-item self-report developed by authors to measure attention, monitoring, strategies	The stimulated recall interview and self-report questionnaire (CPS) were significantly correlated across subscales and with the total interview score. The only exceptions were comparing monitoring understanding from the interview and specific strategy use from the self-report and specific strategy use from the interview and	SRI/Monitoring (CPQ) τ = .55 SRI/Strategies (CPQ) τ = .76 Monitoring (CPQ)/ Strategies (CPQ) τ = .35 Monitoring (SRI)/ Monitoring (CPQ) τ = .23 Strategies (SRI)/ Strategies (CPQ) τ = .19

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
					monitoring understanding from the self-report. Tau (τ) was used to calculate the correlations.	Monitoring (SRI)/ Strategies (CPQ) τ = .07 Strategies (SRI)/ Monitoring (CPQ) τ = .11
Porumb & Manasia, 2015	Off-line (COMEGA-M-ro) Prospective: COMEGAM--ro Concurrent: NA Retrospective: NA	Metacognitive knowledge (person, task, strategy): COMEGAM-ro Metacognitive regulation (planning, monitoring, control, evaluation): COMEGAM-ro	575 Romanian students in secondary school	COMEGAM-ro – 36 item self-report questionnaire measuring all aspects of metacognition	The subscales of metacognition all correlated significantly with one another. KP – person, KS – strategy, KT – task, MP – planning, MCM – monitoring & control, MR – evaluation	KP/KS r = .717 KP/KT r = .715 KS/KT r = .534 KP/MP r = .630 KP/MCM r = .585 KP/MR r = .550 KS/MP r = .483 KS/MCM r = .486 KS/MR r = .454 KT/MP r = .524 KT/MCM r = .537 KT/MR r = .458 MP/MCM r = .606 MP/MR r = .536 MCM/MR r = .497
Sarac & Karakelle, 2012	Off-line (Jr MAI) and On-line (TAP, JOL) Prospective: NA Concurrent: TAP, JOL Retrospective: Jr MAI	Metacognitive knowledge (declarative, procedural, conditional): Jr MAI Metacognitive regulation (orienting, planning, evaluating, elaborating): Jr MAI, TAP, JOL	47 students from 6 classes in 3 state schools in Istanbul. 20 girls and 27 boys aged 9-11 Mean age = 10.0	Self-report questionnaire – Jr Metacognitive Awareness Inventory (Jr MAI) Think Aloud Protocol (TAP) on nonfiction text about balloons Judgment of Learning (JOL)	The questionnaire only correlated significantly with the teacher ratings. The think aloud protocols significantly negatively correlated with the confidence judgment. Nothing else correlated significantly.	Jr MAI/TAP r = .12 Jr Mai/JOL r = .07 JOL/TAP r = -.30

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
Schellings, 2011	Off-line (self-report) and On-line (TAP) Prospective: NA Concurrent: TAP Retrospective: questionnaire	Metacognitive regulation (Orientation, planning, monitoring, elaborating, evaluating): TAP, self-report questionnaire	Study 1: 16 9th grade students from 2 different history classes	Think Aloud Protocol (TAP) on a history text and an author created 58-item task-specific self-report questionnaire	The overall questionnaire and task scores did significantly correlate. One set of subscales correlated significantly, the other did not. No other correlation reported. Note: poor internal consistency of the subscales (not used here), 2nd study done with 190 students, but correlations not reported, internal consistency worsened	Overall score r = 0.51 Elaboration & Evaluation r = 0.60 Orientation & Planning r = 0.24
Schellings et al, 2013	Off-line (self-report) and On-line (TAP) Prospective: NA Concurrent: TAP Retrospective: questionnaire	Metacognitive regulation (Orientation, planning, monitoring, elaborating, evaluating): TAP, self-report questionnaire	4 boys and 16 girls – all 15-year-olds from five different schools in the Netherlands	Think Aloud Protocol (TAP) from a history text Questionnaire created to match the skills used for the think aloud task. The taxonomy used to create the questionnaire was also used to score the TAP. The 58-item survey was task-specific	The questionnaire and task overall scores did correlate significantly. Subscales correlations varied in significance. The study does go on further to break down activities into specifics. Note: poor internal consistency of 3 of the subscales – not used here	Overall r = 0.63 Elaboration & Evaluation r = 0.50 Orientation & Planning r = 0.10
Schraw, 1994	Off-line (pre-test judgment of metacognitive knowledge)	Metacognitive knowledge (predicted accuracy): pre-test judgment Metacognitive	115 students – 68 females and 47 males, enrolled in educational psychology course in	Pre-test self-report of monitoring ability Confidence Judgments (CJ) on accuracy of	Off-line ratings of metacognitive ability (pre-test ratings) correlated significantly	Pre-test/CJ r = .45 Pre-test/Overall r = .46 CJ/Overall r = .53

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
	and On-line (judgments of accuracy) Prospective: Pre-test Concurrent: CJ and judgments of accuracy Retrospective: NA	regulation (monitoring, reflection): CJ and accuracy judgments	the midwestern United States	items for each section of items Overall estimate of accuracy upon completion	with on-line ratings of metacognition (judgements of accuracy of items and overall) Met knowledge score comes from offline judgments. Met monitoring score looks at local and global monitoring skills.	Met Knowledge/ Local Monitoring r = -.24 Met Knowledge/ Global Monitoring r = -.29
Schraw, 1998	Off-line (GMSC) and On-line (CJ) Prospective: GMSC Concurrent: CJ Retrospective: NA	Metacognitive knowledge (strategies): GMSC Metacognitive regulation (monitoring, reflection): GMSC, CJ	95 undergraduates, 59 women, 36 men in introduction to ed psych class	General Metacognitive Skills Checklist (GMSC) – self-report measure of monitoring strategies and knowledge Confidence judgments (CJ) for each maths assessment	The GMSC correlated significantly with all confidence judgments. Confidence judgments also all significantly intercorrelated.	GMSC/CJs r's ranged from .27 to .28 CJs r's ranged from .30 to .62
Schraw & Dennison, 1994	Off-line (Pre-judgment, MAI) and On-line (CJ, MAI) Prospective: Pre-judgment, MAI Concurrent: CJ Retrospective: NA	Metacognitive knowledge (declarative, procedural, conditional): Pre-judgment, MAI/KOC Metacognitive regulation (planning, monitoring, control, debugging, evaluation): CJ, MAI/ROC	Study 1 – 197 undergraduates in Nebraska, 85 males and 112 females Study 2 - 110 undergraduates in Nebraska, 69 females and 41 males, For both studies, all students were enrolled in an introductory Ed. Psych class	Metacognitive Awareness Inventory (MAI) – 52-item self-report created by authors measuring Knowledge of Cognition (KOC) and Regulation of Cognition (ROC) Confidence Judgments (CJ)	Statistically significant relationships were found between the two factors (KOC/ROC) of the MAI, KOC and the pre-judgment of monitoring ability, KOC and CJ, ROC and CJ, the pre-judgment of monitoring ability and CJ, and the prejudgment of	KOC/ROC & Pre-judgment r = 0.31 r = 0.12 KOC/ROC & CJ r = 0.23 r = 0.21 KOC/ROC & Monitoring Accuracy both r = 0.09

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
				Pre-judgment of monitoring ability – rating scale	monitoring ability and monitoring accuracy, and CJ and monitoring accuracy. The pre-judgment of monitoring ability and ROC did not significantly correlate. Similarly, neither factor of the MAI significantly correlated with monitoring accuracy. Correlations are only reported from the two subscales of MAI, no overall MAI score is correlated.	Monitoring Accuracy & Pre-judgment r = -0.19 Monitoring Accuracy & CJ r = .32 KOC/ROC Study 1: r = 0.54 Study 2: r = 0.45, 0.49
*Sperling et al, 2002	Off-line (Jr MAI, MSI, IRA, SPSI) Prospective: Jr MAI, MSI, IRA, SPSI Concurrent: NA Retrospective: NA	Metacognitive knowledge (declarative, procedural, conditional): Jr MAI Metacognitive regulation (planning, monitoring, debugging, evaluating, regulation): Jr MAI, MSI, SPSI, IRA	Study 1: 144 students in grades 3-5 and 200 students in grades 6-9 Study 2: 135 students in grades 3-5 and 264 students in grades 6-8 No ethnic diversity (less than 1%)	Jr Metacognitive Awareness Inventory (Jr MAI) Strategic Problem Solving Inventory (SPSI) Meta-comprehension Strategies Index (MSI) Index of Reading Awareness (IRA)	For the most part, the offline measures correlated with each other. Only the MAI and IRA in younger students failed to reach significance. Overall, correlations at older ages were more significant than younger ages. The authors did note that the correlations were not very strong when	Grades 3-5 KOC/ROC r = .24 MAI/MSI r = 0.30 MAI/IRA r = 0.22 MAI/SPSI r = 0.72 Grades 6-9 KOC/ROC r = .61 MAI/MSI r = 0.23

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
					considering the sample size.	MAI/IRA $r = 0.28$ MAI/SPSI $r = 0.68$
Sperling et al, 2004	Off-line (MAI, LSS, MSLQ) and On-line (CJ) Prospective: MAI, LSS, MSLQ Concurrent: CJ Retrospective: NA	Metacognitive knowledge (declarative, procedural, conditional, strategies): MAI, LSS Metacognitive experience (planning, monitoring, regulating, controlling, debugging, evaluation, strategy use): MAI, LSS, MSLQ	Study 1: 109 undergraduates enrolled in an academic strategies class – most were 1st year students Study 2: 40 education majors in either their sophomore or junior year of college	Study 1: MAI (Metacogntive Awareness Inventory) and LSS (Learning Strategies Survey) – both questionnaires purported to measure aspects of metacognition Study 2: MAI and Motivated Strategies for Learning Questionnaire (MSLQ) – both self-report measures Confidence Judgments on a 20-item objective test Note: for MSLQ, only the Metacognitive self-regulation scale was used for correlations, LSS under study 2 is the learning strategies scale of the MSLQ	Study 1: Within the MAI, the Knowledge of Cognition (KOC) factor correlated significantly with the Regulation of Cognition (ROC) factor. Subscales of the MAI and LSS were also correlated. All were significant correlations except KOC from the MAI and overt strategy use of the LSS. Study 2: Within the MAI, the Knowledge of Cognition (KOC) factor correlated significantly with the Regulation of Cognition (ROC) factor. The MSLQ and subscales of the MAI were also significantly correlated. Correlations with	Study 1 KOC/ROC $r = 0.75$ MAI/LSS $r = .50$ Subscales of MAI & LSS r's ranged from .19 to .53 Study 2 KOC/ROC $r = 0.68$ MAI/MSLQ $r = 0.59$ KOC & ROC/MSLQ $r = 0.59, 0.47$ MAI/LSS of MSLQ $r = 0.60$ KOC & ROC/full LSS scale of MSLQ $r = 0.63, 0.48$ MAI & CJ r's range from -0.28 to 0.16

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
					confidence judgments were small to moderate.	Knowledge of Cognition & Accuracy of CJ r's ranged from -.07 to 0.37 Regulation of Cognition & Predicted Accuracy r's ranged from -0.42 to .04
*Van Kraayenoord & Schneider, 1999	Off-line (WMMTOT, RSCTOT, IRA) and On-line (TAP) Prospective: WMMTOT, RSCTOT, IRA Concurrent: TAP Retrospective: NA	Metacognitive knowledge (declarative, procedural, strategies): WMMTOT, RSCTOT, IRA Metacognitive regulation experience (planning, monitoring, regulating, controlling, debugging, evaluation, strategy use): TAP	140 third and fourth grade students in Germany – 72 in 3rd and 68 in 4th, 75 girls and 65 boys, mean ages were 9.4 (.5) and 10.3 (.4)	Index of Reading Awareness (IRA) – measure of metacognitive knowledge of reading strategies Wurzburg Metamemory Test WMMTOT) Think Aloud Protocols (TAP) on informational text Reading Self-concept Scale (RSCTOT) – measure of metacognitive knowledge of reading	The Index of Reading Awareness (IRA) did not significantly correlate with the think aloud protocols for fourth graders. But, the IRA did significantly correlate with think aloud for third graders. Other measures of metacognition ranged from small to moderate correlations.	Third graders IRA/TAP r = 0.26 IRA/RSCTOT r = .13 IRA/WMMTOT r = .50 RSCTOT/TAP r = -0.09 RSCTOT/ WMMTOT r = 0.20 WMMTOT/TAP r = 0.13 Fourth graders IRA/TAP r = -0.07 IRA/RSCTOT r = .43 IRA/WMMTOT r = .46

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
						RSCTOT/TAP r = -0.03 RSCTOT/ WMMTOT r = 0.35 WMMTOT/TAP r = -0.03
Veenman, 2005 Review of 20 studies. Attempts made to access all articles. Results are detailed here for articles only available in a foreign language and for statistics not reported in the original articles. See * for other studies included in the review	Off-line (questionnaire, ILS) On-line (interview, TAP) Prospective: Questionnaire, ILS Concurrent: TAP Retrospective: Interview, ILS	Metacognitive knowledge (declarative, procedural conditional knowledge): ILS, questionnaire Metacognitive regulation (orientation, systematic orderliness, evaluation, elaboration, strategy use): interview, TAP	2289 participants over 20 studies Artelt, 2000 – 235 9-16 year olds Veenman et al, 2003 (see below) – 33 University students aged 19-22 Veenman & Beishuzen, 2004 (see below) – 23 university students Mean age = 22 Veenman & Elshout, 1999 – 16 university students in psychology courses Veenman et al, 1994 14 freshman psychology students Elshout et al, 1993 17 freshman psychology students	Artelt, 2000 – Questionnaire (unspecified) measuring metacognitive strategy and Interview of metacognitive strategies Veenman, et al, 2003 Inventory of Learning Styles (ILS) given pre and post. Posttest modified to be task specific. Think Aloud Protocols (TAP) Veenman & Beishuizen, 2004; Veenman & Elshout, 1999; Veenman et al, 1994, Elshout et al, 1993 TAP – frequency ratings of behaviors and qualitative analysis of statements while studying forensic text, completing	Artelt, 2000 – Metacognitive strategies self-reported in the questionnaire and the interview did not significantly correlate. Veenman et al, 2003 – moderate correlations were found between the TAP and ILS Veenman & Beishuizen, 2004 – frequency ratings of behavior and qualitative analysis of think aloud data significantly correlated Veenman & Elshout, 1999 – frequency ratings of behavior and qualitative analysis of think aloud data significantly correlated Veenman et al, 1994 – frequency ratings of behavior and qualitative	Artelt, 2000 Offline questionnaire/ Interview r = .02 Veenman et al, 2003 ILS self-regulation scale/Think Aloud r = .22 ILS/ILS r = .49 Think Aloud/ILS adapted r = .31 Veenman & Beishuizen, 2004 r = .80 Veenman & Elshout, 1999 r = .98 Veenman et al, 1994 r = .87 Elshout et al, 1993 r = .95

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
				thermodynamics problems, and completing a learning task on electricity respectively.	analysis of think aloud data significantly correlated Elshout et al, 1993 – frequency ratings of behavior and qualitative analysis of think aloud data significantly correlated	
*Veenman & Beishuizen, 2004	On-line (TAP) Prospective: NA Concurrent: TAP Retrospective: NA	Metacognitive Regulation (planning, orientation, evaluation, elaboration): TAP	46 students in social sciences at Leiden University Mean age = 22 years	Think Aloud Protocol (TAP) – one text on forensic psychology and one on geography	A score of metacognitive skillfulness was calculated. Metacognitive skillfulness across texts significantly correlated with one another.	r = .67
*Veenman et al, 1993a	On-line (Logfiles, TAP) Prospective: NA Concurrent: TAP, Logfiles Retrospective: NA	Metacognitive Regulation (orientation, systematic orderliness, evaluation, elaboration): TAP	28 first year psychology students	Computer logfiles from science problem solving activities Think Aloud Protocol (TAP)	Think aloud scores correlated significantly with metacognitive measures from the computer logfiles. The metacognitive measures from the logfiles also correlated significantly.	TAP/Logfile Orderliness r = .64 TAP/Logfile Monitoring r = .62 Monitoring/ Orderliness r = .73
*Veenman et al, 1993b	On-line (TAP) Prospective: NA Concurrent: TAP Retrospective: NA	Metacognitive Regulation (planning, systematic orderliness, monitoring, elaboration): Logfiles, TAP	28 first year psychology students	Think Aloud Protocol (TAP) with physics and statistics content. The first think aloud preceded the second by two weeks.	Think aloud scores across content correlated significantly.	r = .62

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
*Veenman et al, 2003	Off-line (ILS) and On-line (TAP) Prospective: ILS Concurrent: TAP Retrospective: ILS	Metacognitive Knowledge (learning style, strategy use): ILS Metacognitive Regulation (planning monitoring, regulation, evaluation): ILS, TAP	33 University students aged 19 to 22	Think aloud (TAP) while studying a text about a technical topic with a multiple-choice posttest Inventory of Learning Styles (ILS) self-report questionnaire of regulatory activities and metacognitive knowledge – given before and after, post-ILS adapted to be task specific	Most correlations between scales and performance showed that self-report of learning styles did not match actual performance. Authors noted if outliers were removed, the correlations would have been even smaller.	Think Aloud/ILS subscales r's ranged from -0.18 to 0.29
*Veenman et al, 2004	On-line (TAP, Logfiles) Prospective: NA Concurrent: TAP, Logfiles Retrospective: NA	Metacognitive experience (orientation, planning, evaluation, elaboration): TAP, Logfiles	113 students from the Amsterdam area – 28 4th graders (age M=9.5), 28 6th graders (age M=11.6), 30 8th graders (age M=14.1), 27 university students (age M=22.5)	Computer simulated problems in geography and science. Logfiles record eye gaze, clicking, and other behaviors Think aloud (TAP) during computer problem solving	Think aloud significantly correlated with logfile recorded behaviors for both the science related and geography related computer tasks.	Science Logfile/ TAP $r = 0.85$ Geography Logfile/ TAP $r = 0.84$
*Veenman et al, 2005	On-line (TAP, Observation) Prospective: NA Concurrent: TAP, Observation Retrospective: NA	Metacognitive experience (orientation, planning, evaluation, elaboration): TAP	41 secondary school students in the Netherlands aged 12-13	Think Aloud (TAP) on 6 maths word problems Behavioral observations during the Think Aloud	Observers coding of behaviors and Think Aloud statements significantly positively correlated.	Observation/Think Aloud $r = 0.89$
Veenman & Van Cleef, 2007	Off-line (MSLQ, ILS, Questionnaire)	Metacognitive Regulation (regulation and monitoring): MSLQ,	30 secondary students in math class	Think Aloud Protocols (TAP) during	The Cognitive strategy use and Self-regulation scales from the MSLQ	TAP & MSLQ/ILS r's averaged 0.11

Table 2 (continued)

Authors	Type of Measures	Components Evaluated	Participants	Measures	Results	Effect Size
(as reported in Schellings et al, 2013)	and On-line (TAP) Prospective: MSLQ, ILS Concurrent: TAP Retrospective: Questionnaire	ILS, TAP, Questionnaire		mathematical problem solving Motivated Strategies for Learning Questionnaire (MSLQ) – metacognitive scale Inventory of Learning Styles (ILS) – metacognitive scale Retrospective questionnaire written by authors	and the Self-regulation scale from the ILS did not significantly correlate. Scores on the retrospective questionnaire had a moderate correlation with TAP.	TAP & retrospective questionnaire r = 0.28
Winne & Jamieson-Noel, 2002	Off-line (STQ) and On-line (PrepMate, CJ) Prospective: NA Concurrent: PrepMate, CJ Retrospective: STQ	Metacognitive knowledge (procedural and conditional knowledge): STQ Metacognitive Regulation (strategy use, monitoring, evaluating): PrepMate, CJ	69 undergraduate students from a Canadian University 18 males and 51 females Mean age = 21.73 SD = 5.02 Ages ranged from 17-43 Final sample 62 due to lost assessments	Confidence Judgment (CJ) on Achievement items Study Tactics Questionnaire (STQ) – measure of monitoring and strategy use PrepMate a computerized note taker that tracks students "metacognitive studying" as they fill in the sections	CJ was not correlated with either PrepMate or the STQ The STQ did significantly correlate with behaviors on PrepMate. CJ scores significantly correlated with most subscales of calibrated STQ	STQ/PrepMate r = .34 Subscales of strategy comparing STQ to PrepMate r's ranged from .00 to .72 Calibrated Subscales STQ/CJ r's ranged from -.57 to -.31

*study also reported in Veenman, 2005

R (R Core Team 2018) and statistically analyzed using a random effects model and Hunter and Schmidt (2004) method with the metafor package (Viechtbauer 2010). Because of the small number of studies, Knapp and Hartung's (2003) adjustment was also applied.

For the purposes of this study, all measures were labeled by their factor and/or subcomponent (e.g. metacognitive knowledge, planning), the timing of the measure (prospective, concurrent, retrospective), and assessment type (on-line, off-line). These labels allowed for analysis of moderators where it was necessary, and for meta-analysis of specific variables. Off-line is defined as a measure occurring before or after the learning task (Veenman 2005). Accordingly, overall confidence judgments made after the completion of the entire task were categorized as off-line. Confidence judgments made after completing each problem or question were classified as on-line since the learning was still occurring in a way that could effect the next judgment. Using the same reasoning, confidence judgments were also labeled as retrospective for overall and concurrent for judgements made after each problem or question.

Results

Knowledge and regulation

Thirteen articles analyzed correlations between knowledge and regulation, contributing 20 correlations for the meta-analysis. Measures of knowledge evaluated declarative, procedural, conditional, person, task, and/or strategy knowledge as defined by Flavell (1979) and Schraw and Dennison (1994). Knowledge was assessed by prospective judgments of metacognitive abilities that occured prior to commencing a task, interviews, the Index of Reading Awareness (IRA; Van Kraayenoord and Schneider 1999), Wurzburg Metamemory Test (WMMTOT; Van Kraayenoord and Schneider 1999), and the total score or metacognitive subscale scores of self-report questionnaires (see Table 2 for a complete list of measures). Regulation was evaluated by metacognitive tasks involving orientation, planning, prediction, organization, monitoring, regulation, control, systematic orderliness, debugging, evaluation, and reflection. Regulation was assessed through retrospective interviews, confidence judgments (CJ), think aloud protocols (TAP), PrepMate (Winne and Jamieson-Noel 2002), Index of Reading Awareness (IRA; Van Kraayenoord and Schneider 1999), the Meta-comprehension Strategies Index (MSI; Sperling et al. 2002), Cognitive Developmental aRithmetics (CDR; Desoete 2009), and the total score or metacognitive subscale scores of self-report questionnaires (see Table 2). All questionnaires reported good internal consistency except for 3 subscales of the task specific questionnaire employed in both of Schellings' studies (Schellings 2011; Schellings et al. 2013). Correlations for subscales with poor Cronbach's alpha scores were included in neither Schellings' articles nor this meta-analysis.

The 13 studies amassed a total of 2697 participants that varied in age from primary (604) and secondary (1317) to university students (776). Participants also varied nationally as research was conducted in America, Canada, Germany, the Netherlands, Nigeria, and Turkey. Pearson's r correlations ranged widely from −0.03 to 0.93. A positive correlation indicates that greater knowledge of metacognition was associated with more accurate metacognitive regulation, in other words, greater metacognitive knowledge related to better metacognitive skills. The pooled effect size estimate for the data is $r = 0.34$ (95% CI, 0.22–0.46; see Table 3 for full meta-analysis results). However, interpretations of this value are not feasible because of the elevated heterogeneity ($I^2 = 96.26\%$). Due to the heterogeneity of the data, measures of

Table 3 Meta Analyses of factors and subcomponents of metacognition

Relationship	Number of Correlations	Pooled Effect Size (CI)	Heterogeneity (I^2)	Significant Moderators	Moderator Direction
Factor Relationships					
Knowledge & Regulation	21	0.34 (0.22-0.46)	96.26%	Measure – Interview (CPQ/Retrospective)	positive
Off-line & On-line	23	0.22 (0.14-0.31)	58.78%	Measure – TLSI	negative
				Age - University	positive
Within Factor Relationships					
Person & Task	6	0.41 (0.15-0.68)	89.44%	Age – Secondary	positive
Person & Strategies	5	0.43 (0.13-0.72)	76.06%	Age - Secondary	positive
Task & Strategies	5	0.51 (0.43-0.59)	0%	Time – Retrospective	positive
Planning & Monitoring	5	0.63 (0.46-0.81)	73.67%	none	
Planning & Evaluation	7	0.48 (0.39-0.58)	28.86%	Age – Secondary & University	positive
					positive
Monitoring & Evaluation	7	0.42 (0.23-0.62)	73.36%	Age – Secondary & University	positive
					positive
Between Factor Relationships					
Person & Planning	3	0.40 (-0.27-1.08)*	70.96%	none	
Task & Planning	3	0.48 (0.18-0.77)	27.60%	none	
Strategies & Planning	3	0.32 (-0.29-0.92)*	63.99%	none	
Person & Monitoring	4	0.37 (-0.05-0.79)*	80.89%	none	
Task & Monitoring	4	0.42 (0.13-0.70)	61.70%	none	
Strategies & Monitoring	4	0.38 (0.11-0.64)	50.84%	none	
Person & Evaluation	5	0.47 (0.29-0.64)	51.14%	none	
Task & Evaluation	3	0.46 (0.41-0.52)	0%	none	
Strategies & Evaluation	3	0.46 (0.35-0.56)	0%	none	

*$p > .05$

regulation, timing of the assessment, type of assessment, age, and nationality were evaluated as moderators. The moderators lowered the heterogeneity to 37.07%, 72.96%, 91.66%, 92.04%, and 92.61% respectively. Of particular note, the instruments used to measure knowledge were responsible for 100% of the heterogeneity, leaving 0% residual heterogeneity (see Fig. 7). Additionally, measuring knowledge with an interview was a significant positive moderator indicative of higher effect sizes than other measures. Retrospective instruments (Timing) and the CPQ (measure of regulation) were also significant positive moderators. However, the Pearson's correlation between the CPQ and a retrospective interview was $r = 0.93$. Therefore, timing (retrospective), measures of regulation (CPQ), and interviews are moderators because they are responsible for an extreme outlier. Since the outlier did not affect measures of knowledge, the results indicate that the choice of assessment instrument for measuring knowledge is most responsible for effect size variations.

Knowledge and regulation as off-line and on-line

Brown (1987) posited that all off-line measures of metacognition are actually measures of knowledge, even if statements are querying regulation. This supposition has merit as participant's skills are not being measured in a questionnaire, rather it is awareness or knowledge of regulation that is evaluated. Consequently, a new set of data was selected following the hierarchy detailed above that looked for any correlation between on-line (regulation) and off-line (knowledge) instruments. This alternate classification yielded 21 studies that contributed 23 correlations. The studies were comprised of 1691 American, Canadian, Chinese, Dutch, German, Greek, and Turkish participants. Similar instruments were employed apart

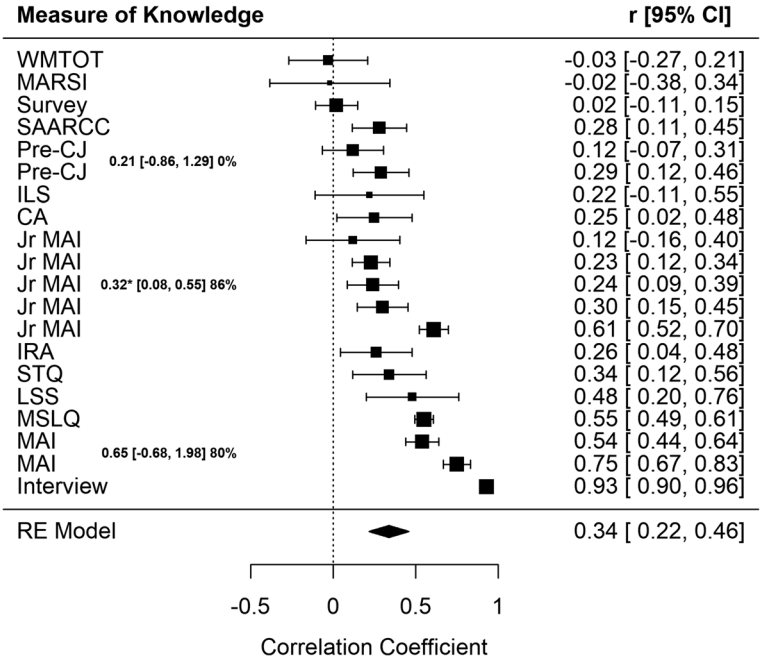

Fig. 7 Forest Plot of knowledge and regulation correlations by the measure of knowledge I2 listed as percentage
*$p < 0.05$

from the IRA, and with the addition of the Interactive Multi-Media Exercises (IMMEX; Cooper et al. 2008) and Sokoban tasks (Li et al., 2015). Primary (390), Secondary (156), and University (1145) students volunteered to take part in research that found correlations ranging from −0.39 to 0.63. This selection of studies resulted in a pooled effect size estimate of $r = 0.22$ (95% CI, 0.14–0.31) with heterogeneity of $I^2 = 58.78\%$. Due to the moderate amount of heterogeneity, a meta-regression was also run on this data. Similar to the previous results, measures of knowledge were responsible for 100% of the variation, left 0% residual heterogeneity, and was a significant moderator. Measures of regulation lowered the heterogeneity to 22.34% and nationality and timing of the instruments to 38.14% and 43.78%. Age was a significant moderator revealing that, correlation coefficients of students at the university level significantly increase the pooled effect size estimate and lower the heterogeneity to 32.93%. When evaluated as subgroups, age was not significant for primary and secondary. Additionally, secondary and university still revealed moderate heterogeneity (see Fig. 8). Thus, in general, older participants have stronger correlations between knowledge and regulation, but the results still vary widely based on the instrument used to measure knowledge. Taken together, then, self-reports of metacognitive knowledge and metacognitive regulation poorly relate to actual performance on metacognitive tasks. Of note, some self-reports appear to correlate more strongly than others (Fig. 7).

Subcomponents of knowledge and regulation

Few studies examine the relationship between the subcomponents of regulation and knowledge. The studies that explore those relationships are often correlating subscales instead of

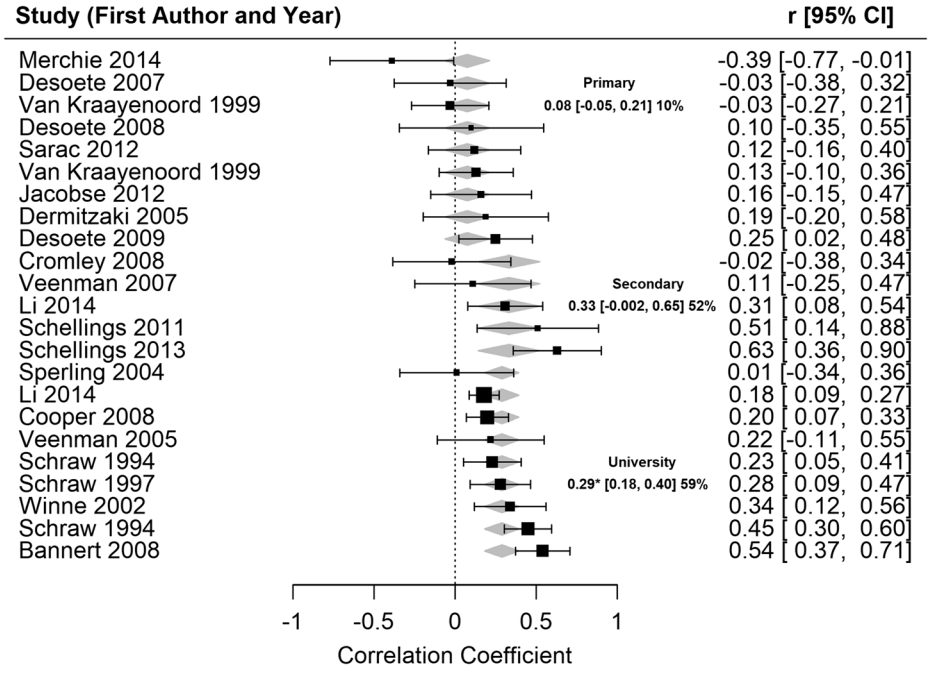

Fig. 8 Forest plot of online and offline correlations moderated by age I2 listed as percentage *$p < 0.001$

overall instrument scores. Because the subcomponents of metacognition operate jointly in the completion of a task, it is rare to see one subcomponent evaluated by one instrument. The studies found for this meta-analysis reflect this rarity, as all of the studies used subscale measures to evaluate relationships between subcomponents of metacognition. Thirteen studies employing 2278 participants compared two different measures evaluating subcomponents of knowledge and regulation. Participants ranged in age from primary (403) and secondary (1270) to university (605). Like the previous analyses, the measures varied widely and included both on-line tasks and off-line questionnaires. Additionally, measures were given across time and in a variety of countries including America, Canada, Germany, Greece, the Netherlands, and Romania.

Meta-analyses on subcomponents of knowledge revealed pooled effect sizes that ranged from 0.41 to 0.43. Pooled effect sizes for subcomponents of regulation ranged from 0.42 to 0.63 (see Table 3). Four of the six estimates displayed elevated heterogeneity. Meta-regressions revealed that in all but one case, measures of knowledge accounted for 100% of the heterogeneity. The five correlations between planning and monitoring came from five different measures, therefore measures of knowledge could not be evaluated as a moderator in the sixth study. Instead, nationality was responsible for 100% of the heterogeneity. Also of note, is that in four of the six meta-regressions, age was a significant moderator indicating that older participants had significantly stronger effect sizes than primary-aged participants. While age was a significant moderator, it did not meaningfully lower the heterogeneity. Meta-analyses of subcomponents across factors found pooled effect sizes that varied from 0.32 to 0.48 (see Table 3). Three of nine meta-analyses found non-significant pooled effect sizes. Pooled effect sizes that were significant had moderate to no heterogeneity. Because of the small number of studies examining these relationships, meta-regressions either could not be run, or moderators did not meaningfully decrease the heterogeneity.

Three other subcomponents of metacognition were evaluated at a subscale level in three studies. Elaboration (Muis et al. 2007) obtained moderate to strong effect sizes with other subcomponents of regulation (Planning 0.38–0.67; Monitoring 0.34–0.70; Evaluation 0.42–0.66). Prediction (Desoete et al., 2008) obtained small effect sizes with subcomponents of knowledge (Declarative 0.16; Procedural 0.10; Conditional 0.18) and small to strong effect sizes with other subcomponents of regulation (Planning 0.12–0.55; Monitoring 0.39–0.84; Evaluation 0.08–0.89). Finally, Attribution (Desoete et al. 2001) was characterized by small to moderate effect sizes with subcomponents of knowledge (rs 0.01 to 0.24) and small effect sizes with subcomponents of regulation (rs − 0.04 to 0.18). Because each study evaluated only one of these components and thus utilizing only one population, meta-analyses could not be run. Taking all the meta-analyses into consideration, it appears that subscales relate more strongly to behavior across and within measures than the overarching factors (knowledge and regulation) of metacognition.

Discussion

Results of the meta-analyses within the factors of knowledge and regulation (Table 3: Within Factor Relationships) reveal moderate to large effect sizes, confirming the existence of the two overarching factors. Conversely, the data shows only small to moderate pooled effect size estimates between knowledge and regulation, and confirm previous research finding that on-line and off-line measures do not strongly correlate. The smaller pooled effect size of 0.22

from measures categorized as on-line and off-line is not dissimilar to (Veenman and Van Hout-Wolter's 2002) estimated average of $r = 0.17$ (as reported in Jacobse and Harskamp 2012). The pooled effect size is greater ($r = .34$) when measures aren't categorized as on-line and off-line assessments. Thus, the data indicates that while self-reports consistently provide a broad overview of participants' understanding of their own metacognitive knowledge and metacognitive regulation, the reports only weakly correlate with participants' metacognitive behavior.

It is important to note that the resulting estimates in this study must be treated with caution because of the high heterogeneity. The heterogeneity can be explained by the wide range and variety of measures used to assess knowledge. One may therefore question whether the measures of knowledge are assessing the same underlying construct making their ability to predict behavior on a metacognitive task variable. Similarly, measures of regulation also meaningfully decrease heterogeneity, though it does not have as significant an impact as measures of knowledge. Consequently, the effect size varies based on the instruments chosen to measure metacognition. This may be due to the fact that tasks tend to measure one specific metacognitive skill (e.g. monitoring) while self-reports give an overview of many metacognitive skills. Thus, the data appears to reinforce the importance of carefully choosing an appropriate measure.

Sorting the data by measures of knowledge and running another meta-analysis still finds some heterogeneity within the results (see Fig. 7). The MAI, as an example, revealed multiple factor structures in the systematic review. Similarly, correlational results are wide ranging when employing the MAI (r's 0.07 to 0.70). This may be explained by age, as it was a significant factor for the on-line versus off-line meta-regression. Age also shows up frequently as a significant modifier among the subcomponents. Meta-regressions with age as a modifier, in general, suggest that older participants achieve stronger effect sizes. But again, forest plots and meta-analyses show heterogeneity still exists when data is sorted by age (see Fig. 8). Thus, both age and choice of instrument appear to meaningfully impact results, reinforcing the import of carefully choosing a self-report as well as lending support to Ning's suggestion that questionnaire factor structure is related to self-reported metacognitive ability.

Meta-analyses assessing components of knowledge and regulation, find strong correlations that lack heterogeneity (rs 0.46–0.51; Table 3: Between Factor Relationships). This supports the existence of two factors. Only attribution failed to have substantial relationships with other possible subcomponents and, like the systematic review, discounts the presence of a third factor based on motivation or attribution. In addition, the meta-analyses suggest that the subscale level of self-reports may strongly relate to behavior on metacognitive tasks. Thus, self-reports of knowledge and regulation may be useful for corelating to behavior at the subcomponent level, more so than at the factor level.

However, like the factor level, many of results must be interpreted with caution. Here again, variation in the instruments used to measure knowledge were most responsible for the wide range of results. Age also appeared as a significant moderator, but again, had less impact than the diversity of measures of knowledge. Thus, subcomponent meta-analysis reinforces the import of choosing the best instrument for the study's specific questions. Furthermore, choice of instrument appeared more critical than timing or type of instrument. The studies varied widely in their use of on-line and off-line assessments and in the timing of the assessments (prospective, retrospective, and concurrent). Yet, timing appeared only once as a significant moderator, and type did not significantly moderate the results at all. This does not mean researcher's emphasis (Sperling et al. 2004; Van Hout-Wolters & Schelllings 2009; &

Veenman 2005) on the need for both on-line and off-line assessments across time should be ignored. Rather, the data seems to indicate that as multi-method approaches are being utilized widely across studies, there is not a superior type or timing of the assessments. Thus, multi-method assessments will provide a more detailed picture of metacognition.

General discussion

Current research that analyzes the factor structure of self-reported metacognition varies widely, from reporting a unidimensional structure to a structure with nine components. The first systematic review of factor analyses indicates that self-reports of metacognition are best suited to measure two factors characterized as regulation and knowledge but does not support the distinct measurement of additional factors or subcomponents of metacognition. Likewise, the second systematic review and associated meta-analysis did not support the inclusion of additional factors, as shown by weaker fit indices and small effect sizes between attribution and subcomponents of knowledge and regulation. Meta-analyses of subcomponents (person, task, strategies, planning, monitoring, evaluation, elaboration) tend towards moderate and strong pooled effect size estimates, again supporting the ability of self-reports to measure a two-factor structure of regulation and knowledge. It is important to note that this review is not evidence that only two factors of metacognition exist, rather that two broad factors of metacognition are robustly found from available self-reports measures.

Overall, the meta-analyses indicate that subcomponents of knowledge correlated with subcomponents of regulation result in considerably stronger estimates than the pooled effect sizes found between the broad factor measurements of knowledge and regulation (Table 3), indicating that subcomponents may better relate to each other and to behavior than the overall factors. Thus, it would appear Van Hout-Wolters and Schelling's (2009) contention that metacognitive relationships should be measured at the subscale level has strong merit. Additionally, it lends support to the presence and importance of the subcomponents. The lack of heterogeneity in some of the pooled estimates of subcomponent relationships lends further credibility to the supposition that choice of measure may be a contributing factor to the wide range of somewhat contradictory results. Of note, every pooled estimate that lacked heterogeneity included the COMEGAN-ro as one of the instruments involved in the correlational analysis. The systematic review also found the COMEGAN-ro to report some of the strongest fit indices of a two-factor model.

While self-reports do not adequately measure the nuances of metacognitive behaviours, there is still a place for them in metacognitive research. Due to the variation among self-reports, the systematic reviews and meta-analyses do not indicate one specific self-report as the "gold" standard. Thus, choice of instrument and how the resulting data is used to measure metacognitive knowledge must be carefully considered. The data does suggest that self-reports are useful in obtaining a broad overview of participants' knowledge and regulation. To correlate with metacognitive behavior, self-reports should be chosen carefully according to the subscales the research is evaluating. Furthermore, self-reports provide a broad understanding of how participants view their own metacognitive abilities. Therefore, the strength of self-reports may lie in their inability to reflect behaviour, allowing researchers to explore why participants tend towards inaccurate self-reporting. For example, research questions such as; Are those with autism or anxiety more accurate self-reporters than neurotypicals or healthy controls? or Do participants with more accurate metacognitive skills on tasks self-report less

metacognitive ability than their peers?, would be valuable explorations for which self-reports are necessary assessment instruments.

It is important to note that choice of instrument could not explain 100% of the heterogeneity in every instance. Age also had a meaningful impact on the results, but like choice of instrument, cannot account for all of the heterogeneity. Ning's 2016 study, described in Study 1, poses an alternative interpretation based on respondents' self-reported metacognitive abilities. It is plausible that heterogeneity found throughout the meta-analyses is due to participant metacognitive capabilities. In other words, Ning's study suggests that those with stronger metacognitive expertise utilize multiple strategies that are more sophisticated, thus employing multiple factors and subcomponents of metacognition. Those with weaker or minimal metacognitive capabilities may only utilize one or two simple strategies, revealing a simplified, or unidimensional, structure of metacognition. Under this hypothesis, it may be possible to adequately measure subcomponents with a self-report, but only in those with strong metacognitive skills.

The difference in nuance of metacognitive skills caused by expertise could effect the relationships between subcomponents, and account for the widely ranging scores that appear across instruments and even within instruments. The interpretation of differences in expertise are supported by the results showing age as a significant moderator while also continuing to show a range of results within each age cohort. Future studies collecting self-report data may want to divide the results by participant capabilities to explore the possibility of stronger relationships and a more complex underlying structure due to more developed metacognitive skill. Accordingly, it may be possible to determine weak metacognitive areas based on differences in structure (unidimensional versus two-factors) and the ability of subcomponents to relate to metacognitive behavior. Metacognitive skill can be taught (Perry et al. 2018). Under this supposition, it may also be possible to train individuals in specific subcomponents of metacognition in pursuit of academic achievement as well as better health and well-being.

Strengths and limitations

Study 1 and Study 2 are the first to comprehensively evaluate the use of self-reports to measure metacognition. Because the term metacognition came into use in the 1970s (Flavell 1979), there are 40 years of available research to analyse. Hence, given the range of studies analysed, the results are likely to be fairly representative of the general population and provide a rich pool of data from which an understanding of a metacognition can be evaluated. In addition, because measuring metacognition in the general population is not dependent on randomization, order of measures, or even participant sample characteristics – as evidenced by the wide range of results within age groupings, there is little risk of bias within the studies included for both reviews. Bias could result from participant response bias on the self-report questionnaires. But this concern is analyzed when comparing on-line versus off-line methods of measuring metacognition. The studies selected for both reviews are certainly subject to publication bias. However, as analysis of factor structure is not dependent on specific thresholds of findings and correlational analysis between metacognitive measures and subscales is generally part of a larger statistical question, a substantial quantity of both insignificant and robust results was reported within and across studies. A funnel plot would serve to further analyze publication bias, but the elevated heterogeneity, due to the wide range of results, renders funnel plot data unreliable (Terrin et al. 2003).

As stated throughout the analysis and discussion the amount of heterogeneity found within the meta-analyses does limit firm conclusions based on statistical analyses. This review was also limited to published studies that appeared in English. While we greatly appreciate the help of authors in providing some of these studies in an accessible format, we were unable to acquire all the inaccessible studies. In addition, the substantial volume of correlational data that had to be eliminated due to the constraint of preventing oversampling of participant populations is also a limitation. It is possible that an alternate hierarchy would obtain different results for the meta-analysis. The study tried to mitigate the effects of the volume of data by establishing deference to measures created specifically based on a theory of metacognition and giving lesser status to measures designed for specific venues (e.g. the classroom or therapeutic setting). The results clearly revealed that choice of instrument to measure metacognitive knowledge has a meaningful impact. Thus, it is probable that a hierarchy with an alternative focus could find significantly different results. To explore this concern, a meta-analysis was run with the entirety of statistical results culled from the systematic review. A meta-analysis of all results provided very similar pooled estimates to the ones reported in Study 2.

Conclusion

Self-reports can be problematic for a variety of reasons, such as effects of participant mood at the time the report is completed, social desirability bias, and central tendency bias with Likert scale responses. Furthermore, the correlations between participant self-reports and participants' corresponding quantifiable behaviour are generally weak (Carver and Scheier 1981; Veenman 2005). Metacognitive self-reports are not exempted from these challenges, as seen in the fact that self-reports analysed for this review cannot adequately measure the nuances of metacognitive behaviour. However, metacognitive self-reports can still be used purposefully in research. Current self-reports can provide a general overview of knowledge and regulation skills. The relationships between subscales of self-reports and participant behaviour can be measured. Furthermore, the act itself of completing a self-report requires metacognition, and as such, can give researchers insights into how metacognitive knowledge can differ from metacognitive behaviour.

The studies analysed in this review support the use of self-report to measure participants' general metacognitive abilities in knowledge and regulation as two distinct, albeit relatively basic, metacognitive factors. However, metacognitive knowledge measured as a broad factor is not strongly related to behavior on metacognitive tasks. Both factors can be divided into subcomponents that work jointly to achieve a goal or complete a task. However, self-reports cannot reveal the complex processes that occur at the subscale level. In contrast, self-reports do seem able to strongly correlate with behavior when subscales are used. However, data exploring the relationships between factors and components varies widely. This appears to be caused predominantly by choice of instrument to measure knowledge, and secondarily by age and choice of instrument to evaluate regulation. Thus, it is imperative that future research using self-reports systematically identify the purpose of the self-report and choose the report carefully based on that purpose. For example, if only a broad measure of knowledge and regulation are needed, then a variety of self-reports are effective. However, to evaluate the relationship between self-report and behavior, the method of self-report should align closely to the skills being measured by an experimental task. Alternatively, self-report may be used to

further understand when or what type of participant is more accurate in predicting or understanding their own metacognitive behavior.

A challenge for researchers is to determine whether metacognitive capabilities effect the underlying structure of metacognition, and how the findings from this exploration can help inform venues such as schools and therapeutic environments where metacognitive skills are essential. Metacognition can be taught. If, as one interpretation of the data suggests, self-reported weak metacognitive skills function as a broad unidimensional construct, then it is feasible that teaching metacognition aimed at specific components prior to academic instruction or mental health therapy can allow individuals to more fully access both learning and the benefits of therapeutic interventions. Future research should look towards establishing a framework of metacognition that can be utilized across settings for advances in achievement and mental health and well-being, and then define how self-reports are best used towards that purpose.

Funding information We wish to draw the attention of the Editor to the fact that this paper was funded by a James Watt Scholarship awarded by Heriot-Watt University.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Sample Searches.

Systematic Review.

EbscoHost, ERIC, PsycINFO, PsycArticles, Scopus, Web of Science, and WorldWideScience.org (for all terms, there were no limits of any kind imposed):

1. metacognit*
2. model
3. 1 and 2
4. Title screening
5. meta-cognit*
6. 2 and 5
7. Title screening
8. "factor analy*"
9. 1 and 8
10. Title screening
11. 5 and 8
12. Title screening
13. Duplicates removed
14. Abstract screening

Meta-Analysis.

EbscoHost, ERIC, PsycINFO, PsycArticles, Scopus, Web of Science, and WorldWideScience.org (for all terms, there were no limits of any kind imposed):

1. metacognit*
2. on-line
3. 1 and 2, Title screening
4. off-line
5. 1 and 4, Title screening
6. multi-method
7. 1 and 6, Title screening
8. Meta-cognit*
9. 2 and 8, Title screening
10. 4 and 8, Title screening
11. 6 and 8, Title screening
12. online
13. 1 and 12, Title screening
14. 8 and 12, Title screening
15. offline
16. 1 and 15, Title screening
17. 8 and 15, Title screening
18. Duplicates removed
19. Abstract screening

References

Akin, A., Abaci, R., & Cetin, B. (2007). The validity and reliability of the Turkish version of the metacognitive awareness inventory. *Educational Sciences: Theory and Practice, 7*(2), 671–678.

Akturk, A. O., & Sahin, I. (2011). Literature review on metacognition and its measurement. In *Procedia - Social and Behavioral Sciences* (Vol. 15, pp. 3731–3736). https://doi.org/10.1016/j.sbspro.2011.04.364.

Allen, B. A., & Armour-Thomas, E. (1993). Construct validation of metacognition. *The Journal of Psychology, 127*(2), 203–211. https://doi.org/10.1080/00223980.1993.9915555.

Altindağ, M., & Senemoğlu, N. (2013). Metacognitive skills scale. *Hacettepe University Journal of Education, 28*(1), 15–26.

Artelt, C. (2000). Wie prädiktiv sind retrospektive Selbstberichte über den Gebrauch von Lernstrategien für strategisches Lernen? Zeitschrift Fur Padagogische Psychologie, 14(2–3), 72–84. https://doi.org/10.1024//1010-0652.14.23.72.

Aydin, U., & Ubuz, B. (2010). Turkish version of the junior metacognitive awareness inventory: An exploratory and confirmatory factor analysis. *Education and Science, 35*(157), 32–47.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*(1), 39–58. https://doi.org/10.1007/s11409-007-9009-6.

Beran, M. J. (2012). Foundations of metacognition. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199646739.001.0001.

Bonner, J. (1988). Implications of cognitive theory for instructional design: Revisited. *Educational Communication and Technology, 36*(1), 3–14. https://doi.org/10.1007/BF02770012.

Bong, M. (1997). Congruence of measurement specificity on relations between academic self-efficacy, effort, and achievement indexes In *AERA 1997*.

Brown, A. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in Instructional Psychology. Volume 1* (Vol. 1, pp. 77–165). Mahwah, NJ: Erlbaum.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Wernert (Ed.), *Metacognition, motivation and understanding* (pp. 65–116). Mahwah, NJ: Erlbaum.

Bryce, D., Whitebread, D., & Szűcs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition and Learning, 10*(2), 181–198. https://doi.org/10.1007/s11409-014-9120-4.

Carver, C.S., Scheier, M. F. (1981). Relationship between self-report and behavior. In: *Attention and Self-Regulation*. SSSP springer series in social psychology (pp. 269-285). New York: Springer.

Çetinkaya, P., & Erktin, E. (2002). Assessment of metacognition and its relationship with Reading comprehension achievement and aptitude. *Bogazici University Journal of Education, 19*(1), 1–11.

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 79–92. https://doi.org/10.1016/j.lindif.2003.08.003.

Cooper, M. M., Sandi-Urena, S., & Stevens, R. (2008). Reliable multi method assessment of metacognition use in chemistry problem solving. *Chemistry Education Research and Practice, 9*(1), 18–24. https://doi.org/10.1039/b801287n.

Core Team, R. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from https://www.r-project.org/.

Cromley, J. G., & Azevedo, R. (2006). Self-report of reading comprehension strategies: What are we measuring? *Metacognition and Learning, 1*(3), 229–247. https://doi.org/10.1007/s11409-006-9002-5.

Dermitzaki, I. (2005). Preliminary investigation of relations between young students' self-regulatory strategies and their metacognitive experiences. *Psychological Reports, 97*, 759–768.

Desoete, A. (2007). Electronic journal of research in Educational Psychology. *Electronic Journal of Research in Educational Psychology, 5*(3), 705–730.

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: How you test is what you get. *Metacognition and Learning, 3*(3), 189–206. https://doi.org/10.1007/s11409-008-9026-0.

Desoete, A. (2009). Metacognitive prediction and evaluation skills and mathematical learning in third-grade students. *Educational Research and Evaluation, 15*(5), 435–446. https://doi.org/10.1080/13803610903444485.

Desoete, A., Roeyers, H., & Buysse, A. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities, 34*(5), 435–449.

Elshout, J. J., Veenman, M. V. J., & Van Hell, J. G. (1993). Using the computer as a help tool during learning by doing. Computers and Education, 21(1–2), 115–122. https://doi.org/10.1016/0360-1315(93)90054-M.

Favieri, A. G. (2013). General metacognitive strategies inventory (GMSI) and the metacognitive integrals strategies inventory (MISI). *Electronic Journal of Research in Educational Psychology, 11*(3), 831–850. https://doi.org/10.14204/ejrep.31.13067.

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.

Georghiades, P. (2004). From the general to the situated: Three decades of metacognition. *International Journal of Science Education, 26*(3), 365–383. https://doi.org/10.1080/0950069032000119401.

Hadwin, A. F., Winne, P. H., Stockley, D. B., Nesbit, J. C., & Woszczyna, C. (2001). Context moderates students' self-reports about how they study. *Journal of Educational Psychology, 93*(3), 477–487. https://doi.org/10.1037/0022-0663.93.3.477.

Harrison, G. M., & Vallin, L. M. (2018). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning, 13*(1), 15–38. https://doi.org/10.1007/s11409-017-9176-z.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the state metacognitive inventory. *Educational and Psychological Measurement, 68*(4), 695–709. https://doi.org/10.1177/0013164407313366.

Jacobse, A. E., & Harskamp, E. G. (2012). Towards efficient measurement of metacognition in mathematical problem solving. *Metacognition and Learning, 7*(2), 133–149. https://doi.org/10.1007/s11409-012-9088-x.

Kim, B., Zyromski, B., Mariani, M., Lee, S. M., & Carey, J. C. (2017). Establishing the factor structure of the 18-item version of the junior metacognitive awareness inventory. *Measurement and Evaluation in Counseling and Development, 50*(1–2), 48–57. https://doi.org/10.1080/07481756.2017.1326751.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693–2710. https://doi.org/10.1002/sim.1482.

Lai, E. R. (2011). Metacognition: A literature review research report. *Pearson's Research Reports*, (April), 41. https://doi.org/10.2307/3069464.

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*(1), 260–271. https://doi.org/10.1037/pas0000019.

Livingston, J. A. (1997). Metacognition: An overview. *Psychology*. https://doi.org/10.1080/09500690320001194 01.

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning, 5*(2), 137–156. https://doi.org/10.1007/s11409-010-9054-4.

Meijer, J., Sleegers, P., Elshout-Mohr, M., van Daalen-Kapteijns, M., Meeus, W., & Tempelaar, D. (2013). The development of a questionnaire on metacognition for students in higher education. *Educational Research, 55*(1), 31–52. https://doi.org/10.1080/00131881.2013.767024.

Merchie, E., & Van Keer, H. (2014). Learning from text in late elementary education. *Comparing Think-aloud Protocols with Self-reports. Procedia - Social and Behavioral Sciences, 112*, 489–496. https://doi.org/10.1016/j.sbspro.2014.01.1193.

Minnaert, A., & Janssen, P. J. (1997). Bias in the assessment of regulation activities in studying at the level of higher education. *European Journal of Psychological Assessment, 13*(2), 99–108. https://doi.org/10.1027/1015-5759.13.2.99.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7). https://doi.org/10.1371/journal.pmed.1000097.

Muis, K. R., Winne, P. H., & Jamieson-Noel, D. (2007). Using a multitrait-multimethod analysis to examine conceptual similarities of three self-regulated learning inventories. *British Journal of Educational Psychology, 77*(1), 177–195. https://doi.org/10.1348/000709905X90876.

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.

Ning, H. K. (2017). The Bifactor model of the junior metacognitive awareness inventory (Jr. MAI). *Current Psychology*, 1–9. https://doi.org/10.1007/s12144-017-9619-3.

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

Ofodu, G. O., & Adepipe, T. H. (2011). Assessing ESL students' awareness and application of metacognitive strategies in comprehending academic materials. *Journal of Emerging Trends in Educational Research and Policy Studies (JETERAPS), 2*(5), 343–346.

Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed Strategies for Learning: A program to improve children's reading awareness and comprehension. *Journal of Educational Psychology, 76*(6), 1239–1252. https://doi.org/10.1037/0022-0663.76.6.1239.

Pedone, R., Semerari, A., Riccardi, I., Procacci, M., Nicolo, G., & Carcione, A. (2017). Development of a self-report measure of metacognition: The metacognition self-assessment scale (MSAS) instrument description and factor structure. *Clinical Neuropsychiatry, 14*(3), 185–194.

Pena-Ayala, A., & Cardenas, L. (2015). Personal self-regulation, self-regulated learning and coping strategies, in university context with stress. In A. Peña-Ayala (Ed.), *Metacognition: Fundaments, applications, and trends* (Vol. 76, pp. 39–72). London: Springer. https://doi.org/10.1007/978-3-319-11062-2_9.

Perry, J., Lundie, D., & Golder, G. (2018). Metacognition in schools: What does the literature suggest about the effectiveness of teaching metacognition in schools? Educational Review, 1911, 1–18. https://doi.org/10.1080/00131911.2018.1441127.

Peterson, P. L., Swing, S. R., Braverman, M. T., & Buss, R. R. (1982). Students' aptitudes and their reports of cognitive processes during direct instruction. *Journal of Educational Psychology, 74*(4), 535–547. https://doi.org/10.1037/0022-0663.74.4.535.

Porumb, I., & Manasia, L. (2015). A Clusterial conceptualization of Metacognization in students. In O. Clipa & C. R. A. M. A. R. I. U. C. Gabriel (Eds.), *Educatia in Societatea Contemporana.Aplicatii* (pp. 33–44). London: Lumen Publishing House.

Pour, A. V., & Ghanizadeh, A. (2017). Validating the Persian version of metacognitive awareness inventory and scrutinizing the role of its components in IELTS academic Reading achievement. *Modern Journal Of Language Teaching Methods, 7*(3), 46–63.

Saraç, S., & Karakelle, S. (2012). On-line and off-line assessment of metacognition improving metacognitive monitoring accuracy in the classroom. *International Electronic Journal of Elementary Education, 4*(2), 301–315.

Schellings, G. (2011). Applying learning strategy questionnaires: Problems and possibilities. *Metacognition and Learning, 6*(2), 91–109. https://doi.org/10.1007/s11409-011-9069-5.

Schellings, G. L. M., Van Hout-Wolters, B. H. A. M., Veenman, M. V. J., & Meijer, J. (2013). Assessing metacognitive activities: The in-depth comparison of a task-specific questionnaire with think-aloud protocols. *European Journal of Psychology of Education, 28*(3), 963–990. https://doi.org/10.1007/s10212-012-0149-y.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary Educational Psychology, 19*, 143–154.

Schraw, G. (1998). On the development of adult metacognition. In C. M. Smith & T. Pourchot (Eds.), *Adult learning and development: Perspectives from educational psychology* (pp. 89–106). Mahwah, NJ: Erlbaum.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*(4), 460–475. https://doi.org/10.1006/ceps.1994.103.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351–371. https://doi.org/10.1007/BF02212307.

Semerari, A., Cucchi, M., Dimaggio, G., Cavadini, D., Carcione, A., Battelli, V., Nicolò, G., Pedone, R., Siccardi, T., D'Angerio, S., Ronchi, P., Maffei, C., & Smeraldi, E. (2012). The development of the metacognition assessment interview: Instrument description, factor structure and reliability in a non-clinical sample. *Psychiatry Research, 200*(2–3), 890–895. https://doi.org/10.1016/j.psychres.2012.07.015.

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., DuBois, N., Howard, B. C., & Staley, R. (2004). Metacognition and self-regulated learning constructs. *Educational Research and Evaluation, 10*(2), 117–139. https://doi.org/10.1076/edre.10.2.117.27905.

Teo, T., & Lee, C. B. (2012). Assessing the factorial validity of the metacognitive awareness inventory (MAI) in an Asian country: A confirmatory factor analysis. *International Journal of Educational and Psychological Assessment, 10*(2), 92–103.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113–2126.

The Cochrane Collaboration (2011). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. Higgins J. P. T., Green S. (Eds). Available from http://handbook.cochrane.org.

van der Stel, M., & Veenman, M. V. J. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences, 20*(3), 220–224. https://doi.org/10.1016/j.lindif.2009.11.005.

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 Hout-Wolters, B. & Schellings, G. (2009). Measuring learning strategies: Different measurement methods and their usability in education and research. *Pedagogische Studien, 86*.

Van Kraayenoord, C. E., & Schneider, W. E. (1999). Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. *European Journal of Psychology of Education, 14*(3), 305–324.

Van Zile-Tamsen, C. M. (1996). *Metacognitive self-regulation and the daily academic activities of college students*. The State University of New York: University at Buffalo.

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und metakognition: Implikationen fur forshung und praxis* (pp. 77–99). Munster: Waxmann.

Veenman, M. V. J. (2013). International handbook of metacognition and learning technologies. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (Vol. 28, 28th ed.). London: Springer. https://doi.org/10.1007/978-1-4419-5546-3.

Veenman, M. V. J., & Beishuizen, J. J. (2004). Intellectual and metacognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learning and Instruction, 14*(6), 621–640. https://doi.org/10.1016/j.learninstruc.2004.09.004.

Veenman, M. V. J., & Elshout, J. J. (1994). Differential effects of instructional support on learning in simultation environments. *Instructional Science, 22*(5), 363–383. https://doi.org/10.1007/BF00891961.

Veenman, M., & Elshout, J. J. (1999). Changes in the relation between cognitive and metacognitive skills during the acquisition of expertise. *European Journal of Psychology of Education, 14*(4), 509–523. https://doi.org/10.1007/BF03172976.

Veenman, M. V. J., & 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.

Veenman, M. V. J., Elshout, J. J., & Busato, V. V. (1994). Metacognitive mediation in learning with computer-based simulations. *Computers in Human Behavior, 10*(1), 93–106. https://doi.org/10.1016/0747-5632(94)90031-0.

Veenman, M. V. J., Elshout, J. J., & Groen, M. G. M. (1993a). Thinking aloud: Does it affect regulatory processes in learning? *Tijdschrift Voor Onderwijsresearch, 18*(6), 322–330.

Veenman, M. V. J., Elshout, J. J., & Hoeks, J. C. J. (1993b). Determinants of learning in simulation environments across domains the electrophysiology of language comprehension: A Neurocomputational model view project. In D. M. Towne, T. de Jong, & S. H. Spada (Eds.), *Simulation-based experiential learning* (pp. 235–248). Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-642-78539-9_17.

Veenman, M. V. J., Prins, F. J., & Verheij, J. (2003). Learning styles: Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology, 73*, 357–372.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*(1), 89–109. https://doi.org/10.1016/j.learninstruc.2003.10.004.

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

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences, 29*, 123–130. https://doi.org/10.1016/j.lindif.2013.01.003.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36* (3), 1–48. https://doi.org/10.1103/PhysRevB.91.121108.

Walker, D. A. (2003). JMASM9: Converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical Methods, 2*(2), 525–530. https://doi.org/10.22237/jmasm/1067646360.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a Knowledge Base for school learning. *Review of Educational Research, 63*(3), 249–294. https://doi.org/10.3102/00346543063003249.

Wells, A. (2011). *Metacognitive therapy for anxiety and depression*. New York: Guilford Press.

Winne, P. H., & Jamieson-Noel, D. (2002). Exploring students' calibration of self reports about study tactics and achievement. *Contemporary Educational Psychology, 27*(4), 551–572. https://doi.org/10.1016/S0361-476X(02)00006-1.

Yildiz, E., Akpinar, E., Tatar, N., & Ergın, Ö. (2009). Exploratory and confirmatory factor analysis of the metacognition scale for primary school students. *İlköğretim Öğrencileri İçin Geliştirilen Biliş Üstü Ölçeği'nin Açımlayıcı ve Doğrulayıcı Faktör Analizi, 9*(3), 1591–1604.

Affiliations

Kym Craig[1] · Daniel Hale[1] · Catherine Grainger[2] · Mary E. Stewart[1]

Daniel Hale
d.hale@hw.ac.uk

Catherine Grainger
catherine.grainger@stir.ac.uk

Mary E. Stewart
M.E.Stewart@hw.ac.uk

[1] Psychology, School of Social Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

[2] University of Stirling, Airthrey Road, Stirling FK9 4LA, UK