Adv in Health Sci Educ (2018) 23:749–765
https://doi.org/10.1007/s10459-018-9826-1

Subjective awareness of ultrasound expertise development: individual experience as a determinant of overconfidence

Jordan Richard Schoenherr[1] · Jason Waechter[2] · Scott J. Millington[3]

Received: 23 October 2017 / Accepted: 6 April 2018 / Published online: 24 April 2018

Abstract Medical decision-making requires years of experience in order to develop an adequate level of competence to successfully engage in safe practice. While diagnostic and technical skills are essential, an awareness of the extent and limits of our own knowledge and skills is critical. The present study examines clinicians' subjective awareness in a diagnostic cardiac ultrasound task. Clinicians answered diagnostic and treatment related questions for a range of pathologies. Following these questions, clinicians indicated their level of confidence in their response. A comparison of response accuracy and confidence revealed that clinicians were generally overconfident in their responses. Critically, we observed that a clinician's overconfidence was negatively correlated with prior experience: clinicians that had more prior experience expressed less overconfidence in their performance such that some clinicians were in fact underconfident. We discuss the implications for training in medical education and decision-making.

Keywords Medical decision-making · Ultrasound · Overconfidence bias · Subjective awareness · Expertise development

Introduction

Medical expertise requires extensive experience and prior knowledge in order to diagnose pathophysiologies, determine a course of treatment, and communicate with other health care professionals (e.g., Ericsson 2007; Lingard 2012; Norman 2006; Norman et al. 2006). An essential part of the process of expertise development is monitoring the quality of information, the decision-making process, and the critical features of the current situation in

✉ Jordan Richard Schoenherr
jordan.schoenherr@carleton.ca

[1] Department of Psychology, Carleton University, Ottawa, Canada

[2] Departments of Critical Care and Anesthesiology, University of Calgary, Calgary, Canada

[3] Division of Critical Care, University of Ottawa, Ottawa, Canada

order to sustain an adequate degree of performance in their activities in the face of uncertainty (Katz 1988; Lingard et al. 2003; Fox 2000; Spafford et al. 2006; Timmermans and Angell 2001). In the outpatient clinic, although there are time constraints, clinicians have a comparatively extended period of time to collect and use information that they obtain from the patient and diagnostic tests. In emergency settings or acute care units, clinicians work in time- and information-limited conditions. How information is gathered, analyzed, and used thus requires an understanding of how these cognitive functions occur in a wide variety of conditions. Thus, not only is medical knowledge and clinical experience critical, a clinician's ability to monitor their own performance and assess the vulnerabilities of the decision-making process in a given situation is also a vital aspect of expertise in the health care professions (e.g., Berner and Graber 2007; Croskerry and Norman 2008; Moulton et al. 2007).

Subjective performance monitoring requires that a decision-maker consider both the kind of information that they have available to them, the nature of the decision-making processes, as well as likelihood of the accuracy of these processes. It has often been studied using post-decisional confidence reports, requiring the assignment of a given response (e.g., diagnosis or treatment) to a given semantic category (e.g., "Certain", "Sure", "100%"; Baranski and Petrusic 1994). Confidence reports have been used extensively in a variety of domains from perceptual tasks, general knowledge question, and memory studies (e.g., Baranski and Petrusic 2001; Kvidera and Koustaal 2008; Keren 1991; Lichtenstein et al. 1982). Subjective assessments of performance have also been used in studies of medical education and decision-making with performance monitoring cited as a critical skill in terms of self-directed learning and subjective confidence (Croskerry and Norman 2008). However, these measures have often been used as alternative or supplementary measures of performance (e.g., Halm et al. 2010; Hogg and Miller 2016; Westbrook et al. 2005) which is problematic given the limited correspondence between subjective measures relationship with objective measures of performance (e.g., Lichtenstein and Fischhoff 1977; Keren 1991). Given claims that overconfidence is a common feature of medical decision-making (Croskerry 2009; Croskerry and Norman 2008) and a major source of diagnostic error (Berner and Graber 2007; Friedman et al. 2005) more studies are needed to assess and clarify the relationship between competency development, task performance, and subjective confidence in order to determine factors that affect overconfidence bias (Barnsley 2004; Morgan and Cleave-Hogg 2002). In the present study, we present a method to examine subjective awareness of performance in a task that assesses cardiac ultrasound competency. By measuring a clinician's prior experience in cardiac and noncardiac ultrasound, we demonstrate how their domain-specific and domain-general knowledge are related to overconfidence bias.

Medical expertise and judgment and decision-making processes

A number of characteristics define judgment and decision-making in the health care professions (e.g., Dowie and Elstein 1988; Gambrill 1990; Norman et al. 2006). Practitioners and diagnosticians are presented with perceptual information (e.g., dermatological conditions, sonography, frequency charts, feedback from physical examinations) and abstract declarative knowledge (e.g., the meaning of a patient's vital signs or symptoms presented verbally by a colleague) that vary in the extent to which information can be verbalized. Clinicians

must integrate this information into a diagnosis and identify an appropriate course of treatment. Understanding how these kinds of knowledge are acquired, stored, and used is critical to understanding medical expertise development as well as the metacognitive abilities used by practitioners to judge the limits of their competency and to seek out additional training and advice from other members of a health care team (Berner and Graber 2007; Croskerry and Norman 2008; Friedman et al. 2005).

Researchers have claimed that there are two qualitatively different kinds of cognitive processes typically used to acquire knowledge and perform tasks (e.g., Ashby et al. 1998; Logan 1988; Norman and Shallice 1980; Shiffrin and Schneider 1977; for reviews, see Evans 2008; Kahneman 2011; Stanovich 2004). Similar proposals have been made within medical decision-making (e.g., Croskerry 2009; Moulton et al. 2007; Norman and Brooks 1997; Reyna 2008). Early in training, learners will engage in extensive deliberation as they attempt to learn abstract categories from cases presented in textbooks, the laboratory, the clinic, and during their preceptorship. Clinicians must also actively monitor their performance using feedback to improve their diagnostic categories, treatment procedures, and interpersonal skills. In general, these explicit, effortful processes are referred to as Type 2 Processes. After extensive training, clinicians will retain representations of typical and atypical features of human anatomy and physiology in memory. With greater experience, these representations will become increasingly accessible to practitioners such that after considerable experience, they will become available automatically thereby dominating response selection (e.g., Ashby et al. 1998). These relatively effortless, rapid processes are referred to as Type 1 Processes. Similarly, rather than thinking faster, the automatic processes that define expertise in fact imply less competition between alternative hypotheses or behavioural responses resulting in faster behaviour (Saling and Philips 2007). Type 1 processes are thus a key aspect of clinical expertise development and performance (e.g., Ericsson 2007). However, such automatic processing can lead experienced clinicians and diagnosticians to overlook information (e.g., Drew et al. 2013; Krupinski et al. 2006; Norman and Brooks 1997). Thus, neither Type 1 or Type 2 processes alone provide an adequate basis for medical expertise.

Any clinical or nonclinical tasks require the contributions of both Type 1 and Type 2 processes. For instance, while certain symptoms might be immediately recognizable as a disease (Type 1 processes), unique combinations of features might require weighing multiple treatment options (Type 2 processes). Thus, both Type 1 and Type 2 processes can contribute to any behaviour (Jacoby 1991). Similarly, even when specific information is available to health care professionals, they might nevertheless rely on gist information (i.e., the general, essential information; Reyna 2008). Research in applied decision-making supports such an account of expertise. For instance, Klein's (1993, 1998) recognition-primed model of decision-making assumes that decision-makers use information to identify a pattern in terms of its overall appearance or, failing that, using partially recognizable features of the situation to reason about its parts. The utility of this model in medical decision-making has also been noted such that experienced clinicians should be able to identify decision-points in the performance of a procedure that require more deliberation (e.g., Moulton et al. 2007). However, stress, attentional division, and lack of task-relevant experience will reduce the availability of Type 2 processes. For instance, in the context of critical care, practitioners will likely need to rely on rapid pattern detection and identification in order to diagnose a patient as well as rapid retrieval of treatment alternatives from long-term memory. Medical decision-making is likely well described by interaction of these two kinds of processes working in parallel.

Medical diagnoses and expertise development

Empirical studies examining diagnostic reasoning and medical decision-making appear to support dual-process accounts of medical decision-making (e.g., Norman and Brooks 1997; Norman 2006). The effect of information accessibility is evidenced in studies of anchoring on medical decision-making wherein an initial presentation of evidence influences subsequent processing. For instance, early studies in medical decision-making such as those conducted by Elstein et al. (1978) found that physicians retained hypotheses generated early in a task regardless of whether disconfirming information was presented later. Still other studies have found that information from previous events anchors our performance in the current event. In a study conducted by Brooks et al. (1991), physicians were presented with two slides in succession that were defined by the presence or absence of cellular pathologies that were indicative of cancer. If the first slide presented diagnostic evidence of cancer, physicians were more likely to respond that the subsequent slide also contained evidence of cancer regardless of whether in fact it was present (i.e., a false positive). When evidence of cancer was absent from the first slide, physicians were more likely to respond that the subsequent slide did not contain cancer (i.e., a miss). Rather than making two independent decisions, the accumulated evidence and response from the previous experience appears to anchor their performance closer to a given assessment (for related results, see Hatala et al. 1999; Kulatunga-Moruzi et al. 2004; Regehr et al. 1994). Thus, the accessibility of information in terms of its availability in long-term memory, anchoring based on recent information, and confirmation bias can influence clinical judgment and decision-making. In that these biases can lead to errors, methods for reducing their influence are essential.

In addition to general decision-making processes, expertise requires domain-specific knowledge. Researchers must therefore identify domain-specific knowledge and aspects of prior experiences that are associated with competency development in a specific area of practice. One area of increased interest is point-of-care ultrasound (POCUS). POCUS scans are complex and difficult to evaluate as they generate 2D representations of 3D anatomy, leaving the physician to create a complex 3D visuospatial mental model. The resultant mental models of the anatomy likely depend on the learner's ability to generate and manipulate accurate mental images (e.g., Corballis 1997; Shepard and Cooper 1982; Shepard and Metzler 1971), to generate these images from distinct parts, to inspect patterns, and to mentally rotate these images (for a review of these issues, see Mast et al. 2003). Spatial visualization and reasoning abilities of learners appear to be strong determinants of anatomical competency (Guillot et al. 2007; Hoyek et al. 2009). Consequently, visuospatial skills are associated with considerable individual differences (Hegarty et al. 2006). Indeed, assessment instruments in cardiac (Millington et al. 2016, 2017b) and lung (Millington et al. 2017a) POCUS have included items assessing specific features of visuospatial knowledge (e.g., viewing angles) that are separable from diagnostic knowledge (i.e., anatomy, pathophysiologies). However, while these instruments allow for the identification of learners' competencies, they leave unexamined the extent to which learners are aware of these competencies. Thus, greater consideration needs to be given to the relationship between domain-specific expertise (i.e., a competency) and a clinicians' subjective awareness.

Metacognitive assessments of performance and clinical decision-making

Metacognitive assessments of performance provide an independent source of information about factors that influence clinical practice: clinicians might have medical knowledge but be incapable of accessing it due to time limitations, stress, or fatigue. In medical decision-making, researchers have suggested that medical errors might be reduced by developing metacognitive strategies for monitoring and regulating performance (Croskerry 2009) and to address issue of overconfidence (Berner and Graber 2007; Croskerry and Norman 2008; Friedman et al. 2005). Metacognitive assessments of performance have been used in a variety of tasks and have either emphasized the processes involved in subjective assessment (Baranski and Petrusic 1998; Pleskac and Busemeyer 2010; Vickers 1979; Vickers and Packer 1982) or the cues that are used to determine confidence (Hart 1965; Koriat 1993).

While early accounts of confidence reports assumed that subjective confidence was based on a direct scaling of available evidence from the primary decision to a confidence report (Ferrel and McGooey 1980), subsequent accounts noted that under- and overconfidence biases suggested that other kinds of information influenced metacognitive assessments of performance (Kvidera and Koustaal 2008; Gigerenzer et al. 1991; Koriat et al. 2002). For instance, task difficulty is associated with systematic variation in subjective confidence such that hard questions produce overconfidence whereas easy questions producing underconfidence, referred to as the Hard–Easy Effect (Lichtenstein and Fischhoff 1977). Subsequent studies have also found that the relationship between accuracy and confidence changes with experience. In memory studies, greater underconfidence is associated with increased practice, referred to as the underconfidence-with-practice effect (UWP; e.g., Finn and Metcalfe 2007; Koriat et al. 2002; Koriat 1993). Consequently, medical decision-makers likely frequently use at least some cues that contain non-diagnostic information relative to the primary decision process (e.g., basing one's confidence on the source of the information rather than amount of information).

While many studies in medical education and decision-making do not use numeric labels for confidence categories (e.g., "guessing" compared to "50% confidence") thereby making it difficult to compare accuracy and subjective assessments of performance (e.g., Millington et al. 2009), there is some evidence that suggests the presence of overconfidence bias. For instance, some studies have found that health care professionals report greater confidence in incorrect responses, a pattern consistent with the Hard–Easy Effect. Leopold et al. (2005) observed that higher levels of confidence were negatively correlated with performance in a knee injection task embedded in an instructional course. Westbrook et al. (2005) observed a more complex relationship with greater confidence for correct responses relative to incorrect responses. However, relative to nurses, physicians were found to exhibited greater overconfidence in their incorrect responses. Moreover, subjective confidence has also been found to increase with experience but does not appear to depend on performance (e.g., Morgan and Cleave-Hogg 2002). Along with these results, some studies have failed to find a significant correlation between performance and confidence (Barnsley 2004; Elzubeir and Rizk 2001; Fox et al. 2000; Mavis 2001; Turner et al. 2009). Taken together, this suggests that confidence reports do not simply reflect a by-product of the decision-making processes of health care professionals. Rather confidence reports might be dependent on the evidence used to complete a task, personal experience, and other individual differences.

Present study

The present study examines the relationship between prior experience, domain-specific competency, and subjective confidence. Using cardiac ultrasound case studies, we examined the relationship between a clinician's experience with a given area of practice and their subjective awareness. Following from research on confidence processing (Baranski and Petrusic 2001; Kvidera and Koustaal 2008), we predicted that clinicians would express overconfidence when presented with hard questions and underconfidence when presented with easy questions (Lichtenstein and Fischhoff 1977). However, given the underconfidence-with-practice effect (e.g., Koriat et al. 2002) and anecdotal evidence from senior clinicians, we assumed that overconfidence would decrease with the amount of experience clinicians self-reported. Given the possibility that mental rotation ability might be related to both prior experience (e.g., a self-selection bias for seeking out this training) and ultrasound competency, we attempted to control for mental rotation (i.e., Mental Rotation Test, MRT; Peters et al. 1995) and handedness (Waterloo Handedness Inventory, WHQ; Bryden 1977), as well as domain-general ultrasound and domain-specific cardiac ultrasound experience.

Methods

Participants

Participants were enrolled in an Acute Critical Events Simulation (ACES) training course for clinical care cardiac ultrasound. Forty participants started the study, and 12 successfully completed all cases online. However, participant experience was sufficiently variable that correlations could be obtained with their performance. The ACES course participants received credit in terms of the Royal College of Physician and Surgeons of Canada (RCPSC) maintenance of competency credit for completing the ACES course.

Procedures

The online study was conducted concurrently with the ACES training course. The ACES course trained participants how to generate and interpret cardiac ultrasound images over a 2-day period. Clinicians completed a battery of individual difference measures containing a measure of visuospatial abilities (MRT; Peters et al. 1995) and handedness (WHQ; Bryden 1977). Clinicians also provided information concerning their prior experience performing ultrasounds. They indicated (1) how many training sessions they have completed, (2) the total hours of experience learning/practicing ultrasound, (3) the total number of cardiac ultrasound studies they performed (clinical and practice), and (4) the total number of non-cardiac ultrasound studies they have performed in the past. We additionally obtained their age in order to control for overall experience.

The competencies assessed in our cardiac ultrasound task were adopted from an international consensus exercise (Expert Round Table on Ultrasound in ICU 2011; Mayo et al. 2009) and our own previous research (Millington et al. 2016, 2017a, b). Twenty-five ultrasound cases consisted of materials previously prepared for a prior iteration of the course. Cases were presented online through www.TeachingMedicine.com. Cases were presented

as a parallel task while learners participated in the multi-day training course. Five case types presented pathophysiologies consisting of left-ventricle, right-ventricle, pericardium, volume, and sepsis. Cases were selected in order to conform to the international consensus on ultrasound examinations (Expert Round Table on Ultrasound in ICU 2011; Mayo et al. 2009). A single "normal" case was also included to establish a baseline. Each case had 6 diagnostic multiple-choice questions and 1 treatment question presented alongside the animated ultrasound image (see Fig. 1). Clinicians entered their answers into the computer and after submitting their answers, were provided with immediately feedback and a case discussion. All their responses were recorded. Questions assessing diagnostic knowledge required that participants report specific features of each case. The number of response alternatives varied (i.e., 3–6). Variation in the number of response alternatives were the result of the presence of multiple pathophysiologies in specific cases. After all diagnostic questions were answered, participants were asked to report their confidence that they had correctly identified the case. Following the treatment question, participants were again asked to provide confidence in their treatment recommendation. In both cases, confidence was rated on a 6-point scale from 50 (guess) to 100 (certain). An overconfidence index was created by taking the mean confidence for a condition and subtracting the mean accuracy, with positive values corresponding to overconfidence and negative values corresponding to underconfidence (e.g., Baranski and Petrusic 1998).

Analyses

As the cases used in the current study had not been formally analyzed as an assessment instrument, our initial analysis examined participants' accuracy and confidence in assessing case type. Prior to analyzing the data, the scores for each case were grouped into two sets based on the kind of information that was assumed to underpin the judgments. Thus, scores for questions of diagnostic features (Questions 1 and 6) were averaged together to create a diagnostic score. We collapsed across diagnostic question type given that (1) the

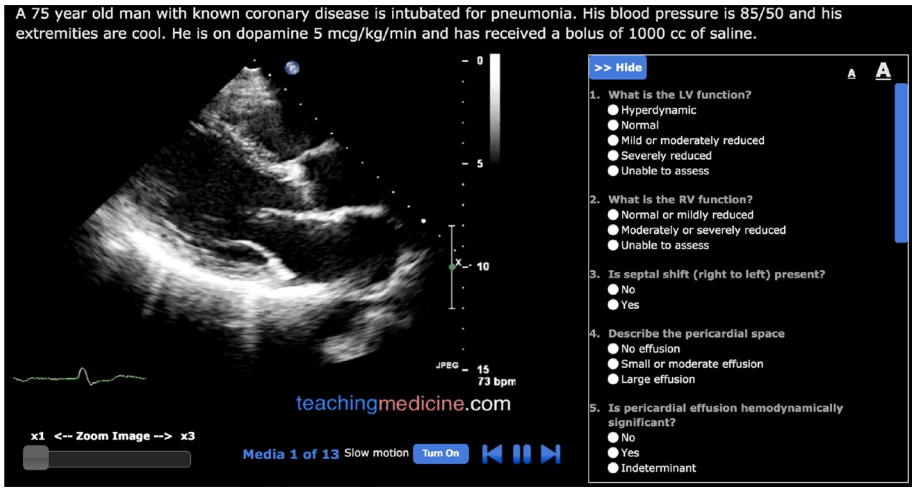

Fig. 1 Sample from the question set used in the study. Case study text is presented at the top of the display, ultrasound image is presented on the left-hand side of the display with questions and response alternatives presented on the right-hand side of the display

MCQs were a means to direct the learners' attention to specific features of the cases and varied in terms of the number of questions that were relevant to a specific case, and (2) to reduce variability in our analysis. The score for Question 7 was examined alone.

A Posteriori *Case Difficulty*. While cardiac ultrasound cases were not initially selected on the basis of their difficulty, we re-classified the cases in order to examine the effect of difficulty on subjective assessments of performance. Due to the variability in responses to the normal case in comparison to other case types, we excluded it from further analysis. Thus, twenty-four cases served as the basis for the correlational analyses. Using a similar rationale to that of Lichtenstein and Fischhoff (1977), cases were then re-classified as "hard", "intermediate", or "easy" on the basis of mean performance in Questions 1 to 6 in order to have equal sets to assess overconfidence bias. Thus, 8 cases were assigned to each of the difficulty conditions creating equivalent groups: hard cases [p (cor) = .69; range .63–.72], intermediate cases [p (cor) = .78; range .74–.81], and easy cases [p (cor) = .85; range .82–.94].

Results

We analyzed the data in two steps. First, we considered participant accuracy for each kind of ultrasound cases used in the online assessment task in order to establish the difficulty of each case (Case Analysis). We then considered the determinants of performance and confidence after partitioning the cases based on their post hoc difficulty (Correlational Analysis).

Case type analysis

A preliminary analysis was conducted to determine whether a "normal case" should be included in our correlational analysis. Using the Greenhouse–Geisser adjusted degrees of freedom, our ANOVA obtained marginally significant results of case type, $F(5,60) = 3.60$, $MSE = 0.025$, $p = .058$. Bonferroni post hoc comparisons, revealed that cases examining left-ventricle pathophysiologies were significantly easier than those cases examining pathophysiologies of right-ventricle function, pericardium, and volume (all p's $< .05$). As Table 1 demonstrates, we additionally observed the greatest variation

Table 1 Proportion correct and mean confidence by case-type for diagnostic questions and treatment question

Case type	Diagnostic questions		Treatment question	
	p (COR)	Confidence	p (COR)	Confidence
Left-ventricle	.835 (.018)	.722 (.033)	.556 (.057)	.718 (.037)
Right-ventricle	.777 (.019)	.756 (.031)	.417 (.058)	.731 (.030)
Pericardium	.696 (.030)	.754 (.043)	.718 (.083)	.697 (.049)
Volume	.753 (.020)	.735 (.031)	.487 (.056)	.717 (.036)
Hyperdynamic	.782 (.025)	.727 (.029)	.529 (.053)	.719 (.031)
Normal	.742 (.052)	.723 (.039)	.385 (.140)	.692 (.038)

Standard error of the mean in parentheses

in performance for the normal case. This is not surprising given that clinicians were only presented with a single normal case to establish a baseline. No significant differences were observed in confidence reports over question type, $F(5,60) = .53$, $p = .604$. This suggests that if there are differences in confidence they are not dependent on the domain-specific knowledge associated with the clinical questions used here.

Unlike the results of the diagnostic questions, we neither observed an effect of case type in performance of the treatment question, $F(5,60) = 2.23$, $p = .123$, nor differences in confidence, $F(5,60) = .62$, $p = .524$. However, given the variability in mean performance across case type, this is likely attributable to the use of a single treatment question. Our failure to find any differences in confidence reports in either the diagnostic questions or treatment question might suggest that confidence is determined by something other than case type.

Correlational analysis

Using a posteriori case difficulty, we conducted a correlational analysis with our individual differences measures after partialling out the effect of clinician age. We did so based on the assumption that age might be related to ultrasound training experience. All variables were analyzed using Pearson's correlation coefficients using two-tailed tests.

Response accuracy

We first considered factors affecting accuracy in diagnostic questions. Not surprisingly, a posteriori case difficulty correlated with accuracy, $r = -.682$, $p < .001$: as case difficulty increased, clinicians' performance decreased. This finding serves as verification that our re-classification procedure for the cases was appropriate.

Performance on diagnostic questions was affected by both prior cardiac, $r = .319$, $p = .051$ and noncardiac ultrasound experience, $r = .449$, $p = .005$. Thus, clinicians' ability to correctly respond to the diagnostic questions was determined by prior ultrasound experience regardless of the type of ultrasound experience. Importantly, cardiac and noncardiac ultrasound experience were only marginally correlated with one another, $r = .292$, $p = .075$, suggesting differences in patterns of ultrasound training that might independently contribute to overall ultrasound competency development.

In terms of general measures of visuospatial ability, only one relationship proved significant. While handedness was not significantly related to performance, $r = .08$, $p = .63$, mental rotation ability was positively correlated with diagnostic question accuracy, $r = .331$, $p = .042$. Thus, the clinicians' general ability to mentally visualize and manipulate objects appears to be related to interpretation of ultrasound images used as the basis for diagnosis.

Our analysis of treatment question accuracy only obtained some of the same trends as those obtained in the analysis of diagnostic question accuracy. We found that a posteriori question difficulty was negatively correlated with performance, $r = -.344$, $p = .034$. We also found that cardiac ultrasound experience was correlated with treatment question accuracy, $r = .382$, $p = .018$.

Importantly, our analysis also suggests that diagnostic and treatment question accuracy was related, $r = .422$, $p = .008$. This suggests that the same kind of knowledge was used for both of these kinds of questions. This would be expected if these kinds of information were associated together in memory in terms of representations of specific pathophysiologies.

Response confidence

Prior to considering overconfidence bias, we first sought to examine the factors that affected mean confidence. Neither confidence in diagnostic nor treatment questions were correlated with question difficulty. Thus, subjective confidence was not influenced by factors other than prior ultrasound knowledge and task demands. However, unlike diagnostic or treatment question accuracy, confidence in diagnostic, $r = .474$, $p = .003$, and treatment question responses, $r = .500$, $p = .001$, were correlated with hours of reported ultrasound experience. However, we observed a negative correlation between previous ultrasound experience obtained through training session and confidence in both diagnostic, $r = -.381$, $p = .018$, and treatment responses, $r = -.506$, $p = .001$. Non-cardiac ultrasound experience was also correlated with confidence in both diagnostic, $r = .404$, $p = .012$, and treatment question responses, $r = .448$, $p = .005$. Thus, while the estimated number of hours of experience tend to increase subjective confidence but did not affect performance, the number of sessions attended by clinicians reduced their overall confidence.

We also observed a positive correlation between confidence in diagnostic question responses and mental rotation ability, $r = .317$, $p = .053$. This might suggest that the ability to visualize the anatomical and physiological features of a case increases certainty and, in turn, might be due to their accessibility to subjective awareness. Importantly, confidence in both diagnostic and treatment questions were strongly correlated, $r = .820$, $p < .001$. While the strength of this relationship likely results from the aforementioned correspondence in accuracy, the comparatively stronger relationship might suggest that confidence processing might utilize a distinct set of cognitive processes.

Overconfidence

In order to disambiguate the factors affecting performance and confidence, we computed an index of overconfidence bias by obtaining the difference between mean confidence and mean accuracy for each condition, i.e., Overconfidence = Mean (Confidence) − Mean

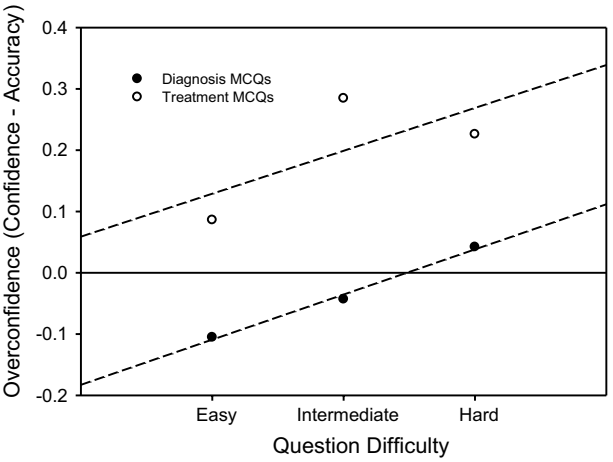

Fig. 2 Overconfidence for MCQs across difficulty level and question type

(Proportion Correct). Importantly, overconfidence for both diagnostic, $r = .447$, $p = .005$, and treatment questions, $r = .287$, $p = .080$, increased with question difficulty. Thus, replicating the Hard–Easy Effect (Lichtenstein and Fischhoff 1977), clinicians exhibited greater overconfidence for difficult questions as compared to easy questions (Fig. 2). If we additionally consider that performance in treatment question was lower than that of diagnostic questions, the Hard–Easy Effect is also evidenced in the overconfidence associated with the harder treatment questions relative to the easier diagnostic questions. Replicating the general pattern observed in the analysis of confidence, we observed decreases in overconfidence bias with previously reported training sessions for both diagnostic, $r = -.290$, $p = .077$, and treatment, $r = -.376$, $p < .020$, questions (Fig. 3). Similarly, we also observed decreased overconfidence bias for treatment questions with increased cardiac ultrasound experience, $r = -.327$, $p < .045$, although we did not observe the same for diagnostic questions.

Discussion

While our results were obtained from a limited number of clinicians, our study provides some insight into factors that are related to clinical knowledge and subjective assessment of performance. Our results suggest that there might be differences in identifying certain pathophysiologies during ultrasound training. Specifically, we found that cases involving LV function abnormalities were associated with better performance relative to cases involving abnormalities in RV function, the pericardial space, and volume. In contrast, accuracy of treatment questions for cases demonstrating pericardial space abnormalities were associated with the greatest accuracy. Importantly, we observed considerable variability in a case that was selected to reflect normal function. For instance, this finding might simply reflect the fact that the normal case represented only a single case relative to the twenty-four other cases that contained evidenced of pathophysiologies. Clinicians might have expected an abnormality given the nature of the other cases. Namely, much like studies wherein clinicians were first presented with cases evidencing pathophysiologies (e.g.,

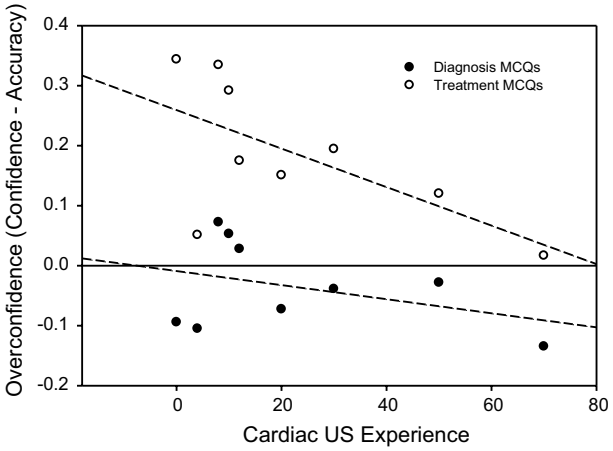

Fig. 3 Experience and overconfidence

Brooks et al. 1991), clinicians might simply perceive abnormalities where none are evidenced, i.e., they are primed to detect features associated with pathology. For this reason, future studies might wish to examine the order in which normal and abnormal cases are presented in or vary the proportion of normal cases presented to clinicians.

Metacognition: determinants and strategies

The metacognitive strategies used by clinicians were also examined in the present study. When cases were classified in terms of a posteriori case difficulty, we observed a number of interesting relationships with mental rotation ability and subjective confidence. We found that mental rotation ability was positively correlated with both response accuracy and response confidence, suggesting that the ability to visualize a problem is related to a clinician's ability to assess ultrasound imagery as well as increasing their confidence. However, by computing an index of overconfidence bias that accounts for both accuracy and confidence, no relationship was observed between mental rotation ability and overconfidence suggesting that clinicians' calibration in judging their own ability was not directly determined by their ability to visualize the anatomical and physiological processes present in the ultrasound image.

In terms of question difficulty, we observed a number of relationships that can inform health care professions education and medical decision-making. Specifically, subjective confidence was not affected by question difficulty suggesting that clinicians' assessment of their performance was not entirely dependent of the specific knowledge used to complete the task. However, an examination of overconfidence bias did find that difficult questions were associated with greater overconfidence relative to easy questions (Lichtenstein and Fischhoff 1977), a pattern observed in some studies of medical decision-making (e.g., Leopold et al. 2005). Thus, clinicians' subjective awareness of their performance appears to be influenced by more than the clinical knowledge used to perform a task. Moreover, the difference in results between our analysis of mean confidence and the index of overconfidence bias indicates the importance for accounting for the relationship between accuracy and confidence prior to interpreting results.

Our study also found that prior experience was an important determinant of both accuracy and overconfidence bias. Unsurprisingly, the amount of prior experience clinicians had was positively related to their performance in the task even after controlling for age. More importantly, we observed decreases in overconfidence bias with increased cardiac ultrasound experience when responding to diagnostic questions. Thus, we find support for the underconfidence-with-practice effect (UWP) observed in studies of memory (Koriat et al. 2002) in medical judgment and decision-making. In the present study, greater experience reduced a clinicians' propensity to be overconfident such that significant experience produced underconfidence. Importantly, given that the effect was strongest for cardiac ultrasound, it suggests that specific training (i.e., cardiac ultrasound experience) is required for clinicians to assess their performance rather than general training (i.e., any ultrasound experience). This also supports the conceptualization of expertise in the health care professions as domain-specific (e.g., Ericsson 2004; Norman et al. 2006).

Limitations

Our study has a number of limitations that should be acknowledged. First, the limited number of clinicians that were available for the study as well as the subset that completed all

cases that qualified them for the analysis might suggest a shared set of characteristics for these clinicians, i.e., self-selection bias. For instance, our clinicians might have personality traits that have complex associations with task performance (e.g., conscientious; e.g., Chen et al. 2001; Martocchio and Judge 1997), and this might have affected the relationship between experience and UWP effect. Similarly, it might also be the case that the self-reported measure of experience could reflect an over- or underestimation of ultrasound training experience that correlated with under- or overconfidence. Second, the features of the website did not permit the random presentation of cases to learners resulting in each learner receiving the same random sequence of cases. Consequently, order-dependencies might have affected responses in some unforeseen way. For instance, the single "normal" comparison cases might have caused learners to identify pathophysiologies that were not present leading to reduced performance for this case. Finally, the present study was limited to the consideration of cardiac ultrasound. Further studies should examine whether the relationship between domain-specific experience generalize to other areas of ultrasound practice (e.g., lung, abdominal, obstetric). Thus, while suggesting the presence of overconfidence bias and the UWP effect, additional controls will be required in futures studies to further reduce the influence of extraneous variables.

Conclusions

Health care professionals must account for the inherent uncertainty of clinical judgments (Katz 1988; Lingard et al. 2003; Fox 2000; Spafford et al. 2006; Timmermans and Angell 2001). Studies of medical decision-making and medical education have started to recognize the importance of understanding metacognitive processes involved in learning, decision-making, and clinical practice (e.g., Croskerry 2009; Moulton et al. 2007; Reyna 2008). Metacognition requires both the monitoring and regulation of one's performance. In the context of medical education and continuing professional development, these processes are critical to understanding self-assessment and self-directed learning as well as reducing diagnostic and treatment errors (Croskerry 2009). Indeed, some of these errors might be attributable to overconfidence in clinical knowledge (Berner and Graber 2007; Croskerry and Norman 2008; Friedman et al. 2005). In the current study, we found that subjective confidence and overconfidence bias were a result of a lack of experience within a specific domain (i.e., cardiac ultrasound) rather than more general visuospatial abilities that might be relevant to the task (e.g., mental rotation).

In contrast to studies that use the confidence ratings of health care professionals as a proxy for comprehension, our study suggests that performance and subjective confidence can differ markedly. Specifically, replicating previous results, clinicians in our sample demonstrate overconfidence for hard questions and underconfidence for easy questions (Lichtenstein and Fischhoff 1977) with the amount of prior cardiac ultrasonography leading to underconfidence (Koriat et al. 2002). Thus, while some studies have been directed toward measuring increases in learners' confidence (e.g., Westbrook et al. 2005), interpreting such outcome measures appears to be problematic without first considering how these subjective assessments of performance relate to a clinician's accuracy. Namely, as we have demonstrated, increases in confidence are not necessarily attributable to increases in diagnostic or treatment accuracy. Confidence processes appear to be influenced by information beyond that used to respond to diagnostic and treatment questions. Attention to nondiagnostic information (Schoenherr et al. 2010) and use of erroneous cues (Koriat 1993; Koriat

et al. 2002) represent possible sources of overconfidence bias. Consequently, confidence reports appear to require a separate, resource-demanding set of cognitive processes (Baranski and Petrusic 1998). Thus, while program evaluators might wish to increase the subjective confidence of health care professions in performing their duties, they should not do so without taking into account learners' performance on a task. Indeed, a focus on increased confidence by educators and clinicians, might suggest to learners that confidence in performing a procedure is more important than assessing their clinical competence.

While a number of interesting findings pertained to the amount and kind of ultrasound experience our clinicians brought into the task, the greatest determinant of their overconfidence in clinical management in the present study was the amount of prior cardiac ultrasound experience that they had. Importantly, as experience increased, the degree of overconfidence bias decreased. This has important implications for patient safety: while medical educators often focus on improvements in clinical knowledge, medical education must also focus on developing resources and skills that allow clinicians to increase awareness of discrepancies between confidence and competency.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. *Perception and Psychophysics, 55,* 412–428.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 929–945.

Baranski, J. V., & Petrusic, W. M. (2001). Testing the architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology, 55,* 195–206.

Barnsley, L. L. (2004). Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Medical Education, 38,* 358–367.

Berner, E. S., & Graber, M. L. (2007). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine*, *121*, S2–S23.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal Experimental Psychology: General, 120,* 278–287.

Bryden, M. P. (1977). Measuring handedness with questionnaires. *Neuropsychologia, 15,* 617–624.

Chen, G., Casper, W. J., & Cortina, J. M. (2001). The roles of self-efficacy and task complexity in the relationships among cognitive ability, conscientiousness, and work-related performance: A meta-analytic examination. *Human Performance, 14,* 209–230.

Corballis, M. C. (1997). Mental rotation and the right hemisphere. *Brain and Language, 57,* 100–121.

Croskerry, P. A. (2009). Universal model of diagnostic reasoning. *Academic Medicine, 84,* 1022–1028.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 12,* S24–S29.

Dowie, J., & Elstein, A. (1988). *Professional judgment: A Reader in clinical decision-making*. Cambridge: Cambridge University Press.

Drew, T., Võ, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science, 24,* 1848–1853.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge: Harvard University Press.

Elzubeir, M. A., & Rizk, D. E. (2001). Assessing confidence and competence of senior medical students in an obstetrics and gynaecology clerkship using an OSCE. *Education for Health, 14,* 373–382.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *79*, S70–S81.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Medical Education, 41,* 1124–1130.

Evans, J. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

Expert Round Table on Ultrasound in ICU. (2011). International expert statement on training standards for critical care ultrasonography. *Intensive Care Medicine*, *37*, 1077–1083.

Ferrel, W. R., & McGooey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behaviour and Human Performance*, *26*, 32–53.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33,* 238–244.

Fox, R. C. (2000). Medical uncertainty revisited. In G. L. Albrecht, R. Fitzpatrick, & S. C. Scrimshaw (Eds.), *The handbook of social studies in health and medicine* (pp. 409–425). London: SAGE Publications.

Fox, R. A., Ingham Clark, C. L., Scotland, A. D., & Dacre, J. E. (2000). A study of pre-registration house officers' clinical skills. *Medical Education, 34,* 1007–1012.

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., et al. (2005). Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *Journal of General Internal Medicine, 20,* 334–339.

Gambrill, E. (1990). *Critical thinking in clinical practice: Improving the accuracy of judgments and decisions about clients*. San Francisco: Jossey-Bass Inc., Publishers.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Guillot, A., Champely, S., Batier, C., Thiriet, P., & Collet, C. (2007). Relationship between spatial abilities, mental rotation and functional anatomy learning. *Advances in Health Sciences Education, 12,* 491–507.

Halm, B. M., Lee, M. T., & Franke, A. A. (2010). Improving medical student toxicology knowledge and self-confidence using mannequin simulation. *Hawaii Medical Journal, 69,* 4–7.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–216.

Hatala, R., Norman, G. R., & Brooks, L. R. (1999). Impact of a clinical scenario on accuracy of electrocardiogram interpretation. *Journal of General Internal Medicine, 14,* 126–129.

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence, 34,* 151–176.

Hogg, G., & Miller, D. (2016). The effects of an enhanced simulation programme on medical students' confidence responding to clinical deterioration. *BMC Medical Education, 16,* 1–8.

Hoyek, N., Collet, C., Rastello, O., Fargier, P., Thiriet, P., & Guillot, A. (2009). Enhancement of mental rotation abilities and its effect on anatomy learning. *Teaching and Learning in Medicine, 21,* 201–206.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, Giroux.

Katz, J. (1988). Why doctors don't disclose uncertainty. In A. Elstein & J. Dowie (Eds.), *Professional judgment: A reader in clinical decision making* (pp. 544–565). Cambridge: Cambridge University Press.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Norwood, NJ: Ablex.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experiment Psychology: General, 131,* 147–162.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., et al. (2006). Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology, 37,* 1543–1556.

Kulatunga-Moruzi, C., Brooks, L. R., & Norman, G. R. (2004). The diagnostic disadvantage of having all the facts: using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30,* 563–572.

Kvidera, S., & Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual decisions. *Journal of Behavioral Decision Making, 21,* 253–281.

Leopold, S. S., Morgan, H. D., Kadel, N. J., Gardner, G. C., Schaad, D. C., & Wolf, F. M. (2005). Impact of educational intervention on confidence and competence in the performance of a simple surgical task. *The Journal of Bone and Joint Surgery, 87,* 1031–1037.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how, much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., Fischhoff, B., & Phillips, S. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–344). Cambridge: Cambridge University Press.

Lingard, L. (2012). Rethinking competence in the context of teamwork. In B. D. Hodges & L. Lingard (Eds.), *The question of competence: Reconsidering medical education in the twenty-first century* (pp. 131–154). Ithaca: Cornell University Press.

Lingard, L., Garwood, K., Schryer, C. F., & Spafford, M. M. (2003). A certain art of uncertainty: Case presentation and the development of professional identity. *Social Science and Medicine, 56,* 603–616.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences on self-deception and self-efficacy. *Journal of Applied Psychology, 82,* 764–773.

Mast, F. W., Ganis, G., Christie, S., & Kosslyn, S. M. (2003). Four types of visual mental imagery processing in upright and tilted observers. *Cognitive Brain Research, 17,* 238–247.

Mavis, B. (2001). Self-efficacy and OSCE performance among second year medical students. *Advances in Health Science Education, 6,* 93–102.

Mayo, P., Beaulieu, Y., Doelken, P., et al. (2009). American College of Chest Physicians/La Société de Réanimation de Langue Française statement on competence in critical care ultrasonography. *Chest*, *135*, 1050–1060.

Millington, S. J., Wong, R. Y., Kassen, B. O., Roberts, J. M., & Ma, I. W. Y. (2009). Improving internal medicine residents' performance, knowledge and confidence in central venous catheteterization using simulators. *Journal of Hospital Medicine*, *4*, 410–416.

Millington, S. J., Arntfeld, R. T., Guo, R. J., Koenig, S., Kory, P., Noble, V., et al. (2017a). The Assessment of Competency in Thoracic Sonography (ACTS) scale: Validation of a tool for point-of-care ultrasound. *Critical Ultrasound Journal, 9,* 1–8.

Millington, S. J., Arntfeld, R. T., Hewak, M., Hamstra, S. J., Beaulieu, Y., Hibbert, B., et al. (2016). The Rapid Assessment of Competency in Echocardiography scale. *Journal of Ultrasound in Medicine*, *35*, 1457–1463.

Millington, S. J., Hewak, M., Arntfeld, R. T., Beaulieu, Y., Hibbert, B., Koenig, S., et al. (2017b). Outcomes from extensive training in critical care echocardiography: Identifying the optimal number of practice studies required to achieve competency. *Journal of Critical Care, 40,* 99–102.

Morgan, P. J., & Cleave-Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. *Medical Education, 36,* 534–539.

Moulton, C.-A. E., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: A new model of expert judgment. *Academic Medicine, 82,* S109–S116.

Norman, G. (2006). Building on experience—the development of clinical reasoning. *New England Journal Medicine, 355,* 2251–2252.

Norman, G. R., & Brooks, L. R. (1997). The non-analytic basis of clinical reasoning. *Advances in Health Science Education Theory and Practice, 2,* 173–184.

Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in medicine and surgery. In K. A. Ericsson (Ed.), *The Cambridge handbook of expertise and expert performance* (pp. 339–353). New York: Cambridge University Press.

Norman, D. A. & Shallice, T. (1980). *Attention to action: Willed and automatic control of behavior*. CHIP Report 99, University of California, San Diego.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect perfor mance. *Brain and Cognition*, *28*, 39–58.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117,* 864–901.

Regehr, G., Cline, J., Norman, G. R., & Brooks, L. (1994). Effect of processing strategy on diagnostic skill in dermatology. *Academic Medicine, 69,* S34–S36.

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy-trace theory. *Medical Decision Making, 28,* 829–833.

Saling, L. L., & Philips, J. G. (2007). Automatic behaviour: Efficient not mindless. *Brain Research Bulletin, 73,* 1–20.

Schoenherr, J. R., Leth-Steensen, C., & Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. *Attention, Perception, & Psychophysics, 72,* 353–368.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge: MIT Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171,* 701–703.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Spafford, M. M., Schryer, C. F., Lingard, L., & Hrynchak, P. K. (2006). What healthcare students do with what they don't know: The socializing power of "uncertainty" in the case presentation. *Communication and Medicine, 3,* 81–92.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.

Timmermans, T., & Angell, A. (2001). Evidence-based medicine, clinical uncertainty, and learning to doctor. *Journal of Health and Social Behavior, 42,* 342–359.

Turner, N. M., Lukkassen, I., Bakker, N., Draaisma, J., & ten Cate, O. T. (2009). The effect of the APLS-course on self-efficacy and its relationship to behavioural decisions in paediatric resuscitation. *Resuscitation, 80,* 913–918.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica, 50,* 179–197.

Westbrook, J. I., Gosling, S., & Coiera, E. W. (2005). The impact of an online evidence system on confidence in decision making in a controlled setting. *Medical Decision Making, 12,* 315–321.