Mind the Gap between Comprehension and Metacomprehension: Meta-Analysis of Metacomprehension Accuracy and Intervention Effectiveness

Chunliang Yang[1,2], Wenbo Zhao[3], Bo Yuan[4], Liang Luo[1,3], David R. Shanks[5]

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, China.

[2] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, China.

[3] Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, China.

[4] Department of Psychology, Ningbo University, Ningbo, China.

[5] Division of Psychology and Language Sciences, University College London, London, the UK.

Author Note

All data and analysis scripts are publicly available via the Open Science Framework (OSF) at https://osf.io/ed4ac/. Correspondence concerning this article should be addressed to Liang Luo (luoliang@bnu.edu.cn), Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, 19 Xinjiekouwai Street, Haidian District, Beijing 100875, China.

Acknowledgement

This research was supported by the Natural Science Foundation of China (32000742; 32171045), the Fundamental Research Funds for the Central Universities (2019NTSS28), and the UK Economic and Social Research Council (ES/S014616/1).

Abstract

Research has consistently demonstrated that learners are strikingly poor at metacognitively monitoring their learning and comprehension of texts. The current meta-analysis aims to explore three important questions about metacomprehension: (1) To what extent can people accurately discriminate well-learned texts from less well-learned ones? (2) What are the (meta)cognitive causes of poor metacomprehension accuracy? (3) What interventions improve the accuracy of metacomprehension judgments? In total, the meta-analysis integrated 502 effects and data from 15,889 participants across 115 studies to assess these questions. The results showed a weighted *MC* = 0.178 for non-intervention effects. Many interventions were shown to be effective, such as delayed-summary-writing and delayed-keyword-generation. In addition, combining different interventions tended to generate additive benefits. The findings support the transfer-appropriate monitoring account, the situation model framework, and the poor-comprehension theory as explanations for why metacomprehension accuracy is typically poor. Practical implications are discussed.

Keywords: metacomprehension accuracy; intervention effectiveness; situation model; transfer-appropriate monitoring; poor comprehension

Learning from texts plays a fundamental role in education and learning, and reading is one of the primary approaches through which people gain knowledge (Garner, 1987; Otero & Graesser, 2014; Rayner, Pollatsek, Ashby, & Clifton Jr, 2012). Frequently learners need to read multiple texts during a limited period, such as in a school or college class (Griffin, Mielicki, & Wiley, 2019). To regulate learning activities (i.e., when, what, and how to study) in an optimal way, learners need to accurately monitor their ongoing comprehension status and discriminate fully mastered texts from those requiring further study (Thiede, Anderson, & Therriault, 2003). Monitoring the quality of one's text learning is referred to as *metacomprehension* (Maki & Berry, 1984), a key component of metacognition (Flavell, 1976).

Over the last four decades, hundreds of studies have been conducted to assess to what extent people can accurately monitor their learning and understanding of text materials (for reviews, see Dunlosky & Lipko, 2007; Maki & McGuire, 2002; Prinz, Golke, & Wittwer, 2020a; Thiede, Griffin, Wiley, & Redford, 2009; Zhao & Linderholm, 2008). These studies include three key component parts: participants first study multiple texts, then make a judgment about their comprehension of each text, and finally take a criterion test on these texts, typically probing retention and comprehension of their content.

Two measures have been widely employed to quantify metacomprehension accuracy. The first is *absolute accuracy*, which is calculated as the signed difference between perceived (measured as subjective comprehension judgments) and actual comprehension (measured as objective test performance). Positive and negative deviations between perceived and actual comprehension are referred to as overconfidence and underconfidence, respectively. Overconfidence is often referred to as confidence bias or the illusion of knowing (Koriat & Bjork, 2005). The magnitude of the deviation reflects the extent to which comprehension judgments are biased in one direction. The other popular measure is *relative accuracy*, which is typically quantified as the intra-individual correlation between comprehension judgments and test performance across different texts. Relative accuracy indexes to what extent a given individual accurately discriminates well comprehended texts from poorly comprehended ones.

Absolute accuracy is largely dependent on test performance, with people often being overconfident when their test performance is poor and underconfident when their performance is good (Zhao & Linderholm, 2008). Hence, absolute accuracy is thought to be influenced by non-

metacognitive factors (Linderholm, Zhao, Therriault, & Cordell-McNulty, 2008), such as test difficulty, exposure duration, study-test interval, level of prior knowledge, and so on. Hence absolute accuracy may not adequately reflect a given individual's metacomprehension ability (Nelson, 1984). By contrast, relative accuracy has been shown to be largely immune to these non-metacognitive factors (for discussion, see Jaeger, 2012; Jaeger & Wiley, 2014). Therefore, Nelson (1984) and many other researchers (e.g., Jaeger, 2012; Jaeger & Wiley, 2014) have recommended that relative accuracy, instead of absolute accuracy, should be adopted as the standard measure of metacomprehension. Accordingly, the target measure employed in the current meta-analysis is relative accuracy. For the sake of brevity, below we use "metacomprehension accuracy" to refer to measures of the relative accuracy of metacomprehension (Thiede & Anderson, 2003). We comment on the important role of absolute metacomprehension accuracy in the General Discussion.

The classic research paradigm for assessing metacomprehension accuracy was developed by Maki and Berry (1984) and Glenberg and Epstein (1985), who conducted the earliest studies on metacomprehension (Eakin & Moss, 2018). In this task design, participants are instructed to read multiple texts. Immediately following reading each text or after reading all texts, they make judgments about their learning or understanding, or predict their future test performance for each text. These judgments or predictions are typically made on a Likert scale, such as a scale ranging from 1 (*I did not understand the text at all*) to 7 (*I understood the text very well*). After that, they complete a test on all texts. To quantify metacomprehension accuracy, intra-individual correlations between judgments and test performance are calculated across texts for each participant, which are then averaged across participants to generate a group average representing overall metacomprehension accuracy, as recommended by Nelson (1984).[1]

Importance of accurate metacomprehension

Thiede et al. (2003) demonstrated why accurate metacomprehension is important for text learning. In this study, Thiede et al. instructed three groups of participants to study six expository texts. After reading all texts, a no-keyword (control) group made a comprehension judgment for each

[1] It is problematic to directly average raw correlation scores across participants (Fisher, 1915), and a more suitable approach is to apply Fisher's *Z* transformation to aggregate correlation coefficients (Silver & Dunlap, 1987). However, to our knowledge, no previous metacomprehension studies applied this *Z* transformation when averaging correlations. Without access to their raw correlation scores, the current meta-analysis is unable to utilize the *Z* transformation method to calculate the mean and variance of correlations.

text. An immediate-keyword-generation group performed the same learning and judgment tasks, except that, immediately after reading each text, they were required to generate five keywords to capture the gist of the text. A delayed-keyword-generation group also generated five keywords for each text, but the keywords were generated after a delay (specifically, after all texts were read). Then all three groups undertook a first comprehension test on all texts. Following this test, participants were offered an opportunity to select some texts for restudy, restudied the selected texts, and then completed a second test on all texts.

Thiede et al. observed that metacomprehension judgments were much more accurate in the delayed-keyword-generation group (intra-individual Gamma (*G*) correlation between judgments and test performance in the first test = 0.70) than those in the immediate-keyword-generation (*G* = 0.29) and no-keyword (*G* = 0.37) groups. More importantly, participants in the delayed-keyword-generation group were more likely to select the objectively less-well comprehended texts to restudy (correlation between test performance in the first test and restudy choices *G* = -0.79) than those in the immediate-keyword-generation (*G* = -0.35) and no-keyword (*G* = -0.36) groups. Because the delayed-keyword-generation group regulated their restudy decisions more effectively, their test performance on the second test was better than those in the other two groups, even though there was no difference in test performance among the three groups in the first test.

Thiede et al.'s findings clearly demonstrate that accurate metacomprehension is related to efficient regulation of study activities, which in turn produces superior learning outcomes. Many subsequent studies have observed similar findings in different languages and populations (e.g., Ackerman & Goldsmith, 2011; Q. S. Chen, 2008, 2009; de Bruin, Thiede, Camp, & Redford, 2011; Engelen, Camp, van de Pol, & Anique, 2018; Little & McDaniel, 2015; Ni, 2019; Shiu & Chen, 2013; Thiede & Anderson, 2003; Thiede, Redford, Wiley, & Griffin, 2012; Thiede, Redford, Wiley, & Griffin, 2017; Xu & Shi, 2008).

These findings are consistent with *discrepancy-reduction* models of self-regulated learning (Dunlosky & Hertzog, 1997; Nelson, Dunlosky, Graf, & Narens, 1994; Verkoeijen, Rikers, & Schmidt, 2005). These models hypothesize that, before studying, learners set a learning goal, and during study they continuously monitor their ongoing learning progress. When the perceived learning level reaches their desired goal, they terminate encoding. Otherwise, further efforts are expended to reduce the gap between perceived and desired learning level (Little & McDaniel, 2015). For instance,

before reading a text, a reader may set a target of correctly answering about 90% of questions in a later test on this text. During reading, the reader monitors her comprehension status, and continues studying the text until she thinks that she will be able to correctly answer this number of questions in the later assessment. In brief, discrepancy-reduction models propose a close linkage between metacognition and learning: metacognitive monitoring affects metacognitive control, which in turn influences learning gains (Thiede et al., 2009).

Overall, both empirical findings and theoretical models suggest that accurate metacomprehension is critical for text learning.

Poor metacomprehension accuracy

Given the critical role of accurate metacomprehension in text learning, numerous studies have asked to what extent individuals can accurately gauge how well they understand what they read (Maki & Berry, 1984). Below we briefly summarize empirical results on metacomprehension accuracy. Before continuing, it is worth noting that previous studies calculated different types of intra-individual correlations to measure metacomprehension accuracy, such as the most widely-used Gamma (*G*) correlation (Griffin, Wiley, & Thiede, 2019; Lin, Zabrucky, & Moore, 2002), the less widely-used Pearson (*r*) correlation (Jaeger & Wiley, 2014; Sarmento, 2018), and the infrequently-used point-biserial (r_{pb}) correlation (Glenberg & Epstein, 1985) (for a comparison of these correlation measures, see Nelson, 1984). Hence, to avoid any potential misunderstanding that all previous studies conducted Pearson *r* correlation analyses, the current review uses the term *mean correlation* (henceforth *MC*), rather than *r* as Prinz et al. (2020a) did, to represent the mean of intra-individual correlations, regardless of how they were calculated.

The dismaying conclusion from prior research – as already evident from the control group in Thiede et al.'s (2003) study discussed above – is that people's metacomprehension accuracy is strikingly poor (e.g., Dunlosky & Lipko, 2007; Lin & Zabrucky, 1998; Miller & Geraci, 2014; Sarac & Tarhan, 2017; Thiede et al., 2009). For instance, Maki (1998c) found that the *MC*, calculated across 25 metacomprehension studies conducted in her laboratory, was only about *MC* = 0.27. The exact same estimate was obtained by Dunlosky and Lipko (2007), who averaged correlations across 36 published experiments from Dunlosky's laboratory.

In a narrative literature review, Thiede et al. (2009) averaged intra-individual correlations across 57 studies published before 2009, which also yielded a correlation value of *MC* = 0.27. More recently,

Prinz et al. (2020a) conducted a meta-analysis aggregating 145 correlations extracted from 66 studies, finding a weighted *MC* = 0.24. Readers are warned to be cautious when interpreting these values because the current meta-analysis documents a smaller estimate of metacomprehension accuracy under normal non-intervention conditions.

What is the practical meaning of a correlation at 0.27? To facilitate readers' interpretation of this correlation value, it can be translated into judgment accuracy, which represents the probability that a given individual correctly makes high comprehension judgments to well-comprehended texts and low judgments to less-well comprehended ones. To do this, it is necessary to estimate the relationship between correlations and judgment accuracy values. A correlation-accuracy simulation was performed to achieve this, and details of the simulation method are available in the Supplemental Information (SI) file.

The simulation showed that the peak of the density distribution of judgment accuracy values corresponding to $0.26 < G < 0.28$ was at 0.569. This means that when a given individual's metacomprehension accuracy is at $G = 0.27$, the probability that she will correctly offer high comprehension judgments to well-comprehended texts and low judgments to less-well comprehended ones is only about 56.9%, against a chance level of 50% (i.e., $G = 0$). If she makes restudy decisions solely according to her comprehension judgments, the accuracy of these decisions (i.e., correctly deciding to restudy less-well comprehended texts in preference to well-comprehended ones) will only be 56.9%, barely better than a coin toss. Stated differently, if she studied 40 textbook sections in a course, and then wanted to restudy the 20 she had comprehended most poorly, she would be expected to select about 11 of the 20 poorly-comprehended ones together with 9 of the 20 well-comprehended ones, rather than the 20 poorly-comprehended sections. Her restudy strategy regulation would be extremely inefficient.

In summary, even though accurate metacomprehension is critical for text learning, learners' ability to monitor their understanding is far from impressive, and they are generally poor at metacognitively evaluating how well they understand what they read. Such poor accuracy is likely to have consequences not only for learning in the school and college classroom but in many other contexts as well, such as doctors' awareness of their levels of understanding of information about the efficacy and side-effects of new medicines (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007).

Putative mechanisms underlying poor metacomprehension accuracy

An important goal of previous metacomprehension studies was to understand the theoretical bases of poor accuracy (Wiley, Griffin, & Thiede, 2005). Understanding its underlying mechanisms is a prerequisite to developing effective interventions to improve metacomprehension accuracy and to boost learning efficiency (Thiede & Anderson, 2003). Indeed, a variety of explanations have been proposed. Below we introduce three popular accounts, which are empirically tested in the current meta-analysis.

Transfer-appropriate monitoring

The *transfer-appropriate monitoring* (TAM) account assumes that the key determinant of judgment accuracy is the extent to which the contexts in which judgments are made match those in the tests (Dunlosky & Nelson, 1997; Dunlosky, Rawson, & Middleton, 2005). According to TAM, poor monitoring accuracy mainly arises from a mismatch in contexts (or processes) between the judgment and test conditions. For instance, learners may lack knowledge about what kind of cues they should use as a basis to form their judgments because they do not know what kind of or how text information will be tested later (Thiede, Wiley, & Griffin, 2011). If the cues employed to construct judgments mismatch those presented in the test, poor accuracy will emerge.

A major prediction of TAM is that the more similar the contexts (or cues) between judgment and test, the more accurate the comprehension judgments will be (Dunlosky et al., 2005; Thiede et al., 2011). This prediction has been corroborated by some findings. For instance, Glenberg, Sanocki, Epstein, and Morris (1987) instructed participants to take a practice test on each text, which was either identical or dissimilar to the criterion test, before making comprehension judgments. The results showed superior monitoring accuracy when the practice test was identical to the criterion test than when it was dissimilar.

The TAM account has also been tested by manipulating test expectancy (Griffin, Wiley, et al., 2019; Thiede et al., 2011). For instance, Thiede et al. (2011) found that when participants were informed that they would undertake inference tests on the to-be-studied texts but they were in fact given content tests, their monitoring accuracy was much lower than when they were actually given inference tests. The converse was also found: judgments were less accurate when participants were informed that they would be tested on text content but the final test actually evaluated inference. These findings have recently been conceptually replicated by Griffin, Wiley, et al. (2019). Overall,

the above-discussed studies imply that test expectancy consistency (i.e., consistency between the expected and actual test) is a critical moderator of metacomprehension accuracy, in line with the TAM account.

Dunlosky and colleagues, however, have questioned the TAM account (Dunlosky, Rawson, & McDonald, 2002; Dunlosky et al., 2005). These researchers speculated that poor metacomprehension accuracy might result from a mismatch in the relative granularity between judgments and tests. Specifically, in most previous studies, participants made a *global* judgment for each text, but the final tests presented a set of questions testing *specific* pieces of information within each text (e.g., Maki & Serra, 1992b). Even though readers might accurately monitor which pieces of information were mastered better than others, it was challenging for them to translate their within-text monitoring into a global judgment for the entire text. Therefore, Dunlosky and colleagues speculated that item-specific judgments (i.e., judgments related to specific pieces of information within the text) should, according to TAM, be more accurate than global judgments (i.e., judgments related to the entire text).

To test this assumption, Dunlosky, Rawson, and Hacker (2002) instructed participants to study multiple expository texts, with each one giving definitions for four concepts. After reading each text, participants first made a global judgment (*How well will you be able to complete a test over this material?*) and then made an item-specific judgment for each of the four terms (*How well do you think you will be able to define XXX?*). Then they undertook a cued recall test on all four definitions (i.e., recalling definitions when prompted with concept terms). Strikingly, Dunlosky et al.'s results contradicted the TAM prediction by showing no statistically detectable difference in accuracy between item-specific and global judgments (for related findings, see Dunlosky et al., 2005). However, it is worth noting that other studies observed that item-specific judgments were more accurate than global ones (e.g., Han, 2010). In brief, the above-discussed studies, which assessed the effect of granularity match on metacomprehension accuracy, obtained inconsistent findings. This casts some doubt on the adequacy of the TAM account.

In summary, the validity of the TAM account is still under debate, and further tests are called for. Aggregating all available data may allow a clearer verdict on the TAM account.

Situation-model approach to metacomprehension

Researchers have attempted to combine theories of comprehension with theories of metacognition to jointly account for poor metacomprehension accuracy (Dunlosky & Lipko, 2007;

Griffin, Mielicki, et al., 2019; Rawson, Dunlosky, & Thiede, 2000). According to the *cue-utilization* model of metacognitive monitoring (Koriat, 1997), people search for a variety of cues to make metacognitive judgments, and these cues are either predictive or misleading (Thiede, Griffin, Wiley, & Anderson, 2010; Yang, Huang, & Shanks, 2018; Yang, Sun, & Shanks, 2018). The reason why comprehension judgment accuracy is generally poor may be that learners frequently make judgments based on cues (e.g., processing fluency) which are not diagnostic of comprehension performance (Thiede & Anderson, 2003).

According to Kintsch's *construction-integration* framework, text information is mentally represented at three levels: a lexical (surface) level, a text-base level, and a situation model level (Kintsch & Walter, 1998). The lexical level simply represents the surface features of the text, such as the exact word-forms and syntax presented in the text, which can be constructed without requiring any comprehension of the text (Kintsch & Walter, 1998). The text-base level goes slightly deeper and comprises parsing surface text segments into individual propositions or idea units. Importantly, the text-level representation only involves abstracting from the exact words and translating them into propositional forms, which requires minimal inference processes. The highest level of text representation is the situation model representation (i.e., mental representation of the situations described in the text). To construct a situation model, readers need to integrate different propositions into a coherent whole and connect them with their prior knowledge (Zwaan & Radvansky, 1998).

According to the construction-integration framework, comprehension performance is largely determined by the quality of the situation model representation because (1) comprehension tests are typically administered after a delay (instead of immediately following reading), and (2) surface features fade quickly from memory whereas the situation model representation is relatively more resistant to forgetting (Jaeger, 2012; Thiede, Dunlosky, Griffin, & Wiley, 2005). Accordingly, accuracy of comprehension judgments should be dependent on the extent to which the cues that readers utilize to construct their judgments appropriately reflect the quality of their situation model representation (Jaeger, 2012).

For the sake of brevity and following precedents (e.g., Wiley et al., 2005), we refer to the combination of theories of metacognitive monitoring and theories of comprehension as the *situation-model approach to metacomprehension* (SMAM). Based on SMAM, a variety of interventions have been developed to enhance metacomprehension accuracy, and the effectiveness of these interventions

has also been taken as a means to test the SMAM framework (Wiley, Thiede, & Griffin, 2007). The most widely-studied interventions include keyword-generation, summary-writing, self-explaining, and a variety of other interventions which involve concept organization and knowledge integration processes (e.g., concept-mapping).

When keywords and summaries are generated immediately after reading, they will mainly be extracted from the lexical level representation because these features are highly accessible immediately after reading (Thiede & Anderson, 2003). However, these features are not reflective of the situation model representation. Consequently, surface cues, activated by the requirement to generate immediate keywords or to write immediate summaries, should not be highly predictive of subsequent test performance, because tests on studied texts are typically administered after a delay (Thiede et al., 2003; Thiede et al., 2005). By contrast, keywords and summaries generated after a delay will mainly be extracted from the situation model representation. Hence, the SMAM framework predicts that delayed-keyword-generation and delayed-summary-writing should more effectively enhance accuracy of comprehension judgments than immediate-keyword-generation and immediate-summary-writing (Thiede et al., 2005).

This prediction has been verified by many studies reporting superior monitoring accuracy in delayed- compared to immediate- or no-keyword-generation conditions (Q. S. Chen & Li, 2008; Thiede et al., 2003; Zhang, Sun, & Li, 2011). Along the same lines, delayed-summary-writing has been shown to be more beneficial than immediate- and no-summary writing (Anderson & Thiede, 2008; Thiede & Anderson, 2003; Thiede et al., 2010; Xu & Shi, 2008).

Self-explaining is hypothesized to force readers to focus on their situation model representation because generating self-explanations (e.g., generating explanations to oneself about the logical and causal relations among idea units within the text) requires readers to connect different propositions, which in turn activates cues related to the mental model corresponding to the situations described in the text. Therefore, instructing readers to generate self-explanations is expected to increase the saliency of cues related to the situation model representation and in turn improve judgment accuracy. Indeed, many studies have observed that self-explaining can enhance comprehension judgment accuracy (e.g., Fukaya, 2013; Griffin, Wiley, & Thiede, 2008; Griffin, Wiley, et al., 2019; Jaeger, 2012; Ni & Xu, 2019).

Based on SMAM, other interventions, which involve cognitive processes of concept organization and knowledge integration, have also been developed to improve metacomprehension accuracy. These include concept-mapping, concept-diagram-drawing, and mind-mapping (Mi, 2020; Redford, Thiede, Wiley, & Griffin, 2012; van de Pol, de Bruin, van Loon, & van Gog, 2019; van Loon, de Bruin, van Gog, van Merrienboer, & Dunlosky, 2014). For instance, Redford et al. (2012) found that, for 7th grade students who were required to construct concept maps before making comprehension judgments, their judgments were more accurate than those of other students who were not required to construct concept maps (for related findings, see Thiede et al., 2010).

Even though all of these findings support the SMAM framework, there are others which do not support or indeed even challenge it. For instance, Thule (2005) failed to replicate the delayed-keyword-generation effect, finding equivalent levels of metacomprehension accuracy among the delayed-, immediate-, and no-keyword-generation groups. Jaeger (2012) found that instructing participants to generate self-explanations failed to enhance metacomprehension accuracy. These inconsistent findings highlight the potential value of a meta-analysis to resolve these uncertainties through integrating results across studies to increase statistical power and permit potential moderators to be evaluated.

Poor-comprehension

Dunlosky and Lipko (2007) noted that "people will have difficulties in monitoring their learning and comprehension if they have minimal understanding of the text in the first place". We term this explanation the *poor-comprehension* theory. Somewhat like a floor effect, it assumes that poor comprehension provides few valid cues to inform judgments and hence leads to poor metacomprehension accuracy (for related discussion, see Maki & Serra, 1992a; Weaver & Bryant, 1995).

A major prediction of this theory is that the better the comprehension, the more accurate the metacomprehension judgments should be. Consistent with this prediction, many studies have reported a positive relationship between test performance (a measure of text comprehension) and metacomprehension accuracy (Commander, Zhao, Li, Zabrucky, & Agler, 2014; Griffin et al., 2008; Maki, 1998a; Maki & Berry, 1984; Maki, Jonas, & Kallod, 1994; Ni, 2019; Pilegard & Mayer, 2015; Rawson, Dunlosky, & McDonald, 2002; Zabrucky, Agler, & Moore, 2009). For instance, Maki and

Serra (1992a, p. 6) observed that "As Ss [subjects] gained more information about texts, the correlations between predictions and performance increased".

Individual differences findings also provide indirect support for the poor-comprehension theory. For instance, previous studies found that individuals with high working memory capacity (S. Chen, 2010; Chiang, 2007; Chiang, Therriault, & Franks, 2010; Ni, 2019), high reading ability (Ozuru, Kurby, & McNamara, 2012), and high comprehension ability (Griffin et al., 2008) made more accurate comprehension judgments, and these individuals also performed better on criterion tests. These findings jointly imply a positive relationship between comprehension and metacomprhension accuracy.

It is, however, worth noting that there are other findings that challenge this theory. For instance, Prinz et al.'s (2020a) recent meta-analysis found no relationship between text difficulty and metacomprehension accuracy: more difficult texts (associated with poorer comprehension) did not lead to poorer monitoring accuracy. In addition, Lin, Moore, and Zabrucky (2001) observed that metacomprehension accuracy did not correlate with test performance, implying that good comprehenders are not necessarily good monitors. Furthermore, unlike other individual differences studies (e.g., Ni, 2019), Thule (2005) documented no relationship between working memory capacity and metacomprehension accuracy. In brief, whether poor comprehension is responsible for poor metacomprehension accuracy remains unclear.

In summary, three popular frameworks (TAM, SMAM, and poor-comprehension) have been proposed to explain why metacomprehension accuracy is typically poor, and while instructive, previous tests of their predictions have been inconclusive or even conflicting. A comprehensive meta-analysis is required to directly evaluate these theories.

Effectiveness of different interventions

Given that metacomprehension plays a critical role in text learning and that people are generally not very proficient at monitoring their comprehension, failing to provide remedies is likely to lead to poor metacomprehension accuracy, inefficient regulation of study activities, and poor learning outcomes (Thiede et al., 2003). Therefore, many studies have sought to develop effective interventions to improve metacomprehension accuracy.

Table 1 summarizes 17 categories of interventions implemented in previous studies, including interventions (e.g., delayed-keyword-generation) discussed in the previous section. Evaluating the

effectiveness of these interventions bears both practical and theoretical importance. From a practical perspective, as discussed above, identifying effective interventions for enhancing metacomprehension accuracy is a promising avenue to boost text learning competence. From a theoretical perspective, because many interventions were motivated by different theoretical accounts (see the above discussion for details), assessing their effectiveness can also help to test the validity of these accounts.

Another reason why a meta-analysis is needed is that, as discussed above, previous research findings regarding the effectiveness of different interventions are inconsistent, which is unhelpful for practitioners. In addition, it is reasonable to assume that not all interventions are equally effective to boost monitoring accuracy, hence it is also important to determine which ones are most effective (and under what circumstances). This important question has been underexplored in previous research, with a few exceptions (e.g., Griffin et al., 2008). Another equally (if not more) important question is whether different interventions can be combined to produce additive benefits to enhance metacomprehension accuracy, and this question has also been underexplored (Griffin, Wiley, et al., 2019).

Rationale of the current meta-analysis

Even though hundreds of metacomprehension studies have been conducted over the past four decades, only three meta-analyses have been reported. The first was conducted by Fukaya (2010), was published over ten years ago, is not available in English, and only included 39 studies. Two more recent meta-analyses were undertaken by Prinz et al. (2020a, b), including more recent studies and larger datasets (k = 145 effects extracted from 66 studies in Prinz et al., 2020a, and k = 28 from 17 studies in Prinz et al., 2020b). The current meta-analysis integrates a much larger dataset than either of Prinz et al.'s (2020a, b). Specifically, k = 508 effects (data from 15,889 participants extracted from 115 studies) were included here. Note that the studies in Prinz et al.'s (2020a, b) meta-analyses were also included in the current one, except for a few which did not report sufficient data for computing the variances of *MC*s (e.g., Rawson et al., 2000). A possible explanation for the substantial difference in the numbers of included effects is that Prinz et al. only searched three electronic databases to identify eligible studies, whereas 30 electronic databases were searched in the current meta-analysis to ensure comprehensiveness. Such a large dataset should enable the current meta-analysis to generate more robust results and reach more reliable conclusions.

It is also noteworthy that the current meta-analysis evaluated 17 different types of interventions, compared to the 7 examined by Prinz et al. (2020b). In addition, it assessed the impact of intervention combinations, which Prinz et al. (2020b) were unable to do with their small sample. Prinz et al. (2020a) focused on moderators of metacomprehension accuracy (e.g., text difficulty/length/genre) and did not evaluate interventions. As shown below, because they included effects from both intervention and non-intervention studies, their meta-analytic estimate noted previously (*MC* = 0.24) cannot be taken as an estimate of metacomprehension accuracy under standard non-intervention conditions.

Overall, many key questions about metacomprehension accuracy and how to improve it have not been fully answered in previous meta-analyses. The current meta-analysis aims to address these questions.

Research questions and overview of the current meta-analysis

The meta-analysis aims to address three major questions concerning metacomprehension accuracy: (1) To what extent is metacomprehension accurate in standard (non-intervention) conditions? (2) What are the mechanisms responsible for poor metacomprehension accuracy? (3) What interventions (both individually and in combination) are effective in improving metacomprehension accuracy?

Metacomprehension accuracy in standard non-intervention conditions

As discussed above, metacomprehension accuracy has been estimated in a few previous reviews. Maki (1998c) and Dunlosky and Lipko (2007) reported that the average correlation calculated across studies conducted in their laboratories was *MC* = 0.27, the same figure obtained by Thiede et al. (2009) who averaged intra-individual correlations across studies published before 2009. We note that these reviews did not employ meta-analytic methods to integrate research results. In addition, as shown in Thiede et al.'s (2009, p.89) Table 1, these estimates were calculated based on results from both non-intervention and intervention studies. It is possible that *MC* = 0.27 might overestimate metacomprehension accuracy in standard non-intervention situations, because, as shown below, many interventions implemented in previous studies effectively enhanced monitoring accuracy.

The same limitation applies to Prinz et al.'s (2020a, p.7) meta-analysis, which also synthesized intervention effects. Specifically, even though Prinz et al. (2020a) endeavored to only include non-intervention effects in their meta-analysis, some of their effects in fact came from intervention conditions. For instance, Prinz et al.'s meta-analysis included studies in which participants were

explicitly informed about the nature of the upcoming tests before making their comprehension judgments, and participants' test expectancy was consistent with the criterial tests (see Prinz et al., 2020a, Table 1). Previous studies have found that inducing consistent test expectancy is an effective intervention to boost metacomprehension accuracy (e.g., Griffin, Wiley, et al., 2019; Thiede et al., 2011), and the power of the test expectancy manipulation was re-confirmed by Prinz et al.'s subsequent meta-analysis (Prinz et al., 2020b) as well as the current one.

It is not clear why Prinz et al. treated test expectancy manipulation as an intervention in one meta-analysis (see Prinz et al., 2020b, Table 1) but not another (see Prinz et al., 2020a, Table 1). Whatever the rationale, the consequence is that the weighted *MC* = 0.24 observed by Prinz et al. (2020a) is likely to overestimate metacomprehension accuracy in standard non-intervention situations. If the goal is to estimate metacomprehension accuracy under standard conditions, intervention effects must be excluded.

In summary, it remains largely unknown to what extent metacomprehension is accurate in non-intervention conditions, and the first aim of the current meta-analysis is to fill this gap.

Mechanisms underlying poor metacomprehension accuracy

As discussed above, a variety of theories have been proposed to account for why people are not proficient at monitoring their comprehension, but previous findings on this important issue are inconsistent. Hence, the second aim of the current meta-analysis is to uncover the (meta)cognitive underpinnings of poor metacomprehension accuracy.

The TAM theory hypothesizes that poor metacomprehension derives from mismatch in contexts or processes between judgments and tests (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Wang, 2015). We test the TAM theory by asking whether test expectancy consistency moderates metacomprehension accuracy. TAM predicts superior accuracy in consistent expectancy conditions (when the expected test and the criterion test are consistent) and lower accuracy in inconsistent expectancy conditions (when the expected and the criterion test mismatch).

A variety of interventions have been developed based on the SMAM framework, such as delayed-keyword-generation, delayed-summary-writing, self-explaining, and concept-mapping (Wiley et al., 2007). Given that previous studies have reported inconsistent findings about the effectiveness of these interventions, the high statistical power achieved in the current meta-analysis provides an opportunity to resolve these divergences and to test the SMAM framework.

The poor-comprehension theory assumes that weak metacomprehension accuracy results from poor comprehension, and predicts a positive correlation between metacomprehension accuracy and level of text comprehension. To test this prediction, we take test performance as an index of comprehension to determine the relationship between metacomprehension accuracy and test performance (Yang, Huang, et al., 2021).

The logic of taking test performance as an index of comprehension is straightforward: the better the level of comprehension, the superior the test performance. Indeed, in almost all previous studies, test performance was taken as the objective measure of comprehension. It should be acknowledged that test performance is an imperfect measure of comprehension, because, besides comprehension, other factors which varied across studies might also affect test performance, such as test format and study-test interval. Below we explain how these issues are mitigated in the current meta-analysis.

Effectiveness of different interventions

As discussed above, identifying effective interventions to enhance metacomprehension accuracy is critical for enhancing text learning. It is hence important to determine which interventions are effective and which are not, and of those that are effective, what their relative degree of enhancement is. Furthermore, whether different interventions can be combined to produce additive benefits needs to be determined. To investigate these questions, a multilevel multivariate random-effects meta-regression analysis was performed to measure and compare the effectiveness of each intervention.

Method

Literature search

To obtain a comprehensive set of eligible studies, we conducted a systematic search in 30 electronic databases, including Web of Science, ProQuest (composed of 26 databases, including PsychArticles, PsychInfo, Psychology Database, Education Database, ProQuest Dissertations & Theses Global Database, Ebook Central, Business Market Research Collection, and others), China National Knowledge Infrastructure (CNKI), Wanfang Database, and Google Scholar. The search terms were [metacomprehension OR meta-comprehension OR judgment* of comprehension OR comprehension judgment* OR calibration of comprehension]. In addition, the reference lists of 21 review articles, identified in the search process, were screened for additional studies (e.g., Dunlosky & Lipko, 2007; Griffin, Mielicki, et al., 2019; Lin & Zabrucky, 1998; Prinz et al., 2020a; Thiede et al., 2009; Wiley et al., 2005; Yan & Huo, 2013).

Inclusion and exclusion criteria

1. Because the main focus of the current meta-analytic review is metacomprehension accuracy of text learning, only studies employing text materials as their principal stimuli were included. Studies which measured metacomprehension accuracy of other types of materials (such as lecture videos, chess endgames, and single sentences) were excluded (e.g., de Bruin, Rikers, & Schmidt, 2007). A few studies inserted decorative or supportive images/pictures into texts to explore their effects on metacomprehension accuracy. Such studies or effects were excluded in order to maintain a focus on metacomprehension accuracy for plain texts.
2. Only empirical studies reporting intra-individual correlations were included. Studies employing other techniques such as the error detection paradigm (e.g., Zabrucky & Moore, 1994) were excluded. In addition, those only reporting absolute but not relative accuracy were excluded (e.g., Lauterman & Ackerman, 2014).
3. Only studies measuring accuracy of prospective judgments were included. Studies which measured accuracy of retrospective judgments (e.g., confidence judgments about answer correctness) were excluded.
4. Studies reporting insufficient data for effect size calculation were excluded (e.g., Ackerman & Goldsmith, 2011; Rawson et al., 2000).
5. Studies employing participants with neurological diseases or physical disabilities (e.g., hearing impairments) were excluded. It should be noted that the meta-analysis included studies from four age groups: elementary children, secondary school adolescents, young adults, and older adults. We did not limit the samples to young adults because including a larger set of studies permitted the current meta-analysis to achieve greater statistical power and more reliable results. In addition, we included participant sample as a control variable to mitigate potential confounding effects (see the SI for detailed results related to participant sample).
6. Most previous studies asked participants to make global judgments for entire texts, with only a few instructing participants to provide item-specific judgments (e.g., Dunlosky, Rawson, & Hacker, 2002). Given that the number of available effects for item-specific judgments was too small to generate a reliable conclusion, results of item-specific judgments were excluded in order to focus on standard global judgments.

7. Studies involving different experimental design methods were included, such as randomized control trial (RCT) studies (e.g., Thule, 2005), quasi-experimental studies (e.g., Thiede et al., 2012), within-subjects design studies (e.g., Anderson & Thiede, 2008), and non-intervention studies (e.g., Commander et al., 2014). The current meta-analysis took experimental design as a control variable to mitigate its potential confounding effects, and the corresponding results are reported in the SI.
8. Given the authors' language proficiency, only English and Chinese studies were considered.

The screening procedure and results are depicted by a flowchart in Figure 1.

Data extraction, coding and analysis methods

The first author and a research assistant independently performed data extraction and moderator coding. The research assistant was trained at the beginning of the project. All divergences were settled through discussion.

Metacomprehension accuracy was calculated as *MC*s, and their variances were calculated using the formula:

$$V_{MC} = {SE_{MC}}^2 = \frac{SD^2}{N} \quad (1)$$

where V_{MC} is the variance of the mean of intra-individual correlations, SE_{MC} is the standard error of *MC*, and *SD* is the standard deviation of the observed intra-individual correlations. If *SD* or *SE* were not reported in a given study, V_{MC} was calculated from other reported measures such as *t* and *N* (Borenstein, Hedges, Higgins, & Rothstein, 2009).

To assess intervention effectiveness, we coded the effects into 18 categories, including non-intervention effects and those from the 17 intervention categories (see Table 1 for details). We note that, for interventions which have been tested in fewer than five studies (i.e., $k \leq 5$), they were combined into a single category to boost the reliability of the effect size estimates. For instance, concept-mapping, concept-diagram-completion, concept-diagram-drawing, graph-drawing, and mind-mapping were combined into a single category because these interventions share similar mental processes (e.g., concept organization and knowledge integration), and each of them has been studied in few (≤ 5) experiments.

To assess the poor-comprehension theory, test performance corresponding to each *MC* was extracted from the original reports. Note that, for $k = 70$ effects, test performance was not reported,

making it difficult to include them in a multilevel multivariate random-effects meta-regression analysis. To solve this problem, we implemented a linear interpolation method to estimate missing test performance scores (Noor, Al Bakri Abdullah, Yahaya, & Ramli, 2015). Specifically, we first conducted a multilevel random-effects meta-regression analysis on the other $k = 432$ effects to obtain the intercept and regression slope between test performance and *MC*s, then we imputed the missing test scores for the remaining $k = 70$ effects.

We scored potential risks of bias in the included studies. According to the Cochrane Risk of Bias (RoB) tool 2 (J. A. C. Sterne et al., 2019) and the What Works Clearinghouse (WWC) Study Review guideline (Version 4.1), we coded three bias variables: (1) bias arising from the randomization process (low *vs*. high); (2) bias due to missing outcome data, which was coded according to attrition rates: low (no attrition) *vs*. concern (overall attrition rate < 10% and differential attrition rate < 10%) *vs*. high (overall attrition rate ≥ 10% or differential attrition rate ≥ 10%); and (3) bias in selection of the reported result, which was coded according to baseline equivalence: low (balanced baseline or no need of baseline balancing) *vs*. concern (baseline equivalence information was not reported) *vs*. high (imbalanced baseline between groups in any characteristics, such as gender, age, working memory capacity, reading ability, and others).[2]

It is common that a metacomprehension study was composed of several experiments, and each experiment included several intervention and control groups (e.g., Griffin, Wiley, et al., 2019). Hence, some effects were extracted from the same studies and populations. To mitigate the influence of dependencies among effects, all meta-analyses were performed using multilevel random-effects models via the R *metafor* package (Pastor & Lazowski, 2018; Van Den Noortgate & Onghena, 2003), except where stated otherwise.

Results

In total, 115 studies were identified as eligible (marked with an asterisk in the References). From these studies, $k = 502$ effects based on data from 15,889 participants were extracted.

The Results section is organized as follows. We first report results on the extent to which metacomprehension is accurate, especially its accuracy under standard non-intervention conditions.

[2] Because most (if not all) included studies did not report whether their participants and/or examiners were blinded to the research aims, the current meta-analysis is unable to measure bias due to deviations from intended interventions and bias in measurement of the outcome. Hence, these two kinds of bias were not assessed.

Then, we report the results of a multilevel multivariate random-effects meta-regression analysis to investigate the effectiveness of different interventions and to quantify the relationship between metacomprehension accuracy and test performance. In this meta-regression analysis, a variety of other variables (such as participant sample, test format, and study-test interval) were included to control their potential confounding effects. Next, we assessed risks of bias. Finally, we report the findings from six methods to determine whether the included studies reveal signals of publication bias.

Metacomprehension accuracy

A multilevel random-effects meta-analysis found that, across the $k = 502$ effects, the weighted *MC* was 0.242 [0.220, 0.265], $p < .001$, indicating that although metacomprehension is somewhat accurate, its level is far from impressive. To put this *MC* value into perspective, it implies that a 1 *SD* increase in comprehension judgments is associated with an increase of about 0.24 *SD*s in actual comprehension, and that actual comprehension only accounts for about $R^2 = 5.9\%$ of the variance in comprehension judgments. The correlation-accuracy simulation described in the Introduction showed that the accuracy value corresponding to $MC = 0.242$ is about 56.1%, which is barely better than a coin toss (50%).

Heterogeneity among the effects was substantial, $Q(501) = 3279$, $p < .001$, indicating the need to explore potential moderators of the overall effect. The weighted $MC = 0.242$ is similar to the values estimated by other reviews (e.g., Maki, 1998c), but it might overestimate normal accuracy because it was generated from a combination of non-intervention and intervention effects. In addition, as shown in Figure 2A, the density distribution of the effects is somewhat right skewed, skewness = 0.065. A possible explanation for the skewness is that the included effects came from two distinct sub-samples: intervention and non-intervention studies. Indeed, as shown in Figure 2B, intervention *MC*s tended to be larger overall than non-intervention ones.

A multilevel random-effects meta-regression analysis was conducted to determine whether non-intervention and intervention effects differed from each other and to what extent metacomprehension is accurate in standard non-intervention circumstances. The results showed that non-intervention and intervention effects were heterogeneous, $Q(1) = 259$, $p < .001$, with intervention effects ($k = 259$, $MC = 0.332$ [0.309, 0.356], $p < .001$) larger than non-intervention effects ($k = 243$, $MC = 0.178$ [0.155, 0.200], $p < .001$). As shown in Figure 2B, when non-intervention and intervention effects are separated, their distributions are approximately symmetric.

Of considerable importance, the weighted *MC* = 0.178 for standard non-intervention effects is smaller than the estimates reported in previous reviews (e.g., 0.24 reported by Prinz et al., 2020a), as reflected by the upper bound (0.200) of its 95% CI. It equates to accuracy of 54.5% according to the simulation described previously. Such poor accuracy in standard conditions further highlights the necessity for exploring its (meta)cognitive causes and identifying effective interventions.

Multilevel multivariate random-effects meta-regression analysis

A multilevel multivariate random-effects meta-regression analysis was performed to assess the effectiveness of different interventions and to determine the relationship between comprehension performance and metacomprehension accuracy. In this analysis, participant sample (e.g., elementary children, middle school students, high school students, young adults, and older adults), test format (e.g., recall tests, recognition tests), study-test interval (e.g., no interval, interval), correlation type (e.g., Gamma correlation, Pearson *r* correlation, point-biserial correlation), and design quality (e.g., quasi-experimental, RCT, within-subjects, non-interventional design) were included to control their potential confounding effects (see the SI for details about the coding procedures of these variables). Because these variables were not the focus of our research interest and for the sake of brevity, their corresponding results are reported in the SI.

Besides these controlled variables, the meta-regression analysis also included bias from missing outcome data as a control variable. As we will show below, this variable also accounted for some of the heterogeneity among the effects.

The multilevel multivariate random-effects meta-regression analysis showed that non-intervention effects and those from the 17 intervention categories were heterogeneous, $Q(17) = 430$, $p < .001$. Table 2 lists the difference in *MC* between each intervention category and non-intervention effects. The differences in *MC*s between intervention and non-intervention effects represent the influence of each intervention on metacomprehension accuracy, with a positive value representing an enhancing effect and a negative value indicating a detrimental effect.

Test expectancy manipulation

We first report results about the effectiveness of test expectancy manipulations. Recall that the TAM theory predicts more accurate monitoring when the expected and actual tests are consistent than when they are inconsistent or when participants have no specific expectancy about the type of test in standard non-intervention conditions.

Consistent with these predictions, the meta-analysis found greater accuracy when test expectancy was consistent than in non-intervention conditions overall (in which no test expectancy manipulation was implemented), difference in *MC* = 0.090, *Z* = 2.67, *p* = .008. In addition, inconsistent test expectancy effects were associated with lower monitoring accuracy than in non-intervention conditions overall, difference in *MC* = -0.117, *Z* = -2.34, *p* = .019. Furthermore, monitoring accuracy was better when test expectancy was consistent with the criterion test compared to when it was inconsistent, difference in *MC* = 0.207, *Z* = 4.38, *p* < .001. These findings jointly support the TAM theory.

Keyword interventions

Studies have implemented three kinds of keyword interventions: delayed-keyword-generation (generating keywords after reading all texts), immediate-keyword-generation (generating keywords immediately after reading each text), and keyword-reading (reading keywords provided by the experimenter). Keyword-reading was typically implemented as a comparison intervention to determine whether active keyword generation is necessary to boost monitoring accuracy (e.g., Thiede et al., 2005).

The meta-analysis found that delayed-keyword-generation produced a statistically detectable enhancement compared to non-intervention effects, difference in *MC* = 0.184, *Z* = 9.87, *p* < .001. By contrast, immediate-keyword-generation failed to enhance metacomprehension accuracy relative to non-intervention effects, difference in *MC* = 0.009, *Z* = 0.37, *p* = .710. Surprisingly, keyword-reading seemed to produce poorer accuracy over non-intervention effects, difference in *MC* = -0.096, *Z* = -1.96, *p* = .050.[3]

Importantly, metacomprehension judgments were more accurate in the delayed-keyword-generation condition than in the immediate-keyword-generation condition, difference in *MC* = 0.175, *Z* = 7.38, *p* < .001, supporting the main proposal of the SMAM framework. Delayed-keyword-generation also produced superior monitoring accuracy than keyword-reading, difference in *MC* = 0.280, *Z* = 5.74, *p* < .001, confirming that active keyword generation is required to improve metacomprehension accuracy, compared to passive reading of experimenter-provided ones.

[3] There is little reason to expect a detrimental effect of keyword-reading on metacomprehension accuracy. This marginal result might be due to sampling error. Because this result is not strong and the number of effects for keyword-reading was small, we do not discuss it further.

Summary interventions

Studies have implemented summary-writing interventions at different timepoints: immediate-summary-writing (writing a summary immediately after reading each text) *vs*. delayed-summary-writing (writing a summary for each text after reading all texts). According to the SMAM framework, delayed- should be more beneficial than immediate-summary-writing.

Indeed, the meta-analysis found that delayed-summary-writing produced a statistically detectable enhancement over non-intervention effects, difference in $MC = 0.210$, $Z = 9.00$, $p < .001$. By contrast, the enhancement effect for immediate-summary-writing over non-intervention effects was not statistically detectable, difference in $MC = 0.030$, $Z = 1.21$, $p = .225$. More importantly, metacomprehension judgments were more accurate in the delayed- than in the immediate-summary-writing condition, difference in $MC = 0.179$, $Z = 6.88$, $p < .001$. Thus, delayed-summary-writing, but not immediate-summary-writing, is beneficial for improving metacomprehension accuracy, a finding in line with the SMAM framework.

Self-explaining

Glenberg and Epstein (1985) assumed that "subjects may accurately assess the number of isolated facts or propositions from the text in memory, but fail to consider relations among the propositions". Studies have suggested that instructing participants to generate self-explanations (e.g., generating explanations to oneself about what new information the just-read sentence or paragraph conveys) enables them to develop a more complete situation model representation (Chi, 2000; Chi, De Leeuw, Chiu, & Lavancher, 1994). In addition, because self-explaining enhances the salience of cues related to the situation model representation, it is also expected to boost metacomprehension accuracy (Griffin et al., 2008).

Consistent with previous research (e.g., Baker, 2008; S. Chen, 2010; Fukaya, 2013; Griffin et al., 2008; Ni, 2019; Ni & Xu, 2019), the meta-analysis found that self-explaining produced more accurate comprehension judgments than the average of non-intervention effects, difference in $MC = 0.179$, $Z = 5.08$, $p < .001$, again supporting the main proposal of the SMAM framework.

Concept-mapping/concept-diagram-completion/concept-diagram-drawing/graph-drawing/mind-mapping

A variety of mapping and drawing interventions have been administered, including concept-mapping, concept-diagram-completion, concept-diagram-drawing, graph-drawing, and mind-

mapping. According to the SMAM framework, to construct a concept map or diagram, readers have to identify connections among different idea units stated in the text (Ainsworth & Th Loizou, 2003), which then activate cues related to the situation model representation and promote judgment accuracy (Redford et al., 2012). We found that these mapping and drawing interventions indeed produced superior monitoring accuracy than non-intervention effects on average, difference in $MC = 0.196$, $Z = 5.89$, $p < .001$, supporting the SMAM framework.

Rereading after a short delay

Rereading after a short delay is another widely-studied intervention. Rawson et al. (2000) proposed that during the initial reading phase, learners may mainly engage in surface or text level processing. In contrast, when they are allowed to reread a text after a short delay, more resources will be allocated to constructing a situation model to represent the text (Millis, Simon, & Tenbroek, 1998). Hence, based on the SMAM framework, Rawson et al. expected that rereading after a short delay should enhance metacomprehension accuracy. Their results supported their expectation by showing more accurate comprehension judgments when participants reread all texts after a short delay than when they were only allowed to read each text once (for related findings, see Dunlosky & Rawson, 2005; Griffin et al., 2008). However, it should be noted that Chiang et al. (2010) found no enhancement effect of rereading after a short delay on metacomprehension accuracy (for related findings, see Wang, 2015).

The current meta-analysis integrated dozens of effects for rereading after a short delay and found a small but statistically detectable enhancing effect of this intervention over non-intervention effects, difference in $MC = 0.063$, $Z = 2.56$, $p = .010$. Again, this finding is consistent with the SMAM framework.

Noteworthy is that the timepoint of rereading might moderate its effect on metacomprehension accuracy. Millis et al. (1998) found that rereading increased integration of text propositions when the interval between initial reading and rereading was short (i.e., when participants reread the texts after they read all of them), but rereading failed to enhance integration when the interval was long (i.e., when participants reread the texts one week later). Most studies administered rereading after a short delay. For instance, researchers instructed participants to reread the texts after they read each text (Jiang, 2017), after they read all texts (Rawson et al., 2000), or after they completed a brief (6 min) filler task (Serra, 2007). Only Dunlosky and Rawson (2005) included a long (1 week) interval

between initial reading and rereading, and they observed that rereading after a short delay enhanced metacomprehension accuracy while rereading after a long delay did not. [Data from Dunlosky and Rawson's (2005) long delay condition were excluded from the analysis reported above and were instead allocated to an other interventions category (see below)].

Test interventions

Undertaking a practice test before making comprehension judgments is expected to improve metacomprehension accuracy in a range of ways (Glenberg & Epstein, 1985). For instance, the experience of taking practice tests may induce awareness of the gap between perceived and actual learning, and practice test performance may provide informative feedback to bring judgments and test performance into closer alignment (Dunlosky, Rawson, & McDonald, 2002; Kelemen, Winningham, & Weaver, 2007). Inferential questions presented in practice tests may make cues related to the situation model available, in turn improving judgment accuracy (Glenberg & Epstein, 1985). In addition, practice tests may inform participants about the nature of the upcoming test (i.e., inducing an expectancy about the criterion test), and, as shown above, consistent test expectancy boosts monitoring accuracy. Consistent with these assumptions, the meta-analysis found that practice tests enhanced monitoring accuracy over non-intervention effects, difference in $MC = 0.169$, $Z = 8.61$, $p < .001$.

In addition to exploring the effects of practice tests on monitoring accuracy, some studies also included another group of participants who were provided with test questions and answers to read before making their judgments (e.g., Wang, 2015). These groups were included to determine whether simply reading test questions and answers without the requirement to actively generate answers to test questions is sufficient to enhance monitoring accuracy. The answer is negative: the difference in *MC* between reading questions and answers and non-intervention effects was not statistically detectable, difference in $MC = -0.095$, $Z = -1.85$, $p = .065$. Critically, practice tests produced a larger benefit than reading questions and answers, difference in $MC = 0.264$, $Z = 5.15$, $p < .001$. Overall, these findings imply that covertly answering practice questions, but not passively reading questions and answers, boosts metacomprehension accuracy.

Letter deletion

Maki, Foley, Kajer, Thompson, and Willert (1990) hypothesized that deeper levels of text processing should increase the relationship between judgments and test performance. Specifically,

Maki et al. (1990, p.610) suggested that "If subjects are more active in their reading, as they must be in order to figure out words with deleted letters, then they should also develop a better idea of which facts will be accessible on the test and which will not be accessible." Therefore, Maki et al. expected that reading texts with deleted letters might produce more accurate judgments than reading intact texts, because letter deletion should induce deeper processing.

Maki et al.'s results supported their prediction by showing that letter deletion improved judgment accuracy (for related findings, see Rawson & Dunlosky, 2002). However, Ikeda and Kitagami (2012) failed to detect an overall beneficial effect of letter deletion on metacomprehension accuracy. Consistent with Maki et al. (1990), the current meta-analysis observed an enhancing effect of letter deletion over non-intervention effects, difference in $MC = 0.112$, $Z = 2.43$, $p = .015$.

Question generation

Ide (2010) proposed that having readers generate questions about studied texts should induce more elaborative processing and increase the number of cues relevant to mastery of studied texts, which should in turn lead to more accurate judgments. However, their results ran counter to their expectation by showing minimal influence of question generation on monitoring accuracy. By contrast, Zeng (2009) observed that question generation enhanced monitoring accuracy.

The meta-analysis found that the effect of question generation on monitoring accuracy was not statistically detectable, difference in $MC = 0.042$, $Z = 0.83$, $p = .406$, reflecting limited value of question generation as a means of improving metacomprehension accuracy.

Analogy provision

Providing analogies can either enhance or impair metacomprehension accuracy (Wiley, Jaeger, Taylor, & Griffin, 2018). From a negative perspective, texts with analogies are typically perceived as more interesting and easier to understand than ones without analogies (Jaeger & Wiley, 2015). Hence, analogies may induce readers to rely heavily on misleading feelings of familiarity and processing fluency (i.e., ease of processing), at the cost of focusing less on their true mastery when making judgments (Jaeger & Wiley, 2010; Maki & Serra, 1992a; Rawson & Dunlosky, 2002; Serra & Dunlosky, 2010; Vössing & Stamov-Roßnagel, 2016). Therefore, analogy provision may yield poorer monitoring accuracy. By contrast, from a positive perspective, analogical examples may prompt readers to consider how the main concepts included in the text relate to those examples, and hence analogy provision may stimulate readers to identify more predictive cues relevant to their

comprehension judgments, leading to superior monitoring accuracy (Ainsworth & Th Loizou, 2003; Jaeger, 2012).

Wiley et al. (2018) recently demonstrated that analogy provision had minimal influence on metacomprehension accuracy. Their main findings were re-confirmed by the meta-analysis, which found that analogy provision did not reliably affect metacomprehension accuracy, difference in MC = 0.010, Z = 0.18, p = .855.

Intervention combinations

As discussed above, many interventions (though by no means all) are effective in enhancing metacomprehension accuracy when they are implemented individually. Some studies have also combined two or more interventions simultaneously to measure their joint influence (e.g., Griffin, Wiley, et al., 2019; Martin, Nguyen, & McDaniel, 2016). In these studies, participants in the experimental condition received more than one kind of intervention, such as rereading plus self-explaining (S. Chen, 2010), delayed-keyword-generation plus collaborative learning (Pao, 2014), test expectancy manipulation plus self-explaining (Griffin, Wiley, et al., 2019), and so on (Martin et al., 2016). These effects were classified into a single category – intervention combinations – as each of these combinations has been exposed to few investigations.

As shown in Table 2, monitoring accuracy for intervention combinations is greater than that for non-intervention effects, difference in MC = 0.316, Z = 13.25, p < .001, suggesting that combining different interventions can facilitate metacomprehension accuracy. To further explore whether combinations are more effective than individual interventions, the k = 502 effects were re-classified into three categories: control (without any intervention), single interventions (with a single kind of intervention implemented), and intervention combinations (with more than one kind of intervention implemented). A new multilevel multivariate random-effects meta-regression analysis showed statistically detectable heterogeneity among the three categories, $Q(2)$ = 202, p < .001. Critically, although both single interventions (difference in MC = 0.101, Z = 9.56, p < .001) and intervention combinations (difference in MC = 0.299, Z = 12.94, p < .001) improved judgment accuracy, intervention combinations were more effective than single interventions, difference in MC = 0.198, Z = 8.92, p < .001.

Overall, these findings imply that greater enhancement in metacomprehension accuracy can be achieved when different interventions are combined, and that combinations of interventions generate additive benefits compared with simply implementing them individually.

Other interventions

Several other interventions have been investigated in small numbers (fewer than five) of studies, such as motivation manipulation (Linderholm, Wang, Therriault, Zhao, & Jakiel, 2012), expectancy of self-explaining (Fukaya, 2013), highlighting (Gier, Kreiner, & Natz-Gonzalez, 2009), and rereading after a long delay (Dunlosky & Rawson, 2005). These were combined into an other interventions category. They did not yield a statistically detectable effect on metacomprehension accuracy over non-intervention effects, difference in *MC* = 0.023, *Z* = 1.30, *p* = .194. Readers are warned to be cautious when interpreting these results because the interventions included in the other interventions category were somewhat heterogeneous.

Comprehension level (test performance)

The results from the multilevel multivariate random-effects meta-regression analysis showed a positive relationship between test performance and metacomprehension accuracy, *b* = 0.298 [0.246, 0.350], *Z* = 11.25, *p* < .001 (see Figure 3), indicating that every increase of 10% in test performance increases *MC* by 0.030.

The above finding should be interpreted cautiously, because, as explained above, there were *k* = 70 effects for which test performance scores were imputed through linear interpolation. To mitigate this problem, another multilevel multivariate random-effects meta-regression analysis was performed, in which these *k* = 70 effects were excluded, leaving final data from *k* = 432 effects. The results again showed a reliable positive relationship between test performance and *MC*s, *b* = 0.482 [0.370, 0.595], *Z* = 8.42, *p* < .001.

It is worth noting that the *k* = 432 effects included both non-intervention (*k* = 202) and intervention (*k* = 230) effects. Previous studies demonstrated that many interventions enhanced monitoring accuracy but had little influence on test performance (e.g., Rawson et al., 2000; Thiede et al., 2012). To reduce the potential confounding effects of interventions on the relationship between comprehension and metacomprehension accuracy, we conducted another multilevel multivariate random-effects meta-regression analysis, in which only the *k* = 202 non-intervention effects were

included. This analysis again returned a positive relationship between test performance and *MC*s, b = 0.329 [0.155, 0.502], Z = 3.72, p < .001.

It should be acknowledged that test performance is an imperfect index of comprehension, likely to be affected not only by text comprehension but also other task characteristics, such as test format and the time interval between study and test. For instance, it is well-known that recall tests are more difficult and associated with lower test performance than recognition tests (Maki, Willmon, & Pietan, 2009; Yang, Luo, Vadillo, Yu, & Shanks, 2021). In addition, previous studies have found that judgment accuracy tends to be associated with test format, although they are inconsistent about the nature of this relationship. For instance, Miesner and Maki (2007) observed that metacomprehension was more accurate when the criterion test was recognition (e.g., multiple-choice) than when it was recall (e.g., short answer) (for related findings, see Maki et al., 2009). By contrast, Prinz et al.'s (2020a) recent meta-analysis found the reverse pattern: Recall tests (e.g., free recall: *MC* = 0.34; cued recall: *MC* = 0.20) were associated with greater metacomprehension accuracy than recognition tests (e.g., old/new recognition: *MC* = 0.03; true/false judgment: *MC* = 0.14).

For the k = 202 non-intervention effects, their test formats varied substantially across studies, including multiple-choice, true/false judgment, short answer, free recall, cued recall, and so on. The majority (69.3%; 140 out of 202) of these non-intervention effects administered multiple-choice tests. Hence, to mitigate the potential confounding effects of test format, another multilevel multivariate random-effects meta-regression analysis was conducted, in which only the k = 140 multiple-choice test effects were included. This analysis again returned a positive relationship between test performance and *MC*s, b = 0.472 [0.242, 0.702], Z = 4.02, p < .001.

Besides test format, another variable that relates to test performance is the study-test interval. Numerous studies have established that, due to forgetting, test performance is lower when the interval is long than when it is short (Averell & Heathcote, 2011; Renken, 2001). Furthermore, researchers have found that judgment accuracy might be affected by study-test interval. For instance, Maki and Berry (1984) and Maki (1998b) observed that comprehension judgments were more accurate when the interval was short than when it was long, although it should be noted that Prinz et al.'s (2020a) meta-analysis showed that this interval did not affect monitoring accuracy.

For the k = 140 multiple-choice test effects, k = 133 of them administered the criterion tests immediately after participants read the texts and made their comprehension judgments, and the

remaining $k = 7$ effects came from studies in which either a short (5 min) or long (1 week) study-test interval was included. To avoid any potential confounding effect of study-test interval, a new multi-level random-effects meta-regression analysis was conducted, in which only the $k = 133$ effects which administered immediate multiple-choice tests were included. Again, the positive relationship was detected, $b = 0.454$ [0.218, 0.690], $Z = 3.77$, $p < .001$.

The $k = 133$ immediate multiple-choice test effects came from different age groups: most (86.5%) employed young adults (e.g., college students) as their participants. To allay potential concern about confounding effects of populations, we conducted a final meta-regression analysis, in which only the $k = 115$ effects with young adults as participants were included. The positive relationship survived, $b = 0.500$ [0.252, 0.747], $Z = 3.96$, $p < .001$.

Overall, regardless of how we restricted the data sample, there was always a positive relationship between comprehension (test performance) and metacomprehension accuracy, which is consistent with the main proposal of the poor-comprehension theory.

Risk of bias

This section assesses potential risks of bias in the included studies. Regarding bias arising from the randomization process, there were $k = 22$ effects which employed problematic randomization procedures, such as allocating participants according to order of appearance (e.g., Thiede & Anderson, 2003), and these effects were assigned into the high category because the experimenters might have expectations based on knowledge of participant assignment. The other $k = 480$ effects were allocated into the low category. There was no statistically detectable difference in *MC* between these two categories, $Q(1) = 1.70$, $p = .192$, with numerically greater accuracy in the high category (*MC* = 0.321) than in the low category (*MC* = 0.239). Excluding intervention effects from the analysis did not alter the results, indicating little bias arising from the randomization process.

Regarding bias due to missing outcome data, heterogeneity among the low (no attrition) *vs*. concern (overall attrition rate < 10% and differential attrition rate < 10%) *vs*. high (overall attrition rate ≥ 10% or differential attrition rate ≥ 10%) categories was statistically detectable, $Q(2) = 17.57$, $p < .001$. *MC*s gradually decreased as bias due to missing outcome data increased: low (*MC* = 0.289), concern (*MC* = 0.228), and high (*MC* = 0.191). To be more cautious, a new analysis was performed, in which all intervention effects were excluded. The results showed exactly the same pattern, $Q(2) = 13.14$, $p = .001$, with *MC*s steadily decreasing across the low, concern, and high categories. Hence,

readers should be cautious about bias arising from missing outcome data when interpreting results of the current meta-analysis, although it is noteworthy that *MC*s decreased as bias increased.

Regarding bias in selection of the reported result, heterogeneity among the low (balanced baseline or no need of baseline balancing) *vs*. concern (baseline equivalence information was not reported) *vs*. high (imbalanced baseline between groups in any characteristics, such as gender, age, working memory capacity, reading ability, and others) categories was not statistically detectable, regardless of whether intervention effects were included ($Q(2) = 2.68$, $p = .262$) or excluded ($Q(2) = 3.09$, $p = .213$).

Overall, the above results imply that the included studies might be contaminated by bias from missing outcome data, while there is little evidence of bias arising from the randomization process or the selection of the reported result.

Publication bias

The above results demonstrate that non-intervention and intervention effects are heterogenous. In addition, the effectiveness of different interventions varies substantially (see Table 2). To reduce heterogeneity among the effects, below we focus on the $k = 243$ non-intervention effects to test whether the included studies suffered from publication bias. Before reporting the results, we highlight that there is currently no perfect method to assess and correct publication bias, and all existing methods have limitations (Carter, Schönbrodt, Gervais, & Hilgard, 2019).

In total, we implemented six methods to detect potential publication bias. The first is to test the moderating role of publication status (Franco, Malhotra, & Simonovits, 2014). It is well-known that studies with null results face a higher publication barrier than those with statistically significant findings. Hence, if the included effects suffered from publication bias, the magnitude of published effects should be larger than that of unpublished ones (Franco et al., 2014). The $k = 243$ non-intervention effects were separated into two categories according to their publication status: published ($k = 185$, including 171 from journal articles and 14 from book chapters) *vs*. unpublished ($k = 58$, including 57 from student dissertations and 1 from a conference report). A sub-group meta-analysis showed no statistically detectable difference in *MC* between the published (*MC* = 0.189 [0.164, 0.214], $p < .001$) and unpublished (*MC* = 0.153 [0.112, 0.195], $p < .001$) effects, $Q(1) = 2.12$, $p = .145$, reflecting little evidence of publication bias.

The second method is to determine the relationship between publication year and effect size (Borenstein & Cooper, 2009). The logic behind this approach is that, if a given effect is spurious and the large effect sizes reported in early studies simply resulted from selective publication, it is unlikely that subsequent studies would replicate the large effect sizes reported in early studies. Hence, the reported size of a spurious effect should decrease as a function of year of publication (Munafò, Matheson, & Flint, 2007). A multilevel random-effects meta-regression analysis, regressing *MC*s on publication year (ranging from 1985 to 2020), showed no statistically detectable relationship between these two variables, b = -0.001 [-0.003, 0.002], Z = -0.521, p = .603, again indicating little risk of publication bias.

The third method measures the relationship between effect sizes and their corresponding *SE*s or variances (precision-effect test and precision-effect estimate with standard errors, PET-PEESE; Gervais, 2015; Stanley & Doucouliagos, 2014; J. A. Sterne & Egger, 2005). If publication bias contaminates the included effects, the PET-PEESE method is expected to observe a positive relationship between effect sizes and their *SE*s or variances (i.e., the larger the *MC*, the larger the *SE* or variance). The reason is that it is easy for large effect sizes to be published even though they are associated with large *SE*s, variances and small sample sizes (J. A. Sterne & Egger, 2005).

Because both the PET and PEESE analyses showed a positive intercept (ps < .001), we focused on the results from the PEESE analysis (Stanley & Doucouliagos, 2014). A multilevel random-effects meta-regression analysis showed a negative relationship between *MC*s and their corresponding variances, b = -2.860 [-5.475, -0.245], p = .032. The corrected effect (i.e., the intercept of the meta-regression) was *MC* = 0.201 [0.172, 0.230], p < .001, which was slightly larger than the original weighted *MC* (i.e., 0.178).

The fourth method is Duval and Tweedie's Trim-and-Fill method (Duval, 2005; Duval & Tweedie, 2000), which gradually "trims" effects with large *SE*s until the funnel plot is symmetric, and then "fills" the removed effects and their missing counterparts to the funnel plot to maintain its symmetry. The expected missing effects should be allocated at the left-lower corner of the funnel plot, because small effects with large *SE*s are less likely to be published if publication bias exists (Hilgard, 2017).

Because, to our knowledge, the Trim-and-Fill analysis is incompatible with a multilevel random-effects meta-analysis in the R *metafor* package, we applied it to a conventional random-effects meta-

analysis. The Trim-and-Fill analysis detected $k = 42$ missing effects, which were mainly allocated at the right side of the funnel plot (see Figure 4). The corrected effect was $MC = 0.229$ [0.208, 0.249], $p < .001$.

The fifth method is to apply a three-parameter selection model (3PSM), which has been shown to be more reliable to assess and correct publication bias than many other conventional methods (McShane, Böckenholt, & Hansen, 2016; Vevea & Woods, 2005). Given that, as with Trim-and-Fill, 3PSM is incompatible with multilevel meta-analysis, we applied this analysis to a conventional random-effects meta-analysis via the R *weightr* package. 3PSM detected little evidence of publication bias, $\chi^2(1) = 0.50$, $p = .481$, and the corrected *MC* was 0.196 [0.167, 0.224], $p < .001$.

Overall, two out of five methods (i.e., PET-PEESE and Trim-and-Fill) found potential evidence of publication bias, but the corrected effect sizes generated from these two methods were if anything slightly greater than the uncorrected one. These results suggest that if publication bias really exists, the missing effects are ones with larger *MC*s.

Even though the results from PET-PEESE and Trim-and-Fill run counter to the pattern that would be expected if publication bias is present (the adjusted effects are larger, not smaller), there is a possible explanation. Many of the non-intervention effects came from studies in which these effects were taken as control comparisons to explore the effectiveness of different interventions. As shown above, intervention effects were overall larger than non-intervention effects. If monitoring accuracy in the non-intervention condition is poor, there would be more scope left for interventions to boost monitoring accuracy. Hence, a possible explanation for the publication bias detected by PET-PEESE and Trim-and-Fill is that intervention studies with poor monitoring accuracy in the non-intervention control condition might be selectively reported, making it easier to detect an enhancement effect of an intervention. To test this possibility, we conducted a multilevel random-effects sub-group meta-analysis, in which we explored whether the non-intervention effects extracted from intervention studies were smaller than those extracted from studies which did not implement any interventions – the sixth method to detect publication bias.

We divided the $k = 243$ non-intervention effects into two categories according to whether they came from an intervention study or not. For the studies which implemented at least one kind of intervention, their non-intervention control effects ($k = 112$) were allocated into the intervention study category. For the studies which did not implement any interventions, their non-intervention effects (*k*

= 131) were allocated into the non-intervention study category. The results showed a small difference in the predicted direction between intervention (*MC* = 0.169 [0.139, 0.199], $p < .001$) and non-intervention studies (*MC* = 0.190 [0.160, 0.221], $p < .001$), but their heterogeneity was not statistically detectable, difference = -0.022 [-0.065, 0.022], $Z = -0.98$, $p = .330$. This pattern thus lends little support to the hypothesis that intervention studies might be biased to report poor accuracy in the control condition.

Overall, we report six methods to assess whether the included studies suffered from publication bias. Four of them showed minimal evidence of such bias. Even though PET-PEESE and Trim-and-Fill showed weak evidence of publication bias, the results were in the reverse direction to what would be expected, with some suggestion of missing larger (instead of smaller) effects. In addition, the results showed that studies that included intervention conditions were not reliably biased to report poor accuracy in their non-intervention control conditions. Hence, there should be little need to worry about publication bias in the studies included in the meta-analysis (for related findings, see Prinz et al., 2020a).

General Discussion

This meta-analysis integrated $k = 502$ effects from 15,889 participants, derived from 115 studies, to explore (1) how accurately readers can metacognitively monitor their text comprehension, especially in standard non-intervention conditions, (2) what (meta)cognitive mechanisms are responsible for poor metacomprehension accuracy, and (3) what interventions are effective in enhancing metacomprehension accuracy. Below we briefly summarize the main findings, discuss their practical and theoretical implications, and illuminate some directions for future research.

Poor metacomprehension accuracy

The current meta-analysis observed a weighted mean of intra-individual correlations at *MC* = 0.242, which is similar to previous estimates such as 0.27 reported by Maki (1998c), Dunlosky and Lipko (2007), and Thiede et al. (2009). However, the weighted *MC* for the non-intervention effects was only 0.178 [0.155, 0.200], appreciably smaller than 0.27. These findings confirm the hypothesis that the estimates provided in some previous reviews tend to overestimate metacomprehension accuracy in standard non-intervention conditions.

The correlation-accuracy simulation showed that the peak of the density distribution of judgment accuracy values corresponding to $0.177 < G < 0.179$ was at 0.545. This means that when a given

individual's metacomprehension accuracy is at $G = 0.178$, the probability that she will correctly offer high comprehension judgments to well-comprehended texts and low judgments to less-well comprehended ones is only about 54.5%. If she makes restudy decisions solely according to her comprehension judgments, her restudy strategy regulation would be extremely inefficient.

Of course, measurement reliability issues have to be considered when interpreting the observed $MC = 0.178$, because it is well-known that the observed correlation between two variables is affected by both the true latent correlation and the reliability of measurement (Spearman, 1961; Wiernik & Dahlke, 2020). For instance, if the measures of comprehension judgments and actual comprehension are imperfect in a given study (which is always true), the observed correlation between these two variables would be attenuated to:

$$r_{obs} = \rho \times \sqrt{r_{jj}} \times \sqrt{r_{pp}}$$

where r_{obs} denotes the observed correlation (i.e., the observed metacomprehension accuracy) between comprehension judgments and actual comprehension (i.e., test performance), ρ represents the true correlation at the latent level, and r_{jj} and r_{pp} are the measurement reliabilities of comprehension judgments and comprehension performance, respectively. Thus, if r_{jj} and r_{pp} are known, the true metacomprehension accuracy can be estimated as:

$$\rho = \frac{r_{obs}}{\sqrt{r_{jj}} \times \sqrt{r_{pp}}}$$

What is the likely effect of measurement error on our estimate of the overall correlation? In two experiments, Kelemen, Frost, and Weaver (2000) measured the reliabilities of comprehension judgments and comprehension performance with a 1-week test-retest interval. In Experiment 1, the observed reliabilities for comprehension judgments and comprehension performance were $r_{jj} = 0.69$ and $r_{pp} = 0.63$, respectively, and in Experiment 2 were $r_{jj} = 0.59$ and $r_{pp} = 0.37$, respectively. We then conducted random-effects meta-analyses to collapse results across experiments, which yielded a weighted $r_{jj} = 0.638$ [0.529, 0.727], $p < .001$, and a weighted $r_{pp} = 0.509$ [0.212, 0.720], $p = .002$. By psychometric standards these are poor to moderate reliabilities, but by no means atypical of cognitive measures (Maloney, Risko, Preston, Ansari, & Fugelsang, 2010; Parsons, Kruijt, & Fox, 2019; Waechter, Nelson, Wright, Hyatt, & Oakman, 2014).

Using $r_{jj} = 0.638$ and $r_{pp} = 0.509$, we can roughly estimate that the "true" intra-individual correlation between comprehension judgments and comprehension performance is $MC_{true} = 0.312$.

Using the simulation method described previously, the estimated judgment accuracy corresponding to $G = 0.312$ is about 58.0%, still quite poor.

Theoretical implications

Three theories provided guidance for the current meta-analysis, and some suggestive implications emerge. For instance, a test-expectancy congruency effect on metacomprehension accuracy was observed. Specifically, test-expectancy manipulations reliably altered monitoring accuracy, with consistent expectancy producing more accurate monitoring and inconsistent expectancy producing less accurate monitoring. Such findings are in line with the TAM theory: greater congruency between processes engaged in judgments and those engaged in the tests leads to better monitoring accuracy (Griffin, Wiley, et al., 2019; Thiede et al., 2011). Put differently, metacomprehension accuracy is greater when participants engage in judgments about text inferences and then undertake inference tests (i.e., when judgment and test processes are matched) than when they engage in judgments about retention of text details but then undertake inference tests (i.e., when judgment and test processes are mismatched).

It is worth noting that the test-expectancy congruency effect can also be accounted for by the SMAM framework. For instance, participants expecting inference tests might use cues related to the situation model representation to inform their comprehension judgments. Hence, their judgment accuracy would be higher when the test is in fact inferential than when it is factual. We highlight that the TAM and SMAM accounts are not mutually exclusive, and the mechanisms proposed by both accounts may contribute to the test-expectancy congruency effect.

The enhancement effect of practice tests on metacomprehension accuracy provides additional support for the TAM account. That is, experience obtained from practice tests might inform participants about the nature of the criterion test and induce a consistent test expectancy, which in turn improves monitoring accuracy. Readers should regard support from practice tests as suggestive, however, because there are other ways through which practice tests might boost metacomprehension accuracy. For instance, practice test performance provides informative feedback and induces awareness of the gap between perceived and actual learning, which then calibrate subsequent comprehension judgments.

The current meta-analysis observed a variety of lines of support for the SMAM framework. For instance, delayed-keyword-generation was more beneficial than immediate-keyword-generation,

delayed-summary-writing produced a larger enhancement than immediate-summary-writing, and self-explaining boosted metacomprehension accuracy. In addition, interventions involving cognitive processes related to concept organization and knowledge integration (e.g., concept-mapping) successfully enhanced monitoring accuracy. Furthermore, rereading after a short interval also produced a weak but statistically detectable enhancement. These findings jointly support the SMAM framework.

The poor-comprehension theory assumes that low monitoring accuracy derives from poor comprehension, and predicts a positive relationship between comprehension and metacomprehension accuracy. This hypothesis was supported by the positive relationship between test performance and *MC*s. Furthermore, regardless of whether intervention effects were included or excluded, whether test format was constrained to multiple-choice or not, whether only immediate criterion tests were included or not, and whether the participant samples were restricted to young adults or not, this positive relationship persisted. These consistent findings provide support for the poor-comprehension theory.

The positive relationship between test performance and *MC*s documented here is somewhat inconsistent with the null relationship between text difficulty and monitoring accuracy reported by Prinz et al. (2020a). Prinz et al. took Flesch-Kincaid-grade-level scores as a measure of text difficulty, but texts with the same Flesch-Kincaid-grade-level scores may not be equally easy or difficult for different participants or populations employed in different studies. We suggest, by contrast, that test performance is a relatively more sensitive measure of text comprehension across studies.[4] Indeed, almost all previous studies took test performance as a measure of text comprehension.

Overall, given that the current analysis included a much larger dataset and the documented findings were strong and consistent, we hence propose that the relationship between text comprehension and metacomprehension accuracy is approximately linearly positive, and in line with the poor-comprehension theory.

[4]Another reason why the current meta-analysis used test performance (rather than Flesch-Kincaid-grade-level scores) to test the poor-comprehension theory is that many previous studies reported Flesch-Kincaid-grade-level scores in a range (rather than a specific value) for multiple texts (e.g., Thiede & Anderson, 2003; Thiede et al., 2003), which means that the exact grade-level score for each study is unavailable. In addition, there are many studies which did not report any information about Flesch-Kincaid-grade-level scores (Q. S. Chen & Li, 2008; Engelen et al., 2018; van Loon et al., 2014). By contrast, most studies provided test performance results.

The TAM, SMAM, and poor-comprehension theories are not mutually exclusive, and the mechanisms proposed by these theories might combine to affect metacomprehension accuracy. Consistent with this idea, we observed that intervention combinations produced additive benefits to boost monitoring accuracy, suggesting that different interventions might influence distinct underlying (meta)cognitive processes (for related discussion, see Griffin, Wiley, et al., 2019).

Readers should bear in mind that meta-analysis only provides a blunt instrument to test theories. The included studies differed in many respects which might confound the findings. Hence, the theoretical implications are suggestive rather than conclusive. Further experimental research is required to directly test metacomprehension theories, which will further profit our understanding of the mechanisms responsible for poor metacomprehension accuracy.

Intervention effectiveness and practical implications

The above findings demonstrate that metacomprehension accuracy (*MC* = 0.178) under typical non-intervention conditions is even poorer than has been estimated by previous reviews, highlighting the need to develop and evaluate effective interventions to improve monitoring accuracy. Indeed, many interventions have been developed and assessed in previous studies. The current meta-analysis found that combining interventions was the most effective technique to improve metacomprehension accuracy, followed (in order of estimated effectiveness) by delayed-summary-writing, concept-mapping/concept-diagram-completion/concept-diagram-drawing/graph-drawing/mind-mapping, delayed-keyword-generation, self-explaining, practice tests, letter deletion, consistent test expectancy, and rereading after a short delay.

Of course, not all interventions enhance metacomprehension accuracy. For instance, the meta-analysis found that question-generation, immediate-summary-writing, analogy provision, immediate-keyword-generation, and reading questions and answers failed to reliably enhance monitoring accuracy. Needless to say, these are statistically non-detectable results and should be interpreted with statistical power issues borne in mind, especially as some are based on small numbers of effects. Inconsistent test expectancy even reduced monitoring accuracy.

Overall, these findings point to the recommendation that practitioners should consider employing delayed-summary-writing, interventions involving processes of concept organization and knowledge integration (e.g., concept-mapping), delayed-keyword-generation, self-explaining, practice tests, letter deletion, consistent test expectancy, and rereading after a short delay to increase metacomprehension

accuracy. In addition, combining different interventions tends to produce additive benefits. Other interventions (e.g., immediate-summary-writing, immediate-keyword-generation, question-generation) are relatively less effective.

Limitations and future research directions

Even though the current meta-analysis integrated a large set of data, it still suffered from several limitations. For example, it is uncertain whether all relevant studies were identified and included (Flather, Farkouh, Pogue, & Yusuf, 1997; Lyman & Kuderer, 2005), even though we exerted considerable effort to search dozens of electronic databases and found many unpublished studies (e.g., k = 166 effects extracted from student theses). Another limitation is that most of the included studies were conducted in laboratory settings and hence might lack ecological validity (Yang, Luo, et al., 2021). For instance, even though all included studies employed educationally-relevant materials (e.g., expository texts), these materials were not part of students' coursework. This limitation also applies to previous studies and meta-analyses (Fukaya, 2010; Prinz et al., 2020a). It will be important for future research to measure metacomprehension accuracy in the classroom, test the effectiveness of different interventions in educational settings, and determine the generalizability of the findings (Wiley et al., 2016).

The current meta-analysis particularly concentrated on assessing relative accuracy of metacomprehension, leaving its absolute accuracy unexplored. This limitation also applies to the other three meta-analyses described in the Introduction (Fukaya, 2010; Prinz et al., 2020a, 2020b). Absolute metacomprehension accuracy is of considerable importance for text learning (Glenberg, Wilkinson, & Epstein, 1982; Son & Metcalfe, 2000). For instance, according to discrepancy-reduction models of self-regulated learning, overconfidence may drive learners to terminate their studying prematurely, in turn leading to underachievement (Dunlosky & Rawson, 2012; Thiede et al., 2003). Conversely, underconfidence may lead learners to repeatedly and unnecessarily study well-mastered texts (Sarac & Tarhan, 2017), leading to poor learning efficiency. Therefore, future systematic reviews should seek to measure absolute accuracy of metacomprehension, explore the factors that constrain it, and evaluate the effectiveness of different interventions (Ghatala, Levin, Foorman, & Pressley, 1989; Glenberg et al., 1982; Miller & Geraci, 2014; Pressley, Snyder, Levin, Murray, & Ghatala, 1987; Serra & Dunlosky, 2010).

Although some interventions have been repeatedly studied (e.g., delayed-keyword-generation, practice tests) and their effectiveness is robust, others have received less attention (e.g., analogy provision, reading questions and answers). In addition, although the timepoint of rereading is assumed to be a key moderator of the effect of rereading on metacomprehension accuracy, to our knowledge only one study has explored this important issue (Dunlosky & Rawson, 2005). These interventions and their boundary conditions deserve more attention in future research.

Our meta-analysis observed that combining different interventions tends to be more beneficial than many kinds of single interventions. However, each kind of intervention combination has received fewer than five assessments. Furthermore, the enhancing effects of individual interventions are relatively small. Even though combining different interventions produces a larger benefit, the *MC* for intervention combinations is still far from perfect (*MC* = 1). Therefore, future research could usefully develop more effective interventions and explore how to combine different interventions (e.g., consistent test expectancy plus self-explaining plus delayed-summary-writing) to further improve metacomprehension accuracy. We have described the effects of combining interventions as "additive", but in reality they may be interactive in the sense that the benefits of combining two interventions are larger (or smaller) than the sum of their effects. High-powered studies will be needed to identify any cases of such interactivity.

Although many studies have demonstrated that accurate monitoring is related to efficient regulation of restudy choices, which in turn produces superior knowledge gains (e.g., Thiede et al., 2003), there are many other studies showing that rereading does not always improve text comprehension (e.g., Rawson et al., 2000), which is especially true if learners do not restudy effectively (Phillips, Mills, D'Mello, & Risko, 2016). For instance, Phillips and colleagues observed that participants' minds wandered more frequently during the rereading phase than during the initial reading phase, and these researchers proposed that increased mind wandering might be responsible for the limited effectiveness of rereading. More research is needed to explore how to improve the effectiveness of rereading (Rawson & Kintsch, 2005).

Lastly but importantly, many effective interventions (e.g., self-explaining, delayed-keyword-generation, delayed-summary-writing, concept-mapping) require readers to exert additional mental effort, which may discourage learners from actively employing them during self-regulated learning (Kirk-Johnson, Galla, & Fraundorf, 2019). Hence, another important direction for future research is to

develop and evaluate techniques to boost learners' motivation to apply these interventions during self-regulated learning. In addition, instructors need to find effective ways to teach students to independently apply these intervention strategies.

Concluding Remarks

In summary, the takeaway messages from the current meta-analytic review are as follows. Metacomprehension accuracy is strikingly poor (*MC* = 0.242), especially in standard non-intervention conditions (*MC* = 0.178). Many interventions are effective for enhancing metacomprehension accuracy, such as intervention combinations, delayed-summary-writing, concept-mapping/concept-diagram-completion/concept-diagram-drawing/graph-drawing/mind-mapping, delayed-keyword-generation, self-explaining, practice tests, letter deletion, consistent test expectancy, and rereading after a short delay. By contrast, there are many others which are less effective or even ineffective, such as question-generation, immediate-summary-writing, immediate-keyword-generation, analogy provision, and reading questions and answers. Several explanations of poor metacomprehension accuracy, in particular the TAM, SMAM, and poor-comprehension theories, garner support from the current meta-analysis.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, *17*, 18-32. doi:http://dx.doi.org/10.1037/a0022086

Ainsworth, S., & Th Loizou, A. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, *27*, 669-681. doi:10.1207/s15516709cog2704_5

*Agler, L.-M. L., Noguchi, K., & Alfsen, L. K. (2019). Personality traits as predictors of reading comprehension and metacomprehension accuracy. *Current Psychology*, 1-10 doi:http://dx.doi.org/10.1007/s12144-019-00439-y

*Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta psychologica*, *128*, 110-118. doi:https://doi.org/10.1016/j.actpsy.2007.10.006

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*, 25-35. doi:https://doi.org/10.1016/j.jmp.2010.08.009

*Baker, J. M. (2008). The effects of cue diagnosticity on accuracy of judgments of text learning: Evidence regarding the cue utilization hypothesis and momentary accessibility. (Ph.D. dissertation). Kent State University, Ann Arbor. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=kent1216127791

*Baker, J. M., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review*, *13*, 60-65. doi:https://doi.org/10.3758/BF03193813

*Baker, J. M., Dunlosky, J., & Hertzog, C. (2010). How accurately can older adults evaluate the quality of their text recall? The effect of providing standards on judgment accuracy. *Applied Cognitive Psychology*, *24*, 134-147. doi:https://doi.org/10.1002/acp.1553

Borenstein, M., & Cooper, H. (2009). The handbook of research synthesis and meta-analysis (Vol. 2). New York: Russell Sage Foundation.

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

*Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and

answering for remembering expository text. *Journal of Educational Psychology, 104*(4), 922-931. doi:http://dx.doi.org/10.1037/a0028661

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*, 115-144. doi:10.1177/2515245919847196

*Chen, G. C. (2011). Exploring the benefits of restudying: Metacognitive judgments during massed versus spaced study sessions. (Master dissertation). California State University, Long Beach, Ann Arbor.

*Chen, Q. S. (2008). Metacomprehension monitoring: Cues, criteria, and accuracy. (Ph.D. dissertation). The Chinese University of Hong Kong (Hong Kong), Ann Arbor.

*Chen, Q. S. (2009). Metacomprehension monitoring and regulation in reading comprehension. *Acta Psychologica Sinica*, *41*, 676-683.

*Chen, Q. S. (2011). The delay effect of metacomprehension monitoring accuracy and its mechanism. *Journal of Psychological Science*, *34*, 828-833. doi:https://www.psysci.org/CN/Y2011/V34/I4/828

*Chen, Q. S., & Li, L. (2008). Rating comprehension and predicting performance: Clarifying two forms of metacomprehension monitoring. *Acta Psychologica Sinica*, *40*, 961-968. doi:http://dx.doi.org/10.3724/SP.J.1041.2008.00961

*Chen, S. (2010). The impact of deep-processing on senior high school students' metacomprehension monitoring accuracy with different working memory capacity. (Master dissertation). Zhejiang Normal Univeristy. Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10345-2010241199.htm

Chi, M. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology* (pp. 161-238).

Chi, M., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439-477. doi:https://doi.org/10.1016/0364-0213(94)90016-7

*Chiang, E. S. (2007). Selective rereading versus rereading in promoting greater metacomprehension accuracy. (Ph.D. dissertation). University of Florida, Ann Arbor.

*Chiang, E. S., Therriault, D., & Franks, B. A. (2010). Individual differences in relative metacomprehension accuracy: Variation within and across task manipulations. *Metacognition and Learning*, *5*, 121-135. doi:http://dx.doi.org/10.1007/s11409-009-9052-6

*Cliburn, R. (2012). The effect of annotation on metacomprehension of text. (Bachelor dissertation). Baylor University, Retrieved from http://hdl.handle.net/2104/8398

*Commander, N. E., Zhao, Y. L., Li, H. L., Zabrucky, K. M., & Agler, L. M. L. (2014). American and Chinese students' calibration of comprehension and performance. *Current Psychology*, *33*, 655-671. doi:10.1007/s12144-014-9235-4

de Bruin, A. B., Rikers, R. M. J. P., & Schmidt, H. G. (2007). Improving metacomprehension accuracy and self-regulation in cognitive skill acquisition: The effect of learner expertise. *European Journal of Cognitive Psychology*, *19*, 671-688. doi:10.1080/09541440701326204

*de Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, *109*, 294-310. doi:https://doi.org/10.1016/j.jecp.2011.02.005

*Dunlosky, J., Baker, J. M., Rawson, K. A., & Hertzog, C. (2006). Does aging influence people's metacomprehension? Effects of processing ease on judgments of text learning. *Psychology & Aging*, *21*, 390-400. doi:http://dx.doi.org/10.1037/0882-7974.21.2.390

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quartly Journal Experimental Psychollogy (Hove), 64*, 467-484. doi:10.1080/17470218.2010.502239

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology: Series B, 52B*, 178-186. doi:10.1093/geronb/52B.4.P178

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228-232. doi:http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language*, *36*, 34-49. doi:https://doi.org/10.1006/jmla.1996.2476

*Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy?

Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, *40*, 37-55. doi:10.1207/s15326950dp4001_2

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271-280. doi:10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., & Hacker, D. J. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J. Otero, J. A. León, & A. Graesser (Eds.), The psychology of science text comprehension (pp. 255-279, Chapter xii, 459 Pages): Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), Applied metacognition (pp. 68-92, Chapter xi, 297 Pages): Cambridge University Press, New York, NY.

*Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *53*, 551-565. doi:http://dx.doi.org/10.1016/j.jml.2005.01.011

Duval, S. (2005). The trim and fill method. In Publication Bias in Meta-Analysis:: Prevention, Assessment and Adjustments (pp. 127-144):John Wiley & Sons, Ltd, West Sussex, England.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463. doi:10.1111/j.0006-341X.2000.00455.x

Eakin, D. K., & Moss, J. (2018). The methodology of metamemory and metacomprehension. In Handbook of Research Methods in Human Memory (pp. 125-153): Routledge.

Engelen, J. A. A., Camp, G., van de Pol, J., & Anique, B. H. d. B. (2018). Teachers' monitoring of students' text comprehension: Can students' keywords and summaries improve teachers' judgment accuracy? *Metacognition and Learning*, *13*, 287-307. doi:http://dx.doi.org/10.1007/s11409-018-9187-4

Feng, Y. (2014). The influence of speech and belief of the speech on metacomprehension monitoring accuracy. (Master dissertation). Zhejiang Normal University, Retrieved from

http://gb.oversea.cnki.net/KCMS/detail/detail.aspx?filename=1014388322.nh&dbcode=CMFD&dbname=CMFDREF

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*, 507-521. doi:10.2307/2331838

Flather, M. D., Farkouh, M. E., Pogue, J. M., & Yusuf, S. (1997). Strengths and limitations of meta-analysis: Larger studies may be more reliable. *Controlled Clinical Trials*, *18*, 568-579. doi:https://doi.org/10.1016/S0197-2456(97)00024-X

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), The nature of intelligence (pp. 231-235). Hillsdale: Lawrence Erlbaum Associates.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502. doi:10.1126/science.1255484

*Fulton, E. K. (2019). How well do you think you summarize? Metacomprehension accuracy in younger and older adults. *The Journals of Gerontology: Series B*, *gbz142*. doi:https://doi.org/10.1093/geronb/gbz142

Fukaya, T. (2010). Factors affecting the accuracy of metacomprehension: A meta-analysis. *Japanese Journal of Educational Psychology*, *58*, 236-251. doi:http://dx.doi.org/10.5926/jjep.58.236

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning*, *8*, 1-18. doi:http://dx.doi.org/10.1007/s11409-012-9093-0

Garner, R. (1987). Metacognition and reading comprehension. Cognition and literacy. Westport CT: Praeger.

Gervais, W. (2015). Putting PET-PEESE to the test. Retrieved from http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

Ghatala, E. S., Levin, J. R., Foorman, B. R., & Pressley, M. (1989). Improving children's regulation of their reading PREP time. *Contemporary Educational Psychology*, *14*, 49-66. doi:https://doi.org/10.1016/0361-476X(89)90005-2

*Gier, V. S., Kreiner, D. S., & Natz-Gonzalez, A. (2009). Harmful effects of preexisting inappropriate highlighting on reading comprehension and metacognitive accuracy. *Journal of General Psychology*, *136*, 287-300. doi:https://doi.org/10.3200/GENP.136.3.287-302

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping

doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53-96. doi:10.1111/j.1539-6053.2008.00033.x

*Glenberg, A., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 702-718. doi:http://dx.doi.org/10.1037/0278-7393.11.1-4.702

*Glenberg, A., & Epstein, W. (1986). Inexpert calibration of comprehension. *Memory & Cognition*, *15*, 84–93. doi:https://doi.org/10.3758/BF03197714

*Glenberg, A., Sanocki, T., Epstein, W., & Morris, C. C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*, 119-136. doi:http://dx.doi.org/10.1037/0096-3445.116.2.119

Glenberg, A., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, *10*, 597-602. doi:https://doi.org/10.3758/BF03202442

*Griffin, T. D., Jee, B., & Wiley, J. (2006). Expertise and the illusion of comprehension. Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

*Griffin, T. D., Jee, B., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*, 1001-1013. doi:https://doi.org/10.3758/MC.37.7.1001

Griffin, T. D., Mielicki, M., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. A. Rawson (Eds.), The Cambridge handbook of cognition and education (pp. 619-646, Chapter xviii, 729 Pages): Cambridge University Press, New York, NY.

*Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*, 93-103. doi:https://doi.org/10.3758/MC.36.1.93

*Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 1066-1092. doi:http://dx.doi.org/10.1037/xlm0000634

*Gündogdu, N. (2016). The effects of visual detail in images on metacomprehension accuracy. (Master dissertation). Tilburg University, Retrieved from http://arno.uvt.nl/show.cgi?fid=141810

*Han, T. (2010). Influences of different tasks and judgment formats on relative accuracy of metacomprehension. (Master dissertation). Henan University. Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10475-2010154316.htm

*Harten, A. C. M. (1999). An investigation of calibration of comprehension: Text processing variables that affect college students' evaluation of their comprehension. (Ph.D. dissertation). The University of Texas at Austin, Ann Arbor. Retrieved from https://search.proquest.com/docview/304528748?accountid=8554

Hilgard, J. (2017). Trim-and-fill just doesn't work. Retrieved from http://crystalprisonzone.blogspot.com/2017/05/trim-and-fill-just-doesnt-work.html

*Huo, J. (2013). Metacomprehension monitoring competence in EFL textual reading: Development and constraining factors. (Master dissertation). Beijing International Studies University, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10031-1013213367.htm

*Ide, M. E. (2010). The impact of a question generation strategy on calibration of comprehension. (Master dissertation). Northern Illinois University, Ann Arbor. Retrieved from https://www.proquest.com/docview/755259943?pq-origsite=gscholar&fromopenview=true

*Ikeda, K., & Kitagami, S. (2012). The effect of working memory capacity and mental effort on monitoring accuracy in text comprehension. *Psychologia*, *55*, 184-193. doi:https://doi.org/10.2117/psysoc.2012.184

*Ikeda, K., & Kitagami, S. (2013). The interactive effect of working memory and text difficulty on metacomprehension accuracy. *Journal of Cognitive Psychology*, *25*, 94-106. doi:http://dx.doi.org/10.1080/20445911.2012.748028

*Jaeger, A. J. (2012). Can self-explanation improve metacomprehension accuracy for illustrated text? (Masters dissertation). University of Illinois at Chicago. Retrieved from https://indigo.uic.edu/articles/thesis/Can_Self-Explanation_Improve_Metacomprehension_Accuracy_for_Illustrated_Text_/10820189/1

Jaeger, A. J., & Wiley, J. (2010). Seductive images and metacomprehension of science texts. Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

*Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction*, *34*, 58-73. doi:http://dx.doi.org/10.1016/j.learninstruc.2014.08.002

Jaeger, A. J., & Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension.

Discourse Processes, *52*, 376-405. doi:10.1080/0163853x.2015.1026679

*Jiang, S. (2017). The effect of different monitoring forms and the learning strategies on the metacomprehension monitoring and self-regulated study. (Master dissertation). Zhejiang Normal University. Retrieved from http://cpfd.cnki.com.cn/Article/CPFDTOTAL-ZGXG201711001351.htm

*Keener, M. C. (2011). Integration of comprehension and metacomprehension using narrative texts. (Ph.D. dissertation). The University of Utah, Ann Arbor. Retrieved from https://search.proquest.com/docview/883489144?accountid=8554

*Kelemen, W., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*, 92-107. doi:10.3758/BF03211579

Kelemen, W., Winningham, R. G., & Weaver, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology*, *19*, 689-717. doi:10.1080/09541440701326170

Kintsch, W., & Walter, K. (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237. doi:https://doi.org/10.1016/j.cogpsych.2019.101237

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349-370. doi:10.1037/0096-3445.126.4.349

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 187-194. doi:https://doi.org/10.1037/0278-7393.31.2.187

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, 35, 455-463. doi:http://dx.doi.org/10.1016/j.chb.2014.02.046

*Lefèvre, N., & Lories, G. (2004). Text cohesion and metacomprehension: Immediate and delayed judgments. *Memory & Cognition*, *32*, 1238-1254. doi:https://doi.org/10.3758/BF03206315

*Li, A. (2010). The experimental study on the impact of anchoring effect on metacomprehension monitoring accuracy. (Master dissertation). Zhejiang Normal University, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10345-2010241493.htm

*Lin, L. M., Moore, D., & Zabrucky, K. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology*, *22*, 111-128. doi:https://doi.org/10.1080/02702710119125

Lin, L. M., & Zabrucky, K. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*, 345-391. doi:https://doi.org/10.1006/ceps.1998.0972

*Lin, L. M., Zabrucky, K., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology*, *115*, 187-198. doi:https://doi.org/10.2307/1423434

*Linderholm, T., Wang, X. S., Therriault, D., Zhao, Q., & Jakiel, L. (2012). The accuracy of metacomprehension judgements: The biasing effect of text order. *Electronic Journal of Research in Educational Psychology*, *10*, 111-128.

Linderholm, T., Zhao, Q., Therriault, D., & Cordell-McNulty, K. (2008). Metacomprehension effects situated within an anchoring and adjustment framework. *Metacognition and Learning*, *3*, 175-188. doi:http://dx.doi.org/10.1007/s11409-008-9025-1

*Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, *43*, 85-98. doi:10.3758/s13421-014-0453-7

*Lu, J. (2013). Form of metacomprehension monitoring and its delay effect. (Master dissertation). Jinan University, Retrieved from http://d.wanfangdata.com.cn/thesis/Y2364542

*Luo, Z. (2010). The impact of reading materials on metacomprehension accuracy. (Master dissertation). Shantou University, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10560-2010269916.htm

Lyman, G. H., & Kuderer, N. M. (2005). The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology*, *5*, 14. doi:10.1186/1471-2288-5-14

*Maki, R. H. (1995). Accuracy of metacomprehension judgments for questions of varying importance levels. *The American Journal of Psychology*, *108*, 327. doi:http://dx.doi.org/10.2307/1422893

*Maki, R. H. (1998a). Metacomprehension of text: Influence of absolute confidence level on bias and

accuracy. In D. L. Medin (Ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 38 (pp. 223-248, Chapter ix, 306 Pages): Academic Press, San Diego, CA.

*Maki, R. H. (1998b). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, *26*, 959-964. doi:https://doi.org/10.3758/BF03201176

*Maki, R. H. (2008). Privileged access for general knowledge and newly learned text material. In J. Dunlosky, & R. Bjork (Eds.), A handbook of metamemory and memory (pp. 173-194): Taylor & Francis Group, New York, NY.

Maki, R. H. (1998c). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), Metacognition in educational theory and practice (pp. 117-144, Chapter xiv, 407 Pages): Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 663-679. doi:http://dx.doi.org/10.1037/0278-7393.10.4.663

*Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. H., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 609-616. doi:http://dx.doi.org/10.1037/0278-7393.16.4.609

*Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension ability and metacomprehension. *Psychonomic Bulletin & Review*, 1, 126-129. doi:https://doi.org/10.3758/BF03200769

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), Applied metacognition (pp. 39-67, Chapter xi, 297 Pages): Cambridge University Press, New York, NY.

*Maki, R. H., & Serra, M. J. (1992a). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 116-126. doi:http://dx.doi.org/10.1037/0278-7393.18.1.116

*Maki, R. H., & Serra, M. J. (1992b). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology*, *84*, 200-210. doi:http://dx.doi.org/10.1037/0022-0663.84.2.200

*Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology*, *97*, 723-731. doi:10.1037/0022-0663.97.4.723

*Maki, R. H., Willmon, C., & Pietan, A. (2009). Basis of metamemory judgments for text with multiple-choice, essay and recall tests. *Applied Cognitive Psychology*, *23*, 204-222. doi:10.1002/acp.1440

Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica*, *134*, 154-161. doi:https://doi.org/10.1016/j.actpsy.2010.01.006

*Martin, N. D., Nguyen, K., & McDaniel, M. A. (2016). Structure building differences influence learning from educational text: Effects on encoding, retention, and metacognitive control. *Contemporary Educational Psychology*, *46*, 52-60. doi:http://dx.doi.org/10.1016/j.cedpsych.2016.03.005

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730-749. doi:10.1177/1745691616662243

*Mi, Y. (2020). Studying on the influence of graphic organizers on media on the reading effect, experience, and meta-comprehension of middle school students. (Master dissertation). Tianjin Normal University. Retrieved from https://cdmd.cnki.com.cn/Article/CDMD-10065-1020743267.htm

*Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, *19*, 650-670. doi:http://dx.doi.org/10.1080/09541440701326196

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition*, *29*, 131-140. doi:https://doi.org/10.1016/j.concog.2014.08.008

Millis, K., Simon, S., & Tenbroek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition*, *26*, 232-246. doi:10.3758/BF03201136

*Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 223-232.

doi:http://dx.doi.org/10.1037/0278-7393.16.2.223

Munafò, M. R., Matheson, I. J., & Flint, J. (2007). Association of the DRD2 gene Taq1A polymorphism and alcoholism: a meta-analysis of case–control studies and evidence of publication bias. *Molecular Psychiatry*, *12*, 454-461. doi:10.1038/sj.mp.4001938

*Nelms, K. R. (2000). The impact of hypermedia instructional materials on study self -regulation in college students. (Ph.D. disssertation). Georgia State University, Ann Arbor. Retrieved from https://search.proquest.com/docview/304620189?accountid=8554

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109-133. doi:https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207-213. doi:https://doi.org/10.1111/j.1467-9280.1994.tb00502.x

*Ni, H. (2019). Metacomprehension monitoring and control in reading: The influence of self-explanation, working memory, and monitoring accuracy. (Master dissertation). Guizhou Normal University. Retrieved from https://cdmd.cnki.com.cn/Article/CDMD-10663-1019860168.htm

*Ni, H., & Xu, W. (2019). Reading comprehension: The influences of cognition and metacognition. *Data of Culture and Education*, *31*, 32-34.

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum, 803*, 278-281. doi:10.4028/www.scientific.net/MSF.803.278

*Novakovic, A. A. (2016). Comparison between students and teachers in judgment accuracy and cue-utilization: An explorative approach. (Bachelor dissertation). Retrieved from https://dspace.library.uu.nl/handle/1874/370832

*Olin, J. T., & Zelinski, E. M. (1997). Age differences in calibration of comprehension. *Educational Gerontology*, *23*, 67-77. doi:http://dx.doi.org/10.1080/0360127970230106

Otero, J., & Graesser, A. C. (2014). The psychology of science text comprehension. New York:

Routledge.

*Ozuru, Y., Kurby, C., & McNamara, D. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning*, *7*, 113-131. doi:10.1007/s11409-012-9087-y

*Pao, L. S. (2014). Effects of keyword generation and peer collaboration on metacomprehension accuracy in middle school students. (Ph.D. dissertation). Columbia University, Ann Arbor. doi: https://doi.org/10.7916/D8HX19TV

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*, 378-395. doi:10.1177/2515245919879695

*Peng, K. (2013). Effects of reading materials' character and TOA on the accuracy of metacomprehension monitoring and cognitive fatigue in English reading. (Master dissertation). Henan Univerisity, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10475-1013349955.htm

Phillips, N. E., Mills, C., D'Mello, S., & Risko, E. F. (2016). On the influence of re-reading on mind wandering. *Quarterly Journal of Experimental Psychology, 69*, 2338-2357. doi:10.1080/17470218.2015.1107109

*Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency—Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31-40. doi:http://dx.doi.org/10.1016/j.learninstruc.2016.01.012

*Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, *29*, 62-67. doi:https://doi.org/10.3758/BF03195741

Pilegard, C., & Mayer, R. E. (2015). Within-subject and between-subject conceptions of metacomprehension accuracy. *Learning and Individual Differences*, *41*, 54-61. doi:http://dx.doi.org/10.1016/j.lindif.2015.07.003

*Poulin, C. M. (2013). Do rubrics improve students' metacomprehension accuracy? (Master dissertation). Kent State University. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=kent1374595640

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing

adjunct questions. *Reading Research Quarterly*, *22*, 219-236. doi:https://doi.org/10.2307/747666

Prinz, A., Golke, S., & Wittwer, J. (2020a). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review*, *31*, 100358. doi:https://doi.org/10.1016/j.edurev.2020.100358

Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review*, 32, 917-949. doi:10.1007/s10648-020-09558-6

*Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 69-80. doi:http://dx.doi.org/10.1037/0278-7393.28.1.69

*Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*, 559-579. doi:https://doi.org/10.1080/09541440701326022

*Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology*, *55*, 505-524. doi:https://doi.org/10.1080/02724980143000352

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*, 1004-1010. doi:https://doi.org/10.3758/BF03209348

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*, 70-80. doi:http://dx.doi.org/10.1037/0022-0663.97.1.70

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). Psychology of reading. Sussex, the UK: Psychology Press.

*Redford, J., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction*, *22*, 262-270. doi:http://dx.doi.org/10.1016/j.learninstruc.2011.10.007

Renken, A. E. (2001). Students' predicted and actual forgetting functions for expository text. (Ph.D.

dissertation). Baylor University, Ann Arbor. Retrieved from https://www.proquest.com/docview/250917199?pq-origsite=gscholar&fromopenview=true

*Sarac, S., & Tarhan, B. (2017). Calibration of comprehension and performance in L2 reading. *International Electronic Journal of Elementary Education*, *2*, 167-179.

*Sarmento, D. (2018). Does realism harm metacomprehension accuracy? (Master dissertation). University of Illinois at Chicago, Ann Arbor. Retrieved from https://indigo.uic.edu/articles/thesis/Does_Realism_Harm_Metacomprehension_Accuracy_/10922459

*Serra, M. J. (2007). Is metacomprehension for multimedia presentations different than for text alone? (Ph.D. dissertation). Kent State University. Retrieved from https://chem.ckcest.cn/Degree/Details?id=215886

*Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, *18*, 698-711. doi:http://dx.doi.org/10.1080/09658211.2010.506441

*Shiu, L. P., & Chen, Q. S. (2013). Self and external monitoring of reading comprehension. *Journal of Educational Psychology*, *105*, 78-88. doi:http://dx.doi.org/10.1037/a0029378

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's z transformation be used? *Journal of Applied Psychology, 72*, 146-148. doi:https://doi.org/10.1037/0021-9010.72.1.146

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204-221. doi:10.1037/^)278-7393.26.1.204

*Sun, X. (2011). The study on reading metacomprehension and its relationship with self-regulated study for Grade 5 students. (Mater dissertation). Liaoning Normal University, Retrieved from https://cdmd.cnki.com.cn/Article/CDMD-10165-1013127601.htm

Spearman, C. (1961). The proof and measurement of association between two things. East Norwalk, CT, US: Appleton-Century-Crofts.

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

Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-

analysis. In Publication Bias in Meta-Analysis (pp. 99-110).

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., . . . Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, l4898. doi:10.1136/bmj.l4898

*Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, *28*, 129-160. doi:http://dx.doi.org/10.1016/S0361-476X(02)00011-5

*Thiede, K. W., Anderson, M. C., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66. doi:https://doi.org/10.1037/0022-0663.95.1.66

*Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1267-1280. doi:http://dx.doi.org/10.1037/0278-7393.31.6.1267

*Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*, 331-362. doi:http://dx.doi.org/10.1080/01638530902959927

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. Handbook of metacognition in education, 85-106.

*Thiede, K. W., Redford, J., Wiley, J., & Griffin, T. D. (2012). Elementary school experience with comprehension testing may influence metacomprehension accuracy among seventh and eighth graders. *Journal of Educational Psychology*, *104*, 554-564. doi:http://dx.doi.org/10.1037/a0028660

*Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology*, *87*, 590-605. doi:10.1111/bjep.12166

*Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*, 264-273. doi:https://doi.org/10.1348/135910710X510494

*Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition*, *35*, 668-678.

doi:https://doi.org/10.3758/BF03193305

*Thule, E. J. (2005). Accuracy of metacognitive monitoring and learning of texts. (Master dissertation). University of Toronto (Canada), Ann Arbor.

*van de Pol, J., de Bruin, A. B. H., van Loon, M. H., & van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology*, *56*, 236-249. doi:10.1016/j.cedpsych.2019.02.001

*van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merrienboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica*, *151*, 143-154. doi:10.1016/j.actpsy.2014.06.007

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2005). The effects of prior knowledge on study-time allocation and free recall: Investigating the discrepancy reduction model. *The Journal of Psychology*, *139*, 67-79. doi:10.3200/JRLP.139.1.67-79

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*, 428-443. doi:10.1037/1082-989X.10.4.428

*Vössing, J., Stamov-RoBnagel, C., & Heinitz, K. (2017). Text difficulty affects metacomprehension accuracy and knowledge test performance in text learning. *Journal of Computer Assisted Learning*, *33*, 282-291. doi:https://doi.org/10.1111/jcal.12179

*Vössing, J., & Stamov-Roßnagel, C. (2016). Boosting metacomprehension accuracy in computer-supported learning: The role of judgment task and judgment scope. *Computers in Human Behavior*, *54*, 73-82. doi:http://dx.doi.org/10.1016/j.chb.2015.07.066

*Vössing, J., Stamov-Roßnagel, C., & Heinitz, K. (2016). Images in computer-supported learning: Increasing their benefits for metacomprehension through judgments of learning. *Computers in Human Behavior*, *58*, 221-230. doi:http://dx.doi.org/10.1016/j.chb.2015.12.058

Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2014). Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and Research*, *38*, 313-333. doi:10.1007/s10608-013-9588-2

*Waldeyer, J., & Roelle, J. (2020). The keyword effect: A conceptual replication, effects on bias, and

an optimization. *Metacognition and Learning*, *16*, 37-56. doi:10.1007/s11409-020-09235-7

*Wang, X. F. (2015). The effects of testing and feedback on metacomprehension accuracy. (Ph.D. dissertation). The Chinese University of Hong Kong (Hong Kong), Ann Arbor.

*Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 214-222. doi:http://dx.doi.org/10.1037/0278-7393.16.2.214

*Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, *23*, 12-22. doi:10.3758/BF03210553

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artifacts. *Advances in Methods and Practices in Psychological Science*, *3*, 94-123. doi:10.1177/2515245919885611

*Wiley, J., Griffin, T. D., Jaeger, A., Jarosz, A. F., Cushen, P. J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied*, *22*, 393-405. doi:http://dx.doi.org/10.1037/xap0000096

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology*, *132*, 408-428. doi:https://doi.org/10.3200/GENP.132.4.408-428

*Wiley, J., Jaeger, A. J., Taylor, A. R., & Griffin, T. D. (2018). When analogies harm: The effects of analogies on metacomprehension. *Learning and Instruction*, *55*, 113-123. doi:http://dx.doi.org/10.1016/j.learninstruc.2017.10.001

Wiley, J., Thiede, K. W., & Griffin, T. D. (2007). What does it mean to learn from and understand science text. Paper presented at the annual meeting of the American Education Research Association (AERA), Chicago, IL.

*Xie, J. (2014). An experimental research on the influencing factor of metacomprehension judgment during hypermedia learning. (Ph.D. dissertation). Central China Normal University, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10511-1014239574.htm

*Xu, F., & Shi, J. (2008). Metacomprehension accuracy and its relation to self-regulated learning. *Psychological Science (China)*, *31*, 1162-1166.

*Yan, M. (2018). The effects of time pressure and presentation forms on reading comprehension and

metacomprehension. (Master dissertation). Zhejiang Normal University, Retrieved from http://cdmd.cnki.com.cn/Article/CDMD-10345-1018299771.htm

*Yan, R., & Huo, J. (2013). Meta-comprehension monitoring competence in EFL textual reading: Development and effect of cross language transfer. *Modern Foreign Languages*, *36*, 158-165.

*Yan, R., Li, T., Li, S., & Yu, H. (2015). Impact of time interval for clue encoding and retrieval on EFL learners' meta-comprehension monitoring accuracy. *Journal of Beijing International Studies University*, *37*, 1-6.

Yang, C., Huang, J., Li, B., Yu, R., Luo, L., & Shanks, D. R. (2020). Learning difficulty determines whether concurrent metamemory judgments enhance or impair learning outcomes: Meta-analytic and emperical tests. Submitted for publication.

Yang, C., Huang, T. S. T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language*, *99*, 99-110. doi:https://doi.org/10.1016/j.jml.2017.11.005

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

Yang, C., Sun, B., & Shanks, D. R. (2018). The anchoring effect in metamemory monitoring. *Memory & Cognition*, *46*, 384–397. doi:https://doi.org/10.3758/s13421-017-0772-6

*Yang, C. Q. (2019). The influence of mind mapping on senior high school students' metacomprehension monitoring ability in English reading. (Master dissertation). Southwestern University (China), Retrieved from http://www.51papers.com/lw/69/18/wz4391995.htm

*Zabrucky, K. M., Agler, L. M. L., & Moore, D. (2009). Metacognition in Taiwan: Students' calibration of comprehension and performance. *International Journal of Psychology, 44*, 305-312. doi:10.1080/00207590802315409

Zabrucky, K. M., & Moore, D. (1994). Contributions of working memory and evaluation and regulation of understanding to adults' recall of texts. *Journal of Gerontology*, *49*, 201-212. doi:10.1093/geronj/49.5.P201

*Zeng, X. (2009). Effects of summarizing and self-questioning on metacomprehension accuracy. (Master dissertation). Zhejiang Normal University.

*Zhang, L., Sun, X., & Li, Y. (2011). The study on metacomprehension and its delayed-effect for Grade 5 students. *Psychological Research (China)*, *4*, 78-81.

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review*, *20*, 191-206. doi:10.1007/s10648-008-9073-8

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162-185. doi:10.1037/0033-2909.123.2.162

Table 1. Interventions implemented in previous studies

Intervention	Description	Sample reference
Analogy provision	Participants read texts with analogies (e.g., an analogy about how computers process information, presented in a text describing how the human mind processes information). In the non-intervention condition, participants read texts without analogies.	Wiley et al. (2018)
Concept-mapping/concept-diagram-completion/concept-diagram-drawing/graph-drawing/mind-mapping	After or during reading, participants draw concept maps, complete concept diagrams, draw concept diagrams, draw graphs, or create mind maps to organize the key ideas stated in the text. In the non-intervention condition, participants do not perform such tasks.	Redford et al. (2012); van Loon et al. (2014)
Question generation	Participants are instructed to generate several questions about the text during or after reading. In the non-intervention condition, participants do not do so.	Bugg and McDaniel (2012)
Keyword interventions		
Delayed-keyword-generation	Participants generate several keywords to capture the essence of the texts *after reading all texts*. In the non-intervention condition, they do not generate keywords.	Thiede et al. (2012); Waldeyer and Roelle (2020)
Immediate-keyword-generation	Participants generate several keywords to capture the essence of the texts *immediately after reading each text*. In the non-intervention condition, they do not generate keywords.	de Bruin et al. (2011); Thiede et al. (2005)
Keyword-reading	Participants read keywords provided by the experimenter after reading the texts. In the non-intervention condition, participants do not read keywords.	Q. S. Chen and Li (2008)
Letter deletion	Participants read texts in which letters of some words are deleted. In the non-intervention condition, participants read intact texts.	Ikeda and Kitagami (2012); Maki et al. (1990)
Rereading after a short delay	Participants reread the studied texts after a short delay. In the non-intervention condition, participants only read the texts once.	Rawson et al. (2000)
Summary interventions		
Delayed-summary-writing	Participants generate a summary to capture the gist of the text *after reading all texts*. In the non-intervention condition, participants do not generate summaries.	Anderson and Thiede (2008); Thiede et al. (2010)

Immediate-summary-writing	Participants generate a summary to capture the gist of the text *immediately after reading each text*. By comparison, participants do not generate summaries in the non-intervention condition.	Thiede and Anderson (2003)
Self-explaining	While reading each text, participants generate explanations to themselves about the meaning and relevance of each sentence or paragraph to the overall purpose of the text, or what new information the just-read sentence or paragraph conveys. In the non-intervention condition, participants do not generate self-explanations.	Fukaya (2013); Jaeger (2012)
Test expectancy interventions		
Consistent expectancy	Participants are pre-informed about the nature of the criterion test (e.g., whether it is a memory or inference test), and their expectancy is *consistent* with the criterion test. In the non-intervention condition, participants are not directly informed how they will be tested and what will be evaluated in the criterion test.	Thiede et al. (2011)
Inconsistent expectancy	Participants are pre-informed about the nature of the criterion test, but their expectancy is *inconsistent* with the criterion test. In the non-intervention condition, participants are not directly informed how they will be tested and what will be evaluated in the criterion test.	Griffin, Wiley, et al. (2019)
Test interventions		
Practice tests	After reading but before making judgments, participants undertake practice tests on the studied texts. In the non-intervention condition, participants do not take practice tests.	Little and McDaniel (2015); Maki and Serra (1992b)
Reading questions and answers	After reading but before making judgments, several questions and their corresponding answers are presented for participants to read. In the non-intervention condition, no questions and answers are provided.	Q. S. Chen and Li (2008); Wang (2015)
Intervention combinations	Participants receive at least two kinds of interventions. In the non-intervention condition, participants do not receive any interventions.	Griffin, Wiley, et al. (2019); Martin et al. (2016)
Other interventions	Other kinds of interventions which received fewer than 5 explorations (e.g., motivation manipulation, highlighting, think-about-the-text, rereading after a long delay, and so on).	Linderholm et al. (2012); Poulin (2013)

Table 2. Multilevel multivariate meta-regression analysis results for different interventions

Intervention	Difference from non-intervention effects				
	k	Difference	95% CI	*Z*	*p*
Analogy provision	6	0.010	[-0.098, 0.118]	0.18	.855
Concept-mapping/concept-diagram-completion/concept-diagram-drawing/graph-drawing/mind-mapping	11	0.196	[0.131, 0.262]	5.89	< .001
Question generation	7	0.042	[-0.057, 0.141]	0.83	.4.06
Keyword interventions					
Delayed-keyword-generation	40	0.184	[0.147, 0.220]	9.87	< .001
Immediate-keyword-generation	17	0.009	[-0.037, 0.054]	0.37	.710
Keyword-reading	6	-0.096	[-0.192, 0.000]	-1.96	.050
Letter deletion	7	0.112	[0.022, 0.202]	2.43	.015
Rereading after a short delay	22	0.063	[0.015, 0.111]	2.56	.010
Summary interventions					
Delayed-summary-writing	9	0.210	[0.164, 0.255]	9.00	< .001
Immediate-summary-writing	9	0.030	[-0.019, 0.080]	1.21	.225
Self-explaining	7	0.179	[0.110, 0.248]	5.08	< .001
Test expectancy interventions					
Consistent expectancy	8	0.090	[0.024, 0.157]	2.67	.008
Inconsistent expectancy	6	-0.117	[-0.214, -0.019]	-2.34	.019
Test interventions					
Practice tests	47	0.169	[0.131, 0.208]	8.61	< .001
Reading questions and answers	6	-0.095	[-0.196, 0.006]	-1.85	.065
Intervention combinations	30	0.316	[0.269, 0.362]	13.25	< .001
Other interventions	21	0.023	[-0.012, 0.058]	1.30	.194

Note: *k* = number of effects.

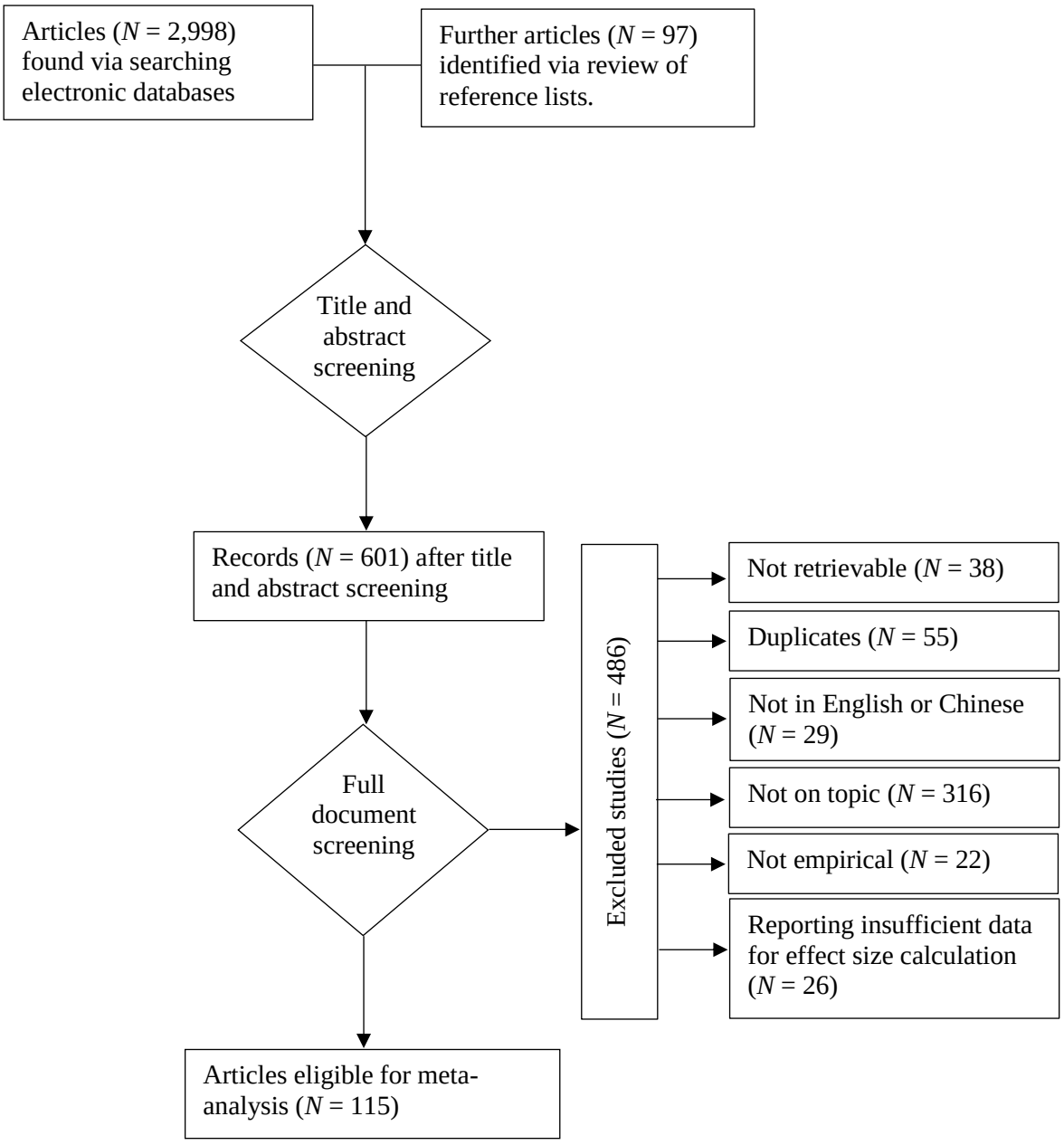

Figure 1. Flowchart depicting the article screening results.

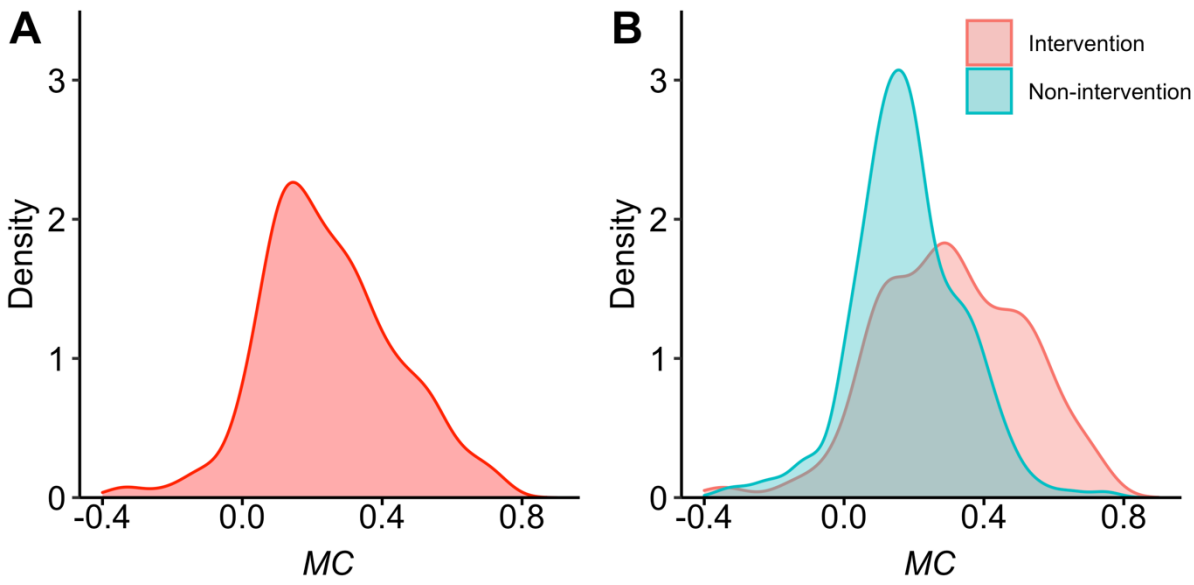

Figure 2. A: Density distribution of all effects. B: Separate density distributions for intervention and non-intervention effects. See the online article for a color version of this figure.

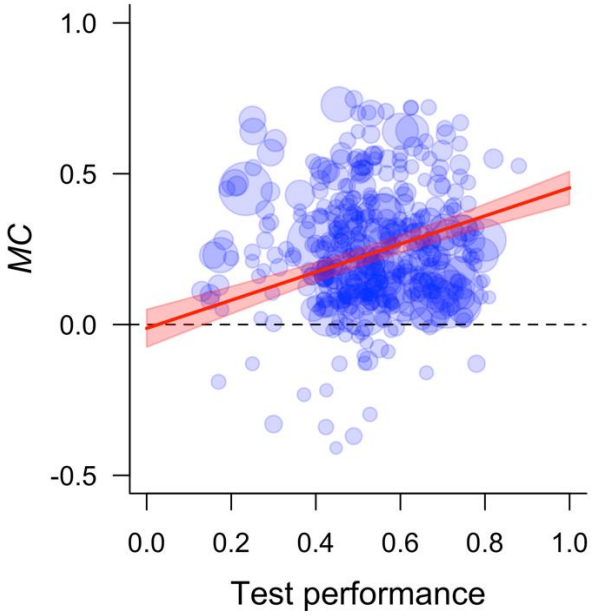

Figure 3. Bubble plots depicting the relationship between *MC*s and their corresponding test performance. Bubble sizes represent the relative weights of the included effects, and error bars represent 95% CI of the regression trend. See the online article for a color version of this figure.

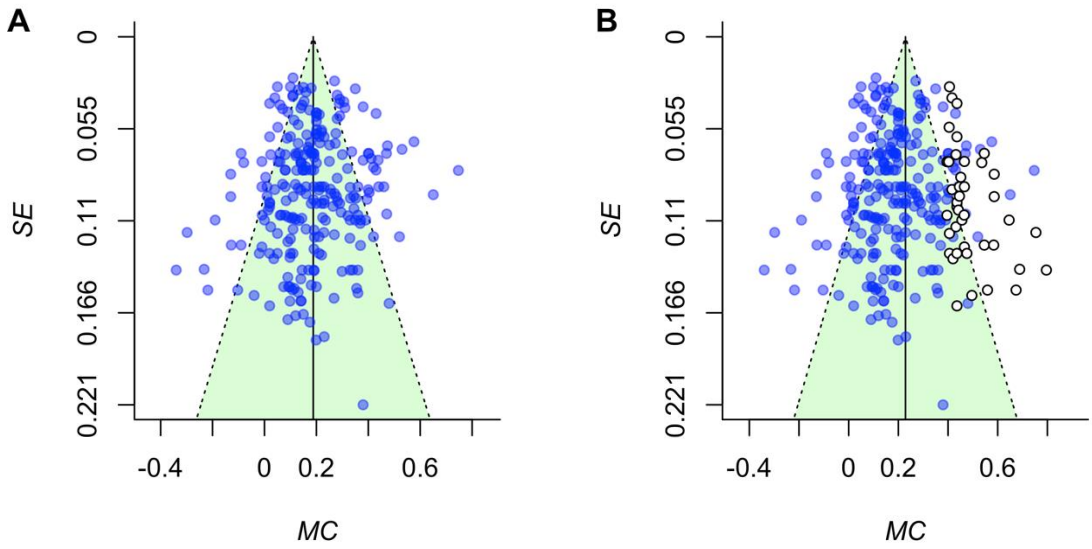

Figure 4. A: Funnel plot depicting the relationship between *MC*s and their corresponding *SE*s. B: Funnel plot depicting the Trim-and-Fill results. Each blue dot represents an included effect, and the unfilled black circles in panel B indicate missing effects detected by the Trim-and-Fill method. See the online article for a color version of this figure.