Metacognition and Learning (2021) 16:57–88
https://doi.org/10.1007/s11409-020-09238-4

Understanding and supporting Chinese middle Schoolers' monitoring accuracy in mathematics

Ying Wang [1] · Rayne A. Sperling [1]

Received: 5 November 2019 / Accepted: 10 August 2020 / Published online: 20 August 2020

Abstract

While many studies of monitoring accuracy have been conducted with college students, less is known about middle school students' monitoring accuracy, especially in Asian countries. Prior research also found discrepancies in students' monitoring accuracy between Western and Asian cultures. To understand and support Chinese middle school students' monitoring accuracy in mathematics, we implemented a three-week long monitoring intervention with 7th grade students in Southwestern China. Students were divided into three conditions: a control condition, a confidence rating condition (CR), and a confidence rating combined with monitoring instructions condition (CR + MI). Findings indicated that Chinese middle school students were slightly underconfident, yet quite accurate when monitoring during task completion in mathematics. Specifically, although students in the CR and CR + MI conditions did not show significant increases in mathematics performance, students in the control condition significantly decreased their performance overtime. In terms of monitoring, students' confidence bias increased overtime in the CR + MI and control conditions while students in the CR condition did not show significant change. Students also demonstrated increased self-reported metacognitive awareness across the three conditions. We also asked students to provide justifications for their confidence judgments. Findings indicated that students tended to consider a single factor when judging their performance during a mathematics task. Regression models of students' justifications indicated potential suggestions for future interventions. Implications and future directions are discussed.

Keywords Metacognitive intervention · Monitoring accuracy · Mathematics learning · Chinese learners · Self-regulated learning

✉ Ying Wang
yqw5386@psu.edu

Rayne A. Sperling
rsd7@psu.edu

[1] Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, State College, PA, USA

Introduction

Metacognition refers to learners' awareness and regulation of their cognition (Flavell 1979). Monitoring is a core component of metacognition that drives learners to detect errors and optimize their performance. Schraw and Moshman (1995) define monitoring as "one's on-line awareness of comprehension and task performance (p. 355)." Specifically, monitoring represents learners' ability to concurrently examine their learning processes and outcomes. Accurate monitoring enables learners to perceive task demands, determine the appropriate selection of strategies, and evaluate and reflect on task performance in ways that improve their future task performance (Winne 1995, 2001; Winne and Hadwin, 1998). In contrast, inaccurate monitoring, can be detrimental to students' academic achievement.

Students' monitoring accuracy has been studied by examining the extent to which students' perceived performance is discrepant from their actual performance, which is also called calibration (Keren 1991; Nietfeld et al. 2006b; Pieschl 2009). Confidence bias and absolute accuracy are two indices that capture this discrepancy (Nelson 1996; Pieschl 2009). Specifically, confidence bias demonstrates the signed difference between students' perceived performance and actual performance. A positive difference indicates students' overconfidence while a negative difference indicates their underconfidence. In comparison, absolute accuracy represents the absolute value of the discrepancy. Monitoring is considered accurate when both indices (i.e., confidence bias and absolute accuracy) are close to zero.

Existing research about these two indices suggests that students often have difficulty rendering accurate judgments relative to their objective academic performance, typically because they are often overconfident (e.g., García et al. 2016; Glenberg and Epstein 1987; Hacker et al. 2000; Miller and Geraci 2011a). As metacognition is related to domain-specific knowledge (Schraw 1998), the phenomenon, in which students' inaccurate appraisals of their own performance hinders them from deploying self-regulated learning strategies and optimizing learning goals, is also considered as an issue in the domain of mathematics.

In the domain of mathematics, students often need to engage in multiple cognitive activities, such as coordinating different strategies during problem-solving tasks (Silver 1987, 1994). To do so, students need to be equipped with both cognitive and metacognitive knowledge of mathematics (Garofalo and Lester 1985). However, this knowledge is not readily acquired by students, perhaps especially as they are transitioning from primary school mathematics to middle school mathematics and confronting more intensive and comprehensive curriculum. Cleary and Chen (2009) reported that middle school students rarely used strategies effectively and efficiently during mathematics problem-solving tasks. One primary reason for students' lack of effective strategy use is a lack of metacognition as students fail to recognize gaps in their mathematics knowledge and their need for further learning to remedy such gaps. Further, even if students have a strong and accurate knowledge base, they may be unable to accurately monitor their learning processes and, as a result, report inaccurate or biased judgments about their own performance (e.g., Dinsmore et al. 2015; Glenberg and Epstein 1987). These deficits in monitoring may create obstacles for middle school students to pursue academic goals in mathematics. Thus, further investigations into interventions to support improvement of middle schoolers' monitoring accuracy in mathematics are warranted.

Researchers (e.g., Nietfeld and Schraw 2002; Zimmerman et al. 2011) have reported positive outcomes from the implementation of metacognitive interventions in mathematics with college students. These interventions typically teach students to judge their performance

accurately by providing monitoring guidelines and/or feedback to improve students' monitoring accuracy as well as improve their mathematics achievement. Interventions in monitoring, however, have not been adequately implemented and tested with middle schoolers.

Further, few studies have examined Asian students' monitoring in any academic domain. In one of the few studies, Yates et al. (1989) compared American and Chinese college students' monitoring accuracy when answering general knowledge questions (e.g., What is the farther north? London or New York). Specifically, the American and Chinese students were asked to answer equally difficult general knowledge questions and indicate the probability that they arrived at the correct answers. Findings suggested that Chinese students were significantly more inaccurate and overconfident when compared to the American students.

However, findings of such cultural differences are not always consistent. For instance, Zabrucky et al. (2009) examined Taiwanese college students' ability to accurately monitor their performance on reading comprehension items. Specifically, students were asked to judge their performance on reading comprehension before (prediction) and after (postdiction) the task. Findings indicated that overall, Taiwanese students' reading comprehension performance was positively associated with accurate performance judgements, corresponding to the findings with students from Western countries (e.g., Dunlosky et al. 2005b; List and Alexander, 2015; Pressley and Ghatala 1988). In contrast, however, to a previously conducted similar study within the United States, Lin et al. (2001) reported discrepant results. As such, American college students were asked to make predictions and postdictions about their performance on reading expository texts. Findings indicated that the American students' postdictions tended to be more accurate than predictions, which was consistent with the Taiwanese students in the later study (Zabrucky et al. 2009). However, American students' monitoring accuracy was not associated with their performance, which was discrepant from Taiwanese students. Taiwanese students also demonstrated more accurate monitoring than American students when comparing these two studies (i.e., Lin et al. 2001; Zabrucky et al. 2009). These findings indicate potential differences in monitoring between Asian and Western students. The current study addressed known gaps in the monitoring literature and targeted Chinese middle school students' monitoring in mathematics.

Collectively, the present study contributes to the literature on monitoring accuracy in several ways. First, with a considerable amount of monitoring literature focused on reading comprehension (e.g., Dabarera et al. 2014; Gillström and Rönnberg 1995; Glenberg et al. 1987; Kinnunen and Varuas 1995; Lauterman and Ackerman 2014; Lin et al. 2002; Ozuru et al. 2012; Schraw et al. 1993; Shiu and Chen 2013; Singer and Alexander 2017; Walczyk and Hall 1989), less is known about monitoring accuracy in mathematics. Further, the present study focused on first-year (7th grade) middle school students' monitoring accuracy, based upon our consideration that these students are likely to experience unique challenges in mathematics learning during the transition between elementary and middle school. Moreover, we asked middle school students to not only judge their performance but to further justify their judgements. Studying middle school students' justifications of their monitoring judgments allows us to identify factors that younger students may consider when making performance judgements and, more importantly, may provide insight into how to improve future monitoring interventions for middle school students in mathematics. Finally, in comparison to most of the reading comprehension monitoring studies that were conducted in Western countries, the present study was conducted in classrooms within China and represents one of very few monitoring intervention studies conducted in China.

Monitoring in metacognition and self-regulated learning

Schraw and Moshman's model of metacognition (1995) explained the importance of monitoring for students' learning. Specifically, they divided metacognition into two metacognitive processes: knowledge of cognition and regulation of cognition. *Knowledge of cognition* refers to one's awareness or knowledge about one's cognition, nature of tasks, and selection of strategies. *Regulation of cognition* refers to a dynamic and proactive self-regulatory process of learning. Regulation of cognition emphasizes the ongoing processes learners deploy in order to reach their learning goals and meet task criteria. Importantly, knowledge of cognition and regulation of cognition dynamically interact with one another. When learners develop accurate awareness of their knowledge, they correspondingly learn to control and regulate their learning. Monitoring falls under regulation of cognition, and not only enables learners to track their learning trajectories but also supports the development of metacognitive awareness.

Winne and Hadwin's (1998) model of self-regulated learning also places metacognitive monitoring as the central component that operates the processes of self-regulated learning. Specifically, they identified four phases of self-regulated learning including the task definition phase, the goal setting and planning phase, the tactics and strategies enacting phase, and the adaptation phase. Monitoring assists learners to move through each of the four phases. In Phase 1 (task definition), learners receive information from the task context as well as their prior task experience to perceive the definition of the task. Monitoring helps learners to compare such information against the current task to establish their perceptions of the task. In Phase 2 (goal setting and planning), monitoring assists learners to set goals and plans through matching with their initial task perceptions from Phase 1. Moving toward Phase 3 (enacting tactics and strategies), learners use monitoring to select and employ tactics and strategies while completing the task. Finally, in Phase 4 (adaptation), monitoring guides learners to evaluate and reflect on their task product and make necessary revisions for future tasks.

Both Schraw and Moshman's and Winne and Hadwin's models demonstrate the crucial function of monitoring in students' learning processes. In other words, students' accurate monitoring spurs improved awareness of their knowledge, effective strategy deployment, and desired learning outcomes (Butler and Winne 1995; Dunlosky et al. 2005a; Flavell 1979; Huff and Nietfeld 2009; Pressley and Ghatala 1990; Winne and Jamieson-Noel 2002).

Monitoring interventions

Although monitoring is critical for learners' academic achievement, prior literature in students' monitoring found that students are typically inaccurate monitors (e.g., Dunlosky and Rawson 2012; Schraw et al. 1995). Some scholars and practitioners have therefore targeted interventions to support students' monitoring abilities. As noted, while the majority of monitoring studies examined reading comprehension (e.g., Dunlosky et al. 2005b; Pressley and Ghatala 1988, 1990), researchers also studied monitoring in a variety of domains and developed interventions to improve students' monitoring. Consistent with results that targeted reading comprehension, across academic domains, high monitoring accuracy was associated with improved academic achievement and students were generally inaccurate monitors (e.g., research methods, Dinsmore and Parkinson 2013; undergraduate educational psychology, Hacker et al. 2000; multiple texts task, List and Alexander 2015; math, Ramdass and Zimmerman 2008).

For instance, Nietfeld and Schraw (2002) implemented a strategy training session to improve college students' monitoring accuracy during mathematics tasks. Students were asked to complete a set of mathematics problems and to provide confidence ratings for each item on a 100-point scale. Students in the training condition received a 2-h training that included strategy instructions for solving mathematical probability problems. In contrast, students in the control condition were not provided with strategy instructions. As a result of this brief intervention, those students who received the strategy training demonstrated improved mathematics problem solving performance as well as better monitoring accuracy. From these positive results, Nietfeld et al. (2006a) extended this research and implemented another monitoring strategy training study. They distributed weekly monitoring exercise sheets that directed undergraduate students to monitor their learning in an introductory educational psychology course. Specifically, students in the treatment condition received weekly monitoring instruction and intensive monitoring exercises over 16 weeks. Students were asked to practice item-by-item monitoring on multiple-choice questions corresponding to the course contents and received feedback from the instructor the following week. Students in the comparison condition did not practice monitoring on a weekly basis. Findings showed that students in the treatment condition demonstrated improved monitoring accuracy as well as academic performance when compared to the comparison condition and indicated that training with monitoring exercises has potential benefit for both students' monitoring accuracy and their academic achievement.

In testing an additional intervention, Bol et al. (2012) provided high school students with monitoring practice guidelines and instruction in a biology class. Specifically, monitoring practice guidelines and instruction directed students to check their answers and understanding during task completion. In addition, students in the experimental condition were further instructed to reflect on their understanding of targeted biology content and to make confidence judgments during task completion. Results indicated that students' monitoring accuracy was improved, as was their biology achievement.

While these positive results suggest support for monitoring interventions, inconsistent findings demonstrate that metacognitive interventions are not always effective. For instance, Bol, Hacker et al. (2005) examined whether having students practice monitoring would improve their monitoring accuracy and academic achievement. Specifically, they asked college students to make predictions and postditctions about their performance on several quizzes across an academic semester in an education major-related course. Students in the control condition were asked to complete the same quizzes but without pre- and postdictions. Unexpectedly, results indicated that students' overt monitoring practice did not improve their monitoring accuracy nor academic performance. This may be due to the lack of explicit monitoring instruction provided to students to inform their monitoring practice. Similar ineffective intervention results were also reported in other studies (e.g., Bol and Hacker, 2001; Nagel and Lindsey 2018). These mixed findings demonstrated the inconsistent intervention effects on students' monitoring.

Factors that influence monitoring

In addition to studying the effectiveness of interventions, other scholars focused on investigating the formation and the multifaceted nature of students' performance judgments. Lin and Zabrucky's (1998) review suggested that external task demands play a role in forming students' performance judgments. Particularly, students' performance judgments were driven by individual-related, task-related, and text-related factors. In a reading comprehension task,

individual-related factors refer to readers' individual differences that influence their reading comprehension and monitoring accuracy. Such factors may include learners' prior knowledge about the text content, reading ability, and related motivational constructs (e.g., interest in the topic, and self-efficacy for reading). Task-related factors refer to possible task characteristics such as types of tests, test item difficulty, task demands, and involvement of feedback or practice. Text-related factors include the genre of texts and text difficulty level. Lin and Zabrucky suggested that accurate performance judgments require learners to take both internal (individual-related) and external (task-related and text-related) factors into consideration simultaneously while rendering performance judgments. When students neglect to consider multiple factors, poorer academic performance and inaccurate monitoring may result.

Drawing from Lin and Zabrucky (1998), Pieschl (2009) argued that traditional studies of students' performance judgments solely focused on students' internal processes, that is, how students capture their individual learning processes when making judgments, and prior investigation neglected external task demands and task complexity. As previous findings indicated that students tend to be overconfident for difficult tasks, it may be that they fail to accurately perceive task demands and terminate improvement on the task (Pressley and Ghatala 1988; Schraw and Roedel 1994). Pieschl (2009) discussed that learners consider more than internal factors in order to render accurate performance judgments and also suggested future research to capture external factors that influence students' performance judgements.

Similar conclusions were voiced by Dinsmore and Parkinson (2013). In their study, college students' monitoring accuracy and academic performance as they read two introductory statistics passages was examined. After reading the passages, students answered a set of multiple-choice questions related to the passage content and then rated their performance on a 100-point scale. Notably, Dinsmore and Parkinson further asked students to justify their performance judgments in open-ended responses explaining reasons for their judgments and identified five categories of factors that students considered when making judgments: prior knowledge, text characteristics, item characteristics, guessing, and 'other'. Besides identifying internal (prior knowledge) and external (text characteristics and item characteristics) similar to Lin and Zabrucky (1998) and Pieschl (2009), Dinsmore and Parkinson also recognized two additional factors: guessing and 'other'. Guessing captured when students stated that their judgments were made based on a guess or a feeling, while the category of 'other' represented those not within previously identified categories [See Dinsmore and Parkinson (2013) for sample responses]. They reported that college students considered both personal and environmental factors when making performance judgments and further verified the multifaceted nature of students' performance judgments.

A recent study (Wang and List 2019) also examined college students' self-evaluations, but in a complex writing composition task. Participants were asked to compose a written product (i.e., a research report or an argument) based on reading of multiple texts. After they composed a product, they evaluated their written responses by assigning themselves a letter grade. They then explained reasons for their grade assignment. Results demonstrated that students considered 12 categories of factors, such as strategies deployment, specific writing mechanics, and personal attributions. Consistent with previous findings, students' justifications were multifaceted but broadly mapped onto personal skills, task context, and strategies that they enacted (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998; Pieschl 2009).

The multifaceted nature of students' monitoring justification has only been established with college students and generally in reading and writing tasks. In the present study, we extended

the work and examined middle school students' justifications about their performance in mathematics problem-solving tasks.

Monitoring in a mathematics context

Although metacognitive monitoring is often considered as domain-general in nature (Gutierrez et al. 2016; Schraw 1998), the motivational (e.g., goal setting) and cognitive components (e.g., strategy use) involved in self-regulated learning may vary across different subject areas. Wolters and Pintrich (1998) asked Grade 7 and 8 students to complete self-report questionnaires regarding motivation and cognition across three subject areas including mathematics, English, and social studies. They found differences in students' reported motivation- and cognition-related constructs across the three subject areas. Specifically, in terms of the motivational components, students demonstrated higher task value for mathematics than English and social studies and students had higher self-efficacy in English than mathematics and social studies. Students also demonstrated varied degrees of cognitive strategy use and students reported that they used strategies more often in social studies when compared to English and mathematics. Given that monitoring plays a critical role throughout the phases of self-regulated learning, it is expected that students may monitor and further regulate their learning differently across domains (Hadwin et al. 2001).

Mathematics is differentiated from other subject areas in a number of ways. First, teachers have different views toward mathematics than other subjects. Particularly, teachers consider mathematics as more structured, sequential, and heavily dependent on previously taught topics, while they consider social studies more open and less sequential (Grossman & Stodolsky 1995; Stodolsky and Grossman,1995). Students also hold different views toward mathematics when compared to other subjects. For instance, compared to social studies, which students may relate to real life during task completion, mathematical problem-solving tasks often require students to confront abstract mathematical concepts and sequential operations (Schoenfeld 1992). Such cognitive activities during mathematics tasks require students to enact specific self-regulatory strategies and accurate monitoring. Yet, students usually fail to do so (e.g., Cleary and Chen 2009; Kramarski and Gutman 2006). For example, García et al. (2016) examined elementary school students' monitoring on mathematics problem-solving tasks using confidence rating scales. Students solved two math word problems and indicated their confidence while also demonstrating their work. These scholars reported that students were inaccurate monitors and were particularly overconfident relative to their actual performance. Their analysis of mathematical problem-solving processes further demonstrated that students who were accurate monitors used strategies more frequently than those who were not accurate. These results reflected students' deficits in monitoring and strategy deployment, which further led to their potential failure in mathematics achievement. Subsequent research by Callan and Cleary (2019) also found that middle school students' mathematics performance was positively associated with metacognitive monitoring and strategy use.

The present study

The present study examined Chinese middle school students' monitoring accuracy in mathematics by implementing a monitoring intervention and exploring students' monitoring judgments. We addressed three primary research questions.

First, what are the associations among absolute accuracy, confidence bias, mathematics performance, and other psychological constructs (i.e., self-regulated learning strategies, metacognitive awareness, and self-efficacy) in Chinese 7th grade students?

According to the theoretical frameworks that guided this study (Schraw and Moshman 1995; Winne and Hadwin 1998), and consistent with previous literature (e.g., Bol et al. 2012) we anticipated that students' improved mathematics performance would be correlated with absolute accuracy and less confidence bias. Meanwhile, self-regulated learning strategies, metacognitive awareness, and self-efficacy were expected to be positively associated with students' mathematics performance and monitoring accuracy. That is, students who use strategies effectively, have high metacognitive awareness, and high self-efficacy in mathematics were also expected to have higher math performance and more accurate monitoring.

Second, within the school setting, to what extent does the monitoring intervention improve 7th grade students' mathematics performance, monitoring accuracy, reported self-regulated learning strategies, metacognitive awareness, and self-efficacy?

Consistent with Nietfeld and Schraw (2002) where students improved their academic performance and monitoring accuracy after receiving both verbal and written monitoring training sessions, we expected that students in the experimental condition, who would receive both explicit monitoring instructions and calibration practice, would improve their mathematics performance and monitoring accuracy more than students in other conditions. Similarly, students' self-regulated learning strategies, metacognitive awareness, and self-efficacy were also expected to improve. We anticipated, however, the magnitude of improvement in monitoring in this study would be constrained for two reasons. First, students were provided only written monitoring instructions without verbal intervention instructions from teachers. Further, the length of three weeks with one session each week was relatively short and limited in comparison to other effective long-term monitoring intervention studies (e.g., 14 sessions: Huff and Nietfeld 2009; 16 sessions: Nietfeld et al. 2006a).

Third, how do 7th grade students justify their performance judgments and do students' performance justifications predict their math performance and monitoring accuracy?

Dinsmore and Parkinson (2013), reported that college students considered multiple factors when making performance judgments. In the present study, we also expected 7th grade students to report a variety of justifications for their performance judgments. However, given the developmental nature of metacognition (Brown 1987), we anticipated that middle school students may not be able to consider as many factors as college students. In other words, middle school students may be more likely to consider a limited number of factors when making performance judgments. We further expected that the specific factors that students commonly considered would significantly predict their math performance as well as monitoring accuracy.

Method

Study design

A three-group pretest/posttest quasi-experimental design with random assignment of classroom to condition was implemented in this study. A pretest was administered to all students to control for potential differences among students and across classrooms. This study examined the effects of the metacognitive intervention on students' monitoring accuracy and

mathematics achievement for a duration of three practice sessions, once a week for three weeks. During each practice session, across conditions, students received the same practice material developed by the teachers, but conditions varied by the inclusion of monitoring directions and confidence rating scales. The three conditions included a control condition, a confidence rating only condition (CR), and a confidence rating with monitoring instructions condition (CR + MI).

Students in the control condition were asked to complete the mathematics practice questions only. In addition to the mathematics practice questions, students in the CR condition were asked to rate their confidence for each item on a 10-point Likert-type scale. Students in the CR + MI condition, were asked to rate their confidence and also were provided with written explicit monitoring directions that instructed them to monitor during the practice session. An example of an explicit monitoring direction was "When you monitor, you ask yourself questions…After you pick an answer you stop and ask yourself if it is the right answer." A detailed description of the study design is presented in Fig. 1.

Participants

Participants were 133 Grade 7 students in a public middle school located in Southwestern China. Of the 133 student participants, 54.14% ($n = 72$) were female, and 45.86% ($n = 61$) were male. Students' average age was 13. The average class size was 44. Data screening was completed to address invalid or missing data and assumption testing was conducted prior to analyses.

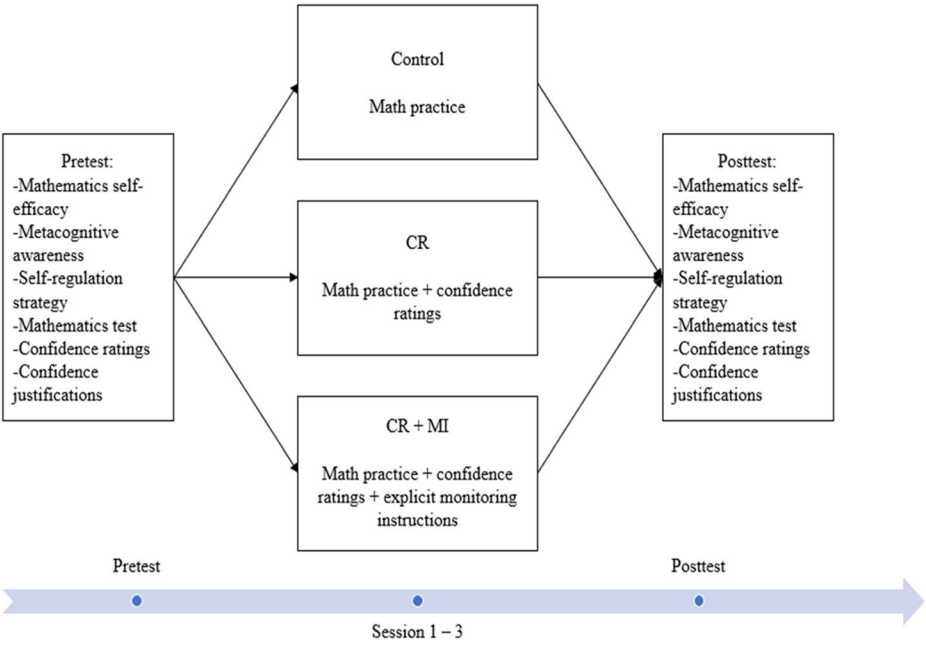

Fig. 1 Study design

Procedures

Students across all conditions received the same practice sets each session during the intervention. Prior to administration of the practice sets, written explicit monitoring directions and/or confidence rating scales designated the three conditions.

The procedures of the study can be briefly described in four steps. First, school permission and students' informed consent were obtained. Second, students completed the pretest. Third, after the pretest, the intervention was implemented across 3 weeks with one session per week. Last, after the intervention phase, all students completed the posttest measures. These measures were identical to those administered in the pretest phase except that a parallel form of the mathematics items was used for the posttest. The total duration of this study was 6 weeks overall including the consent process, pretest, three practice sessions, and posttest. Each practice session took approximately 45 minutes over a regular class period. A timeline is presented in Fig. 1.

Teachers' primary role in this study was to administer the study materials. The first author provided a 2-h training session to the three participating teachers to ensure intervention fidelity in advance of the study including a group session and individual sessions. Contents for the group training session for the three teachers included overall description of the study, specific study procedures, and timeframe. Specifically, the first author met with the three teachers in a conference room and provided general information regarding the study, such as the general instructions that students needed to know when completing the pre- and posttest. The first author also coordinated with the three teachers' schedules, class locations, and other logistical details for conducting the study during the meeting.

After the group session, the first author also held individual meetings with each of the teachers regarding the specific materials they would receive and instructions for administering materials with fidelity prior to the study. Specific study procedures and instructions were provided to each teacher individually corresponding to the designated condition. Typically, students in the target school would complete regular math practice tests independently and teachers would provide feedback regarding students' performance after task completion. As such, during the intervention phase, teachers were asked to provide the practice sets and performance feedback as usual and students were asked to complete the practice independently. Notably, as the written monitoring instructions and monitoring practice were novel for students in the designated conditions, teachers in the CR + MI and the CR conditions were asked to prompt students to read the monitoring instructions and/or take the monitoring practice provided in the materials. The first author also conducted several observations during the study to ensure intervention fidelity.

Materials

Testing and practice materials The materials used for this study included two parts: the pre- and posttest materials and the practice materials for the intervention. Specifically, the pretest and the posttest included a mathematics self-efficacy measure, a metacognitive awareness inventory, a self-regulated learning measure, selected and adapted TIMSS (2011) released math items, confidence rating scales, and open-ended justification questions about students' confidence ratings. In particular, the adapted TIMSS items were shown to the three participating teachers and an expert in middle school mathematics in advance of the study to ensure

the TIMSS items for the pre- and posttest corresponded to the math content students learned in class. Items were only slightly adapted.

The materials for the intervention were teacher-generated math practice sets. These practice sets varied by condition to include CR and CR + MI in the two intervention conditions. Teachers selected the mathematics items from their practice item pool corresponding to the math curriculum taught in their classes with discussions with the first author and the expert in middle school mathematics. A variety of topics were included in these items (e.g., geometry, probability, and algebra). Across conditions, all students completed the same items.

Materials for monitoring instructions The monitoring instructions for the CR + MI condition were developed and adapted from a previous study that demonstrated positive effects on students' monitoring and learning (Sperling et al. 2012). Specifically, the monitoring instructions included three parts: (1) the introduction of monitoring, (2) specific examples for practicing monitoring with instructions, and (3) feedback for the practice examples and instructions that directed students to reflect upon their monitoring process. In Part 1, students learned about the definition of monitoring and its importance in supporting their learning. An example of the introduction was "When we stop and think about what we are doing, it is called monitoring." In Part 2, specific examples of math items were used to demonstrate to students how to monitor. Specifically, monitoring instructions were provided to guide students' monitoring processes during the completion of a specific math item, such as suggesting students to check their work. In Part 3, students rated their certainty about their answer to the math item and were provided feedback regarding the correct answer. Students then were provided with additional instructions that directed them to reflect upon their monitoring process and emphasized the usefulness of monitoring. In total, students were provided with three example math items that included the three instructional scaffolds for each practice session.

Measures

The measures used in this study were all originally published in English. As the students' first language was Chinese and they were not fluent in English, the measures were double translated. The translation process included two steps. First, two translators whose first language is Mandarin Chinese and who are also fluent in English including the first author translated all the instruments administered in this study. Specifically, the two translators translated the English instruments into Mandarin Chinese individually first and then reconciled with each other. Second, a third person who was a Chinese researcher in the field of middle school mathematics and educational theories verified the translation. The translated measures were then finalized to be administered. These procedures were consistent with the Programme for International Student Assessment translation guidelines for double translation and reconciliation (PISA 2018).

Mathematics self-efficacy Student self-efficacy was measured by the *Middle School Mathematics Self-Efficacy Scale* developed and validated by Usher and Pajares (2009). The instrument includes 24 items with a six-point Likert-type scale (1 = definitely false; 6 = definitely true). Original reliability estimates (Cronbach's α) for the measure ranged from $\alpha = .84$ to

$\alpha = .88$ across four subscales and were $\alpha = .86$ for pretest and $\alpha = .82$ for posttest in the present study.

Junior metacognitive awareness (Jr.MAI) Students' metacognitive awareness was assessed by the *Junior Metacognitive Awareness Inventory* Version B developed by Sperling et al. (2002). The Jr.MAI for Grades 6 to 9 includes 18 items on a 5-point Likert scale (1 = Never; 5 = Always). Internal consistency for Jr.MAI was $\alpha = .85$ reported by Sperling et al. (2002). The Cronbach's alphas were $\alpha = .84$ for pretest and $\alpha = .87$ for posttest in the present study.

Self-regulated learning (SRSI-SR) Student self-regulated learning was measured by the *Self-Regulation Strategy Inventory-Self-Report* developed by Cleary (2006). The inventory includes 28 items on a seven-point Likert-type scale (1 = Never; 7 = Always). Cronbach's alpha reported by Cleary (2006) was .92. The original items were designed based on a science-learning context. We adapted the items to a mathematics context. In our study, the Cronbach's alphas were .94 for pretest and .93 for posttest.

Mathematics achievement Released items from an international standardized mathematics achievement test were adapted and administered to measure students' mathematics achievement. Specifically, 10 selected and adapted mathematics items across different difficulty levels and topics from TIMSS (2011) for 8th grade, were administered to all participating students. Parallel forms were used for pretest and posttest. The assessment included three multiple-choice questions, four fill-in-blank questions, and three "show-all-work" questions. Students were asked to show their work and provide written justifications about their confidence for the three show-all-work questions. One point was given for each of the correct dichotomously scored items. The internal consistency for pretest was $\alpha = .63$, and it was $\alpha = .62$ for posttest. Given the assessment was to measure a wide variety of mathematics concepts with a few items, the reliabilities were considered sufficient based on the revised Dutch rating system for test quality (Evers 2001).

Monitoring accuracy Monitoring accuracy was assessed by a confidence rating provided below each mathematics item. The confidence scale ranged from 1 (*not confident at all*) to 10 (*totally confident*). Confidence bias and absolute accuracy were calculated following the formulas suggested by previous literature (e.g., Schraw and Nietfeld 1998), to indicate differences between students' perceived performance and their actual performance. Specifically, students' average confidence scores were divided by nine to arrive to a range of 0 to 1. Students' math scores were then averaged by dividing the number of items. The two indices were obtained by calculating the difference between the averaged rescaled confidence score

Table 1 Calculation for monitoring indices

Calibration index	Calculation	Range
1. Confidence bias	$\frac{1}{N}\sum_{i=1}^{N} c_i - \frac{1}{N}\sum_{i=1}^{N} p_i$	−1 to 1
2. Absolute accuracy	$\left\lvert \frac{1}{N}\sum_{i=1}^{N} c_i - \frac{1}{N}\sum_{i=1}^{N} p_i \right\rvert$	0 to 1

Note. N = number of items, c = confidence scores, p = performance scores. The formulas were adopted from previous research Schraw and Nietfeld (1998).

and the average math score. Thus, confidence bias ranged from −1 to 1, and absolute accuracy ranged from 0 to 1. See Table 1 for the calculation of confidence bias and absolute accuracy.

Confidence justifications After students rated their confidence, for the show-all-work questions, they were asked to justify how they arrived at their judgments. Unlike multiple choice and fill-in-blank items, these three items required students to provide all the work and problem-solving steps. The direction for providing confidence justifications was "Please explain why you would rate your confidence as above."

Coding To capture students' justifications about their confidence judgments, we employed a bottom-up approach to develop a coding scheme. Justifications were coded following four steps. In the first step, the first author read through all of students' justification responses through pre- and posttest and then came up with an initial coding scheme reflecting the variability of students' justifications. In the second step, another researcher who was fluent in Chinese Mandarin joined, as a second coder, read through all of the justifications independently, and then consulted the initial coding scheme with the first author to make sure they shared understanding of the coding categories. In the third step, the first author and the second coder started coding independently. During this phase, the two researchers reconciled any disagreements found among 50% of the students' justifications. Based on the discussion of disagreement, some coding categories were either modified or collapsed into another coding category. Thus, a final coding scheme including ten categories was formed, reflecting various factors that students considered when judging their performance. Overall, these categories represented different dimensions of students' judgments of performance including person-related characteristics, item-related characteristics, and context-related characteristics, corresponding to findings in prior research (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998; Pieschl 2009).

Specifically, three coding categories mapped onto students' person-related characteristics including (1) prior knowledge, (2) confidence, (3) effort. In particular, prior knowledge was identified as a category reflecting how familiar students were about the question in their responses (e.g., *I have done a similar question before*). Another person-related category, confidence, indicated students' consideration of how confident they were about their performance in general without indicating more details (e.g. *I am confident.*). The category of effort was identified when students justified their performance based on perceived effort (e.g., *Because I put effort into this question*).

Furthermore, four categories [i.e., (4) required knowledge, (5) problem-solving process, (6) item difficulty, and (7) calculation] reflected item-related characteristics tied to a particular item. Specifically, the *required knowledge* category was identified when students mentioned the knowledge they had for solving a specific item. For example, they made statements such as: *I learned the unit of triangles well*. In addition, the coding category reflected students' specific *problem-solving process* showing their explanation of how they solved the problem and specific steps of their problem-solving procedures. For example, Item 10 was about calculating the interior angle sum of a pentagon. One student responded *"I drew two lines so that the pentagon becomes three triangles. The interior angle sum of one triangle is 180 degrees. Therefore, adding them up is 540 degrees*." Thus, this student demonstrated the process of how one particular problem was solved. When students took the item difficulty into consideration, we coded it as item difficulty. For example, "*I made my decision based how difficult the question is*." Moreover, the calculation category was identified when students rated their confidence based on their calculation skills or

accuracy. For example, "*I am just not sure about my calculation on this item*." These four categories represented the characteristics of the specific items.

Another category considered context-related characteristics, such as perceived task demands. Specifically, when students rated their performance based on the format that they wrote on the paper, we coded it as (8) *format* reflecting students who were especially concerned about the written format required for the task (e.g., "*I am not sure whether or not the format is right for this task*"). When students stated that checking processes lead to their confidence ratings, we coded it as (9) *checking* (e.g., "*I checked my answer after I completed the item*"). In addition, when students expressed uncertainty or guessing in general (e.g., "*I don't know*"), we coded such statements as (10) *unknown*.

Finally, in the last phase, both coders independently read through all of the justifications again to finalize ratings. Exact agreement between two coders was 92.36% through pretest and posttest. See Table 2 for the coding categories and response examples. Additional examples of students' justifications by study condition are provided in Appendix A.

Results

Data screening

Data were collected via printed paper copies. We first conducted data screening to identify any missing or invalid data. There were very few missing entries across the pre- and post-measures. Missing data analysis explored potential missing patterns and identified only 1.16% missing data across all the variables. The dataset met the assumption of data missing completely at random (MCAR), χ^2 (11084) = 175.485, p = 1.00. An EM (expectation-maximization) estimation method was used to impute missing values according to Peng et al. (2006).

Table 2 Coding scheme for students' justifications about their confidence ratings

Coding category	Example	Percentage	
		Pretest	Posttest
Person-related categories:			
1. Prior knowledge	"I have done this type of question before"	18.94% (n = 25)	21.97% (n = 29)
2. Confidence	"I am very confident about myself"	10.61% (n = 14)	13.64% (n = 18)
3. Effort	"I put my effort into it"	0.76% (n = 1)	0.76% (n = 1)
Item-related categories:			
4. Item-relevant knowledge	"The sum of interior angles for triangles"	66.67% (n = 88)	63.64% (n = 84)
5. Problem-solving process	"The total length of the stick is 40 cm. Therefore, I divided it by three and then calculated the value of the x"	28.03% (n = 37)	19.70% (n = 26)
6. Item difficulty	"This question is very easy"	1.52% (n = 2)	2.27% (n = 3)
7. Calculation	"I calculated very carefully"	14.39% (n = 19)	15.91% (n = 21)
Context-related categories:			
8. Format	"I am not sure about the format"	16.67% (n = 22)	6.06% (n = 8)
9. Checking	"I checked my answer after"	6.82% (n = 9)	5.30% (n = 7)
10. Unknown	"I guessed."	27.27% (n = 36)	18.94% (n = 25)

Note. Students' responses were originally written in Chinese. The examples presented in this table are English translations.

We further tested normality of all the variables across pre- and posttest. Although absolute accuracy did not meet the assumption of normality, as we found similar statistical findings and interpretations after performing both parametric and non-parametric analyses. Therefore, we reported the parametric results here for consistency. Please see the notes in Tables 6 and 7 for the non-parametric results.

Research question 1 What are the associations among absolute accuracy, confidence bias, mathematics performance, and other psychological constructs (i.e., self-regulated learning strategies, metacognitive awareness, and self-efficacy) for Chinese 7th grade students?

The first research question examined the extent to which the key variables were associated with one another. Descriptive statistics showed that students performed well overall on both the math pretest and posttest. Furthermore, overall, students had very low bias and absolute accuracy scores indicating effective metacognitive monitoring for both pre- and posttest. See Table 3 for descriptive statistics.

Students' math scores were negatively associated with their confidence bias scores for both the pre- and posttest. Their math scores were also negatively associated with absolute accuracy for the posttest. As expected, these two negative associations indicated that students who received lower mathematics scores tended to be overconfident and less accurate. As anticipated, students' math scores were also found to be positively associated with metacognitive awareness and self-efficacy in mathematics.

Moreover, students' reported self-regulatory strategy use was also positively associated with their metacognitive awareness and self-efficacy, indicating that students with higher metacognitive awareness and self-efficacy tended to report more self-regulatory strategies. Noticeably, while students' absolute accuracy was associated with the three psychological self-regulation measures (i.e., SRSI-SR, Jr.MAI, and Mathematics Self-Efficacy) on the pretest, there were no significant associations on the posttest. Correlation results are presented in Tables 4 and 5.

Table 3 Descriptive statistics across conditions overtime

		Highest possible score	Condition			Overall
			Control	CR	CR + MI	
Pretest	Math	1	0.85(0.16)	0.80(0.18)	0.81(0.19)	0.82(0.18)
	Bias	–	−0.07(0.20)	−0.07(0.17)	−0.04(0.20)	−0.06(0.19)
	Accuracy	–	0.14(0.16)	0.13(0.13)	0.16(0.13)	0.14(0.14)
	SRSI-SR	196	91.81(13.95)	94.32(10.76)	95.11(15.49)	93.75(13.55)
	Jr.MAI	90	60.72(9.51)	60.51(8.17)	58.58(11.58)	59.92(9.86)
	SE	144	86.74(21.74)	89.33(17.52)	89.43(20.69)	88.50(19.97)
Posttest	Math	1	0.79(0.17)	0.83(0.14)	0.78(0.23)	0.80(0.18)
	Bias	–	−0.02(0.21)	−0.07(0.18)	0.04(0.20)	−0.02(0.20)
	Accuracy	–	0.15(0.15)	0.15(0.12)	0.14(0.14)	0.15(0.14)
	SRSI-SR	196	93.70(11.61)	94.07(10.96)	93.31(16.13)	93.69(12.92)
	Jr.MAI	90	62.42(10.47)	61.71(8.52)	59.71(11.33)	61.26(10.19)
	SE	144	89.86(21.98)	88.80(17.36)	88.38(19.57)	89.00(19.60)

Note. Math scores were averaged. *CR* = confidence rating only; *CR + MI* = confidence rating and monitoring instructions.

Table 4 Pearson correlations among absolute accuracy, confidence bias, mathematics performance, and other measures for pretest

Measure	1	2	3	4	5
1. Math performance	–				
2. Confidence bias	−.27**	–			
3. Absolute accuracy	−.12	−.56**	–		
4. SRSI-SR	.16	.25**	−.29**	–	
5. Jr.MAI	.23**	.23**	−.27**	.68**	–
6. SE	.50**	.26**	−.42**	.63**	.61**

**$p < .01$.

Research question 2 To what extent does the monitoring intervention improve 7th grade students' mathematics performance, monitoring accuracy, reported self-regulated learning strategies, metacognitive awareness, and self-efficacy?

Our second research question examined the effects of intervention on the dependent variables including students' math scores, confidence bias, absolute accuracy, metacognitive awareness, self-regulated strategy use, and mathematics self-efficacy. Prior to analyses, we examined whether significant differences existed before the intervention was implemented. Importantly, results showed no significant pretest differences on mathematics performance [F (2, 131) = 1.17, p = .31), confidence bias [F (2, 131) = 0.26, p = .77], absolute accuracy [F (2, 131) = 0.50, p = .61], metacognitive awareness [F (2, 131) = 0.63, p = .53], self-regulated strategy use [F (2, 131) = 0.71, p = .49], or mathematics self-efficacy [F (2, 131) = 0.25, p = .77].

Mathematics performance

To examine the effect of the intervention on students' mathematics performance, we performed a 3 × 2 (Condition |Control, CR, CR + MI| × Time |pretest, posttest|) repeated measure analysis. Results showed no significant differences among the three conditions (p = .10). Nevertheless, a significant interaction between condition and time was found [F (2,129) = 3.39, $p < .05$, $\eta^2 = .05$]. This indicated that students' changes in mathematics achievement were significantly different over time across conditions. Further, post hoc analyses demonstrated that students in the control condition surprisingly decreased in their mathematics performance overtime [F (1,43) = 6.69, $p < .05$, $\eta^2 = .14$]. However, there were no significant changes for

Table 5 Pearson correlations among absolute accuracy, confidence bias, mathematics performance, and other measures for posttest

Measure	1	2	3	4	5
1. Math performance	–				
2. Confidence bias	−.31**	–			
3. Absolute accuracy	−.22*	−.32**	–		
4. SRSI-SR	.19*	.15	.01	–	
5. Jr.MAI	.25**	.08	.05	.77**	–
6. SE	.42**	.27**	−.16	.55**	.56**

*$p < .05$ **$p < .01$.

Table 6 Results for ANOVA repeated measure for math performance

Effect	*M*	*df*	*F*	*p*	η^2
Condition	0.02	2	0.40	.67	.01
Time	0.03	1	2.71	.10	.02
Condition×Time	0.04	2	3.39	< .05	.05
Error	0.01	129			

Note. Consistently, the nonparametric Wilcoxon Signed ranks test showed a significant decrease in students' mathematics scores for the control condition ($Z = -2.51$, $p < .05$)

the CR ($p = .18$) or CR + MI conditions ($p = .51$) between pre- and posttest measures (See Table 6).

Monitoring accuracy

Monitoring accuracy was assessed by indices of confidence bias and absolute accuracy. Two 3 × 2 (Condition |Control, CR, CR + MI| × Time |pretest, posttest|) repeated measure analyses were conducted separately for confidence bias and absolute accuracy. Results demonstrated significant changes in students' confidence bias over time [F (1, 129) = 7.14, $p < .01$, $\eta^2 = .05$]. Specifically, students in the control condition marginally [F (1, 43) = 3.93, $p = .05$, $\eta^2 = .08$] increased in their underconfidence. Students in the CR + MI condition [F (1,44) = 6.28, $p < .05$, $\eta^2 = .13$], however, significantly changed from underconfident to overconfident between the pretest and the posttest indicating more confidence on the posttest. There were no significant changes over time for the CR condition ($p = .88$). The results are presented in Table 7.

There were no significant results for students' absolute accuracy (between effect $p = .92$, within effect $p = .73$, interaction effect $p = .38$) over time.

Metacognitive awareness, self-regulated strategy use, and self-efficacy

Another three 3 × 2 (Condition |MC, CR, Control| × Time |pretest, posttest|) repeated measures analysis was performed to examine the effects on students' metacognitive awareness, self-regulated strategy use, and mathematics self-efficacy, respectively.

With the assumption of multivariate sphericity being met (Box's M test = 6.95, $p = .34$), repeated measures ANOVA demonstrated significant increases in students' metacognitive awareness following a significant increasing linear trend, F (1,129) = 5.48, $p < .05$, $\eta^2 = .04$ across the three conditions. Nevertheless, there was no significant difference among the three

Table 7 Results for ANOVA repeated measure for confidence bias

Effect	*MS*	*df*	*F*	*p*	η^2
Condition	0.10	2	1.64	.20	.03
Time	0.13	1	7.14	< .01	.05
Condition×Time	0.04	2	2.32	.10	.04
Error	0.02	129			

Note. The nonparametric Wilcoxon Signed ranks tests demonstrated the similar results that the changes in confidence bias were significant for the control ($Z = -2.01$, $p < .05$) and the CR + MI condition ($Z = -2.38$, $p < .05$).

conditions ($p = .23$) indicating all students improved in their metacognitive awareness. In contrast, there were no significant differences among conditions overtime in students' self-report self-regulated strategy use or their mathematics self-efficacy.

Research question 3 How do 7th grade students justify their performance judgments and do students' performance justifications predict their math performance and monitoring accuracy?

For our third research question, we were interested in investigating the factors that students considered when making confidence ratings. The coding scheme captured various dimensions that students reported including categories of person-related characteristics, item-related characteristics, context-related characteristics, and an unknown category. We also explored how many factors students reported when arriving their confidence ratings. Overall, most students only considered a single factor with only a few students considering multiple factors. Specifically, across the six items, students considered from zero to four factors, but only one student considered four factors when justifying their rating.

Among the ten identified factors, item-related categories were more represented than other justifications. For instance, more than half of the students (pretest: 66.67%, $n = 88$; posttest: 63.64%, $n = 84$) made their performance judgments based on the extent to which they knew about item-relevant mathematical conceptual knowledge. Other than item-relevant knowledge, a number of students considered problem-solving procedures as the attribution for their performance judgments (pretest: 28.03%, $n = 37$; posttest: 19.70%, $n = 26$).

Within the person-related characteristics, most students considered prior knowledge as an important factor when justifying their confidence judgments (pretest: 18.94%, $n = 25$; posttest: 21.97%, $n = 29$). In contrast, another person-related factor, effort, was considered by very few students (pretest: 0.76%, $n = 1$; posttest: 0.76%, $n = 1$) when judging their performance. For context-related categories, students considered the format of their responses for pretest justifications (16.67%, $n = 22$).

While many students provided specific justifications for their confidence judgments, some students guessed their answers and did not know whether they were correct (pretest: 27.27%, $n = 36$; posttest: 18.94%, $n = 25$). Table 2 presents the frequencies of the justification categories across pre- and posttest.

We further selected the most frequently cited category from posttest for each of the three dimensions (i.e., person-related, item-related, and context-related) as well as the unknown category. We then performed multiple regression tests to examine the extent to which the four selected attribution categories (i.e., prior knowledge, item-relevant knowledge, formatting, and

Table 8 Multiple regression results for justification Item 1 for math performance

Step and predictor variable	*B*	*SE B*	β	*t*	*p*
Step 1:					
Pretest math scores	.99	.21	.40	4.80	.00
Step 2:					
Prior knowledge	.09	.10	.08	0.93	.36
Item-relevant knowledge	−.01	.08	−.01	−0.16	.87
Format	−.18	.25	−.06	−0.71	.48
Unknown	−.03	.12	−.02	−0.23	.83

unknown) predicted students' math performance and monitoring accuracy (i.e., two indices: confidence bias and absolute accuracy) for the three justification items on the posttest.

Confidence bias and absolute accuracy indices were calculated for each of the three items. Students' pretest math scores were entered at Step 1, prior knowledge, item-relevant knowledge, formatting, and the unknown category for each item were entered at Step 2.

Mathematics performance

Results showed that the multiple regression models for math posttest scores were significant for all three justification items. Specifically, students' consideration of prior knowledge, item-relevant knowledge, formatting, and their uncertainty (i.e., the unknown category) significantly predicted their performance on the justification items (Item 1: F (5, 126) = 5.08, $p < .001$, $R^2_{adj} = .14$; Item 2: F (5, 126) = 5.03, $p < .001$, $R^2_{adj} = .13$; Item 3: F (5, 126) = 2.51, $p < .05$, $R^2_{adj} = .06$).

There were no significant individual predictors for Item 1 ($ps > .36$) and Item 3 ($ps > .06$). In comparison, prior knowledge and the unknown category were significant predictors ($ps < .05$) for students' math performance on Justification Item 2. Specifically, students' consideration of prior knowledge positively predicted their performance on Item 2; and their uncertainty negatively predicted their performance. This indicated that students who took prior knowledge into consideration when judging their performance were likely to perform well on the item, and those who were not sure about their judgments were likely to perform poorly. See Tables 8, 9 and 10 for the reports of the regression models.

Confidence Bias

We next performed regression tests to examine the extent to which the selected justification categories predicted students' confidence bias. Specifically, for students' confidence bias, the regression models were overall significant for Item 1 [F (5, 126) = 2.67, $p < .05$, $R^2_{adj} = .06$] and marginally significant for Item 3 [F (5, 126) = 2.25, $p = .05$, $R^2_{adj} = .05$]. The model was not significant for Item 2 ($p = .64$). The unknown category was a significant predictor ($p < .05$) for both Item 1 and 3 ($p < .05$). In particular, the unknown category negatively predicted students' bias scores, which indicated that students who were uncertain when judging their performance tended to be overconfident. These regression results are presented in Tables 11 and 12.

Table 9 Multiple regression results for justification Item 2 for math performance

Step and predictor variable	*B*	*SE B*	β	*t*	*p*
Step 1:					
Pretest math scores	.49	.22	.18	2.21	.03
Step 2:					
Prior knowledge	.27	.13	.17	2.06	.04
Item-relevant knowledge	.13	.09	.13	1.52	.13
Format	.41	.45	.08	0.92	.36
Unknown	−.33	.14	−.20	−2.39	.02

Table 10 Multiple regression results for justification Item 3 for math performance

Step and predictor variable	*B*	*SE B*	β	*t*	*p*
Step 1:					
Pretest math scores	.44	.18	.22	2.49	.01
Step 2:					
Prior knowledge	−.11	.11	−.09	−1.06	.29
Item-relevant knowledge	−.04	.07	−.06	−0.61	.55
Format	.05	.17	.03	0.28	.78
Unknown	−.21	.11	−.17	−1.89	.06

Absolute accuracy

We performed another three regression models for absolute accuracy with the same four justification categories. None of the models were significant (*ps* > .08). This indicated that the selected justification categories did not predict students' absolute accuracy.

Discussion

In this study, we examined the extent to which Chinese middle school students' monitoring accuracy and mathematics achievement changed through intervention. Absolute accuracy and confidence bias were adopted as the indices of students' monitoring accuracy. We examined the extent to which students' metacognitive awareness, self-regulatory strategy use, and self-efficacy were associated with students' monitoring accuracy and mathematics performance. Finally, we investigated students' justifications for their metacognitive judgments.

Overall, this study contributes to monitoring literature in at least three ways. First, to our knowledge, this is the first monitoring intervention study that targets Chinese middle school students in the domain of mathematics in a school setting. Though monitoring has been widely studied with college students in other domains (e.g., reading comprehension) in Western countries, it remains valuable to extend the research to younger populations in an Asian country.

In general, Asian students have relatively low self-efficacy when compared to Western students (Eaton and Dembo 1997; Klassen 2004), which may be explained by cultural differences in the sources of self-efficacy (Bandura 1994). For example, the target Chinese sample in the present study demonstrated different patterns in regard with the sources of self-efficacy in mathematics when compared to a similar American sample (Usher and Pajares

Table 11 Multiple regression results for justification Item 1 for confidence bias

Step and predictor variable	*B*	*SE B*	β	*t*	*p*
Step 1:					
Pretest math scores	−.54	.22	−.21	−2.45	.02
Step 2:					
Prior knowledge	−.05	.11	−.04	−0.50	.62
Item-relevant knowledge	.07	.08	.08	0.88	.38
Format	−.14	.26	−.05	−0.55	.58
Unknown	−.30	.12	−.22	−2.43	.02

Table 12 Multiple regression results for justification Item 3 for confidence bias

Step and predictor variable	*B*	*SE B*	β	*t*	*p*
Step 1:					
Pretest math scores	.06	.20	.03	.29	.77
Step 2:					
Prior knowledge	.20	.12	.15	1.69	.09
Item-relevant knowledge	.03	.08	.04	-0.38	.71
Format	−.23	.19	−.11	−1.19	.24
Unknown	−.25	.12	−.19	−2.02	.05

2009). Students in both samples were given the same *Middle School Mathematics Self-Efficacy Scale* to assess students' self-efficacy. Data collected from the target Chinese students reported lower scores of mastery experience and social persuasion than did American students as reported in Usher and Pajares (2009). Social persuasion refers to the encouragement students receive regarding their performance from others, such as teachers and parents (Bandura 1994). Thus, the lower scores on social persuasion in the current sample may be attributed to Chinese teachers or parents' high standards and expectations when evaluating students' academic progress and achievement (Kifer 2002; Kifer and Robitaille 1989; Salili 1996), which may lead to less frequent persuasion. Such differences in self-efficacy may affect students' performance judgments as well as the effects of interventions. The included psychological measures (i.e., Mathematics Self-Efficacy, Jr.MAI, and SRSI-SR) were translated in Chinese Mandarin and demonstrated high reliabilities.[1] The administration of these measures in Chinese Mandarin advances their implemental value and generalizability.

Second, findings indicated that 7th grade students were slightly inaccurate in judging their performance, as reflected by the indices of absolute accuracy and confidence bias. While this finding does not directly correspond to previous research in monitoring, which found students were overconfident and inaccurate in monitoring across age groups (e.g., college students: Dunlosky and Rawson 2012; high school students: Bol et al. 2012; primary school students: van Loon et al. 2013), the negative associations found between monitoring indices and mathematics performance is consistent with previous research. That is, students' overconfidence and low accuracy were associated with poor performance. This common finding was consistent across the pretest and posttest in the present study, which confirms the critical role of monitoring in students' academic achievement.

Last, we further examined students' justifications reflecting upon what factors they considered when making performance judgments on mathematics tasks. Such considerations were previously explored in text comprehension-related tasks (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998; Wang and List 2019). The present study extended this exploration to mathematics problem-solving tasks in order to investigate potential factors that 7th grade students consider when making confidence ratings. For instance, Dinsmore and Parkinson (2013) identified that college students considered multiple factors when forming performance judgments. However, the present study found that middle schoolers tended to only take a single factor into consideration when rendering performance judgments in mathematics.

Research question 1 Associations among Absolute accuracy, Confidence Bias, Mathematics Performance, and Other Psychological Measures.

[1] The translated instruments are available upon request.

Our first research question examined associations among key constructs including students' absolute accuracy, confidence bias, mathematics performance, metacognitive awareness, mathematics self-efficacy, and self-regulated learning strategy use. To do so, we asked students to complete a set of psychological measures and to rate their confidence on a 10-point scale after they completed each math item, indicating how confident they were about their given answers.

Interestingly, the mean scores of students' confidence bias and absolute accuracy showed that students tended to be just slightly both underconfident and inaccurate in general, across pretest and posttest. This is distinct from most prior work that reported students are inaccurate monitors and they generally overestimate their performance relative to their actual performance (e.g., Dunning et al. 2003). This inconsistent finding may be due to the ease of the math items included in the current study, as students tend to be more accurate on easier items when compared to more difficult items (Nietfeld et al. 2005; Schraw and Roedel 1994). Moreover, from a cultural perspective, Klassen (2004) suggested that Asian individuals tended to be realistic and generated accurate judgments of their abilities when compared to Western students. Our finding, perhaps, demonstrated a cultural distinction in making performance judgments.

Furthermore, the negative association between mathematics performance and confidence bias showed that students who were overconfident about their actual performance performed poorly on the mathematics tests. This finding is consistent with previous research, which demonstrated low-achieving students are often overconfident relative to their actual performance (e.g., Bol and Hacker 2001; Labuhn et al. 2010; Stone and Opel 2000). Another negative association between students' mathematics scores and absolute accuracy indicated that low-achieving students had low absolute accuracy in monitoring. This finding is also consistent with previous research in other contexts, which has found that higher absolute accuracy is associated with improved objective academic performance (e.g., Hadwin & Webster 2013). These common findings about students' overconfidence and low absolute accuracy leading to less ideal academic performance may be explained by the theoretical framework of metacognition. Specifically, with deficits in accurate monitoring, students are likely unable to detect errors and employ strategies to remedy the corresponding obstacles that they may encounter during task completion (Butler and Winne 1995; Flavell 1979).

Finally, for the pretest, students' absolute accuracy was negatively and moderately associated with mathematics self-efficacy, metacognitive awareness, and self-regulatory strategy use. This finding showed that students who inaccurately monitored their performance were likely to have low self-efficacy about mathematics, low metacognitive awareness, and deficits in using self-regulatory strategies. These findings correspond to theories of self-regulated learning (Winne and Hadwin 1998) and metacognition (Schraw and Moshman 1995), which support that inaccurate monitoring hinders students' awareness of cognition and regulation, resulting in failure to enact appropriate strategies.

Research question 2 The Effects on Students' Mathematics Performance, Monitoring Accuracy, Metacognitive Awareness, Self-regulated Strategy Use, and Self-Efficacy.

We further examined the intervention effects on students' mathematics performance, confidence bias, absolute accuracy, and other psychological outcomes (i.e., metacognitive awareness, self-efficacy, and self-regulatory strategy use). Results suggested that students in the control condition significantly decreased in their mathematics performance overtime. This was not indicated in the other two conditions. Further, although students in the control condition and the CR + MI condition slightly increased their confidence bias over time, there was no significant effect for the CR condition. Students across the three conditions also

showed increases in metacognitive awareness, while there were no significant changes in students' self-efficacy and self-regulatory strategy use.

While there were no increases for the CR and the CR + MI conditions in mathematics performance, the significant decrease for the control condition may indicate that the posttest was more difficult than the pretest and the students did benefit from the intervention conditions (i.e., CR and the CR + MI conditions).This finding deviates from a previous monitoring study by Nietfeld et al. (2006) where students who received feedback and verbal instructions about monitoring produced better learning results. The insignificant increases in the CR and the CR + MI conditions may be due to the low intensity of the intervention in the present study. Specifically, the intervention was delivered in a written format without teachers' verbal directions for performing accurate monitoring. Although we found positive effects for written directions with middle students in United States in a previous study (Sperling et al. 2012), this may not be the case in the current sample. An inclusion of teachers' explicit monitoring instructions may be a focused modification for future interventions.

In addition, in terms of the intervention intensity, one may argue that the insignificant effects may be due to the low frequency and short duration of the intervention. Specifically, students in the present study received three intervention sessions with one session per week. This dosage may not be powerful enough to improve students' monitoring and academic performance. Interestingly, previous monitoring interventions have demonstrated mixed results regarding intervention frequency and duration. For instance, Bol et al.'s (2012) intervention included one treatment session only resulting in students' improved monitoring and academic performance, whereas Huff and Nietfeld's (2009) monitoring intervention included multiple treatment sessions over weeks resulting in insignificant effects on students' performance. A recent systematic review that examined the effective characteristics of mathematics SRL interventions reported no consistent patterns in the nature of effective versus ineffective interventions when comparing effect sizes in learning and monitoring outcomes (Wang and Sperling 2020). Thus, other factors may also likely contribute to the insignificant effects, such as students' ineffective use of the intervention materials.

Furthermore, as students were accurate monitors before the intervention, the intervention effect on students monitoring accuracy were therefore weak. This finding corresponds to Bol et al. (2005). Specifically, Bol and colleagues' work, in which they explicitly asked college students to make predictions and postdictions about their performance as the overt monitoring condition. They reported no increases in students' monitoring accuracy for the overt condition when compared to the control condition. The finding in the present study may indicate the ineffectiveness of monitoring instructions, especially when students are good monitors already.

Moreover, all participating students' metacognitive awareness was improved between pre- and posttest measured by the Jr.MAI. This finding was a little surprising as we only expected main increases in metacognitive awareness for the CR and CR + MI conditions. This may be that the exposure to the stems of the Jr.MAI items served as an intervention and led students to think metacognitively.

Research question 3 Factors Students Considered When Justifying Their Metacognitive Judgments.

Our final research question aimed to understand students' rationales for making performance judgements, which may inform avenues to improve future monitoring interventions.

We expected that students would consider a variety of factors when judging their performance. Nevertheless, most students only considered one factor even though multiple categories were identified among all students. This is not consistent with the prior research that investigated the formation of metacognitive judgments when reading texts (Dinsmore and Parkinson 2013; Lin and Zabrucky 1998; Wang and List 2019). In the previous work, learners were found to consider multiple dimensions or factors when rendering their performance judgments. From a developmental perspective, this inconsistency is likely a result of 7th grade students' limitations in metacognitive awareness that may prevent them from thinking holistically (Flavell et al. 1995). Metacognition develops as students gain exposure to increasing metacognitive and educational experiences (Flavell 1976,1979). In particular, Baker and Brown (1984) suggested developmental differences between child and adult readers, in which children tend to be less aware of their reading processes when compared to college students. Such developmental limitations for children' metacognition in reading also apply to mathematics. For instance, Shilo and Kramarski (2019) examined fifth grade students' metacognitive processes in mathematics through qualitative analyses of recorded math classroom videos. They reported that fifth graders tended to have difficulties verbalizing and justifying their learning processes, further indicating middle school students' limited metacognition. As such, this may explain the Chinese middle school students' limited justification factors for their performance judgments in the present study.

Furthermore, students commonly considered only item-related characteristics. This focus may be due to the perceived nature of mathematics. Unlike open-ended items in other domains (e.g., social science), which students can compose their responses in different ways, these items required specific knowledge and objective answers. In contrast, previous studies have demonstrated students' considerations of multiple factors when judging their performance in a writing composition task based on multiple texts (List and Alexander 2015; Wang and List 2019).

As Mosenthal (1998) suggested, processing reading text tasks requires students to be able to integrate and connect inferential information from texts and match the information found from the texts with the task questions. It may be difficult for students to produce high-quality responses when readings are lengthy and are not cohesive. In contrast, mathematics items tend to avoid syntactic complexity and commonly ask for one objective answer (Martiniello 2009). Such characteristics of math items may directly activate the targeted mathematical knowledge necessary for solving the particular item. This may explain the differences in justifications of performance on essay questions and math problem-solving items. In addition to the prevalence of item-related factors that students considered, a person-related factor (i.e., prior knowledge) was also found to be significant for predicting students' math performance. This finding also suggests the objectivity and specificity of mathematics problem-solving items, that is, they require specific item knowledge as well as specific prior task experience. Thus, students' various perceptions of the task context may be crucial for them to activate relevant knowledge and use strategies, which perhaps help students to justify their performance judgements (List et al. 2019; Wang and List 2019; Wright 1981).

In addition, the Chinese culture may also affect how students justify their performance judgments. For instance, Lundeberg et al. (2000) examined college students' monitoring internationally across five regions including the United States, the Netherlands, Israel, Palestine, and Taiwan. Specifically, students were asked to judge their performance on multiple course exams including subjects such as mathematics, biology, and psychology. Findings suggested that college students from Taiwan demonstrated more accurate monitoring and

underconfidence when compared to students from the United States, the Netherlands, Israel, and Palestine. Such findings may suggest cultural components in students' metacognitive judgments and the factors they consider during monitoring. For example, the justification data in the present study demonstrated that Chinese seventh grade students were mostly accurate about their prior knowledge for a certain item when justifying their performance judgements. In particular, students who justified their performance judgments based on accurate perceptions of their prior knowledge tended to also judge their performance accurately. However, the present study was only conducted with Chinese students and lacked a comparison sample. Future research should further explore potential cultural differences in middle school students' justification factors when making performance judgments in mathematics.

Furthermore, the unknown category negatively predicted students' performance and confidence bias for Item 2 and 3. Specifically, students who were uncertain about how they arrived at their performance judgments were likely to perform poorly on the corresponding item and feel underconfident about their performance. These findings suggest the importance of supporting middle school students' metacognition. Empirical evidence indicates that providing students with metacognitive guidelines and feedback is a viable avenue to improve students' metacognitive awareness as well as monitoring accuracy (e.g., Miller and Geraci 2011b; Nietfeld and Schraw 2002; Shilo and Kramarski 2019). Additional research should explore effective metacognitive guidelines for middle school students in mathematics.

Moreover, as the items in the present study were generally easy, future interventions that include high difficulty math items may unmask more information about possible factors that Chinese middle school students may consider when making performance judgments. Asking students to explain their performance judgments may also prove to be a viable instructional strategy to support students' metacognition and accurate monitoring. In a seminal review, Schraw (1998) encouraged the employment of a strategy evaluation matrix (SEM) and a regulatory checklist (RC) as strategies to promote students' knowledge of cognition and regulation of cognition. Teachers who implemented these strategies found positive increases in students' metacognition. Moreover, according to recent studies that investigated the attributions of performance judgments, teaching students about the multidimensional structure of monitoring may be another approach to improve students' metacognitive awareness and monitoring accuracy. For instance, teachers or educators can guide and encourage students to consider the variated person-, item-, and context-related factors that may influence their performance judgements.

Conclusions and implications

Findings from this intervention study indicated overall Chinese middle school students are accurate in monitoring in mathematics. Consistent with the extant monitoring literature, overconfidence and poorer accuracy were associated with lower mathematics performance. The study explored potential benefits of two interventions designed to scaffold students' metacognition. Findings indicated some support for the interventions but also indicated the potential instrumentation concerns. Given that the control group mathematics scores significantly decreased between pre- and post-measures, we explored potential differences in performance on the pre and post mathematics measures. Findings indicated the posttest was significantly more difficult than the pretest ($z = 12.77$, $p < .001$). As neither of the intervention

scores significantly decreased but the control condition scores did, the viability of the interventions to improve monitoring needs to be further explored.

Further, findings demonstrated consistent relationships among students' math performance, confidence bias, metacognitive awareness, and self-efficacy corresponding to previous studies in mathematics (e.g., Labuhn et al. 2010; Nietfeld and Schraw 2002; Usher and Pajares 2009). Specifically, students' improved mathematics performance was found to be associated with low confidence bias, high metacognitive awareness, and high self-efficacy. While the consistent relationships among these constructs indicates the importance of the given psychological constructs to students' mathematics achievement, one challenge is this stable relationship may result in resistance to intervention. Further investigation into the complexity of the potential effects of interventions that target one or more of these constructs is needed for future research.

Also, for further consideration for future research is the dosage and delivery of metacognitive interventions. In the current study, paper and pencil written interventions did not successfully improve 7th grade students' monitoring. While there are relatively few metacognitive intervention studies for middle school mathematics, Kramarski (2004) reported benefit for a verbal instruction intervention for junior high school students. Consist with recommendations for strategy instruction (e.g., prompt students to use appropriate strategies for problem solving questions) more and longer exposure is likely necessary to realize long term benefits for children's metacognition. Future intervention research should carefully consider dosage implications. Like this study, conducted in classrooms and interventions were administered by teachers, additional future intervention research should strive to further include teachers to integrate even more seamlessly into classroom practice.

In this study we also examined the justifications students made for their performance judgments. Findings indicate that middle school students generally consider only a single factor as they form their performance judgments. These factors were grouped into person-, item- and context factors. Future research should continue to examine the rationales students use to inform their performance judgment to both inform future research but also to inform instructional practice. Although students' performance scores and the monitoring indices demonstrated high monitoring accuracy, these justification categories reveal potential deficits in metacognitive awareness. Continued investigation into effective interventions for improving middle school students' monitoring accuracy in mathematics are needed.

Compliance with ethical standards The authors declare that they have no conflicts of interest and that there was no funding source to declare in association with this project. All human subjects were consented for their participation in the study.

Appendix A

	Sample justifications
Control condition	
Prior knowledge	"I've done this before, but I don't remember exactly"
	"I've learned about similar items"
	"I am familiar about these kinds of items"
	"Because I've learned about the sum of interior for a triangle is 180°…"
Item-relevant knowledge	"Convert the denominators into the same"
	"I can use equations"
	"It was based on the sum of interior angles for quadrilateral and the theorem"
	"I didn't think of a proper way to solve this problem"
Format	"The work could have been written with more details"
	"I think the language I used to describe was not precise"
	"My answer should be correct, but the work shown might be wrong"
	"It was not written well"
Unknown	"I am not sure about whether my answer was correct"
	"I don't know how to solve it…"
	"Feeling correct"
	"I feel it was right, but what if I got it wrong"
CR condition	
Prior knowledge	"My teacher taught this before, and I've practiced"
	"I've learned about similar items"
	"Because my elementary and middle school taught about this. I have done many times"
	"My tutor taught about this"
Item-relevant knowledge	"Because the sum of one trapezoid is 360, two are 720"
	"I know about the knowledge for this item"
	"Use linear equation"
	"Use x to solve this equation"
Format	"I didn't show my work"
	"The format may be wrong"
	"I may have written my work in wrong format"
Unknown	"Feeling"
	"Maybe"
	"I don't how to solve it, so I just scribbled"
	"I am not sure"
CR + MI condition	
Prior knowledge	"I've learned this before"
	"It was taught before"
	"I've done similar items"
	"done it"
Item-relevant knowledge	"Because I used the formula"
	"Because I don't know about this specific knowledge exactly"
	"Common denominator"
	"The sum of interior angles for a triangle is 360 and it is 720 for a quadrilateral"
Format	–
Unknown	"There might be another answer that I didn't think of"
	"Not very sure"
	"I don't know"
	"I am not sure about this item. It might be a different answer"

References

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson & M. L. Kamil (Eds.), *Handbook of reading research* (pp. 353–394). New York: Longman.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior* (Vol. 4, pp. 71–81). New York: Academic Press.

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education, 69*(2), 133–151. https://doi.org/10.1080/00220970109600653.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269–290.

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology, 37*(4), 280–287. https://doi.org/10.1016/j.cedpsych.2012.02.004.

Brown, A. (1987). Metacognition, executive control, self-regulation and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65–116). Hillsdale: Lawrence Erlbaum.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.

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

Cleary, T. J. (2006). The development and validation of the self-regulation strategy inventory–self-report. *Journal of School Psychology, 44*, 307–322. https://doi.org/10.1016/j.jsp.2006.05.002.

Cleary, T. J., & Chen, P. P. (2009). Self-regulation, motivation, and math achievement in middle school: Variations across grade level and math context. *Journal of School Psychology, 47*(5), 291–314. https://doi.org/10.1016/j.jsp.2009.04.002.

Dabarera, C., Renandya, W. A., & Zhang, L. J. (2014). The impact of metacognitive scaffolding and monitoring on reading comprehension. *System, 42*, 462–473. https://doi.org/10.1016/j.system.2013.12.020.

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgements made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001.

Dinsmore, D. L., Loughlin, S. M., Parkinson, M. M., & Alexander, P. A. (2015). The effects of persuasive and expository text on metacognitive monitoring and control. *Learning and Individual Differences, 38*, 54–60. https://doi.org/10.1016/j.lindif.2015.01.009.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.

Dunlosky, J., Hertzog, C., Kennedy, M. R. F., & Thiede, K. W. (2005a). The self-monitoring approach for effective learning. *Cognitive Technology, 10*(1), 4–11.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005b). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235.

Eaton, M. J., & Dembo, M. H. (1997). Differences in the motivational beliefs of Asian American and non-Asian students. *Journal of Educational Psychology, 89*(3), 433–440.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing, 1*(2), 155–182. https://doi.org/10.1207/S15327574IJT0102_4.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence*. Hillsdale: Erlbaum.

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

Flavell, J. H., Green, F. L., Flavell, E. R., Harris, P. L., & Astington, J. W. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Development, 60*, 1–113.

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*(2), 139–170. https://doi.org/10.1007/s11409-015-9139-1.

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16*(3), 163–176.

Gillström, Å., & Rönnberg, J. (1995). Comprehension calibration and recall prediction accuracy of texts: Reading skill, reading strategies, and effort. *Journal of Educational Psychology, 87*(4), 545–558. https://doi.org/10.1037/0022-0663.87.4.545.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93. https://doi.org/10.3758/BF03197714.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116*(2), 119–136. https://doi.org/10.1037/0096-3445.116.2.119.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5–23. https://doi.org/10.3102/0013189X024008005.

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction, 44*, 1–10. https://doi.org/10.1016/j.learninstruc.2016.02.006.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160.

Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgements of confidence. *Learning and Instruction, 24*, 37–47. https://doi.org/10.1016/j.learninstruc.2012.10.001.

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

Huff, J., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgements to improve metacognitive monitoring. *Metacognition and Learning, 4*(2), 161–176. https://doi.org/10.1007/s11409-009-9042-8.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychological Review, 100*, 609–639. https://doi.org/10.1016/0001-6918(91)90036-Y.

Kifer, E. (2002). Students' attitudes and perceptions. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 251–275). New York: Springer.

Kifer, E., & Robitaille, D. (1989). Attitudes, preferences and opinions. In D. Robitaille & R. A. Garden (Eds.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics* (pp. 178–202). London: Pergamon Press.

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high- and low-achieving primary school children's reading. *Learning and Instruction, 5*, 143–165. https://doi.org/10.1016/0959-4752(95)00009-R.

Klassen, R. M. (2004). Optimism and realism: A review of self-efficacy from a cross-cultural perspective. *International Journal of Psychology, 39*(3), 205–230. https://doi.org/10.1080/00207590344000330.

Kramarski, B. (2004). Making sense of graphs: Does metacognitive instruction make a difference on students' mathematical conceptions and alternative conceptions? *Learning and Instruction, 14*, 593–619. https://doi.org/10.1016/j.learninstruc.2004.09.003.

Kramarski, B., & Gutman, M. (2006). How can self-regulated learning be supported in mathematical E-learning environments? *Journal of Computer Assisted Learning, 22*, 24–33. https://doi.org/10.1111/j.1365-2729.2006.00157.x.

Labuhn, A. S., Zimmerman, B., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning, 5*(2), 173–194. https://doi.org/10.1007/s11409-010-9056-2.

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

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implication for education and instruction. *Contemporary Educational Psychology, 23*(4), 345–391. https://doi.org/10.1006/ceps.1998.0972.

Lin, L. M., Moore, D., & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*(2), 111–128. https://doi.org/10.1080/02702710119125.

Lin, L. M., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology, 115*(2), 187–198.

List, A., & Alexander, P. A. (2015). Examining response confidence in multiple text tasks. *Metacognition and Learning, 10*, 407–436. https://doi.org/10.1007/s11409-015-9138-2.

List, A., Du, H., & Wang, Y. (2019). Understanding students' conceptions of task assignments. *Contemporary Educational Psychology, 59,* 1–16. https://doi.org/10.1016/j.cedpsych.2019.101801.

van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction, 24*, 15–25. https://doi.org/10.1016/j.learninstruc.2012.08.005.

Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on confidence: Country and gender. *Journal of Educational Psychology, 92*(1), 152–159. https://doi.org/10.1037/0022-0663.92.1.152.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*, 160–179. https://doi.org/10.1080/10627190903422906.

Miller, T. M., & Geraci, L. (2011a). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 502–506. https://doi.org/10.1037/a0021802.

Miller, T. M., & Geraci, L. (2011b). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7.

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal, 35*(2), 269–307.

Nagel, M., & Lindsey, B. (2018). The use of classroom clickers to support improved self-assessment in introductory chemistry. *Journal of College Science Teaching, 47*(5), 72–79.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item comments on Schraw (1995). *Applied Cognitive Psychology, 10*(3), 257–260.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131–142. https://doi.org/10.1080/00220670209596583.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*(1), 7–28.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006a). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy and self-efficacy. *Metacognition and Learning, 1*(2), 159–179. https://doi.org/10.1007/s10409-006-9595-6.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006b). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*(2), 258–271. https://doi.org/10.1177/0013164404273945.

Ozuru, Y., Kurby, C. A., & McNamara, D. S. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning, 7*(2), 113–131. https://doi.org/10.1007/s11409-012-9087-y.

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Charlotte: Information Age.

Pieschl, S. (2009). Metacognitive calibration-an extended conceptualization and potential applications. *Metacognition and Learning, 4*(1), 3–31. https://doi.org/10.1007/s11409-008-9030-4.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23*(4), 454–464 https://www.jstor.org/stable/747643.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*, 19–33. https://doi.org/10.1207/s15326985ep2501_3.

Programme for International Student Assessment (PISA). (2018). *PISA 2018 translation and adaptation guidelines*. Prague: OECD Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2018-TRANSLATION-AND-ADAPTATION-GUIDELINES.pdf.

Ramdass, D., & Zimmerman, B. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics, 20*(1), 18–41.

Salili, F. (1996). Achievement motivation: A cross-cultural comparison of British and Chinese students. *Educational Psychology, 16*(3), 271–279.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp. 334–370). New York: Macmillan.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*, 113–125. https://doi.org/10.1023/A:1003044231033.

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

Schraw, G., & Nietfeld, J. (1998). A further test of the general monitoring skill hypothesis. *Journal of Educational Psychology, 90*(2), 236–248. https://doi.org/10.1037/0022-0663.90.2.236.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22*(1), 63–69. https://doi.org/10.3758/BF03202762.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*(4), 455–463. https://doi.org/10.1006/ceps.1993.1034.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*(3), 433–444. https://doi.org/10.1037/0022-0663.87.3.433.

Shilo, A., & Kramarski, B. (2019). Mathematical-metacognitive discourse: How can it be developed among teachers and their students? Empirical evidence from a videotaped lesson and two case studies. *ZDM, 51*(4), 625–640. https://doi.org/10.1007/s11858-018-01016-6.

Shiu, L. P., & Chen, Q. (2013). Self and external monitoring of reading comprehension. *Journal of Educational Psychology, 105*(1), 78–88. https://doi.org/10.1037/a0029378.

Silver, E. A. (1987). Foundations of cognitive theory and research for mathematics problem-solving instruction. In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 33–60). Hillsdale: Lawrence Erlbaum.

Silver, E. A. (1994). On mathematical problem posing. *For the Learning of Mathematics, 14*(1), 19–28.

Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155–172. https://doi.org/10.1080/00220973.2016.1143794.

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

Sperling, R. A., Ramsay, C. M., Richmond, A. S., Nietfeld, J. L., Reeves, P. M., & Hood, A. M. (2012). *General monitoring and instructional scaffolds that support metacognition in middle school students*. Vancouver, BC: Presentation at the annual meeting of the American Educational Research Association.

Stodolsky, S. S., & Grossman, P. L. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal, 32*(2), 227–249.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83*(2), 282–309. https://doi.org/10.1006/obhd.2000.2910.

TIMSS (2011). *Assessment. International Association for the Evaluation of Educational Achievement (IEA)*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands.

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology, 34*(1), 89–101. https://doi.org/10.1016/j.cedpsych.2008.09.002.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*(3), 435–437. https://doi.org/10.1037/0022-0663.81.3.435.

Wang, Y., & List, A. (2019). Calibration in multiple text use. *Metacognition and Learning, 14*(2), 131–166. https://doi.org/10.1007/s11409-019-09201-y.

Wang, Y., & Sperling, R. A. (2020). Characteristics of effective self-regulated learning interventions in mathematics classrooms: A systematic review. *Frontiers in Education, 5*(58). https://doi.org/10.3389/feduc.2020.00058.

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*(4), 173–187.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Erlbaum.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah: Erlbaum.

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

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

Wright, G. N. (1981). Cultural and task influences on decision making under uncertainty. *Current Anthropology, 22*(3), 290–291. https://doi.org/10.1086/202669.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D. F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes, 43*(2), 145–171. https://doi.org/10.1016/0749-5978(89)90048-4.

Zabrucky, K. M., Agler, L. M. L., & Moore, D. (2009). Metacognition in Taiwan: Students' calibration of comprehension and performance. *International Journal of Psychology, 44*(4), 305–312. https://doi.org/10.1080/00207590802315409.

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling, 53*(1), 108–127.