Fall 2020

Can Metacognitive Monitoring Ability be Trained?

Erica Abed
Claremont Graduate University

Can Metacognitive Monitoring Ability be Trained?

By

Erica Abed

Claremont Graduate University

2020

APPROVAL OF THE DISSERTATION COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Erica Abed as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Psychology.

Kathy Pezdek, Chair
Claremont Graduate University
Professor of Psychology

Andrew R. A. Conway
Claremont Graduate University
Professor of Psychology

Lise Abrams
Pomona College
Professor of Linguistics and Cognitive Science

John Dunlosky
Kent State University
Professor of Psychology

Abstract

Can Metacognitive Monitoring Ability be Trained?

by

Erica Abed

Claremont Graduate University: 2020

Low performers tend to greatly overestimate their performance on a task, but high performers slightly underestimate their performance; the unskilled-unaware effect (Kruger & Dunning, 1999). Because assessment of one's own skill (monitoring) impacts future decisions, such as selecting information to re-study (control), low performers may be disadvantaged in both what they know and what they are likely to learn. Although most research has attempted to reduce metacognitive errors in low performers by training *cognitive* ability (e.g., teaching them to perform better on a task), training *metacognitive* ability may be both more efficient and more likely to transfer to other tasks. In light of recent findings that suggest the unskilled-unaware effect is the result of a true metacognitive error, this dissertation tests two methods for reducing overconfidence (Experiment 1) and improving monitoring accuracy (Experiment 2) in high and low performers.

Experiment 1 tested whether answering easy rather than hard questions prior to taking a medium-difficulty test reduced trial-by-trial overconfidence in low performers. This hypothesis was not supported. Global, but not local judgments were affected by the difficulty of a preceding task. Experiment 2 tested whether training and feedback improved metacognitive monitoring accuracy, especially for low performers for whom monitoring accuracy is relatively poorer. This hypothesis was also not supported. Across all performance quartiles and experimental conditions, monitoring accuracy remained consistent for the trial-by-trial confidence judgments.

Taken together with previous research, results from both experiments indicate that when making global judgments at the end of a task, people rely on various sources of information, including perceptions of task difficulty, to inform their metacognitive judgments. By contrast, when making trial-by-trial judgments, people more likely rely on information specific to the question itself (e.g., information from memory *or* gut feelings) and not information about the task to inform their confidence judgments.

Table of Contents

Chapter 1: Review of Literature

When people perform well on a task, they tend to slightly underestimate their performance; in contrast, when people perform poorly, they tend to greatly overestimate their performance. This finding is called the unskilled-unaware effect (Kruger & Dunning, 1999). This unskilled and unaware pattern has been replicated across many domains, including medicine (Mehdizadeh, Sturrock, Myers, Khatib, & Dacre, 2014) and eyewitness memory (Grabman, Dobolyi, Berelovich, & Dodson, 2019), and occurs in ecologically valid settings such as the classroom (Callender, Franco-Watkins, & Roberts, 2016). According to most researchers (e.g., Händel & Dresel, 2018; Abed, Nguyen, & Pezdek, 2020), low performers lack task-relevant knowledge necessary to both perform well on a task and accurately estimate their performance. Moreover, errors in metacognitive judgments are proposed to stem from different sources in high than low performers. Whereas high performers tend to rely more on information-based processes (e.g., rely on the content of retrieved information), low performers may rely more on experience-based processes (e.g., gut feelings; Händel & Dresel, 2018). The unskilled-unaware pattern is especially problematic because people use estimates of their performance (metacognitive monitoring) to direct future learning (Dunlosky, 2005). Thus, poor performers are disadvantaged both in what they know and what they are likely to learn.

Although most research has attempted to reduce metacognitive errors in low performers by training *cognitive* ability (e.g., teaching them to perform better on a task), training *metacognitive* ability may be (a) more efficient and (b) more likely to transfer to other tasks. For example, if people become better at distinguishing between what they do and do not know, they could more accurately select information to restudy (e.g., monitoring that influences control; Nelson & Narens, 1990). Moreover, improved monitoring accuracy may affect control of

important real-world judgments, such as deciding what information to share online (Fenn, Kantner, Ramsay, Pezdek, & Abed, 2019) or willingness to testify as an eyewitness (Michael & Garry, 2016). Given the applied importance of accurately monitoring one's own performance, this proposed dissertation tests whether presenting easy questions first (Experiment 1) and providing item-level feedback (Experiment 2) reduces overconfidence and increases monitoring accuracy, especially in low performers.

Mechanism for the Unskilled-Unaware Effect

Various mechanisms have been proposed for the unskilled-unaware effect. One explanation for the asymmetric errors exhibited by high and low performers (i.e., low performers overestimate their performance to a greater extent than high performers underestimate their performance) is that high performers incorrectly assume that other people perform as well as they do, but low performers simply lack the task-relevant knowledge required to make accurate performance estimates. According to Kruger and Dunning (1999), expertise on a task is necessary to judge performance on a task accurately – either one's own or someone else's. Thus, having more task-relevant knowledge enables people to perform well on a task, and subsequently to assess their performance accurately. To support this, Kruger and Dunning (Exp. 4) had participants take a logical reasoning test and then estimate how many questions they answered correctly (an absolute judgment) and how well they performed on the test compared to other students (a relative judgment). Some participants then received logical-reasoning training. Later, all participants reviewed their test again, estimated whether they thought they had gotten each question right or wrong, and then re-assessed how many questions they answered correctly and how well they performed on the test compared to other students. Whereas participants who scored in the bottom quartile significantly overestimated both the number of items they answered

correctly and their percentile rank on the test, participants who scored in the top quartile accurately estimated the number of items they answered correctly and slightly underestimated their percentile rank on the test. However, when participants received logical reasoning training, they adjusted their estimates to be more accurate; high performers slightly raised their estimated number correct and percentile rank, but low performers significantly lowered their estimated number correct and percentile rank. In contrast, participants who did not receive training did not adjust their performance estimates. Kruger and Dunning reasoned that low performers were able to recognize their own relative incompetence only (ironically) after becoming more competent. These results lend initial support to the notion that the unskilled-unaware effect is the result of a true metacognitive error; that is, low performers lack the ability to estimate their performance accurately because they lack the skill to perform well in the first place. Thus, increasing one's ability to perform well subsequently increases monitoring accuracy, even when monitoring accuracy is not itself trained.

If low task-relevant knowledge is the mechanism responsible for the unskilled-unaware pattern, then one way to improve monitoring accuracy would be to improve skill on a task (e.g., Kruger & Dunning, 1999 Exp. 4). However, some researchers have argued that task-relevant knowledge is not the mechanism responsible for the unskilled-unaware effect, and that the unskilled-unaware pattern is simply an artifact of regression or task difficulty. For example, according to Krueger and Mueller (2002), the unskilled-unaware pattern reported by Kruger and Dunning (1999) is inevitable because people tend to rate themselves as slightly better than average regardless of how well they actually perform – the better-than-average effect (BTA). Thus, the observation that low performers' estimates deviate from their actual performance to a much greater extent than high performers is simply because their performance is farther from the

average estimate, and not because they are less capable of monitoring their own skill. Similarly, Burson, Larrick, and Klayman (2006) argued that the unskilled-unaware effect is an artifact of task difficulty, and the unskilled-unaware pattern occurs only on relatively easy tests. According to Burson and colleagues, people tend to rate themselves similarly regardless of skill level. Thus, whereas relatively easy tests result in an unskilled-unaware pattern, more difficult tests should result in a *skilled*-unaware pattern. In other words, when the average performance estimate shifts downwards (as would occur for difficult tests) then low, not high performers would be more accurate.

To test the hypothesis that task difficulty and not performance per se is responsible for the unskilled-unaware effect, Burson et al. (2006) had participants take either an easy or a difficult version of a test about the University of Chicago and then estimate how many questions they had answered correctly and how well they performed compared to other participants. To assess metacognitive accuracy, they calculated the difference between actual and estimated percentile and number of items answered correctly. For the easier test, low performers overestimated their percentile rank to a greater extent than high performers underestimated their percentile rank, replicating the asymmetrical errors reported by Kruger and Dunning. However, on the more difficult test, low and high performers were equally miscalibrated. This pattern was similar for the absolute performance estimates; on the easier test, low and high performers were equally miscalibrated on estimates of number correct, but on the more difficult test, high performers underestimated the number of items answered correctly more than low performers overestimated the number of items answered correctly. These results suggest that task difficulty affects performance estimates equally for high and low performers, and that people use task difficulty, not assessment of ability per se, to evaluate their overall performance.

Recently, several researchers have argued that global (i.e., overall performance) judgments are inaccurate representations of metacognitive ability because global judgments require people to aggregate their performance estimates across a variety of test items (e.g., Dunlosky & Lipko, 2007; Händel & Dresel 2018; Schraw, 2009). Subsequently, environmental features such as overall task difficulty or one's holistic assessment of one's skill may affect global judgments. In contrast, local (i.e., trial-by-trial) judgments pinpoint peoples' assessment of only one item at a time, so each of these confidence judgments should be independent of one another. Thus, researchers studying the unskilled-unaware effect have recently begun to examine local confidence judgments to understand better how monitoring accuracy differs as a function of task performance (Händel & Dresel, 2018; Händel & Fritzsche, 2016; Abed, Nguyen & Pezdek, 2020). Indeed, some researchers acknowledge that their criticisms of the unskilled-unaware effect apply to research on global, but not local judgments (e.g., Gignac & Zajenkowski, 2020)[1]. Whereas global estimates likely reflect a variety of judgments, such as perceived task difficulty, self-evaluations, and attempted recall of item-level metacognitive judgments, local estimates reflect more pure assessments of metamemory for a given test item and thus may be more accurate, even for low performers.

However, according to cue-utilization models of metacognition (e.g., Koriat, 1997; see also Koriat & Levy-Sadot, 1999), people base item-level metacognitive judgments on a combination of cues, and high and low performers may rely on different cues to make their confidence judgments (Händel & Dresel, 2018). For example, high performers have high levels of task-relevant knowledge, and thus should be able to use information-based processing to make

[1] Gignac and Zajenkowski specifically acknowledge that their criticisms do not apply to the findings of Abed et al., 2020, cited in their paper as Nguyen, 2018.

a metamemory judgment. In other words, high performers should be able to rely on the *content* of information retrieved from memory to make a metamemory judgment. In contrast, low performers may lack sufficient task relevant knowledge to use information-based processing and subsequently may utilize more experience-based processing to make a metamemory judgment. In other words, lacking sufficient task-relevant knowledge, low performers may subsequently rely on subjective *feelings* – such as how easy information comes to mind – to make a metamemory judgment.

Consistent with this hypothesis, Händel and Dresel (2018) reported that high and low performers differed on both accuracy of local metacognitive judgments and the reasons they provided for making those judgments. In their study, participants completed a multiple-choice math test. After answering each item, participants gave (a) a performance judgment ("Do you think you picked the correct answer?" yes/no), (b) a reason for their performance judgment ("Why do you think that your provided answer is correct/wrong?"), and (c) a second order judgment (SOJ) about their performance judgment ("How sure are you that your judgment is correct?") on a scale from *absolutely unsure* to *absolutely sure*. Händel and Dresel reported that on average, high performers provided significantly more "yes" performance judgments (i.e., thought they solved items correctly more often) and were more confident (i.e., provided higher SOJs) in their performance judgments than low performers, and they interpreted this as evidence that low performers are somewhat aware of their lack of skill. However, critically, high performers were more confident in their *correct* than *incorrect* performance judgments (i.e., mean confidence was higher for hits and correct rejections than for misses and false alarms), but low performers were equally confident in correct and incorrect performance judgments. This result indicates that high, but not low performers can somewhat accurately distinguish between

information they do and do not know. Further, consistent with the hypothesis that high and low performers rely on different types of processing to make metacognitive judgments, high performers more often reported using prior knowledge to evaluate performance, but low performers more often reported relying on their gut instincts. Taken together, these results indicate that high and low performers rely on different cues to assess performance. Whereas high performers have sufficient task-relevant knowledge to respond accurately and subsequently can use this information (i.e., information-based processing) to evaluate their own performance accurately, low performers lack sufficient task-relevant knowledge to either respond accurately *or* evaluate their performance, and subsequently rely on their gut feelings (i.e., experience-based processing) to assess confidence.

Händel and Dresel's finding that low performers were equally confident when they were correct and incorrect is concerning because it suggests that, although they are less confident than high performers overall, low performers lack the insight to evaluate their own knowledge accurately. If low performers cannot discriminate between what they do and do not know, they also cannot make optimal decisions about what information to re-study, act on, or share with others. However, an important question to consider is whether even at high confidence (e.g., "I am absolutely sure my response is correct"), can low performers accurately assess what they do or do not know? Research from other domains, such as eyewitness memory, have used measures such as Confidence-Accuracy Characteristic (CAC) analysis to plot the proportion of test items answered correctly (hereafter simply referred to as "proportion correct") at each level of confidence. According to this literature, the confidence-accuracy relationship is strongest at high confidence; when people report the highest level of confidence, proportion correct is objectively high, but calibration is weaker at lower levels of confidence. It is unclear from Händel and

Dresel's results whether this pattern would be consistent for low performers because they reported mean confidence as a function of whether participants were correct or incorrect. In contrast, CAC analysis assesses the opposite relationship; the likelihood of being correct as a function of confidence level. This more fine-grained analysis can provide additional insight into how well low and high performers can distinguish between correct and incorrect responses. For example, it is possible that low performers report high confidence less often than high performers, but when they *do* report high confidence, they might be just as likely as high performers to be correct. In other words, low performers may exhibit poor monitoring accuracy overall, but have high accuracy at the extreme high end of the confidence scale. Thus, CAC analysis can provide important insight into the specific differences between low and high performers' metacognitive ability.

To date, only one study has examined high and low performers' metacognitive ability using CAC analysis. In their first experiment, Abed, Nguyen, and Pezdek (2020) had participants answer multiple choice logical reasoning questions and rate their confidence in each answer on a scale from 0% (*not at all confident*) to 100% (*absolutely confident*) (a local confidence judgment). After completing the task, participants then estimated the number of questions they thought they had answered correctly and their percentile rank compared to their peers. Results for global performance estimates replicated the asymmetric errors reported by Kruger and Dunning (1999); low performers overestimated the number of items answered correctly to a greater extent than high performers underestimated the number of items answered correctly. Accuracy of the local (trial-by-trial) metacognitive judgments was assessed by calculating the proportion of items answered correctly at each confidence level. One hypothesis is that high performers have higher metacognitive accuracy than low performers *overall*, but that they are

equally accurate at high confidence. However, Abed and colleagues reported that the confidence accuracy relationship was significantly stronger (i.e., proportion correct was higher at higher levels of confidence) for high than low performers. Critically, this was true even for the highest level of confidence; though for high performers, the proportion correct at high (80-100%) confidence was objectively high (M = .90), for low performers, the proportion correct at high confidence was significantly lower, and objectively low (M = .18). The authors interpreted this as evidence that the unskilled-unaware effect is the result of a true metacognitive deficit in low performers; low performers lack task relevant knowledge necessary to perform well on the task, and subsequently cannot accurately discriminate between what they do and do not know, a metacognitive judgment.

In a second experiment, Abed et al. tested whether, despite having poor metacognitive ability (i.e., the ability to identify correct versus incorrect responses), low performers are metacognitively aware of this deficit. Participants answered multiple choice logical reasoning questions and judged (a) whether they thought they had answered each question correctly and then (b) how confident they were that their correct/incorrect judgment was correct on a scale from 0% (*not at all confident*) to 100% (*absolutely confident*). This design was similar to that of Händel and Dresel (2018), where participants made a first order metacognitive judgment (what Händel and Dresel called a *performance judgment*), and then made a second order judgment (SOJ) about their metacognitive judgment. Thus, whereas in Abed et al.'s Experiment 1, CAC analyses were based on *metacognitive* judgments and proportion correct was calculated on whether the correct answer was selected, in Experiment 2 CAC analyses were based on SOJs, or *meta*-metacognitive judgments and proportion correct was calculated based on whether the appropriate metacognitive judgment was selected (e.g., correct/incorrect). Thus, Experiment 1

tested whether high and low performers can accurately assess their *task* performance, and Experiment 2 tested whether high and low performers can accurately assess their *metacognitive* performance.

Similar to Händel and Dresel (2018), Abed and colleagues (2020) reported that overall, low performers were less confident (i.e., had lower SOJs) than high performers, indicating some awareness of their low *metacognitive* ability. However, as in Experiment 1, the confidence accuracy relationship was significantly stronger for high than low performers. For high performers, proportion correct at high (80-100%) confidence was objectively high (M = .84), but for low performers, proportion correct was objectively low across all levels of confidence, even high confidence (M = .13). This finding indicates that unlike high performers, low performers have limited insight into their *metacognitive* ability. Taken together with results from Abed et al.'s Experiment 1 and Händel and Dresel (2018), these findings paint a bleak picture of low performers' metacognitive skill. The fact that low performers tend to give high first- and second-order metacognitive judgments less frequently than high performers seems to indicate some awareness of their low metacognitive ability. However, it is unlikely that these judgments are based on probative information because they are only weakly related to accuracy of either cognitive or metacognitive judgments, even at high confidence. Rather, CAC analyses (Abed et al.) revealed that for low performers, proportion correct was similar across all levels of confidence, indicating that low performers have limited ability to distinguish between when their cognitive *or* metacognitive judgments are likely to be correct or incorrect. In other words, although low performers seem to understand that they should not be highly confident for all items, this awareness does not help them distinguish between information that they do or do not know. In contrast, it seems that task relevant knowledge helps high performers (a) perform well

on a task, (b) accurately assess their performance on a task, and (c) evaluate the validity of their performance estimates. Given that the ability to monitor one's own performance is critical to directing future behavior (such as selecting information to re-study), these findings indicate that low performers are not only disadvantaged in what they know, but also in what they are likely to learn. Thus, training low performers to evaluate their own performance accurately is an important area for future research.

Improving Monitoring Accuracy: Two Approaches

In light of the importance of metacognitive monitoring to control processes (e.g., directing future study), researchers have long been interested in factors that can improve metacognitive monitoring. However, much of this research either (a) targets monitoring accuracy by improving performance on the task itself (e.g., Kruger and Dunning, 1999) or (b) targets monitoring accuracy but subsequently improves task performance. For example, results from classroom studies suggest that providing feedback on local confidence judgments improves monitoring accuracy (e.g., Callender, Franco-Watkins, & Roberts, 2016; Gutierrez & Schraw, 2015; Renner & Renner, 2001), but these improvements are accompanied by increases in test performance. Although increases in test performance are obviously desirable, from these studies it is impossible to determine whether monitoring accuracy improved test performance, or whether test performance improved monitoring accuracy.

To disentangle the relationship between monitoring and control processes, it is important to determine whether monitoring accuracy can be trained independently of task performance. Such an outcome would be exceedingly useful to practitioners, because it would suggest that increases in monitoring accuracy could transfer to other tasks (or in the field of education, other subjects or classes). Changes in monitoring accuracy that result from improved task performance

(e.g., Kruger & Dunning, 1999; Nietfeld & Schraw, 2002) are by definition task-specific – that is, training math ability does not increase metacognitive monitoring accuracy on an unrelated verbal task. By contrast, changes in monitoring accuracy that result from a shift in processing strategy are more likely to be task-independent. That is, a manipulation that reduces overconfidence could be applied to any task domain, and may even generalize across unrelated tasks, such as math and verbal tests. Generally, researchers have adopted two approaches to improving monitoring accuracy independently of task performance: reducing overconfidence and training calibration accuracy.

Improving Monitoring Accuracy by Reducing Overconfidence

According to dual-process models of metamemory (e.g., Koriat & Levy-Sadot, 1999), people base their confidence judgments on a combination of cues such as how quickly a response comes to mind (experience-based processing) or the content of information associated with a response (information-based processing). Because experience-based processing is based on subjective experiences, manipulations of context, such as fluency (Kelley & Lindsay, 1993) or task difficulty (Michael & Garry, 2016; Weinstein & Roediger, 2010; 2012) can affect metamemory judgments independently of task performance. For example, Weinstein and Roediger (2010) reported that participants were more confident about their test performance (a global judgment) when questions were ordered from the easiest to the most difficult than when the same questions were ordered randomly. The authors suggested that when easy questions were presented first, participants felt that they were performing well on the test and failed to adjust their estimates of their own skill as the test progressed – an anchoring effect.

In a follow up study, Weinstein and Roediger (2012) tested whether perceptions of task difficulty affect local as well as global confidence judgments. Participants answered 100 general

knowledge questions in blocks of ten. Questions were ordered easiest to hardest, or hardest to easiest. Participants answered each question, rated their confidence, and after each block estimated how many questions they had answered correctly (out of 10) within that block. After answering all 100 questions, participants made a final assessment of how many items they had answered correctly in total (out of 100). Overall, participants in the hard-easy condition were less confident than participants in the easy-hard condition. Compared to participants in the easy-hard condition, participants in the hard-easy condition provided lower estimates of number correct within each block (out of 10) and on the overall test (out of 100). However, neither local confidence judgments nor test accuracy differed across groups. The authors interpreted this as evidence that people use their initial judgments of task difficulty to estimate their own ability and do not properly adjust their estimates as a task progresses. Further, they suggested that local confidence judgments are affected less by difficulty order because people rely on a different source of information (i.e., item-level cues) to make such judgments.

Michael and Garry (2016) replicated this finding in a similar study using an eyewitness memory paradigm, and further demonstrated that people use their performance estimates to direct future behavior. Participants in this study watched a video of a mock crime. Later, they answered questions about the crime and rated their confidence in each response. Questions were presented in order from easiest to hardest, or hardest to easiest. Finally, participants estimated the number of questions they thought they had answered correctly and how confident they would be in their memory if asked to testify as an eyewitness. Similar to results from Weinstein and Roediger (2012), question order did not affect either test accuracy or local confidence judgments; mean accuracy and mean item-level confidence judgments were similar for both difficulty conditions. However, participants in the hard-easy test condition estimated they had answered

significantly fewer questions correctly and indicated they would be less confident testifying as an eyewitness than participants in the easy-hard test condition. The authors interpreted this as evidence that people use their initial impressions of task difficulty as an anchor and do not adjust up or down appropriately. Moreover, this finding suggests that difficulty order affects not only global confidence judgments, but potentially future intentions as well, such as willingness to testify.

Results of Weinstein and Roediger (2010, 2012) and Michael and Garry (2016) suggest that manipulations of initial task difficulty may not be a useful approach to increasing monitoring accuracy in low performers. In these studies, global but not local confidence judgments were affected by perceptions of task difficulty. However, there are two reasons why the order of question difficulty might actually affect local confidence for low, if not high performers. First, results from Händel and Dresel (2018) suggest that low performers rely more on experience-based processing than high performers, who rely more on information-based processing. Subsequently, context or environmental manipulations such as initial task difficulty would be hypothesized to reduce overconfidence in low performers but have a comparatively smaller effect on high performers. Second, in these studies, item difficulty was increased or decreased gradually. In other words, whereas these studies manipulated *initial* item difficulty, *change* in difficulty from each items to the next was minimal (e.g., questions were presented in a fixed increasing or decreasing order of normed proportion correct; subsequently, the change in normed difficulty from one item to the next was very small). However, it is the perceived change in difficulty that tends to affect local confidence judgments. For example, according to Kelley and Lindsay (1993; see also Koriat, Lichtenstein, & Fischhoff, 1980), people base their confidence judgments on how quickly and easily information comes to mind (i.e., retrieval fluency). Further,

perceptions of fluency are relative (e.g., Alter & Oppenheimer, 2009; Whittlesea & Williams, 2000), and it is the discrepancy in fluency between to-be-judged items that is critical for achieving these effects.

To illustrate, Pansky and Goldsmith (2014) tested whether the comparative difficulty of an initial task affected monitoring and control judgments on a subsequent task. Participants answered moderately difficult general knowledge questions preceded by a set of very easy or very difficult questions. Unlike previous studies, the perceived increase or decrease in item difficulty was clear to participants, not gradual. Participants answered each question and rated their confidence in each response (forced report). Next, participants reviewed their answers to each question and indicated whether they wanted to submit their answer for scoring (free report). Participants could choose not to submit an answer, but for each answer they submitted, participants received a bonus if they were correct and a penalty if they were wrong. Although accuracy for the medium-difficulty target questions was similar for all participants, those who initially answered easy questions were both less confident and less likely to submit answers for scoring than those who initially answered difficult questions. In other words, people lowered their item-level confidence judgments when questions felt relatively more difficult and raised their confidence judgments when questions felt relatively easier, and this affected subsequent behavior (i.e., willingness to submit a response). These results suggest that fluency manipulations can reduce overconfidence independently of task performance.

Together with results of Weinstein and Roediger (2010, 2012) and Michael and Garry (2016), results from Pansky and Goldsmith (2014) suggest that changes in task difficulty affect both global and local confidence judgments, albeit through different mechanisms. Whereas initial perceptions of task difficulty affect global confidence judgments by anchoring

performance expectations, they affect local confidence judgments by serving as a basis for judging the relative ease of responding on a subsequent task.

Moreover, reducing overconfidence per se would not universally improve metacognitive monitoring accuracy. Rather, manipulations aimed at reducing confidence should provide the most benefit to people who are the most overconfident, such as low performers, but may actually *lower* metacognitive accuracy for people who are less overconfident, such as high performers. True, universal improvements in monitoring accuracy would require people to better discriminate between information they do and do not know.

Improving Monitoring Accuracy with Training

Another method researchers have used to improve monitoring accuracy is calibration training, a type of metacognitive training teaching people to assign high confidence to answers only when they are very likely to be correct and low confidence to answers when they are relatively less likely to be correct. Some researchers have employed effortful calibration training programs, such as 11 one-hour sessions (Lichtenstein & Fischhoff, 1980) or a single two-hour session (Nietfeld & Schraw, 2002). Moreover, most researchers striving to improve monitoring accuracy have simply instructed people to adjust their use of the confidence scale (Lichtenstein & Fischhoff, 1980; Zechmeister, Rusch, & Markell, 1986) or trained them to perform better on a related task, such as logical reasoning (e.g., Nietfeld & Schraw, 2002).

Less commonly, researchers have investigated whether techniques often used in an applied setting (e.g., performance feedback) can be leveraged to improve monitoring accuracy – that is, to distinguish better between information people do and do not know. For example, Rawson and Dunlosky (2007) tested whether providing people with feedback would reduce their overconfidence that their vocabulary definitions were accurate. In their study, participants read

several texts, each including to-be-remembered terms (e.g., “adaptive thermogenesis”). After reading each text, participants attempted to define the key terms, and half were given feedback on the accuracy of their response. Later, participants rated their confidence in the accuracy of each definition. Overall, participants were overconfident in their key term definitions; mean confidence was greater than mean accuracy. However, for items for which people provided incorrect definitions (commission errors), feedback significantly reduced overconfidence. In other words, when people received feedback that a response was incorrect, they significantly lowered their confidence judgments compared to those who did not receive feedback. This finding suggests that being confronted with one’s own errors (e.g., receiving “incorrect” feedback) can increase monitoring accuracy. Critically though, it is unclear how receiving feedback affects future confidence; specifically, after receiving feedback that a commission error was incorrect, are people less overconfident on *subsequent* errors? Because people received feedback *before* making each confidence judgment, it is unclear whether receiving feedback on one trial or set of trials affects later metacognitive judgments; in other words, does this feedback transfer to other judgments or tasks?

Critical questions for practitioners are whether an improvement in monitoring accuracy transfers to other judgment targets (“near” transfer) or tasks (“far” transfer), or persists over time (Barnett & Ceci, 2002). Most recent research has focused on improving (a) global judgment accuracy or (b) task performance along with metacognitive performance, and thus provides limited insight into whether training or feedback can transfer. However, results of Zechmeister, Rusch, and Markell (1986) suggest that monitoring training can transfer to other questions on a similar task (near transfer) even after a delay, and that low performers especially benefitted.

Zechmeister and colleagues (1986) trained college students to improve their confidence calibration on a general knowledge test. Students were categorized according to their performance in an introductory psychology class as "high achievers" (70th percentile) or "low achievers" (30th percentile). High and low achievers were given different general knowledge tests that were matched on perceived difficulty; that is, high achievers were given a more difficult test and low achievers were given an easier test, but mean performance on the general knowledge test was the same (about 70%)[2]. Participants answered two-alternative forced choice questions and gave a confidence judgment for each. After completing the test, half of the participants received a 30-minute calibration training to improve the accuracy of their confidence judgments. During the training session participants were shown calibration curves and instructed to try their best to achieve perfect calibration (i.e., 100% correct at 100% confidence, 90% correct at 90% confidence, etc.) on a later test. Participants were told that most people are overconfident because they only search for evidence that their answer is right, and not that their answer is wrong. Participants were then given sample questions to practice improving their calibration. Two weeks later, all participants completed another general knowledge test and were told to try their best to be perfectly calibrated.

Zechmeister, Rusch, and Markell (1986) reported that prior to training, low achievers were more overconfident (i.e., the magnitude of their calibration scores was greater) than high achievers. Further, calibration training reduced overconfidence overall; participants who

[2] Zechmeister and colleagues used different tests so the mean proportion of test items answered correctly would be similar for high and low achievers. Their analyses were conducted on calibration scores calculated by subtracting proportion correct from mean confidence, where negative values represent underconfidence and positive values represent overconfidence. Thus, according to the authors, the difference between high and low achievers on calibration scores results from differences in metacognition alone (mean confidence) rather than a combination of metacognition and performance (proportion correct).

received training significantly improved their calibration on the second test compared to the first, but participants who did not receive training had similar calibration on both tests. Critically though, the effect of training was significant for low achievers but not high achievers. Although the authors reported only aggregate measures of calibration (i.e., mean confidence – proportion correct), they did provide calibration plots (see Figure 1 for their calibration plot of high and low performers as a function of training). Across all levels of confidence, low performers' calibration improved with training, and their proportion correct at high confidence was objectively high. Moreover, for both high and low achievers, performance on the second test did not improve as a function of training. This finding indicates that metacognitive ability can be trained independently of task ability.

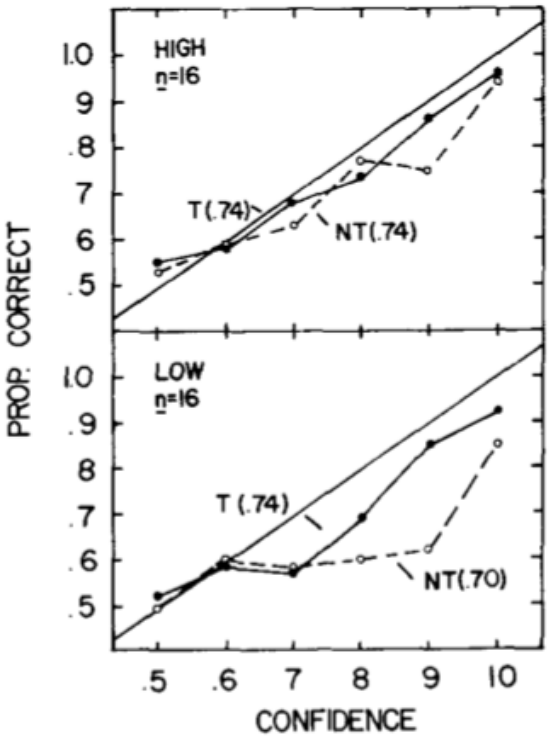


Figure 1. Calibration plots from Zechmeister, Rusch, and Markell (1986) page 15, Figure 5. Proportion correct on the second (post-training) test is plotted as a function of achievement (high achievers in the top panel, low achievers in the bottom panel) and

training (training group depicted with a solid line, no-training group with a dashed line). Numbers in parentheses indicate proportion correct on the second test; the difference between proportion correct for low achievers who did (M = .74) and did not (M = .70) receive training was not significant.

However, the generalizability of this training is unclear because participants in this study may have increased their calibration accuracy simply by decreasing the frequency of their high confidence responses, and not by appropriately distinguishing between when they were more or less likely to be correct. Thus, future research should use signal detection measures of metacognitive judgments (i.e., correct/incorrect) to assess whether training affects confidence (i.e., SOJs) equally for hits and false alarms. For example, after calibration training, do people decrease their confidence judgments similarly for both hits and false alarms (thereby decreasing monitoring accuracy for hits and increasing accuracy for false alarms), or do they decrease confidence for false alarms only (i.e., become better at assessing when they are more versus less likely to be correct)? This question is likely to be of particular interest to practitioners who are concerned with how well people can direct their future study efforts (Gutierrez & Schraw, 2015), how strongly people believe in (or are inclined to share) false information (Fenn et al., 2019), or whether a confident eyewitness is likely to be correct (Nguyen, Pezdek, & Wixted, 2016).

If people can be trained to improve their monitoring accuracy independently of task performance, this would suggest that practitioners could improve low performers' poor metacognitive ability, without having to train them to perform better on each and every task. Moreover, according to prominent models of metamemory (e.g., Nelson & Narens, 1990), improved metacognitive monitoring should result in improved control, such as more optimal

selection of material to re-study. In other words, teaching low performers to monitor their performance more accurately should help them learn more efficiently because they would then be able to allocate their study time optimally to least-learned material – a difficult task for someone who cannot distinguish between information they do and do not know. Thus, understanding whether low performers can improve their monitoring accuracy is an important topic for future research.

The research on metacognitive training is outdated and inconsistent. For example, the training employed by researchers has varied widely in (a) duration (taking between 2-11 hours), (b) setting (lab versus classroom), and (c) general approach (e.g., training use of a confidence scale versus training a related skill). Although the results of each independent study suggest that training metacognitive ability is a promising area for future research, the specific method for how training should be executed has yet to be decisively explored and confirmed.

Aims of Dissertation

Recent research on the unskilled-unaware effect suggests that the unskilled-unaware pattern is the result of a true metacognitive error rather than a statistical artifact (e.g., Abed, Nguyen, & Pezdek, 2020). Low performers lack the task-relevant knowledge to perform well on a task (Kruger & Dunning, 1999), judge their performance on a task (Händel & Dresel, 2018), and distinguish between answers they got correct or wrong (Abed et al., 2020). A likely explanation for the asymmetrical errors exhibited by high and low performers is that high performers rely more on information-based processing, whereas low performers rely more on experience-based processing (Händel & Dresel, 2018; see also Koriat & Levy-Sadot, 1999).

In light of this account of the unskilled-unaware effect, this dissertation investigates two promising methods for reducing overconfidence and improving monitoring accuracy in low

performers. First, results from some studies suggest that overconfidence can be reduced for global (Michael & Garry, 2016; Weinstein & Roediger, 2012) and local (Pansky & Goldsmith, 2014) confidence judgments by altering the context of a task, for example by presenting difficult or easy items first. If low performers rely more on experience-based processing (Händel & Dresel, 2018) then altering the context of a task to make it feel more difficult may reduce overconfidence of both monitoring and control judgments (e.g., Pansky & Goldsmith, 2014). Second, some research suggests that providing people with item-level feedback (Rawson & Dunlosky, 2007) or a brief calibration training (Zechmeister et al., 1986) can improve monitoring accuracy independently of task performance. Such a manipulation would be expected to benefit both high and low performers, but especially low performers because they are relatively more poorly calibrated to begin with (Abed et al., 2020).

The purpose of this dissertation is to test two potential methods for reducing overconfidence in low performers. Experiment 1 tests whether altering the context of a task to make it feel more difficult (i.e., by presenting easy questions first) reduces item-level overconfidence in low performers. Experiment 2 tests whether monitoring accuracy can be trained independently of task performance by providing participants with item-level feedback about the accuracy of their metacognitive judgments.

Chapter 2: Experiment 1

Researchers have reported that task difficulty affects both global and local confidence judgments, albeit through different mechanisms. Weinstein and Roediger (2012; see also Michael & Garry, 2016) reported that people anchor their global performance estimates based on initial task difficulty. Similarly, Burson, Larrick, and Klayman (2006) reported that people base their global performance estimates on their perceived difficulty of a task. According to Burson

and colleagues, participants' estimates of the number of questions answered correctly were similar for all performance levels but varied as a function of task difficulty, resulting in the typical unskilled-unaware pattern for easy tests but a *skilled*-unaware pattern for difficult tests. This pattern would be expected to remain the same even if task difficulty varied across an experimental session, as in Weinstein and Roediger's study. That is, across all skill levels, people likely anchor their global performance estimates based on initial task difficulty (e.g., Weinstein & Roediger; see also Burson et al., 2006). Consequently, low performers should be the most poorly calibrated when an initial task is *easy*; low performers tend to overestimate their performance, and should do so to an even greater extent if they anchor their performance based on an easy task. On the other hand, high performers should be the most poorly calibrated when an initial task is *difficult*; high performers tend to underestimate their performance, and should do so to an even greater extent if they anchor their performance based on a difficult task.

By contrast, previous research suggests that high and low performers rely on different processes when making *local* performance judgments, and subsequently the same manipulation of task difficulty should affect high and low performers differently. Results from Händel and Dresel (2018) suggest that when making local confidence judgments, high performers rely more on information-based processing (e.g., rely on the content of retrieved information), but low performers rely more on experience-based processing (e.g., gut feelings). If high performers rely primarily on information retrieved from memory to make item-level confidence judgments, and not on gut feelings such as how easily information comes to mind, then context manipulations such as initial task difficulty *should not* affect their local confidence judgments.

In contrast, if low performers rely primarily on experience-based cues such as how easily information comes to mind, then context manipulations such as initial task difficulty *should*

affect their local confidence judgments. For instance, researchers have reported that people are more confident in their answers when questions feel relatively easier to answer, and less confident in their answers when questions feel relatively more difficult to answer, for example when medium-difficulty target questions are preceded by very easy or very difficult questions (e.g., Kelley & Lindsay, 1993; Pansky & Goldsmith, 2014). This pattern should be especially true for low performers, who according to Händel and Dresel rely primarily on their gut feelings rather than information from memory to make confidence judgments.

Interestingly, it appears from the literature that the same context manipulation (initially hard or easy questions) has opposite effects on global and local confidence judgments. Whereas Weinstein and Roediger (2012) reported that presenting difficult questions first *lowered* global performance estimates, Pansky and Goldsmith (2014) reported that presenting a difficult test first *increased* local performance estimates. These results suggest that changes in task difficulty affect global and local confidence judgments through different mechanisms. Whereas initial perceptions of task difficulty affect global confidence judgments by anchoring performance expectations, they affect local confidence judgments by serving as a basis for judging the relative ease of responding on a subsequent task.

Experiment 1 in the current study tests whether answering easy questions prior to taking a medium-difficulty test reduces trial-by-trial overconfidence in low performers. In Experiment 1, participants answered a total of 30 four-alternative multiple-choice questions; 10 easy or hard questions followed by 20 medium-difficulty questions. Similar to Pansky and Goldsmith (2014), the first set of 10 questions served as a manipulation of the perceived task difficulty (the independent variable) and the second set of 20 questions served as target questions.

Global Performance Judgments

Similar to results of Weinstein and Roediger (2012) and Michael and Garry (2016), it is predicted that when both low and high performers initially answer 10 hard questions, they will underestimate the number of medium-difficulty target questions answered correctly relative to when they initially answer 10 easy questions – an anchoring effect. In other words, participants are predicted to anchor their *global* performance estimates according to the perceived difficulty of a task. Thus, a main effect of initial question difficulty is predicted; global estimates of number of target questions answered correctly are predicted to be higher in the initially-easy than the initially-hard condition.

Local (Trial-by-Trial) Performance Judgments

For trial-by-trial performance judgments, high performers are predicted to be similarly confident on trial-by-trial judgments when they initially answer easy questions and when they initially answer hard questions. By contrast, low performers are predicted to be less overconfident on trial-by-trial judgments when they initially answer easy questions than when they initially answer hard questions.

If high performers rely more on information-based processing to inform their confidence judgments (as results of Händel & Dresel, 2018, suggest), then local confidence judgments are predicted to be informed by item-specific information (e.g., "I remember hearing about that in class") and subsequently should be affected less by context manipulations such as relative ease of answering a question. Thus, high performers are predicted to be *equally* confident in their answers to medium-difficulty target questions when they initially answer difficult questions and when they initially answer easy questions.

Conversely, if low performers rely more on experience-based processing to inform their confidence judgments, then local confidence judgments *should* be affected by context

manipulations such as relative ease of answering a question (e.g., Kelley & Lindsay, 1993; Pansky & Goldsmith, 2014). Thus, low performers are predicted to be *more* confident in their answers to medium-difficulty target questions when they initially answer difficult questions than when they initially answer easy questions.

Method

Participants and Design

Experiment 1 is a 2 (initial question difficulty: easy, hard) x 4 (performance quartile: 1, 2, 3, 4)[3] between subjects design. The dependent variable is the difference score between the actual and the estimated number of target questions answered correctly for global performance estimates, and confidence-specific accuracy (proportion correct at each confidence level) for local performance estimates.

A power analysis was conducted to determine the sample size needed to detect an interaction between initial question difficulty and performance quartile on the dependent measures of global and local judgment accuracy. A G*Power analysis determined that for this design, the minimum sample size[4] needed to have a 95% chance of detecting a small to medium effect that exists is 331 (η_p^2 = .05, α = .05)[5]. 377 participants were recruited using Amazon's Mechanical Turk (MTurk). Participants were eligible to participate if they were over 18 years old, lived in the U.S., and had an MTurk approval rating of 90% or more. Participants were excluded from the analyses if they (a) were a univariate outlier (+/- 3 standard deviations away from the mean) on the number of easy or hard manipulation questions answered correctly, (b)

[3] Participants were categorized into performance quartiles based on the number of medium-difficulty target questions they answer correctly.
[4] The selected effect size was a conservative estimate based on the interaction effect reported by Pansky and Goldsmith (2014).
[5] A η_p^2 = .05 is analogous to a Cohen's *f* = .23, R^2 = .05.

responded to the 20 target questions with the same confidence judgment 90-100% of the time (18 or more responses), (c) were a univariate outlier (+/- 3 standard deviations away from the mean) or a multivariate outlier (Mahalanobis distance value with an associated p-value < .01) on hit rate or false alarm rate, (d) indicated on the question to this effect, that they did not pay attention to the task, (e) said that their data should not be used in the analyses, or (f) reported using outside sources to answer the questions. After the exclusion criteria were applied, the final sample consisted of 345 participants (*M* age = 38.73, *SD* = 11.33, range = 19-73; 48% female)[6]; 177 participants in the easy condition and 168 participants in the hard condition.

Materials and Procedure

Following Abed et al. (2020), stimuli were four-alternative multiple-choice general knowledge questions selected from a list of 150 normed test items from DeSoto and Simons (2015)[7]. The 10 items with the lowest accuracy (9-16% of participants responded correctly) and the 10 items with the highest accuracy (91-99%) were selected for the hard and easy questions, respectively. The 20 items that fell closest to 50% accuracy (10 items below 50% and 10 items above 50%; 46-58% accuracy) were selected for the medium-difficulty target questions.

This study was conducted using Qualtrics. Participants were told that they would answer 30 questions and would be asked periodically to assess their own performance. First, participants were randomly assigned to answer a set of 10 very easy or very hard questions. For each question, participants were instructed to (1) select the response option (out of 4) that they think is correct, (2) indicate whether they think their answer was correct or incorrect, and then (3) rate

[6] The breakdown for the Easy condition was: (*M* age = 38.37, *SD* = 11.48, range = 20-73; 51% female); the breakdown for the Hard condition was: (*M* age = 39.11, *SD* = 11.18, range = 19-73; 44% female).
[7] Each question from DeSoto and Simons was answered by 97-113 participants.

their confidence in their correct/incorrect judgment on a 6-point scale from 0% (*not at all confident*) to 100% (*absolutely confident*) in 20-point increments (a second order judgment; SOJ). After answering the 10 easy or hard questions, all participants answered the same set of 20 medium-difficulty target questions. After answering all 20 target questions, participants made a global performance judgment and estimated how many questions they thought they answered correctly out of 20.

The 6-point confidence scale used in this study is the same scale that has been used in similar studies that have examined local confidence judgments (Händel & Dresel, 2018; Abed et al., 2020). In these studies, high (but not low) performers appropriately used this confidence scale to estimate the likelihood that their performance judgments were correct. Further, this same confidence scale has been used in many studies on eyewitness memory (e.g., Grabman et al., 2019; Nguyen, Pezdek, & Abed, 2018) that demonstrate that participants can appropriately use this 6-point confidence scale[8] to estimate the likelihood that their judgments are correct.

Results

Categorizing Low and High Performers by Quartiles

[8] Although this specific 6-point confidence scale was selected for this study, results from the eyewitness memory literature suggest that the type of confidence scale does not affect participants' ability to adjust their confidence ratings to account for the strength of their response. For example, Nguyen and colleagues (2016) reported that across three experiments, regardless of the confidence scale (i.e., a 1-5 or a 1-10 scale), participants adjusted their confidence ratings to account for variables that are known to affect recognition memory, such as exposure duration or race of the target face (see Mickes, 2015 for a discussion of the utility of a high-confidence response regardless of the confidence scale used). Moreover, researchers have reported that the confidence-accuracy relationship does not vary as a function of the range of the confidence scale (e.g., 4-point, 5-point, 20-point, or 100-point; Tekin & Roediger, 2017) or whether the confidence scale is verbal only or both verbal and numeric (e.g., "completely confident" versus "4 - completely confident"; Tekin, Lin, & Roediger, 2018). These results indicate that the type of confidence scale used in an experiment does not affect the confidence accuracy relationship.

The mean number of medium-difficulty questions answered correctly did not differ for participants in the easy (M = 11.20, 95% CI [10.67, 11.73]) and hard (M = 11.15, 95% CI [10.62, 11.68]) conditions $t(343) = 0.13$, $p = .898$, $d = .014$. Thus, scores denoting the performance quartile cutoffs were calculated based on all 345 participants.

Performance quartiles were determined based on the number of medium-difficulty target questions answered correctly (out of 20). Overall, the number of questions answered correctly ranged from 2-20, and the number of questions answered correctly for each quartile was: 2-8 (bottom 25%), 9-11 (50th percentile), 12-14 (75th percentile), and 15-20 (top 25%)[9]. The number of target questions answered correctly was significantly different for each of the performance quartiles, $F(3, 341) = 992.55$, $p < .001$, $\eta_p^2 = .897$ (all pairwise comparisons significant, $p < .001$). For the remaining analyses, the bottom and top quartiles will be referred to as low and high performers respectively, and the 50th and 75th percentiles will be referred to as low-middle and high-middle respectively[10].

Accuracy of the *Global* Metacognitive Judgments

All analyses were conducted on data for the 20 medium-difficulty target questions that all 345 participants answered. The initial 10 easy or hard questions served only as a manipulation of task difficulty and were not included in the analyses[11].

[9] The performance quartile cutoffs were in fact the same for both experimental conditions.
[10] The breakdown of demographics for low performers was: (M age = 34.84, SD = 9.62, range = 19-73; 53% female); the breakdown for low-middle performers was: (M age = 37.62, SD = 9.87, range = 20-67; 52% female); the breakdown for high-middle performers was: (M age = 40.12, SD = 10.80, range = 21-73; 43% female); the breakdown for high performers was: (M age = 43.7, SD = 14.20, range = 22-73; 43% female).
[11] The average number of easy manipulation questions answered correctly was 9.59 (SD = 0.66, range = 7-10). The average number of hard manipulation questions answered correctly was 1.82 (SD = 1.68, range = 0-8). All analyses were conducted including and excluding the 10 participants who answered more than 4 hard questions correctly and the results did not differ.

Following Kruger and Dunning (1999; see also Abed et al., 2020), difference scores on the global metacognitive judgements for the 20 target questions were computed by subtracting the actual number correct from the estimated number correct (scores closer to zero indicate more accurate estimation; positive scores indicate overconfidence; negative scores indicate underconfidence).

Do High and Low Performers Anchor their Global Confidence Judgments Based on a Preceding Task?

A 2 (initial question difficulty: easy, hard) x 4 (performance quartile: 1, 2, 3, 4) between subjects ANOVA was conducted on difference scores. There was a significant main effect of performance quartile, $F(3, 337) = 12.044$, $p < .001$, $\eta_p^2 = .097$. Replicating the *skilled*-unaware pattern that Burson et al. (2006) reported for difficult (compared to easy) tests, pairwise comparisons revealed that overall, low performers underestimated their performance ($M = -.67$, 95% CI [-1.49, .15]) to a smaller extent than low-middle performers ($M = -2.14$, 95% CI [-3.93, -2.35]), high-middle performers ($M = -3.41$, 95% CI [-4.17, -2.66]), and high performers ($M = -4.11$, 95% CI [-5.09, -3.12]). No other comparisons were significant.

There was also a significant main effect of initial question difficulty, $F(1, 337) = 14.83$, $p < .001$, $\eta_p^2 = .042$. Overall, participants in the hard condition ($M = -3.66$, 95% CI [-4.26, -3.05]) underestimated their performance to a greater extent than participants in the easy condition ($M = -2.01$, 95% CI [-2.59, -1.43]). This finding does not support the fluency hypothesis of Pansky and Goldsmith (2014), that participants in the easy condition would underestimate their performance (because medium-difficulty questions feel difficult in comparison) and that participants in the hard condition would overestimate their performance (because medium-difficulty questions feel easy in comparison). Rather, replicating findings by Weinstein and

Roediger (2012) and Michael and Garry (2016), this result suggests that participants anchored their performance estimates for the second, medium-difficulty task based on the difficulty of the initial 10 questions.

Finally, the interaction between performance quartile and initial question difficulty was not significant, $F(3, 337) = 0.25$, $p = .860$, $\eta_p^2 = .002$.

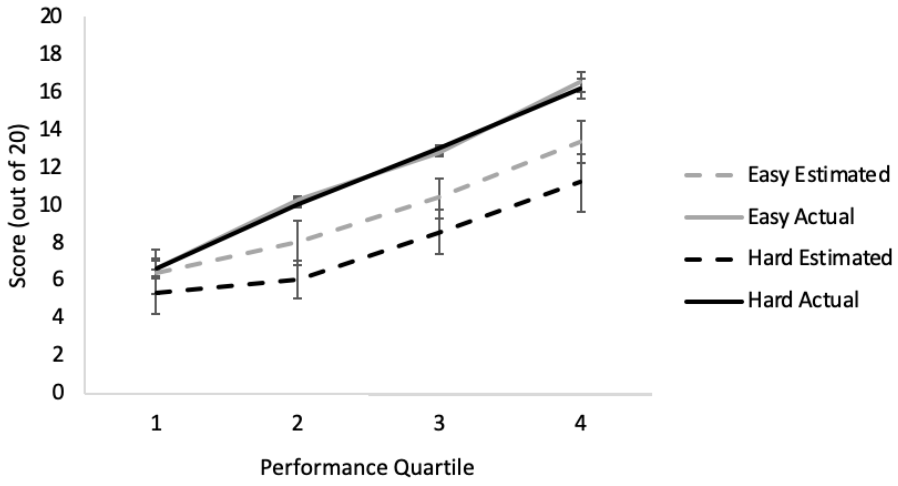


Figure 2. The mean actual and estimated number correct on the 20 medium-difficulty target questions for each of the four performance quartiles as a function of initial question difficulty. Difference scores were computed by subtracting the actual number correct from the estimated number correct, and actual number correct was the same for participants in the easy and hard conditions. Error bars represent 95% confidence intervals.

Accuracy of the *Trial-by-Trial* Metacognitive Judgments

Global metacognitive judgments alone cannot provide insight into the questions of (a) whether high and low performers can accurately distinguish between information they do and do

not know, and (b) whether high and low performers are equally affected by context manipulations such as relative ease of answering a question. To address these questions, analyses were conducted on the trial-by-trial judgments.

First-Order Metacognitive Judgments: Are the Unskilled "Subjectively Aware"?

Handel & Dresel (2018) reported that low performers provided significantly fewer "yes" performance judgments (i.e., thought they solved items correctly less often) than high performers. In other words, to some extent, low performers were aware that they did not perform as well as high performers. To test whether this result replicated in the current experiment, a 2 (initial question difficulty: easy, hard) x 4 (performance quartile: 1, 2, 3, 4) between subjects ANOVA was conducted on the number of "yes/correct" performance judgments. There was a significant main effect of performance quartile, $F(3, 337) = 24.55$, $p < .001$, $\eta_p^2 = .179$. Pairwise comparisons revealed that overall, low performers thought they answered questions correctly less often ($M = 11.32$, 95% CI [10.42, 12.22]) than high performers ($M = 17.25$, 95% CI [16.17, 18.33]) and high-middle performers ($M = 13.96$, 95% CI [13.13, 14.79]), but not low-middle performers ($M = 12.76$, 95% CI [11.90, 13.62]). Additionally, low-middle performers and high-middle performers thought they answered questions correctly less often than high performers; see Figure 3 for significant pairwise comparisons. Neither the main effect of initial question difficulty [$F(1, 337) = 0.32$, $p = .575$, $\eta_p^2 = .001$] nor the interaction [$F(3, 337) = 0.33$, $p = .80$, $\eta_p^2 = .003$] were significant. This result indicates that to some extent, low performers were cognizant of the fact that they did not perform well, and high performers were cognizant of the fact that they did perform well.

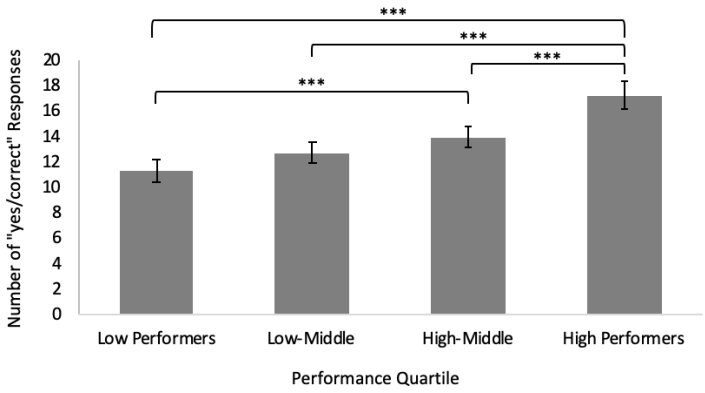


Figure 3. Mean number of "yes/correct" performance judgments for each of the four performance quartiles. Significant pairwise comparisons are denoted with ***p < .001. Error bars represent 95% confidence intervals.

Proportion Correct (CAC Analysis)

To understand the metacognition-performance relationship as a function of initial question difficulty, confidence-specific accuracy and CAC curves (Mickes, 2015) were calculated. Proportion correct was calculated based on hits (accurate "correct" responses) and false alarms (inaccurate "correct" responses) [# $hits_c$/(# $hits_c$ + # false $alarms_c$)] for each level of confidence, where *c* indicates hits and false alarms that were made with a specific level of confidence. The six levels of confidence were binned into low (0-20%), medium (40-60%), and high (80-100%) confidence for all analyses of proportion correct as is commonly done to increase the stability of proportion correct estimates (Wixted et al., 2015).

Can High and Low Performers Similarly Distinguish Between Information They Do and Do Not Know?

To test the prediction that (a) high performers would be similarly confident when they initially answer easy questions and when they initially answer hard questions, but (b) low performers would be less overconfident on trial-by-trial judgments when they initially answer

easy questions than when they initially answer hard questions, three separate 2 (initial question difficulty: easy, hard) x 4 (performance quartile: 1, 2, 3, 4) ANOVAs were conducted on proportion correct at each level of confidence. A single 2 (initial question difficulty: easy, hard) x 3 (confidence: low, medium, high) x 4 (performance quartile: 1, 2, 3, 4) ANOVA was not conducted because there would have been too few cases (124 participants) with non-missing data (i.e., a proportion correct calculated using non-zero data for each confidence level). For this series of tests, a Bonferroni correction of $\alpha = .017$ (.05 / # of tests) was used.

For high confidence responses (80-100% confidence), there was a significant effect of performance quartile on proportion correct, $F(3, 330) = 45.47$, $p < .001$, $\eta_p^2 = .292$. Proportion correct was higher for high performers ($M = .94$, 95% CI [.89, .98]) than for low performers ($M = .60$, 95% CI [.56, .64]) and low-medium performers ($M = .74$, 95% CI [.71, .78]), but not high-medium performers ($M = .87$, 95% CI [.83, .90]). All other comparisons were significant, $p < .001$. Contrary to predictions, neither the main effect of initial question difficulty nor the interaction was significant (F's < 1, p's > .9). See Figure 4 for proportion correct as a function of performance quartile and initial question difficulty.

For medium confidence responses (40-60% confidence), there was a significant effect of performance quartile on proportion correct, $F(3, 315) = 32.12$, $p < .001$, $\eta_p^2 = .234$. Proportion correct was higher for high performers ($M = .73.$, 95% CI [.67, .80]) than for low performers ($M = .34$, 95% CI [.29, .40]), low-medium performers ($M = .45$, 95% CI [.40, .51]), and high-medium performers ($M = .61$, 95% CI [.56, .67]). All other comparisons were significant, $p < .05$. Contrary to predictions, there was not a significant effect of initial question difficulty on proportion correct at medium confidence, $F(1, 315) = 2.29$, $p = .131$, $\eta_p^2 = .007$. The interaction between performance quartile and initial question difficulty was not significant ($F < 1$, $p > .9$).

Finally, for low confidence responses (0-20% confidence), there was a significant effect of performance quartile on proportion correct, $F(3, 124) = 4.83$, $p = .003$, $\eta_p^2 = .105$. Proportion correct was higher for high performers ($M = .69$, 95% CI [.52, .87]) than for low performers ($M = .34$, 95% CI [.23, .44]), but not low-medium performers ($M = .47$, 95% CI [.36, .59]) or high-medium performers ($M = .56$, 95% CI [.44, .67]). Proportion correct was also significantly higher for high-medium performers than low performers ($p < .05$). No other comparisons were significant. Contrary to predictions, neither the main effect of initial question difficulty nor the interaction were significant (F's < 1.5, p's > .25).

Across all three levels of confidence, proportion correct was higher for high than low performers, indicating that compared to low performers, high performers were more accurate discriminating between answers that were more vs. less likely to be correct. This result is consistent with previous research suggesting that high performers are more accurate than low performers discriminating between information they do and do not know (Abed et al., 2020). However, contrary to the prediction that initially answering easy questions would reduce item-level overconfidence in low performers, initial item difficulty did not affect the degree of overconfidence for low performers. Item-level overconfidence was *not* reduced for participants in the easy condition; rather, low performers were similarly overconfident in both the easy and the hard conditions with low proportion correct even at high confidence.

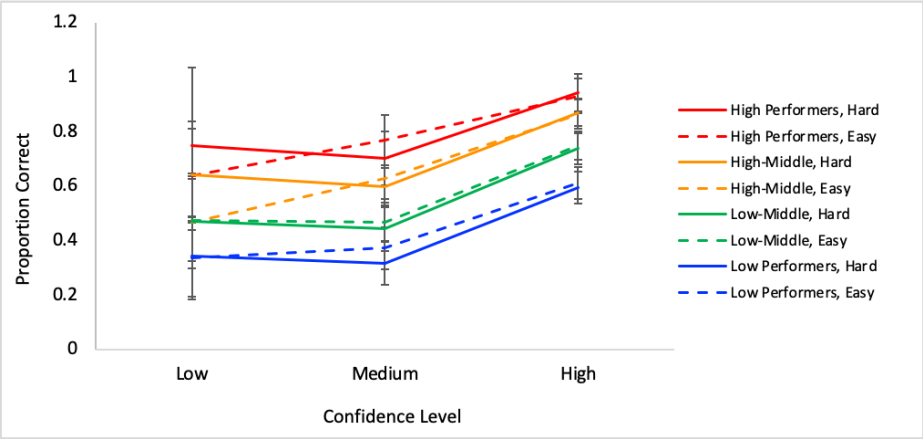


Figure 4. Confidence-accuracy characteristic (CAC) curves (proportion correct on 20 medium-difficulty target questions at low, medium, and high confidence) for the four performance quartiles as a function of initial question difficulty (easy vs. hard). Error bars represent 95% confidence intervals.

Discussion

Findings from the global metacognitive judgments in Experiment 1 replicated results from previous research on the unskilled-unaware effect using difficult tests (e.g., Burson et al., 2006). Unlike the typical finding that low performers overestimate their performance while high performers underestimate their performance, across all performance levels in Experiment 1, participants generally underestimated their performance. Nevertheless, low performers did so to a lesser degree than higher performers. Although the multiple-choice questions were selected for this study with the intent of constructing a medium-difficulty test, the pattern of results replicated Burson et al.'s (2006) finding of a *skilled*-unaware pattern for difficult tests (in contrast to the typical *unskilled*-unaware pattern for relatively more easy tests). This suggests that participants may have perceived the test to be more difficult than "medium" difficulty. In other words, there may be a discrepancy between the operational definition of "medium" difficulty (i.e., 50%

accuracy) and subjective "medium" difficulty. Indeed, when asked the end of the task, "How difficult were the 20 questions you just answered on a scale from 1 (*not at all difficult*) to 6 (*very difficult*)?" the mean response was 4.68.

Nevertheless, the main purpose of Experiment 1 was to test how initial question difficulty affected later global and local metacognitive judgments – not to assess how task difficulty affects the degree of over- and under-confidence. Overall, participants in the hard condition underestimated their performance to a greater extent than participants in the easy condition, supporting the anchoring hypothesis (Michael & Garry, 2016; Weinstein & Roediger, 2012) but not the fluency hypothesis (Pansky & Goldsmith, 2014). That is, this result suggest that participants anchored performance estimates for the medium-difficulty task based on the difficulty of the initial 10 questions. Additionally, all but the lowest performers provided significantly lower estimates of number correct in the hard than the easy condition. One possible explanation for why low performers provided similar estimates for number correct in both the easy and hard condition is that participants may have been reluctant to provide performance estimates close to zero. With already low estimates of number correct (out of 20 questions; M = 6.47), perhaps low performers in the hard condition (M = 5.41) simply were unwilling to provide even lower estimates.

Of particular interest in this study, CAC analyses assessed participant's likelihood of being correct as a function of trial-by-trial confidence level. In other words, could low and high performers distinguish between information they did and did not know, and adjust their confidence judgments accordingly? Across all three confidence levels, proportion correct was higher for high than low performers. The most critical finding is that at high confidence, higher performers were more accurate at distinguishing between information they did and did not know;

proportion correct was significantly greater for high than low performers. This finding replicates previous research that suggests that metacognitive monitoring is more accurate for high than low performers (e.g., Abed et al., 2020; Händel & Dresel, 2018). However, the effect of initial question difficulty and the interaction between initial question difficulty were not significant, and critically, initially answering easy test questions did not reduce overconfidence in low performers. Similar to the results reported in previous studies (e.g., Weinstein & Roediger, 2012, and Michael and Garry, 2016), initial question difficulty affected global, but not local judgments.

Taken together with previous findings, results from Experiment 1 indicate that when making global judgments at the end of a task, people likely rely on various sources of information, including perceptions of task difficulty to inform their metacognitive judgments. By contrast, when making trial-by-trial judgments, people more likely rely on information specific to the question itself (e.g., information from memory *or* gut feelings) and not information about the task to inform their confidence judgments. Similar to previous findings (e.g., Weinstein & Roediger, 2012), participants in this experiment anchored their global performance estimates based on the difficulty of a preceding task, but task difficulty had no effect on trial-by-trial confidence judgments. Results of Experiment 1 provide further evidence that compared to global judgments made at the end of a task, trial-by-trial confidence judgments reflect a purer assessment of peoples' confidence in their own performance.

Chapter 3: Experiment 2

Experiment 1 tested whether a contextual manipulation (i.e., initial question difficulty) would increase metacognitive judgment accuracy by reducing overconfidence. However, training monitoring ability independently of task performance would be exceedingly useful to practitioners who are unable to train people to perform well on all tasks they are expected to

master. This includes, for example, instructors who want their students to better assess their knowledge across a variety of subjects, and not just on a single task. Although results from many studies suggest that monitoring ability and task performance are independent (e.g., Weinstein & Roediger, 2012; Zechmeister et al., 1986), it is less clear whether monitoring ability can be trained independently of task performance, and further, whether such training would transfer to other tasks.

Results of Zechmeister, Rusch, and Markell (1986) suggest that people can be trained to improve their calibration accuracy independently of task performance (i.e., people who receive calibration training improve the accuracy of their confidence judgments *but not* test performance relative to a control group) and that the effect of training persists over at least two weeks. However, it is unclear from these results whether participants simply shifted their use of the confidence scale (e.g., said "100% confident" less often) or whether they actually learned to distinguish better between what they did and did not know. In other words, participants might have reduced their confidence judgments overall, but ideally participants would better calibrate their confidence judgments to distinguish between what they do and do not know; that is, provide high confidence for hits but low confidence for false alarms.

Some research suggests that such an outcome is possible; results from Rawson and Dunlosky (2007) indicate that feedback (e.g., showing people the correct answer) reduces overconfidence specifically for commission errors (i.e., false alarms). However, it is unclear from their study whether effects of feedback can transfer to other judgments within a task ("near" transfer), much less to other tasks ("far" transfer).

Can people (especially low performers) utilize feedback to distinguish better between information they do and do not know, and then apply this training to a later task? Experiment 2

tests this question. First, it is hypothesized that training will increase monitoring accuracy, replicating similar findings (e.g., Zechmeister et al., 1986); proportion correct will be higher in the training than the control condition. Second, given that low performers are less accurate than high performers discriminating between information they do and do not know (e.g., Abed et al., 2020), it is hypothesized that low performers will benefit the most from training; the difference between the proportion correct in the training and control conditions will be greater for low than high performers.

Method

Participants and Design

This experiment is a 2 (training, control) x 4 (performance quartile: 1, 2, 3, 4) between subjects design. The dependent variable is the difference score (between actual and estimated number of target questions answered correctly) for global performance estimates, and confidence-specific accuracy (proportion correct at each confidence level) for local performance estimates. A power analysis was conducted to determine the sample size needed to detect an interaction between initial question difficulty and performance quartile on the dependent measures of global and local judgment accuracy. A G*Power analysis determined that for this design, the minimum sample size[12] needed to have a 95% chance of detecting a small to medium effect that exists is 331 (η_p^2 = .05, α = .05)[13]. 383 participants were recruited using Amazon's Mechanical Turk (MTurk). As in Experiment 1, participants were eligible to participate if they were over 18 years old, lived in the U.S., and had an MTurk approval rating of 90% or more. Participants were excluded from the analyses if they (a) responded to the 60 test questions with

[12] The selected effect size was a conservative estimate based on the interaction effect reported by Nietfeld and Schraw (2002).
[13] A η_p^2 = .05 is analogous to a Cohen's f = .23, R^2 = .05.

the same confidence judgment 90-100% of the time (55 or more responses), (b) were a univariate outlier (+/- 3 standard deviations away from the mean) or a multivariate outlier (Mahalanobis distance value with an associated p-value < .01) on hit rate or false alarm rate, (c) indicated on the question to this effect, that they did not pay attention to the task, (d) said that their data should not be used in the analyses, or (e) reported using outside sources to answer the questions. After the exclusion criteria were applied, the final sample consisted of 358 participants (*M* age = 41.75, *SD* = 12.29, range = 20-82; 48% female)[14]; 176 participants in the control condition and 182 participants in the training condition.

Materials and Procedure

As in Experiment 1, stimuli used in Experiment 2 were selected from a list of normed items from DeSoto and Simons (2015). 60 questions were selected for Experiment 2 by taking the items closest to 50% difficulty and randomly assigning them to one of three blocks. These three blocks had identical mean accuracy (*M* = 50%, range = 31-69%).

This study was conducted using Qualtrics. Participants answered a total of 60 questions, in three blocks of 20 questions; a pre-test block (Block 1), a training block (Block 2), and a post-test block (Block 3). After each block, participants estimated the number of questions they answered correctly out of 20 (a global judgment). The question procedure was the same that described in Experiment 1; for each of the 60 questions, participants selected their response, indicated whether they think their response was correct or incorrect, and rate their confidence in their correct/incorrect judgment using the same 0-100% confidence scale used in Experiment 1.

[14] The breakdown for the Training condition was: (*M* age = 42.07, *SD* = 12.90, range = 20-82; 46% female); the breakdown for the Control condition was: (*M* age = 41.43, *SD* = 11.65, range = 21-74; 49% female).

For participants in the control condition, the procedure for the training block (Block 2) was exactly the same as the procedure for the pre-test block (Block 1). For participants in the training condition, the procedure for the training block varied in two ways. First, participants in the training condition were given instructions to help them improve the accuracy of their confidence judgments[15]. Second, for each of the 20 questions in the training block, after making their confidence judgment, participants were told whether they got the question correct or incorrect. In the instructions administered to participants in the training condition, participants were told to use the correct/incorrect feedback to judge the appropriateness *of each confidence judgment*. Thus, participants in the training condition received feedback about the accuracy of their *confidence* judgment (e.g., "I should not have said I was 100% confident because I turned out to be wrong"). For all participants, the post-test block (Block 3) followed the same procedure as the pre-test block (Block 1), with no feedback. See Figure 5 for an outline of the Experiment 2 procedure.

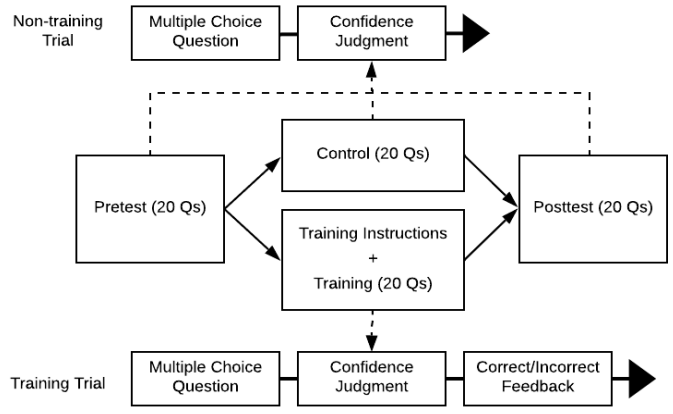


Figure 5. Experiment 2 procedure.

[15] The training used in this study was closely modelled on that of Zechmeister, Rusch, and Markell (1986). The key elements of their training were: (a) instructing participants how to weight evidence for and against a response option and (b) providing response feedback. See Appendix A for the full training procedure.

Results

Categorizing Low and High Performers by Quartiles

The mean number of pre-test questions answered correctly (out of 20) did not differ for participants in the training (M = 11.27, 95% CI [10.80, 11.85]) and control (M = 11.77, 95% CI [11.26, 12.28]) conditions $t(356) = 1.282$, $p = .201$, $d = .136$, and this pattern was the same for the manipulation (training or control) block, the post-test, and for all 60 questions collapsed across block[16]. Thus, scores denoting the performance quartile cutoffs were calculated based on all 358 participants (rather than calculating quartile for training and control blocks separately).

Performance quartiles were determined based on the number of pre-test questions answered correctly (out of 20). Overall, the number of questions answered correctly ranged from 3-20, and the number of questions answered correctly for each quartile was: 3-9 (bottom 25%), 10-12 (50th percentile), 13-14 (75th percentile), and 15-20 (top 25%)[17]. The number of pre-test questions answered correctly was significantly different for each of the performance quartiles, $F(3, 354) = 882.73$, $p < .001$, $\eta_p^2 = .882$ (all pairwise comparisons significant, $p < .001$). As in Experiment 1, for the remaining analyses, the bottom and top quartiles will be referred to as low

[16] The number of questions answered correctly also did not differ for the manipulation (training or control) block (M_t = 10.58, 95% CI [10.03, 11.14], M_c = 11.20, 95% CI [10.70, 11.70]; $t(356) = 1.63$, $p = .109$, $d = .173$), the post-test (M_t = 10.89, 95% CI [10.34, 11.44], M_c = 11.14, 95% CI [10.62, 11.66]; $t(356) = 0.66$, $p = .513$, $d = .069$), or overall for all 60 questions collapsed across block (M_t = 32.75, 95% CI [31.29, 34.20], M_c = 34.11, 95% CI [32.79, 35.44]; $t(356) = 1.37$, $p = .172$, $d = .144$).

[17] The performance quartile cutoffs were also computed for the two experimental conditions separately, as the cutoffs were slightly different. In the Training condition, the number of questions answered correctly for each quartile was: 3-9 (bottom 25%), 10-11 (50th percentile), 12-14 (75th percentile), and 15-20 (top 25%). In the Control condition, the number of questions answered correctly for each quartile was: 3-10 (bottom 25%), 11-12 (50th percentile), 13-14 (75th percentile), and 15-20 (top 25%). However, results of the analyses were the same regardless of how the performance quartile variable was determined.

and high performers respectively, and the 50th and 75th percentiles will be referred to as low-middle and high-middle respectively[18].

Accuracy of the *Global* Metacognitive Judgments

As in Experiment 1, difference scores on the global metacognitive judgements were computed by subtracting the actual number correct from the estimated number correct in the post-test (scores closer to zero indicate more accurate estimation; positive scores indicate overconfidence; negative scores indicate underconfidence).

First, an independent samples t-test was conducted to determine whether training and control groups differed on the differences scores in the pre-test, and they did not. The difference between estimated and actual number of questions answered correctly was similar for participants in the training (M = -1.73, 95% CI [-2.32, -1.14]) and control conditions (M = -1.98, 95% CI [-2.58, -1.39], $t(356) = 0.59$, $p = .554$, $d = .062$). Thus, subsequent analyses were conducted on post-test difference scores.

The first analyses tested whether the unskilled-unaware effect was replicated for the global performance judgments. A 2 (training, control) x 4 (performance quartile: 1, 2, 3, 4) between subjects ANOVA was conducted on post-test differences scores. The main effect of training [$F(1, 350) = 0.11$, $p = .745$, $\eta_p^2 < .001$], the main effect of performance quartile [$F(3, 350) = 2.24$, $p = .084$, $\eta_p^2 = .019$] and the interaction between training and performance quartile were all non-significant, $F(3, 350) = 1.16$, $p = .326$, $\eta_p^2 = .010$. Planned comparisons revealed that there was no significant difference between training and control conditions for low

[18] The breakdown of demographics for low performers was: (M age = 38.88, SD = 11.81, range = 20-69; 60% female); the breakdown for low-middle performers was: (M age = 42.65, SD = 12.44, range = 21-69; 48% female); the breakdown for high-middle performers was: (M age = 40.36, SD = 10.60, range = 24-68; 42% female); the breakdown for high performers was: (M age = 44.84, SD = 12.94, range = 21-82; 36% female).

performers (M_t = -2.15, 95% CI [-3.28, -1.02], M_c = -2.54, 95% CI [-3.62, -1.45], t(101) = -0.48, p = .635, d = -.096), low-middle performers (M_t = -2.72, 95% CI [-3.87, -1.57], M_c = -3.49, 95% CI [-4.41, -2.57], t(111) = -1.06, p = .293, d = -.199), high-middle performers (M_t = -3.00, 95% CI [-4.54, -1.06], M_c = -1.58, 95% CI [-2.84, -.32], t(53) = 1.33 p = .190, d = .365), or high performers (M_t = -1.30, 95% CI [-2.44, -.15], M_c = -2.12, 95% CI [-3.09, -1.14], t(85) = -1.10, p = .275, d = .236). The unskilled-unaware effect was not replicated in Experiment 2; across all performance quartiles, participants generally underestimated their performance, but all participants did so to a similar degree. See Figure 6 for the actual and estimated number of questions answered correctly on the post-test for each performance quartile as a function of training condition.

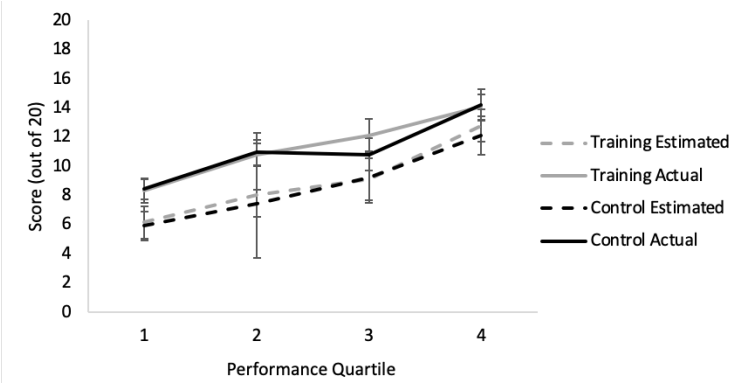


Figure 6. The mean actual and estimated number correct on the post-test for each of the four performance quartiles as a function of training condition. Difference scores were computed by subtracting the actual number correct from the estimated number correct, and actual number correct was the same for participants in the training and control conditions. Error bars represent 95% confidence intervals.

Accuracy of the *Trial-by-Trial* Metacognitive Judgments

First-Order Metacognitive Judgments: Are the Unskilled "Subjectively Aware"?

First, an independent samples t-test was conducted to determine whether training and control groups differed on the number of "yes/correct" performance judgments in the pre-test, and they did not. The number of "yes/correct" performance judgments (out of 20) was similar for participants in the training (M = 13.02, 95% CI [12.34, 13.69]) and control conditions (M = 12.90, 95% CI [12.23, 13.57], $t(356) = -0.24$, $p = .815$, $d = .026$). Thus, subsequent analyses were conducted on the number of "yes/correct" performance judgments in the post-test.

Similar to Experiment 1, a 2 (training, control) x 4 (performance quartile: 1, 2, 3, 4) between subjects ANOVA was conducted on the number of "yes/correct" performance judgments in the post-test. There was a significant main effect of performance quartile, $F(3, 350) = 21.06$, $p < .001$, $\eta_p^2 = .153$. Pairwise comparisons revealed that overall, low performers thought they answered questions correctly less often (M = 9.80, 95% CI [8.84, 10.76]) than low-middle performers (M = 13.08, 95% CI [12.11, 13.90]), high-middle performers (M = 13.24, 95% CI [11.93, 14.54]), and high performers (M = 15.37, 95% CI [14.35, 16.39]). Additionally, low-middle performers thought they answered questions correctly less often than high performers; see Figure 7 for significant pairwise comparisons. The main effect of training condition [$F(1, 350) = 0.42$, $p = .518$, $\eta_p^2 = .001$] and the interaction term [$F(3, 350) = 0.94$, $p = .423$, $\eta_p^2 = .008$] were both nonsignificant. As in Experiment 1, results on the number of "yes/correct" performance judgments indicate that to some extent, low performers were aware that they did not perform well, and high performers were aware that they did perform well.

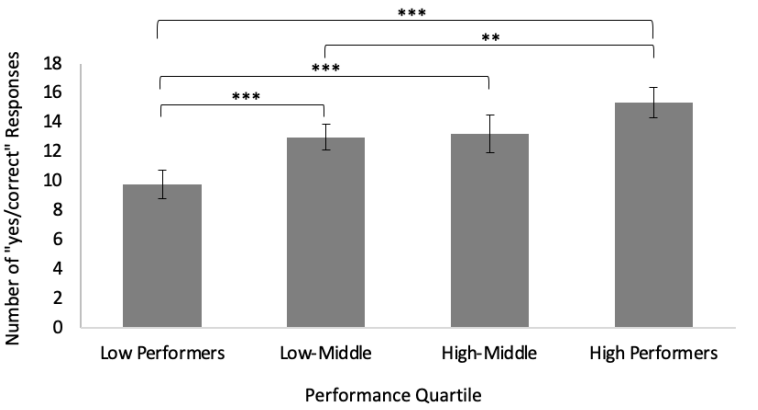


Figure 7. Mean number of "yes/correct" performance judgments for each of the four performance quartiles. Significant pairwise comparisons are denoted with ***p < .001, **p < .01. Error bars represent 95% confidence intervals.

Proportion Correct (CAC) Analysis

The primary question of interest for Experiment 2 is, after training, can people (especially low performers) learn to distinguish between responses they are more vs. less likely to have answered correctly? In other words, can metacognitive training teach people to provide high confidence judgments for "correct" responses only when they are likely to be correct?

As in Experiment 1, proportion correct for the post-test was calculated based on hits (accurate "correct" responses) and false alarms (inaccurate "correct" responses) [# $hits_c$/(# $hits_c$ + # false $alarms_c$)] for each level of confidence, where *c* indicates hits and false alarms that were made with a specific level of confidence. The six levels of confidence were binned into low (0-20%), medium (40-60%), and high (80-100%) confidence for all analyses of proportion correct.

First, an independent samples t-test was conducted to determine whether training and control groups differed on proportion correct in the pre-test. For high confidence responses (80-100% confidence), proportion correct was similar for participants in the training (M = .84, 95%

CI [.81, .87]) and control conditions (M = .86, 95% CI [.84, .89], $t(351) = 0.99$, p = .321, d = .106).

However, for medium confidence responses (40-60% confidence), proportion correct was lower for participants in the training (M = .56, 95% CI [.51, .60]) than the control condition (M = .62, 95% CI [.58, .67], $t(338) = 2.00$, p = .046, d = .217). Similarly, for low confidence responses (40-60% confidence), proportion correct was lower for participants in the training (M = .35, 95% CI [.27, .44]) than the control conditions (M = 54, 95% CI [.45, .62], $t(146) = 3.09$, p = .002, d = .508).

Given that pre-test proportion correct was not similar for the training and control conditions for low and medium confidence responses, the next set of analyses of proportion correct were conducted on proportion correct change scores. For each of the three confidence levels (low, medium, and high) a proportion correct change score was calculated by subtracting the pre-test proportion correct from the post-test proportion correct. Thus, positive change scores indicate that proportion correct *increased* from pre-test to post-test, and negative change scores indicate that proportion correct *decreased* from pre-test to post-test.

Does Training Help People Distinguish Between Information They Do and Do Not Know?

To test the prediction that training increases monitoring accuracy (i.e., positive proportion correct change scores), especially for low performers, three separate 2 (training, control) x 4 (performance quartile: 1, 2, 3, 4) ANOVAs were conducted on proportion correct change scores at each level of confidence. A single 2 (training, control) x 3(confidence: low, medium, high) x 4 (performance quartile: 1, 2, 3, 4) ANOVA was not conducted because there would have been too few cases (86 participants) with non-missing data (i.e., a proportion correct calculated using non-

zero data for each confidence level, for *both* pre-test and post-test). For this series of tests, a Bonferroni correction of α = .017 (.05 / # of tests) was used.

For high confidence responses (80-100% confidence), the main effect of performance quartile was not significant, $F(3, 326) = 2.36$, $p = .072$, $\eta_p^2 = .021$. Contrary to the hypothesis that the change in proportion correct would be greater in the training than the control condition, the main effect of training was not significant, $F(1, 326) = 1.78$, $p = .183$, $\eta_p^2 = .005$. Additionally, contrary to the hypothesis that training would have a larger effect on proportion correct for low than high performers, the interaction between training and performance quartile was not significant, $F(3, 326) = 2.53$, $p = .057$, $\eta_p^2 = .023$. See Figure 8 for proportion correct change scores as a function of performance quartile and training condition.

Planned comparisons were conducted to compare the difference between training and control groups on high confidence proportion correct change scores at each performance quartile. Although these comparisons were all non-significant, the pattern of results was contrary to predictions. Planned comparisons revealed that the difference in proportion correct change scores between the training and control groups was greater for low-middle performers ($M_t = .01$, 95% CI [-.04, .06], $M_c = -.08$, 95% CI [-.14, -.01], $t(108) = -12.18$, $p = .031$, $d = -.417$) and high-middle performers ($M_t = .01$, 95% CI [-.08, .09], $M_c = -.13$, 95% CI [-.21, -.05], $t(52) = -2.22$, $p = .031$, $d = -.619$), than for low performers ($M_t = .00$, 95% CI [-.10, .10], $M_c = .05$, 95% CI [-.05, .15], $t(81) = 0.69$, $p = .495$, $d = .154$) or high performers ($M_t = -.08$, 95% CI [-.13, -.02], $M_c = -.05$, 95% CI [-.09, -.01], $t(85) = 0.74$, $p = .460$, $d = .159$). That is, proportion correct was predicted to *increase* from pre-test to post-test in the training group (i.e., positive change scores) and stay the same in the control group (i.e., change scores close to zero). However, for low-middle and high-middle performers, proportion correct did not change in the training group but

slightly *decreased* from pre-test to post-test in the control group, as evidenced by negative change scores (though again, this pattern was not significant).

For medium confidence responses (40-60% confidence), the main effect of performance quartile was significant, $F(3, 303) = 4.69$, $p = .003$, $\eta_p^2 = .044$. For high performers, proportion correct decreased from the pre-test to the post-test ($M = -.12$, 95% CI [-.21, -.04]), and this change was significantly different than for low performers, for whom proportion correct slightly increased from pre-test to post-test ($M = .08$, 95% CI [-.01, .16], $p = .005$). Low performers also differed significantly on proportion correct change scores from high-middle performers ($M = -.12$, 95% CI [-.23, -.02], $p = .023$). No other comparisons were significant.

Contrary to the hypothesis that the change in proportion correct would be greater in the training than the control condition, the main effect of training was not significant, $F(1, 303) = 2.48$, $p = .116$, $\eta_p^2 = .008$. Additionally, contrary to the hypothesis that training would have a larger effect on proportion correct for low than high performers, the interaction between training and performance quartile was not significant, $F(1, 303) = 0.86$, $p = .462$, $\eta_p^2 = .008$.

For low confidence responses (0-20% confidence), the main effect of performance quartile was not significant, $F(3, 93) = 1.34$, $p = .265$, $\eta_p^2 = .042$. Contrary to the hypothesis that the change in proportion correct would be greater in the training than the control condition, the main effect of training was not significant, $F(1, 93) = 1.01$, $p = .317$, $\eta_p^2 = .011$. Additionally, contrary to the hypothesis that training would have a larger effect on proportion correct for low than high performers, the interaction between training and performance quartile was not significant, $F(1, 93) = 0.84$, $p = .478$, $\eta_p^2 = .026$.

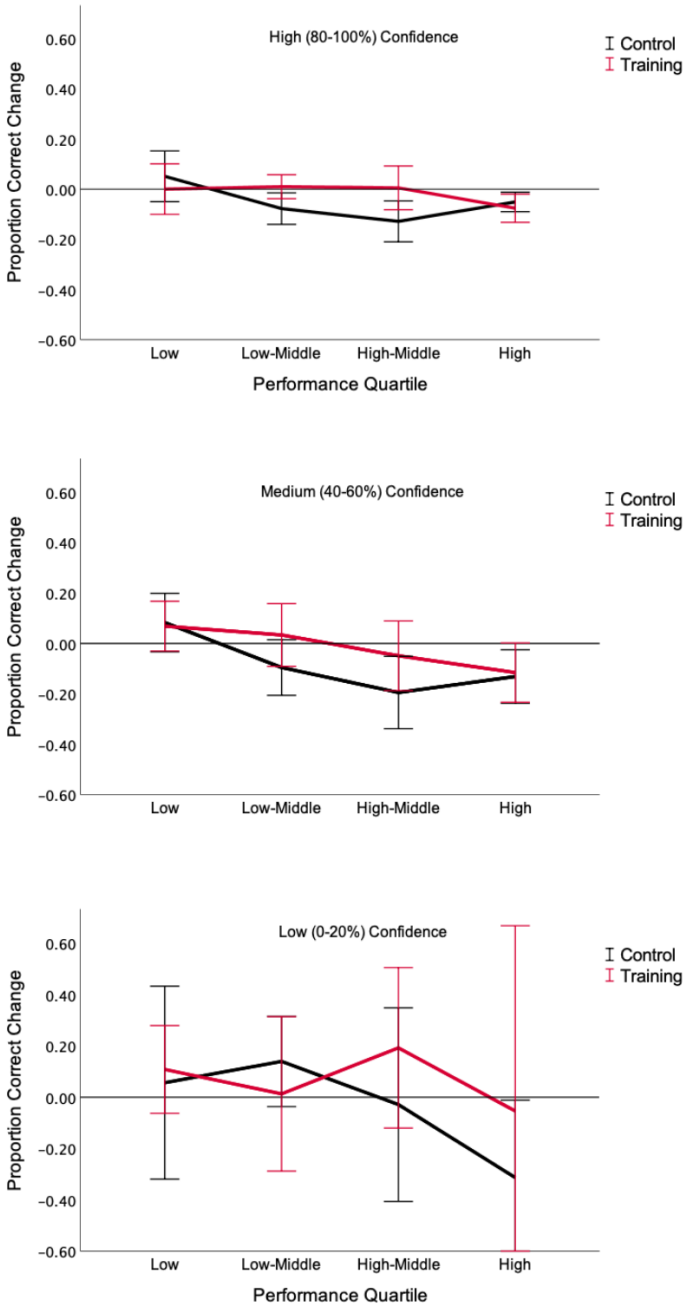


Figure 8. Proportion correct change scores (post-test minus pre-test) for the four performance quartiles as a function of training condition. Positive change scores indicate that proportion correct *increased* from pre-test to post-test, and negative change scores indicate that proportion correct *decreased* from pre-test to post-test. Error bars represent 95% confidence intervals.

Can High and Low Performers Similarly Distinguish Between Information They Do and Do Not Know?

Analyses conducted on the proportion correct change scores provide insight into how the training affected proportion correct across performance quartiles, but not how accurate participants were overall at discriminating between information they did and did not know. Thus, analyses were conducted on proportion correct in the post-test to assess whether high and low performers were similarly accurate discriminating between information they did and did not know.

To test the prediction that high performers are more accurate than low performers discriminating between information they do and do not know, three separate 2 (training, control) x 4 (performance quartile: 1, 2, 3, 4) ANOVAs were conducted on post-test proportion correct at each level of confidence. A single 2 (training, control) x 3(confidence: low, medium, high) x 4 (performance quartile: 1, 2, 3, 4) ANOVA was not conducted because there would have been too few cases (122 participants) with non-missing data (i.e., a proportion correct calculated using non-zero data for each confidence level). For this series of tests, a Bonferroni correction of α = .017 (.05 / # of tests) was used.

For high confidence responses (80-100% confidence), there was a significant effect of performance quartile on proportion correct, $F(3, 328) = 6.71$, $p < .001$, $\eta_p^2 = .058$. Proportion correct was higher for high performers ($M = .87$, 95% CI [.85, .94]) than for low performers ($M = .75$, 95% CI [.70, .80]), but not high-medium performers ($M = .86$, 95% CI [.80, .92]) or low-medium performers ($M = .82$, 95% CI [.78, .86]). No other comparisons were significant.

Although the effectiveness of the training was assessed in analyses of change scores reported above, there was not a significant effect of training on proportion correct at high confidence, $F(1, 328) = 2.97$, $p = .086$, $\eta_p^2 = .009$. Additionally, the interaction between training

and performance quartile was also not significant $F(3, 328) = 2.54$, $p = .054$, $\eta_p^2 = .023$. See Figure 9 for proportion correct as a function of performance quartile and training condition.

For medium confidence responses (40-60% confidence), there was a significant effect of performance quartile on proportion correct, $F(3, 314) = 4.06$, $p = .008$, $\eta_p^2 = .037$. Proportion correct was higher for high performers ($M = .64$, 95% CI [.57, .70]) than for low performers ($M = .47$, 95% CI [.41, .54]), but not high-medium performers ($M = .56$, 95% CI [.47, .65]) or low-medium performers ($M = .54$, 95% CI [.48, .60]). No other comparisons were significant. Again, there was not a significant effect of training on proportion correct at medium confidence, $F(1, 314) = .21$, $p = .645$, $\eta_p^2 = .001$, and the interaction between training and performance quartile was also not significant $F(3, 314) = 1.33$, $p = .265$, $\eta_p^2 = .013$.

For low confidence responses (0-20% confidence), there was no significant effect of performance quartile on proportion correct, $F(3, 131) = 0.36$, $p = .779$, $\eta_p^2 = .008$. Proportion correct was similar for high performers ($M = .44$, 95% CI [.30, .58]), high-medium performers ($M = .53$, 95% CI [.37, .69]), low-medium performers ($M = .45$, 95% CI [.34, .55]), and low performers ($M = .43$, 95% CI [.31, .54]). Again, there was not a significant effect of training on proportion correct at low confidence, $F(1, 131) = 1.32$, $p = .252$, $\eta_p^2 = .010$, and the interaction between training and performance quartile was also not significant $F(3, 131) = 0.57$, $p = .635$, $\eta_p^2 = .013$.

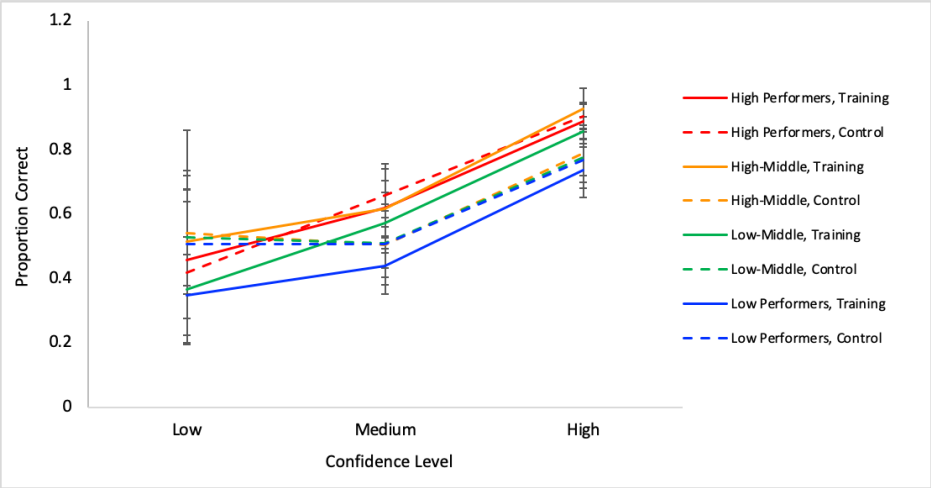


Figure 9. Confidence-accuracy characteristic (CAC) curves (proportion correct on the post-test at low, medium, and high confidence) for the four performance quartiles as a function of training condition. Error bars represent 95% confidence intervals.

Discussion

Unlike results from Experiment 1 and previous findings on the unskilled-unaware effect, there was no difference across performance quartiles in the degree of over- or under-estimation of the number of questions answered correctly on the post-test. Rather, across all four performance quartiles, participants underestimated their performance, and to a similar degree. As with Experiment 1, the perceived difficulty of the task could explain why participants generally underestimated their performance, as previous researchers have found that task difficulty affects peoples' performance estimates (e.g., Burson et al., 2006). Although it is unclear why the typical unskilled-unaware effect was not replicated for global performance judgments, one possible explanation is that the range of test questions varied somewhat widely in difficulty, from 31-69% accuracy. If participants relied upon the context of a task (as Experiment 1 suggests) to make their global performance judgments, then variables such as the presentation order of questions

may have affected estimates of number correct. Although the presentation order of questions was randomized, it is possible that the general distribution of question difficulty was inconsistent across participants in such a way that influenced later estimates of number correct.

Of critical interest in this experiment was whether training increased metacognitive monitoring accuracy, especially for low performers. Contrary to predictions, training did not help participants distinguish between information they did and did not know (i.e., "correct" responses that were accurate vs. inaccurate), and this was true for all four performance quartiles. Analyses conducted on the change in proportion correct from pre-test to post-test indicated that training did not increase metacognitive monitoring accuracy; proportion correct change scores did not differ between the training and control conditions.

One possible explanation for the nonsignificant effect of training on proportion correct is that the training itself was inadequate. Although the training implemented in Experiment 2 was modelled on a similar training conducted by Zechmeister and colleagues (1986), there are a number of differences between their study and the present experiment. First, the training implemented in this experiment was significantly shorter than their 30-minute training; this training was self-paced, taking approximately 5-10 minutes to complete. Second, this training was conducted online, whereas Zechmeister and colleagues conducted in-person group training sessions. It is possible that participants in their study paid more attention to try to compete with their peers. Considering these discrepancies, it is possible that compared to their training, the training implemented in this experiment was weaker, and subsequently insufficient to improve monitoring accuracy.

General Discussion and Conclusion

The goal of this dissertation was to test two potential methods for reducing overconfidence in low performers. Specifically, can low performers learn to more accurately distinguish between what they do and do not know? According to prominent models of metamemory (e.g., Nelson & Narens, 1990), teaching low performers to monitor their performance more accurately should subsequently help them learn more efficiently, and this domain-general ability is likely to transfer to other tasks (Gutierrez & Schraw, 2014). Consequently, if people become better at distinguishing between what they do and do not know, they could be better able to select information to restudy (e.g., monitoring that influences control; Nelson & Narens, 1990). Thus, although poor performers are disadvantaged in what they know, training their monitoring accuracy may improve their ability to learn in the future.

Experiment 1 tested whether answering easy rather than hard questions prior to taking a medium-difficulty test reduced trial-by-trial overconfidence in low performers. Contrary to predictions, global performance judgments made at the end of the 20-item test *but not* trial-by-trial confidence judgments were affected by initial question difficulty. For global performance judgments made at the end of the 20-item test, participants underestimated the number of questions they answered correctly to a greater degree when the test was preceded by hard questions compared to when the test was preceded by easy questions – an anchoring effect. Moreover, there was no significant interaction between initial question difficulty and performance quartile; rather, participants underestimated their performance more in the hard than the easy condition, and this difference was similar across performance quartiles.

Similar to results of previous studies (e.g., Michael & Garry, 2016; Weinstein & Roediger, 2012), participants in Experiment 1 anchored their global performance estimates based

on the difficulty of a preceding task, but task difficulty had no effect on trial-by-trial confidence judgments. Taken together with previous findings, these results indicate that when making global judgments at the end of a task, people likely rely on various sources of information, including perceptions of task difficulty to inform their metacognitive judgments. By contrast, when making trial-by-trial judgments, people more likely rely on information specific to the question itself (e.g., information from memory *or* gut feelings) and not information about the task to inform their confidence judgments.

Experiment 2 tested whether feedback and training could help people (especially low performers) better distinguish between information they did and did not know. Replicating results of Abed et al. (2020), high performers were more accurate than low performers distinguishing between information they did and did not know, even when they were sure they were correct. For high confidence judgments (80-100% confidence), the proportion of post-test questions answered correctly was higher for high than low performers. However, contrary to predictions, there was no effect of training on proportion correct, and this was true for all levels of confidence.

Moreover, for global performance judgments made at the end of the post-test, participants in both the training and control conditions underestimated the number of questions they answered correctly similarly across all performance quartiles. In other words, the typical unskilled-unaware effect was not replicated for global performance judgments. Similar to Experiment 1, the general finding that participants were *under*confident is similar to the *skilled*-unaware pattern that has been reported for difficult tests (Burson et al., 2006). Further, in light of Experiment 1 results indicating that global performance estimates are affected by the context of

the task, it is possible that the random distribution of questions in Experiment 2 was inconsistent across participants in such a way that influenced later estimates of number correct.

Although the main hypotheses for both Experiment 1 and 2 were not supported, results from this dissertation nevertheless add to the growing body of literature revealing two key findings. First, in Experiment 1, the difficulty of a preceding task (a context manipulation) affected global, but not trial-by-trial confidence judgments. Taken together with previous findings (e.g., Michael & Garry, 2016; Weinstein & Roediger, 2012), these results indicate that trial-by-trial judgments reflect a relatively purer assessment of peoples' confidence in their own performance than global judgments. This general finding highlights the importance of trial-by-trial confidence judgments in research on metacognition, and researchers interested in this topic should collect trial-by-trial judgments whenever possible.

Second, in both experiments, high performers exhibited a stronger confidence-accuracy relationship with high proportion correct at high confidence and low proportion correct at low confidence, whereas low performers exhibited a weaker confidence-accuracy relationship with relatively lower proportion correct, even at high confidence. Taken together with previous findings (e.g., Abed et al., 2020), these results indicate that high performers are more accurate than low performers distinguishing between information they do and do not know, and this is true even for judgments made with high confidence. This finding is particularly important given the applied relevance of metacognitive monitoring accuracy. This indicates, for example, that low performers may have more difficulty selecting information to restudy, and subsequently may be disadvantaged both in what they already know and what they are likely to learn.

The goal of this dissertation was to test the question, can low performers learn to more accurately distinguish between what they do and do not know? Results from two experiments fail

to provide any evidence suggesting that either context manipulations *or* training can increase metacognitive monitoring accuracy, especially in low performers. Although results from previous training studies (e.g., Lichtenstein & Fischhoff, 1980; Zechmeister et al., 1986) indicate that metacognitive monitoring *can* be trained, results from this dissertation suggest that such training is unlikely to provide metacognitive performance benefits without the relatively more significant time and effort used in previous trainings – for example, the 30-minute in-person sessions employed by Zechmeister et al. In fact, perhaps future researchers should first seek to replicate the successful training employed by Zechmeister and colleagues before attempting to shorten the training duration, as was done in this study.

Although results of this dissertation fail to provide evidence that people can be trained to discriminate better between information they do and do not know, metacognitive training is nevertheless a noble area for future study. Indeed, the potential societal benefits of such an intervention are potentially far too great *not* to pursue. As an example, researchers have reported that people are more likely to share information they think is true compared to information they think is false, regardless of the *actual* veracity of the information (Fenn et al., 2019). Thus, improving peoples' ability to discriminate between true and false information may reduce the spread of false information online.

Additionally, researchers have reported that people are less likely to act when they are not very confident – for example, less likely to submit an answer for scoring when they are not very confident in a response, and less likely to be willing to testify as an eyewitness when they are not very confident in their memory (see Pansky & Goldsmith, 2014; Michael & Garry, 2016, respectively). Perhaps if people are better able to identify when they know very little about a topic (in a broad sense), they may then be less likely to share such information online. Thus,

researchers interested in applied problems such as the spread of misinformation online may consider identifying methods to reduce overconfidence more broadly before attempting a more elaborate training method to increase confidence calibration accuracy.

References

Abed, E., Nguyen, T. B., & Pezdek, K. (2020). *The Unskilled-Unaware Effect: Are the unskilled always unaware?* Manuscript submitted for publication.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219-235. doi: 10.1177/1088868309341564

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637. doi: 10.1037//0033-2909.128.4.612

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*(1), 60-77. doi: 10.1037/0022-3514.90.1.60

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215-235. doi: 10.1007/s11409-015-9142-6

DeSoto, K. A., & Simons, D. J. (2015, July 15). General knowledge question norming. http://doi.org/10.17605/OSF.IO/Y4B9R

Dunlosky, J., Hertzog, C., Kennedy, M. R., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology, 9*(1), 4-11.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228-232. doi: 10.1111/j.1467-8721.2007.00509.x

Fenn, E., Kantner, J., Ramsay, N., Pezdek, K., & Abed, E. (2019). *Nonprobative photos increase truth, like, and share judgments in a simulated social media environment.* Manuscript submitted for publication.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, 101449. doi: 10.1016/j.intell.2020.101449

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*. doi: 10.1016/j.jarmac.2019.02.002

Gutierrez, A. P., & Schraw, G. (2015). Effects of strategy training and incentives on students' performance, confidence, and calibration. *The Journal of Experimental Education, 83*(3), 386-404. doi: 10.1080/00220973.2014.907230

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265-285. doi: 10.1007/s11409-018-9185-6

Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring and respective awareness in low-performing students. *Memory & Cognition, 44*(2), 229-241. doi: 10.3758/s13421-015-0552-0

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1-24. doi: 10.1006/jmla.1993.1001

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

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 483-502). New York, NY: Guilford.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118. doi: 10.1037/0278-7393.6.2.107

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*(2), 180-188. doi: 10.1037//0022-3514.82.2.180

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

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*(2), 149-171. doi: 10.1016/0030-5073(80)90052-5

Mehdizadeh, L., Sturrock, A., Myers, G., Khatib, Y., & Dacre, J. (2014). How well do doctors think they perform on the General Medical Council's Tests of Competence pilot examinations? A cross-sectional study. *BMJ Open, 4*, 1-8. doi: 10.1136/bmjopen-2013-004131

Michael, R. B., & Garry, M. (2016). Ordered questions bias eyewitnesses and jurors. *Psychonomic Bulletin & Review, 23*(2), 601-608. doi: 0.3758/s13423-015-0933-1

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93-102. doi: 10.1016/j.jarmac.2015.01.003

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.) *Psychology of learning and motivation.* New York: Academic Press.

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. doi: 10.1080/00220670209596583

Nguyen, T.B., Abed, E., and Pezdek, K. (2018). Postdictive confidence (but not predictive confidence) predicts eyewitness memory accuracy. *Cognitive Research: Principles & Implications.* doi: 10.1186/s41235-018-0125-4

Nguyen, T.B., Pezdek, K., & Wixted, J.T. (2016). Evidence for a confidence-accuracy relationship in memory for same- and cross-race faces. *Quarterly Journal of Experimental Psychology*, 1-17. doi: 10.1080/17470218.2016.1246578

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review, 21*(5), 1255-1262. doi: 10.3758/s13423-014-0597-2

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4-5), 559-579. doi: 10.1080/09541440701326022

Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology, 15*(1), 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

Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods, 45*(4), 1115-1143. doi: 10.3758/s13428-012-0307-9

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49. doi: 10.1186/s41235-017-0086-z

Tekin, E., Lin, W., & Roediger, H. L. (2018). The relationship between confidence and accuracy with verbal and verbal+ numeric confidence scales. *Cognitive research: principles and implications, 3*(1), 41. doi: 10.1186/s41235-018-0134-3

Weinstein, Y., & Roediger, H. L. (2010). Retrospective bias in test performance: Providing easy items at the beginning of a test makes students believe they did better on it. *Memory & Cognition, 38*(3), 366-376. doi: 10.3758/MC.38.3.366

Weinstein, Y., & Roediger, H. L. (2012). The effect of question order on evaluations of test performance: How does the bias evolve? *Memory & Cognition, 40*(5), 727-735. doi: 10.3758/s13421-012-0187-3

Whittlesea, B. W., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 547-565. doi: 10.1037//0278-7393.26.3.547

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515-526. doi: 10.1037/a0039510

Zechmeister, E. B., Rusch, K. M., & Markell, K. A. (1986). Training college students to assess accurately what they know and don't know. *Human Learning: Journal of Practical Research & Applications, 5*(1), 3-19.

Appendix A

Experiment 2 Training (Block 2) Instructions and Procedure[19]

For the questions you just answered, you were asked to judge whether your answer was correct or incorrect, and to rate your confidence in your correct/incorrect judgment on a scale from 0% (*not at all confident*) to 100% (*completely confident*).

When people think they answered a question correctly, they are much more confident than they should be. For example, when people say they are 100% confident, they are actually correct much less than 100% of the time.

For this next set of questions, your task is to try to make your confidence judgments match the likelihood that you answered each question correctly. Our primary interest is not the number of questions you answer correctly, but rather **how well your confidence judgments predict the likelihood that you answered each question correctly**. On the next page, you will be given further instructions to help you improve the appropriateness of your confidence judgments. [*page break*]

The purpose of this study is to teach you how to accurately predict whether or not you answered a question correctly. Please pay attention to the following

[19] The contents of these instructions are based on the 30-minute training session described by Zechmeister, Rusch, and Markell (1986).

instructions, because they will help you improve the appropriateness of your confidence judgments. [*page break*]

A great deal of psychological research has demonstrated that people are not very good predictors of what they know.

One reason that people overestimate their likelihood of being correct is that they search for evidence that their answer is *correct*, but fail to search for evidence that their answer is *wrong*. This is called *confirmation bias*: people are primed to see and agree with ideas that fit their preconceptions, and to ignore and dismiss information that conflicts with them.

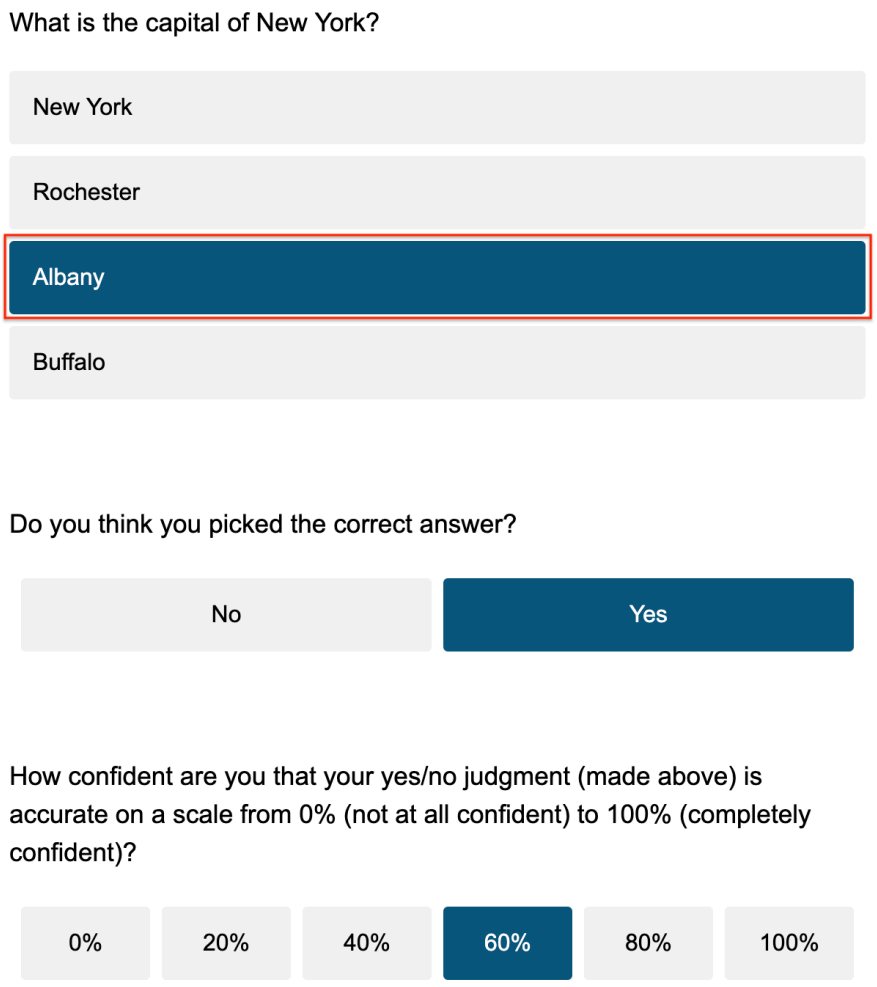
In other words, once people decide that an answer is correct, they stop searching their memory for evidence that the answer is wrong. In the next set of questions, to help you better predict whether you answer each question correctly, try to search for evidence that your answer is *wrong* in addition to evidence that your answer is *correct*. [*page break*]

- When assessing the likelihood that your answer is correct, consider evidence for and against the answer you selected. Can you think of any reasons why that answer might be wrong?

- Next, consider how strong your evidence is to support the answer you selected. Did you remember a specific time that you had learned about the correct answer?
    - If so, this type of evidence is strong, so you should have said that you were really confident in your response.
    - On the other hand, if you simply had a gut feeling that you were selecting the correct answer, then you should NOT have said that you were really confident in your response.

Try really hard to use the confidence scale to indicate the strength of the evidence you have in your memory. To help you understand how to weigh the strength of evidence for an answer, please consider the following example. [*page break*]

Imagine that you have been asked, “what is the capital of New York?” and you have given the following responses. The correct answer is indicated by the red box.



Which of the following reasons do you think is a good justification for a confidence judgment of 60%?

[*participants were presented with the following response options; the correct answer is highlighted in green but was not highlighted for the participant.*]

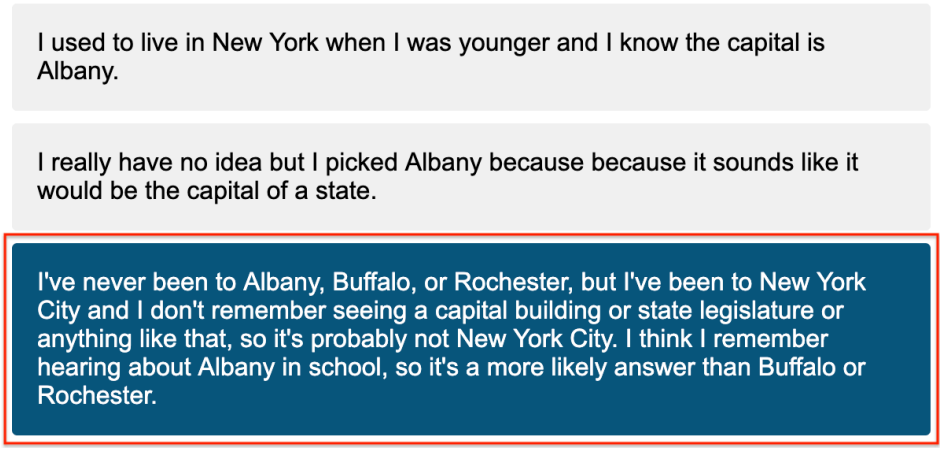
I used to live in New York when I was younger and I know the capital is Albany.

I really have no idea but I picked Albany because it sounds like it would be the capital of a state.

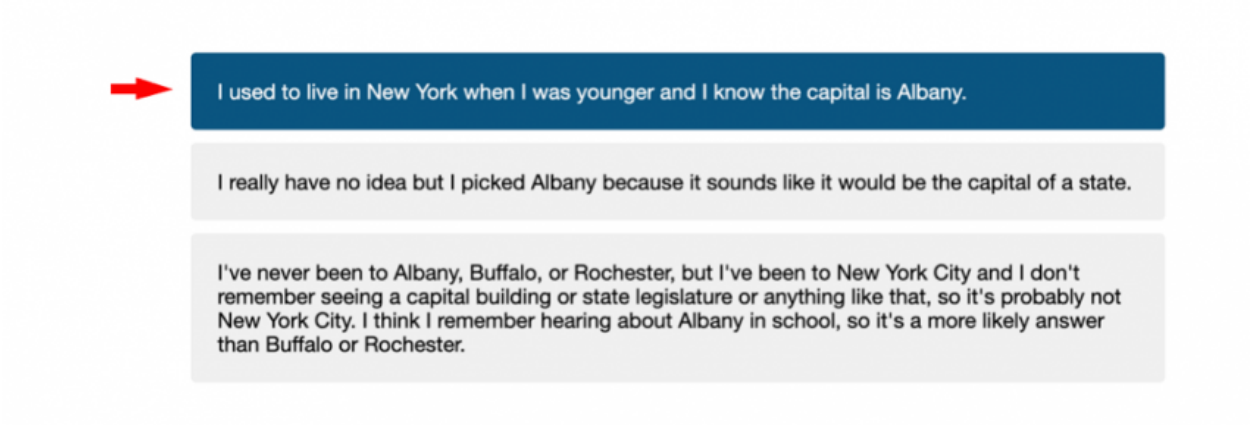
I've never been to Albany, Buffalo, or Rochester, but I've been to New York City and I don't remember seeing a capital building or state legislature or anything like that, so it's probably not New York City. I think I remember hearing about Albany in school, so it's a more likely answer than Buffalo or Rochester.

[*page break*]

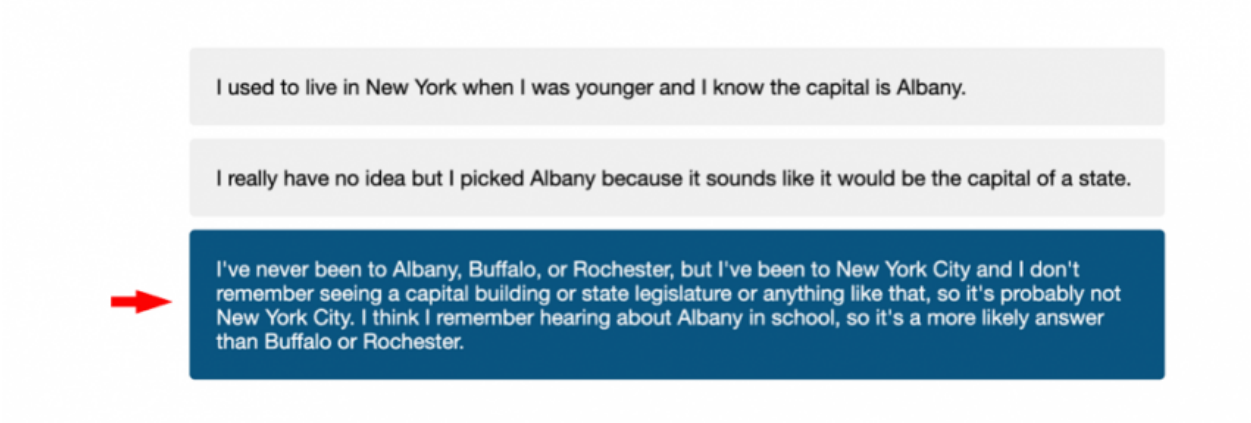
[*If a participant selected the highlighted answer they would see: "*Correct! The answer is:*" and if they select a non-highlighted answer they would see: "*Wrong. The correct answer is:*" along with the following screenshot revealing the correct answer.*]



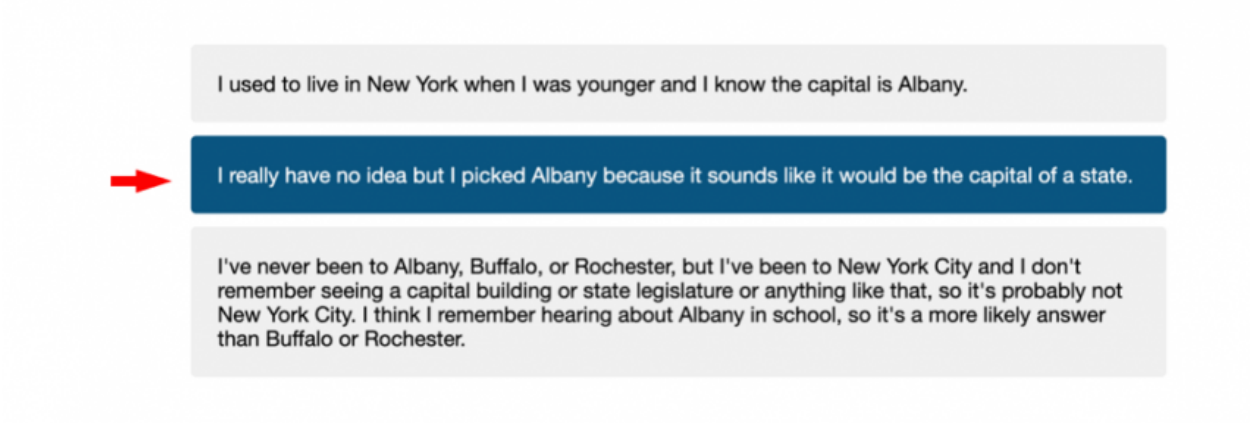
[*page break*]



If you remember living near the state capital, that would be really strong evidence so you should have said you were very confident in your answer (for example, a confidence rating of 100%).



If you just had a gut feeling that Albany is the capital of New York, that would be really weak evidence so you should have said you were not very confident in your answer (for example, a confidence rating of 0%).



If you were able to rule out some options and remember evidence that points to one of the remaining options as the correct answer (for example, learning about Albany in school), then this evidence would be neither very weak nor very strong, so you should have rated your confidence somewhere in the middle of the scale.

[*page break*]

Remember, the purpose of this study is to teach you how to accurately predict whether or not you answered a question correctly.

Now you will answer multiple choice questions just like the ones you answered earlier, but this time you will be told whether the response you selected was correct or incorrect.

The purpose of this feedback is to help you practice the strategies you just learned about:

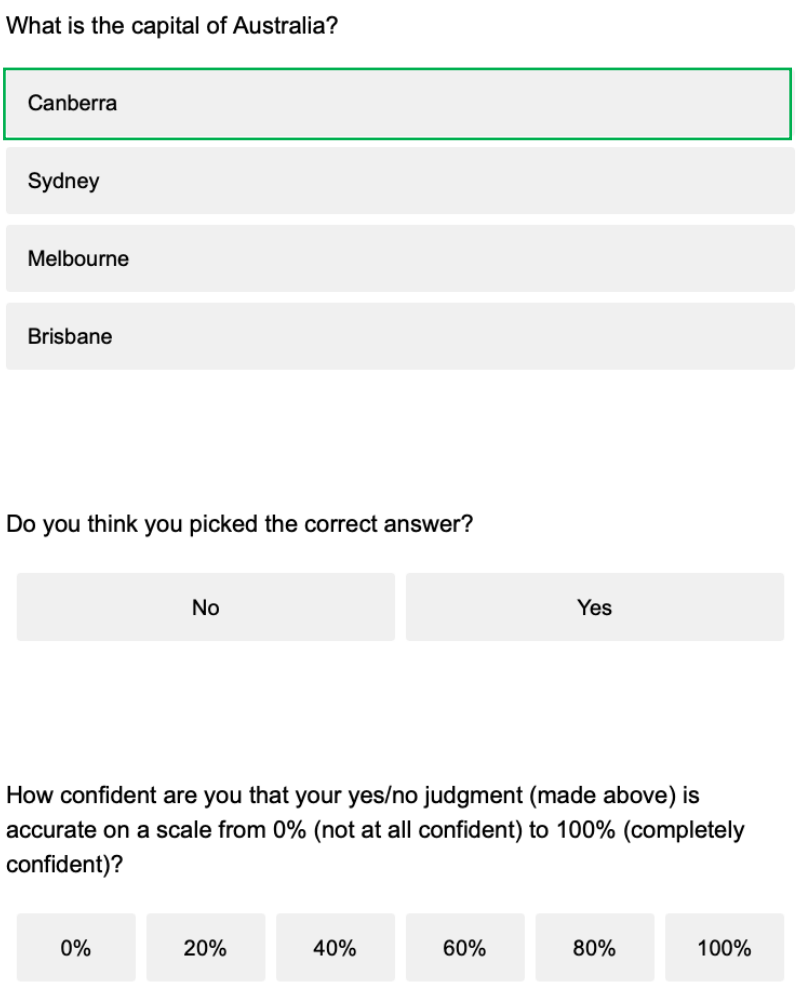
1. Considering evidence that your answer is **wrong**
2. Using the confidence scale to indicate the **strength** of the evidence you have in your memory.

When you get a question *wrong*, think about clues you might have had that your answer was wrong. For example, maybe you only had a gut feeling, but you didn't have strong evidence that your answer was correct. **Use the feedback you receive to evaluate the appropriateness of your confidence judgment and improve your use of the confidence scale on future questions.** After viewing an example question, you will have the chance to improve your confidence judgment accuracy.

Please proceed to the next page for an example question.

[*page break*]

[*Participants responded to the following question. The correct answer is highlighted in green but was not highlighted for the participant.*]



[*If the participant selected the correct answer, the following screen would be displayed*]

Correct! The correct answer is Canberra.

[*If the participant selected an incorrect answer, the following screen would be displayed*]

Wrong. The correct answer is Canberra.

[*After participants completed the practice question they viewed the following, final instructions before completing the 20-question training block.*]

You will now answer 20 questions just like the ones you answered earlier. **Remember, the purpose of this study is to teach you how to accurately predict whether or not you answered a question correctly. For the following questions, practice using the feedback to help you judge the appropriateness of your confidence judgment.**

If you understand your next task, you may continue with the experiment by selecting "I'm ready to continue."

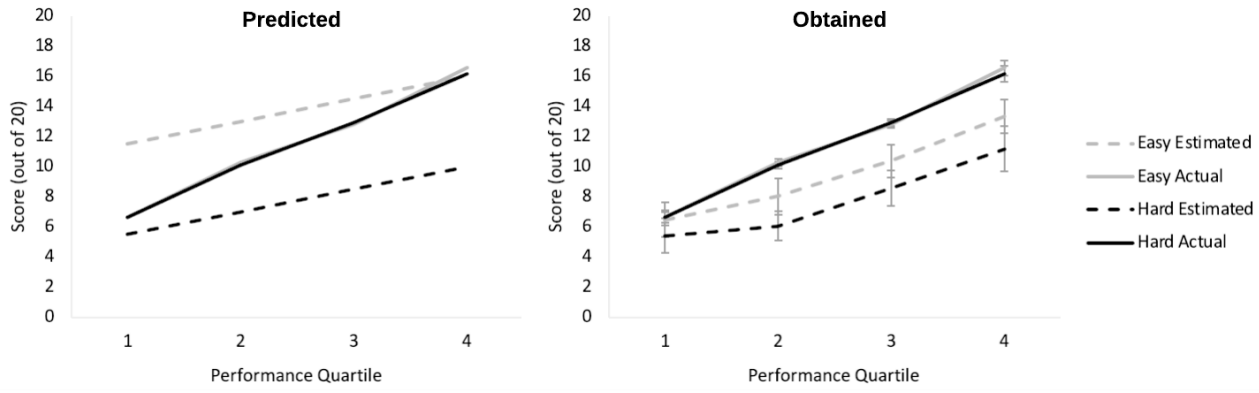
If you do NOT understand your next task, you may go back to the beginning of the instructions by selecting "Take me back."

Take me back | I'm ready to continue

[Participants could choose to review the training instructions as many times as they needed before continuing with the experiment.]

Appendix B

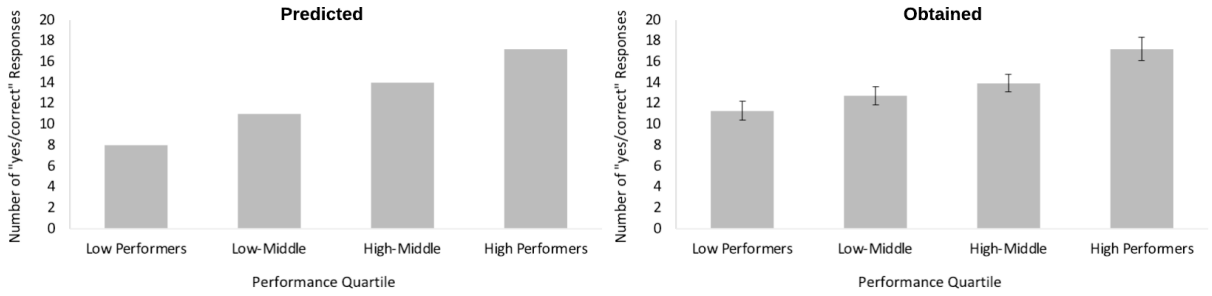
Comparison of Predicted and Obtained Results: Figure 2



Predicted (left panel) and obtained (right panel) Experiment 1 results for global performance judgments. Figures represent the mean actual and estimated number correct on the 20 medium-difficulty target questions for each of the four performance quartiles as a function of initial question difficulty. Difference scores were computed by subtracting the actual number correct from the estimated number correct. Error bars represent 95% confidence intervals.

Appendix C

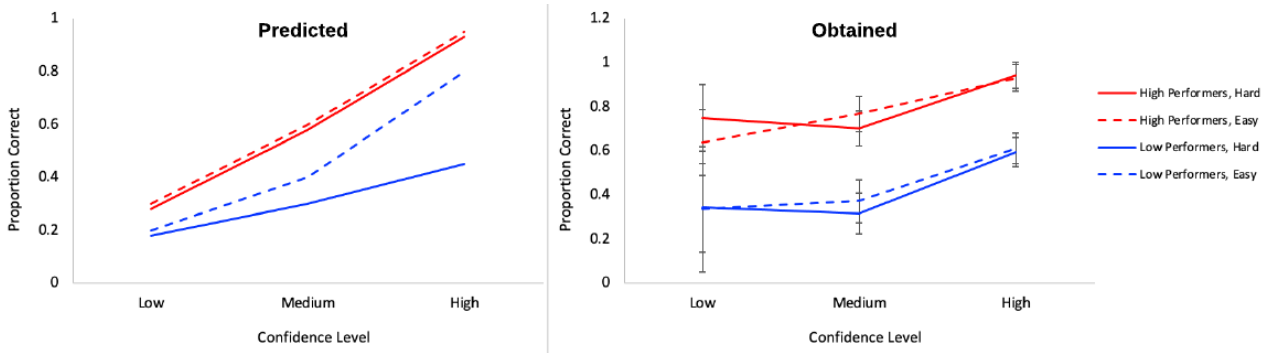
Comparison of Predicted and Obtained Results: Figure 3



Predicted (left panel) and obtained (right panel) Experiment 1 results for trial-by-trial "yes/correct" performance judgments. Figures represent the mean number of "yes/correct" performance judgments for each of the four performance quartiles. Error bars represent 95% confidence intervals.

Appendix D

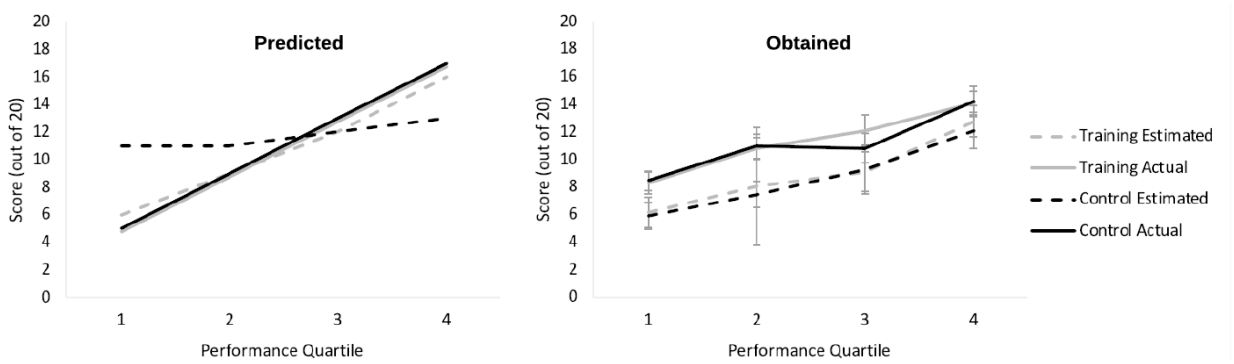
Comparison of Predicted and Obtained Results: Figure 4



Predicted (left panel) and obtained (right panel) Experiment 1 results for trial-by-trial confidence judgments. Figures represent the Confidence-accuracy characteristic (CAC) curves (proportion correct on 20 medium-difficulty target questions at low, medium, and high confidence) for high and low performers as a function of initial question difficulty (easy vs. hard). Error bars represent 95% confidence intervals.

Appendix E

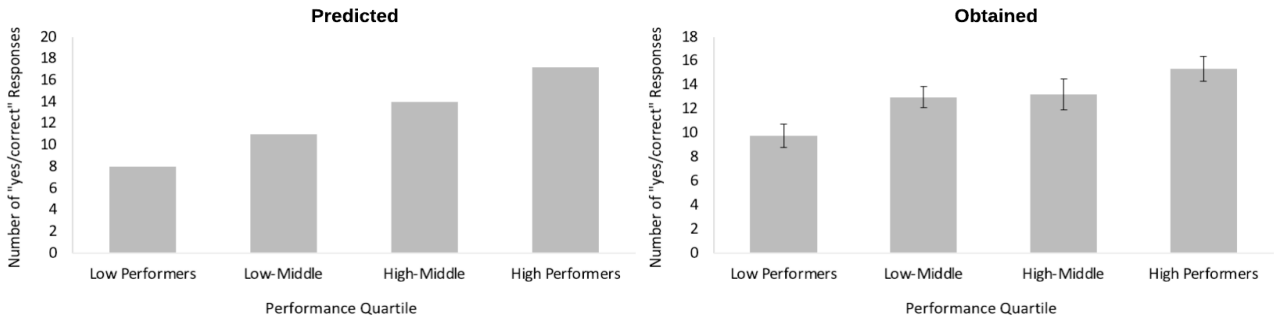
Comparison of Predicted and Obtained Results: Figure 6



Predicted (left panel) and obtained (right panel) Experiment 2 results for global performance judgments. Figures represent the mean actual and estimated number correct on the post-test for each of the four performance quartiles as a function of training condition. Difference scores were computed by subtracting the actual number correct from the estimated number correct. Error bars represent 95% confidence intervals.

Appendix F

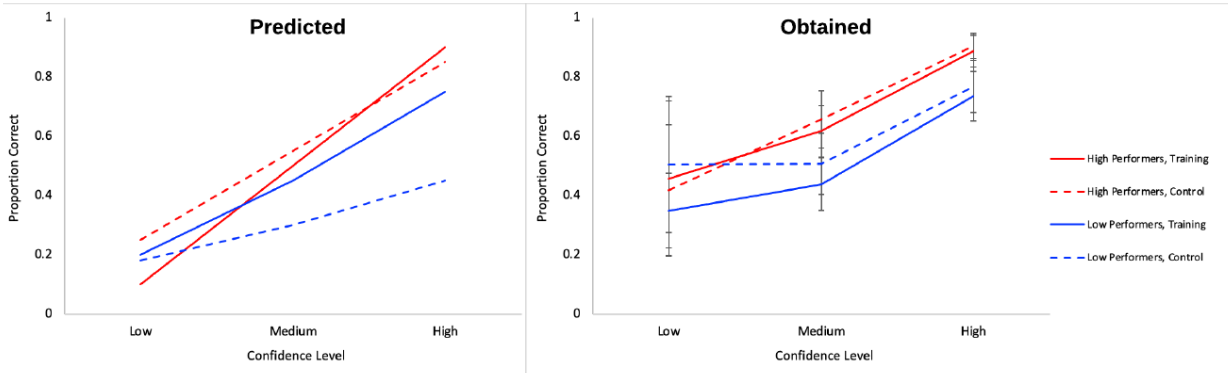
Comparison of Predicted and Obtained Results: Figure 7



Predicted (left panel) and obtained (right panel) Experiment 2 results for trial-by-trial "yes/correct" performance judgments. Figures represent the mean number of "yes/correct" performance judgments for each of the four performance quartiles. Error bars represent 95% confidence intervals.

Appendix G

Comparison of Predicted and Obtained Results: Figure 9



Predicted (left panel) and obtained (right panel) Experiment 2 results for trial-by-trial confidence judgments. Figures represent the Confidence-accuracy characteristic (CAC) curves (proportion correct on the post-test at low, medium, and high confidence) for high and low performers as a function of training condition. Error bars represent 95% confidence intervals.