Research Article

Maria Klar*, Josef Buchner, Michael Kerres

Limits of Metacognitive Prompts for Confidence Judgments in an Interactive Learning Environment

https://doi.org/10.1515/edu-2022-0209
received July 13, 2023; accepted November 3, 2023

Abstract: Metacognitive activities are reported to improve learning but prompts to support metacognition have only been investigated with mixed results. In the present study, metacognitive prompts for confidence judgments were implemented in a learning platform to provide more insights into their effectiveness and their limits. Comparing the prompted group (n = 51) with the control (n = 150), no benefits of the prompts are seen: Performance is not better with prompts, and there is no improvement in metacognitive accuracy over time within the prompted group. Notably, half of the prompted group did not use the metacognitive prompts as intended. Alternative ways to integrate such prompts are discussed.

Keywords: metacognitive prompts, metacognitive accuracy, confidence prompts, self-regulated learning, K-12

1 Introduction

In interactive learning environments, students are often required to regulate their learning, for example, in self-paced courses. To many students, especially lower performing students, this can pose a challenge (DiFrancesca, Nietfeld, & Cao, 2016). There are many models of self-regulated learning, and they usually involve that students must plan and monitor their learning to some degree (e.g., Winne, 1997; Zimmerman & Moylan, 2009). *Monitoring* is the observation of one's own thinking and learning behavior, and together with its counterpart, *control*, which is the regulation of that behavior, they are the two key processes of metacognition (Nelson & Narens, 1990). Therefore, metacognition is an integral part of self-regulated learning (Dinsmore, Alexander, & Loughlin, 2008).

As part of monitoring, learners assess whether they have already understood a concept and can move on, or whether they should review the material. Monitoring helps students find sweet spots in their learning process: Overconfident students invest too little time and effort, while underconfident students might invest too much time (Son & Metcalfe, 2000). Monitoring helps the learner to control their learning behavior effectively, and thus, strengthening metacognition might result in improved performance (Ohtani & Hisasaka, 2018). If we can help learners by supporting their metacognitive skills and activities, the question is what shape this support could have and how this support could be implemented in learning environments. One possible measure of support is prompts for confidence judgments that might help students not to be over- or underconfident and to regulate their learning process effectively. To date, it is yet to be determined whether such metacognitive prompts have a beneficial effect on learning. Hence, this research examines this evidence gap.

1.1 Metacognitive Accuracy

Monitoring can lead to more or less accurate judgments about cognition. Metacognitive accuracy is the degree of correspondence between confidence and performance (Jang, Lee, Kim, & Min, 2020). It is further distinguished into several aspects of metacognitive accuracy, the most common ones being relative accuracy (resolution) and absolute accuracy (calibration [Jang et al., 2020; Schraw, 2009]). Absolute accuracy is low if a student works on a set of tasks and expresses a high confidence, for example, between 80 and 100%, for these tasks but then scores low. However, for the same confidence levels and results, relative accuracy can

*** Corresponding author: Maria Klar,** Chair of Educational Technology & Instructional Design, University of Duisburg-Essen, Essen, 45141, Germany, e-mail: maria.klar@uni-due.de
Josef Buchner: Institute for ICT and Media, St. Gallen University of Teacher Education, St. Gallen, 9000, Switzerland
Michael Kerres: Chair of Educational Technology & Instructional Design, University of Duisburg-Essen, Essen, 45141, Germany

be high if the results are better on the tasks that had a confidence rating of 100% than on the tasks with a confidence of 80%. Conversely, relative accuracy can be low while absolute accuracy is high. Therefore, these two aspects of metacognitive accuracy are distinct. In assessments, both measures should be used to complement each other (Schraw, 2009), and students ideally should be accurate in both dimensions (Schwartz & Efklides, 2012).

Generally, metacognitive accuracy is found to be far from perfect (Glenberg & Epstein, 1985). Overall, students tend to be overconfident in predictions of their performance (Dunning, 2011; Hacker, Bol, Horgan, & Rakow, 2000; Maki & Berry, 1984; Miller & Geraci, 2011). More specifically, Hacker et al. (2000) found that high-performing students were rather accurate in predicting and postdicting (i.e., prediction after task completion) their exam scores. The highest-performing group even showed some underconfidence. In contrast, the lower performing groups showed overconfidence which increased as performance decreased (see also Bol, Hacker, O'Shea, & Allen, 2005; Lingel, Lenhart, & Schneider, 2019; Maki & Berry, 1984).

This is in line with the Dunning–Kruger–Effect (Dunning, 2011), which stipulates that less skilled people are overconfident in their self-judgments because their lack of skill keeps them from correctly assessing what they do not know. Highly skilled people show a weak tendency to underestimate their abilities (Schwartz & Efklides, 2012). Their more comprehensive knowledge of a domain enables them to better assess which aspects they might not know yet. However, since research shows that even high-performing K-12 students can be overconfident (Lingel et al., 2019), support structures for metacognition could benefit K-12 students from all achievement levels.

1.2 Metacognitive Prompts

For students to use metacognition effectively, they need the *ability* to monitor and control (Nelson & Narens, 1990), and they also need to *use* these abilities frequently enough (Bannert & Mengelkamp, 2013). Instruction can help establish the ability, while reminders in the form of prompts can help increase the frequency of metacognitive activities (Moser, Zumbach, & Deibl, 2017). Metacognitive prompts can take on different forms. In some studies, students are prompted to verbalize their thoughts and decisions while learning (think-aloud method [Bannert & Mengelkamp, 2008]), or prompts can ask the students to take notes on how they want to plan their learning (Zumbach, Rammerstorfer, & Deibl, 2020). Prompts can also ask the students how confident they feel in their answers (Feyzi-Behnagh et al., 2014). Such a prompt for a confidence judgment requires the learner to use cues in order to assess their own performance (Koriat, 1997). With practice, students should get more fluent in recognizing cues for understanding or a lack thereof. The more often students recognize a lack of understanding, the more often they have the chance to regulate their learning, for example, by rereading the instructions, and thus to improve their performance.

It is reported that metacognitive prompts can be beneficial for learning because they can make students reflect on their learning more often (Sonnenberg & Bannert, 2015), and with practice, confidence judgments can become more accurate (Feyzi-Behnagh et al., 2014). More accurate judgments help the students to regulate their learning more effectively and thus perform better. However, performance does not improve when students are able to monitor their learning process but fail to take the step of regulating their learning behavior (Dunlosky et al., 2021). There is also evidence that metacognitive prompts evoke neither monitoring nor regulation of the learning process (Johnson, Azevedo, & D'Mello, 2011), and in several studies, students did not use the prompts as intended (Bannert & Mengelkamp, 2013; Lingel et al., 2019; Moser et al., 2017). Therefore, evidence on the usefulness of metacognitive prompts is mixed.

2 Research Questions and Hypotheses

First, as described earlier, there is an evidence gap concerning the effectiveness of such prompts for learning. Second, there is ample research that supports the claim that high-performing students are better at assessing their learning than low-performing students. We aim to further test this claim with prompts in an authentic K-12 setting. Third, there is little research on whether prompts can contribute to improved metacognitive accuracy. Therefore, this study is guided by the following research questions:

RQ1: Does the use of metacognitive prompts lead to higher performance?

H1: The group that uses metacognitive prompts performs better than the group that does not.

RQ2: Is performance related to metacognitive accuracy?

H2: Higher-performing students show better metacognitive accuracy than lower-performing students.

RQ3: Does metacognitive accuracy improve with repeated response to metacognitive prompts?

H3: Metacognitive accuracy is better for the last third of the tasks than for the first third of the tasks.

3 Methods

To test these hypotheses in an applied setting, a quasi-experimental study was conducted. Metacognitive prompts were implemented in a learning platform. The prompts asked for confidence judgments and were added to 33 multiple-choice questions from an introductory course on computational thinking. In the experimental group, 51 secondary school students worked through the course with prompts; 150 secondary school students had completed the same course before the prompts were implemented and functioned as a control group. Data on performance and confidence judgments were used to calculate absolute and relative metacognitive accuracy.

3.1 The Platform and the Course

The study was conducted in cooperation with the German learning platform PearUp,[1] which was founded in 2017 and has now merged with the platform eduki.[2] PearUp was a for-profit learning platform where teachers could create interactive material and design interactive courses. Inherent to the learning experience with PearUp was a gamified meta-narrative: Students started their own "start-up business," and whenever they solved tasks, they collected "PearCoins," which they could invest into their start-up. In terms of feedback, "PearCoins" were also a rough quantitative measure of how many tasks they had completed compared to their peers. There was a leaderboard of the three students scoring highest on the indicators of the start-up. As part of the treatment implementation, two dashboards were developed, showing the rate of correct first attempts and metacognitive accuracy. Because these features were implemented in the live product, it was not possible to conduct randomized A/B-Testing but instead the experimental data were collected after implementation, and data from before the implementation were used for control.

The course titled "Introduction to Computational Thinking" was designed by the content creators of PearUp. The average duration was stated as 2h. There were six units on: "Sequences and Algorithms," "While-Loops," "For-Loops," "Comparing While-Loops and For-Loops," "Conditions," and "Nested For-Loops." The course was designed as a fully online, self-paced course. The majority of the tasks were interactive, such as multiple-choice questions and coding tasks.

1 https://www.pearup.de/
2 https://eduki.com/

3.2 Design of the Prompts

As discussed earlier, in experiments, metacognition is often measured through verbalizations made by the learners after completing a set of tasks. However, online monitoring methods correlate more strongly with higher performance than offline measures (Ohtani & Hisasaka, 2018), so in this study, metacognitive prompts were presented online, with each multiple-choice question.

The prompts consisted of four buttons representing a 4-point confidence scale: "sure, rather sure, unsure, no clue." The buttons took the place of the "Submit"-button, so that in order to submit the task, students had to click one of the four buttons. Figure 1 shows the multiple-choice question design with the metacognitive prompt.

3.3 Sampling

Before the implementation of the prompts, 150 students had completed the course already, and their data were used for the control group. No personal data were gathered but teachers had to give a "class name" when they wanted to use the material for their groups. These class names were provided by PearUp, and most of them implied the students' grade levels. Thus, it was inferred that the students in the control group were aged 11–18. The class names also implied that many students accessed this course as part of an extracurricular activity.

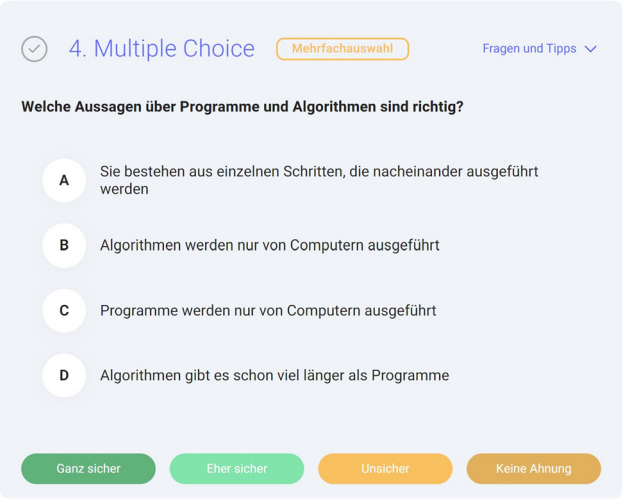

Figure 1: Exemplary MC task with metacognitive prompts below, from the left: "sure," "rather sure," "unsure," and "no clue" – experimental group (screenshot taken by the first author).

Data for the experimental group were gathered in several ways: One of the authors taught the course as part of classes and extracurriculars; 28 students completed the course this way. These were students from grades 7 to 9, so they were aged 12–16. The rest of the data stem from classes which were taught by other teachers. These teachers were recruited through PearUp and call for participation on social media; 23 students from a similar completed the course in this way. The teachers reported a range of ages 13–16. Together, this results in a sample size of 51 for the experimental group.

PearUp complies with the General Data Protection Regulation, and no personal data were gathered. The students were informed that the anonymous interaction data gathered by the platform would be used for research purposes, and they were given the option not to use the material. None of the students chose this option.

3.4 Data Processing and Analysis

PearUp provided the raw data as well as a Jupyter Notebook file with some preliminary pre-processing code written in Python. For further pre-processing, Python code was written in a Jupyter Notebook, using the pandas library. Performance was coded as correct (1.0) and wrong (0.0). Metacognitive judgments were coded as sure (1.0), rather sure (0.66), unsure (0.33), and no clue (0.0).

3.5 Measures for Metacognitive Accuracy

Absolute metacognitive accuracy measures how well students can judge their performance. Schraw (2009, p. 36) suggested the Absolute Accuracy Index:

$$\text{Absolute accuracy index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2. \quad (1)$$

Performance is scored as either 0 or 1. The learners give a confidence rating ranging between 0 and 1. The performance score is subtracted from the confidence score, resulting in a number between −1 and 1. This number is squared, so the result ranges from 0 to 1, with 0 indicating the highest accuracy because the deviation between confidence and performance is the lowest. Division through the number of items results in the mean absolute accuracy.

Relative metacognitive accuracy measures how confidence and performance correlate. Thus, one established measure of relative metacognitive accuracy is Pearson's r (Schraw, 2009).

4 Results

4.1 Descriptive Statistics

4.1.1 Performance

In the control group (n = 150), on average, students solved 52.3% of the tasks on their first attempt (SD = 16.7). In the group with metacognitive prompts (prompted group; n = 51), students solved 54.7% of the tasks on their first attempt (SD = 12.9). According to the Kolmogorov–Smirnov test for normality, performance is not normally distributed in the combined sample of N = 201 (p = 0.011).

4.1.2 Metacognitive Judgments

The highest confidence judgment was given in 88.6% of the tasks. Table 1 gives an overview of the distribution of judgments for all tasks.

For reporting the results, the four confidence judgments are coded ranking from 3.0 = "sure" to 0.0 = "no clue." The average of all confidence judgments is 2.85 (SD = 0.27); 23 of the 51 students only used the button for the highest confidence judgment for all the tasks. When these students are excluded, the average confidence judgment remains high at 2.73.

4.2 Testing the Hypotheses

4.2.1 RQ1: Does the use of metacognitive prompts lead to higher performance?

Because performance values were not normally distributed and because the samples were of unequal size, the non-parametric Mann–Whitney U-test was used (Harwell, 1988; Zimmerman, 1987). The group with metacognitive prompts solved 54.7% of the tasks correctly on their first attempt, while the group without the prompts solved 52.3%

Table 1: Frequency distribution of confidence judgments for all tasks

	Correct	Incorrect	Total	Percentage
Sure	838	653	1,491	88.6
Rather sure	66	77	143	8.5
Unsure	15	26	41	2.4
No clue	1	7	8	0.5
Total	920	763	1,683	

correctly. The exact Mann–Whitney *U*-test yielded no significant difference between the groups (U = 3,591; p = 0.257). This is also the case when only the students who did not use the highest confidence button are considered (performance: 54.4%; U = 1,989; p = 0.33). Therefore, hypothesis 1, stating that the prompted group would outperform the control group, is rejected.

4.2.2 RQ2: Is performance related to metacognitive accuracy?

The sample was split along the median into two groups. Students were ranked by their performance. The lower-performing half (n = 26) was categorized as low-performing students (low). The other half (n = 25) was categorized as high-performing students (high).

4.2.2.1 Absolute Accuracy

The lower-performing group chose slightly lower confidence levels than the higher-performing group ($\bar{x}$ (low) = 2.76, $\bar{x}$ (high) = 2.94), but they performed more poorly on average ($\bar{x}$ (low) = 0.44, $\bar{x}$ (high) = 0.65), and thus, the low-performing group had a lower absolute accuracy (0.49) than the high performing group (0.33). This difference is significant according to the exact Mann–Whitney *U*-test (U = 106, $p < 0.001$).

4.2.2.2 Relative Accuracy

The correlation between confidence and performance is positive, weak, and not significant for the higher-performing group ($r(23) = 0.2$, p = 0.33). It is negative, moderate, and significant for the low-performing group ($r(24) = -0.46$, p = 0.02). This means that in the low-performing group, high confidence is moderately correlated with low performance. The low-performing group has higher confidence judgments when they perform poorly, while the high-performing group has higher confidence judgments when they indeed perform better. Fisher's z transformation shows that the difference is significant (z = 2.689, p = 0.004).

Hypothesis 2 is, therefore, accepted. Low-performing students show a lower absolute accuracy as well as a lower relative accuracy compared to the high-performing group.

4.2.3 RQ3: Does metacognitive accuracy improve with repeated responses to metacognitive prompts?

In order to test the hypothesis that metacognitive accuracy improves over time, it is necessary to define temporal segments of the task events. The students could choose the order of the tasks to some degree, so they did not solve the tasks in exactly the same order. The data were split into three temporal segments. An average was calculated for the first eleven tasks (t1), the second eleven tasks (t2), and the third eleven tasks (t3) for each student.

4.2.3.1 Absolute Accuracy

Absolute accuracy is sensitive to task difficulty, and task difficulty was not standardized here. Still, the results show that absolute accuracy did not improve over time. While confidence levels remained stable from t1 to t3, absolute accuracy varies in accordance with performance, as can be seen in Table 2.

Thus, changes in absolute accuracy can be attributed to changes in performance because confidence levels remain very stable.

4.2.3.2 Relative Accuracy

For each segment, performance values and confidence values were tested for correlation. The results are as follows:

- tasks 1–11 (t1): $r(49) = 0.03$ (p = 0.83)
- tasks 12–22 (t2): $r(49) = 0.033$ (p = 0.81)
- tasks 23–33 (t3): $r(49) = 0.045$ (p = 0.75)

Correlation is very low for all three time segments. There is a marginal increase in relative accuracy, but it is not significant between t1 and t3 ($z = -0.072$, p = 0.47).

Thus, hypothesis 3, stating that metacognitive accuracy would increase, is rejected.

Table 2: Mean performance, confidence, and absolute accuracy across the three time segments

	Tasks 1–11	Tasks 12–22	Tasks 23–33
Mean performance (min = 0, max = 1)	0.686	0.459	0.494
Mean confidence ("no clue" = 0, "sure" = 3)	2.85	2.84	2.86
abs. acc. (min = 1, max = 0)	0.28	0.50	0.47

5 Discussion

The aim of this study was to investigate the potential benefits and limits of metacognitive prompts in an interactive learning environment. For this purpose, a feature was implemented in a learning platform that required students to assess their level of confidence for each answer to a multiple-choice question in a self-paced course. When answering a multiple-choice question, students chose between four confidence levels: "sure," "rather sure," "unsure," and "no clue." Such prompts are used in existing learning platforms, and they are easy to implement, so *if* they showed benefits, it would be efficient to implement them in more learning environments. Since there is little research on this particular type of prompt in an applied setting, this research looked into their effectiveness in a K-12 setting.

The results show that almost half of the students exclusively chose the highest confidence judgment, "sure." This could be regarded as non-compliant use, because if students actually engaged in monitoring and received feedback on their varying performance, some variation in confidence levels would be expected. The other half of the students did show some variation in their confidence judgments. Overall, average self-assessment can be described as overconfident. This level of overconfidence is in line with findings from a study with a similar sample from Lingel et al. (2019): German middle school students took a math test and judged their results as correct and likely correct in 84% of the cases before the task and 73% after the task, while only answering 52% of the questions correctly.

With this high level of overconfidence and non-compliance, there is little leverage for a beneficial influence on performance and hardly any room for improvement.

Consequently, no significant difference in performance between the group with metacognitive prompts and the control group could be found. As expected from previous research, higher-performing students showed better absolute and relative accuracy than lower-performing students in the prompted group (Hacker et al., 2000; Miller & Geraci, 2011). Finally, there was no improvement in relative or absolute metacognitive accuracy across time.

5.1 Empirical Contributions

There is mixed evidence on the effects of metacognitive prompts on students' performance. Metacognitive prompts were shown to lead to better performance in some cases (Renner & Renner, 2001; Veenman, Kok, & Blöte, 2005), though some found an effect only for transfer tasks (Bannert & Mengelkamp, 2013; Lin & Lehman, 1999). Along these lines, Stark and Krause (2009) found improved performance only for complex tasks, not simple ones. As the tasks used in the course for this study were less complex, the lack of performance improvement could be a further indication of this pattern.

There is less research on whether students improve metacognitive accuracy with repeated self-assessment. Here, the students did not improve their metacognitive accuracy, which is in line with studies that saw no improvement in calibration after several training sets of quizzes or practice tests (Bol et al., 2005; Bol & Hacker, 2001). In a study by Hacker et al. (2000) students made predictions and postdictions for three exams and throughout the course, there was instruction and emphasis on the benefits of self-assessment. Here, the high-performing students, but not the low-performing students, showed an increase in accuracy. The present study provides evidence that without additional instruction, repeated exposure to metacognitive prompts does not increase metacognitive accuracy.

5.2 Practical Contributions

When designing interactive learning environments, prompts like the ones used in this study are relatively easy to implement and could be used as part of the default course design.[3] However, as a limit, it should be critically examined whether such prompts by themselves have a beneficial effect. As Schwonke (2015) pointed out, metacognitive processes could be ineffective or even detrimental if they use cognitive resources in a way that does not support the learning process. Schwonke suggested categorizing metacognitive load as a kind of working memory load. It is plausible to assume that mismatched or overly complex metacognitive prompts hinder the learning process.

And yet, metacognitive prompts – even in the simple form that was used in this study – might be beneficial if they are complemented with explicit instruction on metacognition and its role in learning (Kistner et al., 2010) and enough training time for students to engage in monitoring (Bannert & Mengelkamp, 2013). In order to avoid fatigue, these prompts could be used for a limited amount of course time with instruction in the beginning and formative

3 For example, as of January 2024, the platform "Area9 Rhapsode Learner" (https://area9lyceum.com/the-platform/rhapsode-learner/) uses such prompts for every multiple-choice question and reports these data back to the learners as a score for "meta learning."

reflection throughout the course. Furthermore, the learning environment used for this study had a gamified meta-narrative, but students did not receive game benefits if they judged themselves correctly. In a future iteration of the prompting feature, it could be tested whether game benefits for accuracy could provide an incentive for students to invest the required mental effort demanded by monitoring.

5.3 Limitations and Future Research

Above all, the lack of variation in confidence judgments indicates that metacognitive prompts were not used as intended, especially regarding, but not limited to, the students who only used the highest confidence judgment. This non-compliant use presents an issue concerning validity and impedes testing the hypotheses to some degree. Conclusions about the effects of *metacognitive* prompts can only be made to the degree that the selection of the confidence button reflects an actual metacognitive judgment made by the student.

This student behavior might have been caused by some choices in the sampling and research design. The sample of the prompted group was primarily gathered in contexts of extracurricular, voluntary activities where students might not have been motivated to invest the extra mental effort demanded by the prompts. On top of that, the gamified design of the learning platform might have contributed to this by evoking a playful sense of learning which does not ask for the increased mental effort required by metacognitive monitoring and control. However, as described earlier, participants in a study by Lingel et al. (2019), students of similar demography, showed similar overconfidence in a more formal setting. Still, future studies could investigate whether students react to such prompts differently in more formal learning contexts.

6 Conclusion

In this study, we have shown that a simple form of metacognitive prompts without supplementary instruction confers no benefits to student learning and does not result in improved metacognitive accuracy over time. Students showed high levels of overconfidence and non-compliance. When implementing such a feature into an interactive learning environment, it should be critically examined whether the feature brings about the desired results. As such prompts can induce a higher (meta-)cognitive load and might negatively influence motivation and affect, it might be advisable to not use them extensively but intentionally and with supplementary instruction and formative reflection.

Acknowledgments: We thank PearUp for implementing the prompting feature and providing the raw data.

Funding information: No funding was received for conducting this study.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

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

Bannert, M., & Mengelkamp, C. (2013). Scaffolding hypermedia learning through metacognitive prompts. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 171–186). New York: Springer. doi: 10.1007/978-1-4419-5546-3_12.

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. doi: 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. doi: 10.3200/JEXE.73.4.269-290.

DiFrancesca, D., Nietfeld, J. L., & Cao, L. (2016). A comparison of high and low achieving students on self-regulated learning variables. *Learning and Individual Differences*, *45*, 228–236. doi: 10.1016/j.lindif.2015.11.010.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*(4), 391–409. doi: 10.1007/s10648-008-9083-6.

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). Why does excellent monitoring accuracy not always produce gains in memory performance? *Zeitschrift für Psychologie*, *229*(2), 104–119. doi: 10.1027/2151-2604/a000441.

Dunning, D. (2011). Chapter five - The Dunning–Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Bd. 44, pp. 247–296). Cambridge, Massachusetts: Academic Press. doi: 10.1016/B978-0-12-385522-0.00005-6.

Feyzi-Behnagh, R., Azevedo, R., Legowski, E., Reitmeyer, K., Tseytlin, E., & Crowley, R. S. (2014). Metacognitive scaffolds improve self-judgments of accuracy in a medical intelligent tutoring system. *Instructional Science*, *42*(2), 159–181. doi: 10.1007/s11251-013-9275-4.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 702–718. doi: 10.1037/0278-7393.11.1-4.702.

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. doi: 10.1037/0022-0663.92.1.160.

Harwell, M. R. (1988). Choosing between parametric and nonparametric tests. *Journal of Counseling & Development*, *67*(1), 35–38. doi: 10.1002/j.1556-6676.1988.tb02007.x.

Jang, Y., Lee, H., Kim, Y., & Min, K. (2020). The relationship between metacognitive ability and metacognitive accuracy. *Metacognition and Learning*, *15*(3), 411–434. doi: 10.1007/s11409-020-09232-w.

Johnson, A. M., Azevedo, R., & D'Mello, S. K. (2011). The temporal and dynamic nature of self-regulatory processes during independent and externally assisted hypermedia learning. *Cognition and Instruction*, *29*(4), 471–504. doi: 10.1080/07370008.2011.610244.

Kistner, S., Rakoczy, K., Otto, B., Dignath-van Ewijk, C., Büttner, G., & Klieme, E. (2010). Promotion of self-regulated learning in classrooms: Investigating frequency, quality, and consequences for student performance. *Metacognition and Learning*, *5*(2), 157–171. doi: 10.1007/s11409-010-9055-3.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. doi: 10.1037/0096-3445.126.4.349.

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, *36*(7), 837–858. doi: 10.1002/(SICI)1098-2736(199909)36:7<837::AID-TEA6>3.0.CO;2-U.

Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: Do different metacognitive monitoring measures make a difference? *ZDM*, *51*(4), 587–600. doi: 10.1007/s11858-019-01062-8.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 663–679. doi: 10.1037/0278-7393.10.4.663.

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

Moser, S., Zumbach, J., & Deibl, I. (2017). The effect of metacognitive training and prompting on learning success in simulation-based physics learning. *Science Education*, *101*(6), 944–967. doi: 10.1002/sce.21295.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Bd. 26, pp. 125–173). Amsterdam: Elsevier. doi: 10.1016/S0079-7421(08)60053-5.

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179–212. doi: 10.1007/s11409-018-9183-8.

Renner, C., & Renner, M. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology*, *15*, 23–32. doi: 10.1002/1099-0720(200101/02)15:1<23::AID-ACP681>3.0.CO;2-J.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45. doi: 10.1007/s11409-008-9031-3.

Schwartz, B. L., & Efklides, A. (2012). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition*, *1*(3), 145–151. doi: 10.1016/j.jarmac.2012.06.002.

Schwonke, R. (2015). Metacognitive load – Useful, or extraneous concept? Metacognitive and self-regulatory demands in computer-based learning. *Journal of Educational Technology & Society*, *18*(4), 172–184.

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

Sonnenberg, C., & Bannert, M. (2015). Discovering the effects of metacognitive prompts on the sequential structure of SRL-processes using process mining techniques. *Journal of Learning Analytics*, *2*(1), Article 1. doi: 10.18608/jla.2015.21.5.

Stark, R., & Krause, U.-M. (2009). Effects of reflection prompts on learning outcomes and learning behaviour in statistics education. *Learning Environments Research*, *12*(3), 209–223. doi: 10.1007/s10984-009-9063-x.

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211. doi: 10.1007/s11251-004-2274-8.

Winne, P. H. (1997). Experimenting to bootstrap self-regulated learning. *Journal of Educational Psychology*, *89*(3), 397–410. doi: 10.1037/0022-0663.89.3.397.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 299–315). New York: Routledge/Taylor & Francis Group.

Zimmerman, D. W. (1987). Comparative power of student *T* test and Mann-Whitney *U* test for unequal sample sizes and variances. *The Journal of Experimental Education*, *55*(3), 171–174. doi: 10.1080/00220973.1987.10806451.

Zumbach, J., Rammerstorfer, L., & Deibl, I. (2020). Cognitive and metacognitive support in learning with a serious game about demographic change. *Computers in Human Behavior*, *103*, 120–129. doi: 10.1016/j.chb.2019.09.026.