Adv in Health Sci Educ (2011) 16:311–329
DOI 10.1007/s10459-010-9263-2

Exploring the divergence between self-assessment and self-monitoring

Kevin W. Eva · Glenn Regehr

Received: 8 July 2010 / Accepted: 9 November 2010 / Published online: 30 November 2010

Abstract Many models of professional self-regulation call upon individual practitioners to take responsibility both for identifying the limits of their own skills and for redressing their identified limits through continuing professional development activities. Despite these expectations, a considerable literature in the domain of self-assessment has questioned the ability of the self-regulating professional to enact this process effectively. In response, authors have recently suggested that the construction of self-assessment as represented in the self-regulation literature is, itself, problematic. In this paper we report a pair of studies that examine the relationship between self-assessment (a global judgment of one's ability in a particular domain) and self-monitoring (a moment-by-moment awareness of the likelihood that one maintains the skill/knowledge to act in a particular situation). These studies reveal that, despite poor correlations between performance and *self-assessments* (consistent with what is typically seen in the self-assessment literature), participant performance was strongly related to several measures of *self-monitoring* including: the decision to answer or defer responding to a question, the amount of time required to make that decision to answer or defer, and the confidence expressed in an answer when provided. This apparent divergence between poor overall self-assessment and effective self-monitoring is considered in terms of how the findings might inform our understanding of the cognitive mechanisms yielding both self-monitoring judgments and self-assessments and how that understanding might be used to better direct education and learning efforts.

Keywords Self-assessment · Self-monitoring · Professional self-regulation · Self-directed learning

Although many models of professional self-regulation and many health professional training curricula worldwide have incorporated some form of planned self-assessment

K. W. Eva (✉) · G. Regehr
University of British Columbia, Vancouver, BC, Canada
e-mail: kevin.eva@ubc.ca

activity into their educational process (Sargeant et al. 2010), there is now a well established literature that raises doubts about the capacity of individuals to effectively self-assess either personal (Dunning et al. 2004) or professional (Gordon 1991; Boud 1995; Davis et al. 2006) areas of relative strength and weakness. Increasingly it is being recognized that self-assessment as "a process of personal reflection based on an unguided review of practice and experience for the purposes of making judgments regarding one's own current level of knowledge, skills, and understanding as a prequel to self-directed learning activities that will improve overall performance and thereby maintain competence" (Eva and Regehr 2007, p. 81) is inherently flawed. As a result, research in the field has begun to move away from efforts to quantify the self-assessment ability of individuals using "guess your grade" (Colliver et al. 2005) research designs, toward efforts to understand the nature and sources of this inaccuracy (e.g., Kruger and Dunning 1999; Hodges et al. 2001; Eva and Regehr 2005).

As part of this re-emphasis, Eva and Regehr (2007) have drawn a theoretical and methodological distinction between self-assessment as a cumulative evaluation of overall performance, and self-assessment as a process of self-monitoring performance in the moment. Building on findings in the literature that suggest experts spontaneously "slow down" when faced with personally challenging situations or problems (Norman et al. 1989), they presented research participants with 60 trivia questions and asked these participants to answer when they were confident that they could do so accurately or to pass when they felt unable to answer with confidence. After viewing all 60 questions, participants were asked to make their best guess in response to questions they chose not to answer during the first presentation. Participants were timed with regard to the speed with which they made their decision to either answer or pass in the first round. Eva and Regehr found numerous indications that self-monitoring is an importantly different (and substantially more accurate) process relative to self-assessment. Answers were much more likely to be correct for items that participants chose to answer in the first round (relative to questions that were not answered until the second round) and, in particular, when they made the decision to answer/pass quickly. This was true despite the fact that participants' overall judgments of the strength of their knowledge replicated the typically poor correlations with performance seen in other self-assessment studies. Taken together, these findings suggest that at least one source of inaccuracy in overall self-assessments may be an inability to effectively mentally aggregate performance over past events despite an apparent capacity to effectively self-monitor in the moment for each event.

However, a number of questions remain. For example, Eva and Regehr's (2007) evidence of effective self-monitoring in the moment was largely circumstantial based on behavioural patterns. They did not collect data that would allow a determination of whether participants were consciously aware of their likelihood of success in the moment, or whether these behavioural indices were largely measures of unconscious processing. The extent to which participants have conscious access to their ongoing determinations of likely success in the moment has important implications for how they might more effectively use this ability to slow down when they should (Moulton et al. 2007), not only for safer practice in the moment, but also for future improvement (cf. Regehr and Mylopoulos 2008). That is, with the considerable emphasis placed on self-directed learning in the health professions it is important to determine: (1) how self-monitoring might influence continuing professional development activities; (2) whether or not explicit prompts to self-monitor alter those tendencies; and (3) whether or not the learning activities in turn alter self-perceptions in a meaningful way.

In this paper we present the results from two studies aimed at testing:

(a) whether or not individuals are consciously self-aware of their likelihood of success in specific situations (i.e., do explicit confidence ratings mimic the behavioural indications of accurate self-monitoring reported in previous work?)
(b) whether or not any such awareness (i.e., conscious or otherwise) dictates individuals' efforts to search for the information that would help them solve the problems encountered (i.e., do self-monitoring indicators relate to the likelihood of searching for the correct answer to a question when given the opportunity to conduct an internet search), and
(c) whether or not those efforts in turn influence participants' overall self-assessments (i.e., does the accuracy of individuals' overall self-assessments change as a result of a self-monitoring experience?).

In addition, for study 2 we manipulated the response format, utilizing both short answer questions (SAQ) and multiple-choice questions (MCQ) in an effort to begin exploring the extent to which the increased accuracy of self-monitoring relative to self-assessment is better characterized as an indication of the temporal relationship between the performance and the judgment (i.e., overall vs moment-by-moment) or as an indication of the task requirements of the self-monitoring judgment. That is, in deciding whether or not one has the capacity to answer a short answer question one can be guided by the directly relevant information of whether or not a response is generated that seems likely to be accurate. In contrast, if one knows response alternatives will be presented in MCQ format one might feel confident in one's capacity to answer not only by the ability to generate a plausible response, but also by a feeling of knowing leading to anticipation that one could identify the correct answer if given alternatives from which to choose. Indeed, the literature on feeling-of-knowing judgments in psychology would suggest that feeling-of-knowing judgments are based on heuristics that operate automatically and unconsciously (Koriat 2000). These mechanisms would seem to be aligned well with the mechanisms that have been postulated to make self-assessment untrustworthy (i.e., difficulty mentally aggregating past experiences). By manipulating response format as a variable in Study 2 we offer preliminary information regarding when and how self-monitoring is likely to accurately reflect awareness of the limits of one's competence.

Study 1: Methods

Participants

Participants were recruited from undergraduate Psychology courses at McMaster University. Informed consent was collected and, in exchange for their participation, subjects received either bonus credits toward their coursework or a $10 stipend. Ethics approval was granted by McMaster University's Faculty of Health Sciences Research Ethics Board.

Materials

Participants were presented with 60 general knowledge trivia questions, 10 from each of 6 domains. Study 1 used the same set of questions as were used by Eva and Regehr (2007). These questions were drawn from a norming study in which Nelson and Narens (1980) asked a large number of questions to a large number of individuals to determine the probability of answering each question correctly. Questions were selected from this set to

ensure variability in item difficulty with the probability correct lying between 0.2 and 0.8 according to the 1980 norms. The domains included were geography, history, literature, pre-1990s entertainment, science, and sports.

Procedure

A computer-based platform was developed to present each of the 60 questions in turn. The basic design was consistent with that of Eva and Regehr (2007) with modifications made as illustrated in Fig. 1. Participants were told they would be asked to answer a series of trivia questions and that a correction factor would be imposed such that the number of incorrect

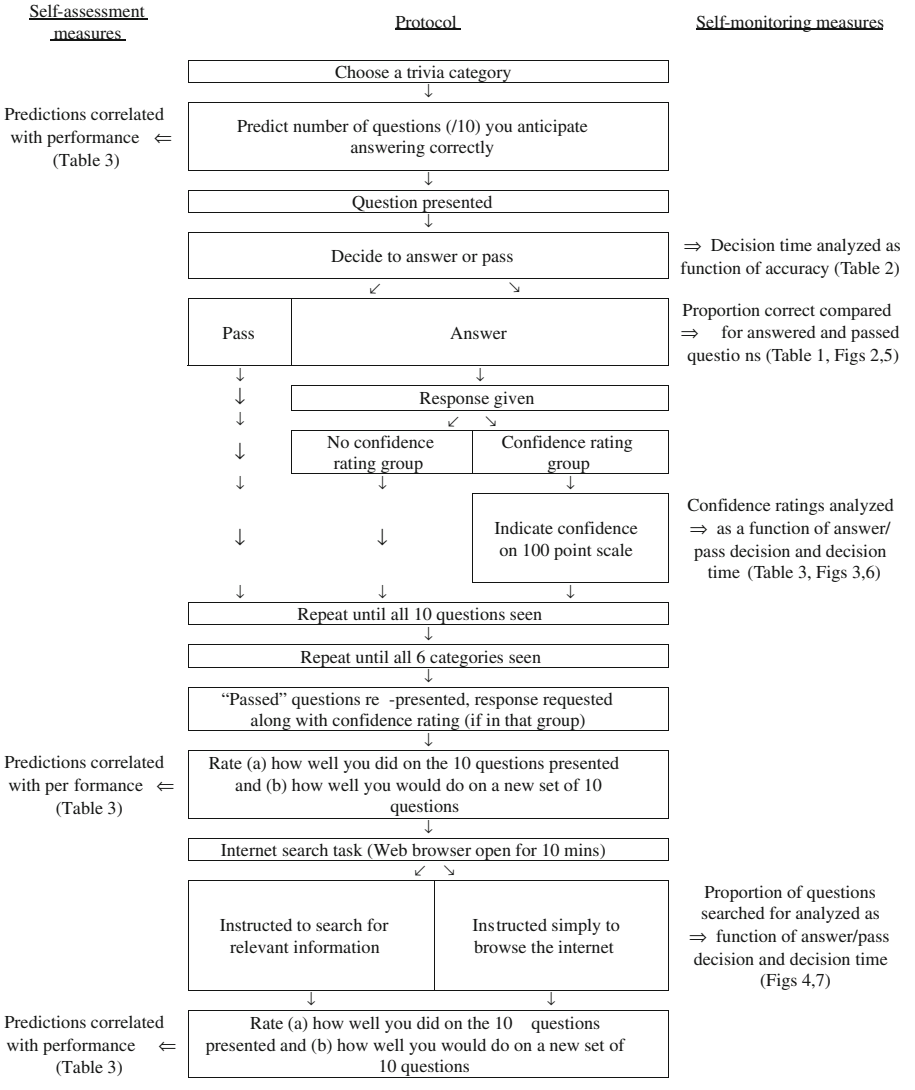

Fig. 1 Flow-chart illustrating the experimental procedure and summarizing the dependent variables used as indicators of self-assessment and self-monitoring

responses would be subtracted from the number of correct responses to determine their total score. As such, they were told they should only answer a question if they felt confident in their ability to do so accurately.

To begin, they were shown a list of the 6 categories, asked to select the one in which they were most confident, then asked to estimate how many questions (out of 10) they expected to get right within that category. After completing this process, the 10 questions from the selected category were presented. When the first question from within that category was presented the amount of time the participant took to decide whether or not to attempt to answer the question (indicated by clicking on an "Answer" or a "Pass" button) was measured. If the participant clicked the "Pass" button, the question disappeared and the next question was presented. If the respondent clicked the "Answer" button, a free text box was presented and the participant was asked to type in a response.

Participants were randomized into one of two groups: a "no confidence rating" condition (group 1) or a "confidence rating" condition (group 2). After answering the question, those in group 1 were simply presented with the next question and asked, again, to make an answer/pass decision. Those in group 2 were additionally prompted, after answering the question, to use a 100-point rating scale indicating how confident they were that the answer they just provided was correct.

The procedure continued in this way (presentation of a question, a decision to answer or not, answering if they chose to try, giving a confidence rating if in group 2, then the next question being presented) until all 10 questions from the selected category had been presented. The procedure was then repeated (select their next best category, estimate their likely number of correct answers, and respond to the 10 questions successively) for each of the remaining 5 categories.

After completing the procedure for all 6 categories, participants were shown, in sequence, all the questions they chose *not* to answer on the first round. For this second round, they were told that the correction factor had been removed and that they should, therefore, take their best guess at the correct response for every item. Again, those in group 2 were asked to assign confidence ratings each time a response was given.

After completing both rounds of the test, participants were asked, for each category, to indicate (a) how many questions they thought they answered correctly and (b) how many questions they would anticipate answering correctly were a new set of 10 questions to be presented to them in the future.

A web browser was then opened to google.com and participants were told that they should spend 10 min browsing the internet while the researcher prepared the final task in the experiment. Half of the participants in the "confidence rating" group and half in the "no confidence rating" group were randomized to receive specific instructions to search for answers to the questions with which they had just been presented. The remaining participants were not directed to search for any particular material. We recorded the google search terms entered and web sites visited, and later coded each web site based on the trivia question to which it likely corresponded (if any).

Finally, after 10 min the browser closed and participants were again asked to respond, for each of the six categories of trivia, to the two questions described earlier (accuracy on this test and anticipated accuracy on a similar test in the future).

Analysis

To address the self-assessment literature Pearson's correlations were used to calculate the relationship between performance and participant's ratings of how well they anticipated

performing or believed themselves to have performed. As measures of self-monitoring we compared decision time, proportion correct, confidence ratings (group 2), and the number of question-relevant websites visited during the internet search task as a function of whether or not participants chose to respond to a given question and the accuracy of the response generated. ANOVA or paired samples t-tests were used as inferential statistics in these instances. The procedure, and a summary of the dependent measures are illustrated in Fig. 1.

Study 1: Results and discussion

The number, gender and average age of participants was 51 (26 male, age = 20.12). For analyses that did not explicitly address the "in the moment" confidence ratings (elicited from group 2), there were no significant effects of this additional task on outcome measures, suggesting that the explicit self monitoring instruction did not alter participants' overall self-assessment accuracy, their self-monitoring behaviours, or their subsequent information seeking behaviour. Thus, for these analyses data from both groups are collapsed and treated as a single group.

Overall self-assessments (groups 1 and 2 together)

Replicating the typical approach to studying self-assessment, we examined participants' overall self-assessments in comparison to their actual performance. Participants' mean performance score (and standard deviation) was 34.0% correct (SD = 16.3%). Their mean predictions regarding anticipated performance was 44.7% correct (SD = 13.6%). Paired sample *t*-tests revealed that performance was statistically lower than predictions ($p < 0.05$).

We also correlated, for each of the 6 trivia categories, participants' actual performance and their predictions regarding how many correct answers they would provide for each category. The mean (and range) of Pearson's r values for the six domains were $r = 0.28$ (0.15–0.42), thereby replicating the typical poor correlation between predictions and performance seen in most self-assessment studies.

Self-monitoring: knowing when to defer (groups 1 and 2 together)

Participants chose to offer an answer the first time they were presented with a question 41.3% of the time. Comparing the percent correct achieved during round 1 (i.e., when participants chose to respond to the questions presented) to that of round 2 (i.e., the questions for which candidates opted not to answer until the correction factor was removed) provides one opportunity to determine whether or not participants were able to self-monitor. Consistent with data reported by Eva and Regehr (2007), participants revealed greater accuracy rates in round 1 (72.6%) relative to round 2 (6.8%). The difference was statistically significant ($p < 0.001$) using a paired samples *t*-test.

Self-monitoring: slowing down at the borders of competence (groups 1 and 2 together)

To determine whether participants were showing appropriate caution and slowing down at the borders of their competence, we calculated the amount of time it took to decide whether

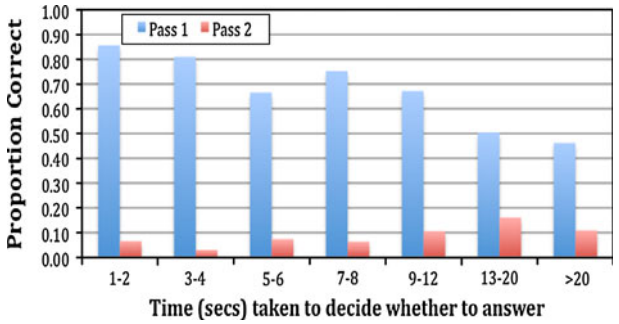

Fig. 2 Mean proportion correct as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 1

or not to answer a given question in relation to the accuracy of the eventual response. If participants are slowing appropriately at the edge of their competence, then for questions that participants chose to answer on the first round, slower decisions (to answer) should be associated with lower accuracy relative to faster decisions. Similarly, for questions that participants decided NOT to answer in the first round, slower decisions (to pass) should be associated with HIGHER accuracy relative to faster decisions. The data confirmed this pattern, again replicating the findings reported by Eva and Regehr (2007). Figure 2 elaborates on this pattern by graphing mean proportion correct as a function of response time (with response time bundles selected to roughly equate the number of observations in each column) and whether participants chose to answer on round 1 or deferred their response to round 2.

Self-monitoring: confidence ratings (group 2 data only)

To test whether or not participants were able to explicitly report their likelihood of success and to determine whether or not being prompted to do so altered their capacity to self-monitor, half of all participants ($n = 26$) were asked to rate their confidence in each response whenever they answered a question. For comparability with earlier analyses we averaged the confidence ratings provided within each category and correlated these ratings with within category performance. These moment-by-moment assessments were more related to performance than were the general self-assessments of participants' confidence in their knowledge of the domain (i.e., their predictions of performance). The mean (and range of) correlations across domain between these question-by-question confidence ratings and performance were $r = 0.81$ (0.66–0.91). Further, confidence ratings were found to map well onto decision times in a pattern very similar to that seen in the accuracy rates. Figure 3 replicates Fig. 2, plotting confidence ratings rather than proportion correct on the y-axis.

Learning activity: the internet search task (groups 1 and 2 together)

After answering all 60 trivia questions participants were given 10 min worth of access to the internet. During this time the rate at which they searched for information relevant to the trivia questions did not reveal a consistent effect of instruction to search for related information. The proportion of questions for which participants searched did, however,

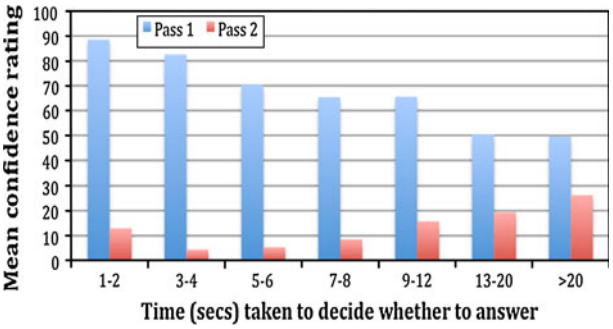

Fig. 3 Mean confidence ratings as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 1

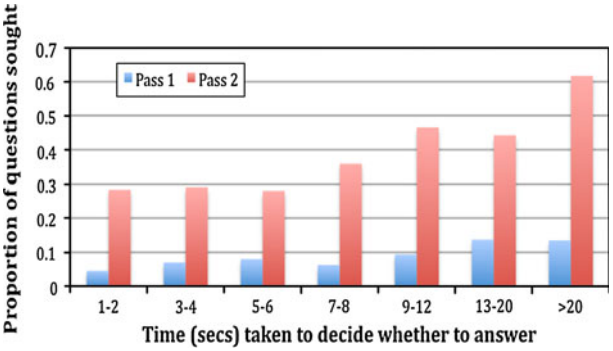

Fig. 4 Proportion of time question-relevant information was sought during a 10-minute internet search task as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 1

reveal a consistent relationship with whether participants had earlier decided to answer or defer responding. Questions on which participants chose to defer were searched for more often (mean = 33.1% of the deferred questions) relative to questions participants opted to answer (mean = 7.9% of the answered questions). These differences were statistically significant ($p < 0.001$) using a paired samples *t*-test. The rate of searching for question-related information was similar for both questions answered correctly and those answered incorrectly. The likelihood of searching for a response to a question was generally greater when participants took longer to make their decision regarding whether to answer or defer as illustrated in Fig. 4.

Post-exercise self-assessments

To determine whether or not the accurate self-monitoring that participants demonstrated in this study translated into improved overall self-assessments of performance we examined the correlation between performance and participants' self-assessments, both immediately following the test and after completing the internet search, by asking (a) how many questions were answered correctly within each category during the study and (b) how many questions would be answered correctly were a new set of 10 questions from within each

category presented in the future. As the responses to these four questions were highly correlated with one another we calculated an average post-performance score for use in our analyses. Consistent with other literature, the self-assessment correlations post-performance were somewhat higher than those offered prior to performance. The mean (and range) of correlations were $r = 0.66$ (0.59–0.72).

Summary

The findings of study 1 replicate and extend the work of Eva and Regehr (2007), demonstrating that moment-by-moment self-monitoring elicits different (and apparently more accurate) indications of awareness of the limits of one's competence than do more traditionally collected overarching estimates of one's ability (i.e., the global self-assessments of ability in each domain). The additional collection of confidence ratings in the current study demonstrates that respondents were indeed consciously aware of the likelihood that they would answer any given question correctly. Interestingly, the requirement to rate one's confidence on a question by question basis did not impact upon the patterns of data seen in the other measures collected.

Study 2 extends these findings by testing for the same pattern of data in a new set of trivia questions and by further by comparing the pattern of results seen in the context of responding to short answer questions to the pattern seen when responding to multiple choice questions (with the decisions to answer or not occurring before the choices are presented).

Study 2: Methods

Study 2 was identical to study 1 with two exceptions. First, to ensure the generalizabilty of the findings, new questions were selected. These questions were drawn from a variety of recently published trivia games using the same categories with the exception that media and the arts replaced pre-1990s entertainment. Second, response format was treated as a variable in study 2. After each question was presented and participants chose to answer, some participants were presented with the same free text response box as was used by participants enrolled in study 1. Others were presented with four response options and were asked simply to indicate which option was the correct answer. Respondents in the "confidence rating" group (group 2) were still asked to assign confidence ratings after making a response and, in all cases, the response format of round 1 was maintained during round 2 (i.e., when participants were asked to respond to questions on which they passed during round 1).

Study 2: Results and discussion

The number, gender and average age of participants was 90 (31 male, age = 18.59). As in study 1, there were no significant effects of the "confidence rating" task on outcome measures leading us to collapse across this variable and examine both groups together when possible to do so.

Overall, the patterns of data for the SAQ condition of this study replicated the patterns seen in study 1. The effects were slightly, but not substantially smaller. By contrast, the patterns of data for the MCQ condition were generally much (significantly) smaller and often were not statistically significant. The details of the results are presented below.

Overall self-assessments (groups 1 and 2 together)

Mean performance scores (and standard deviations) for the SAQ and MCQ conditions were 25.6% (SD = 9.4%), and 42.9% (SD = 9.0%), respectively. Mean candidate predictions regarding their anticipated performance were 38.1% (SD = 11.9%), and 44.7% (SD = 13.7%), respectively. Paired sample *t*-tests revealed that performance was statistically lower than predictions ($p < 0.05$) in the SAQ condition, but not in the MCQ format (i.e., the format in which performance attributable to chance would be highest).

The correlations between participants' actual performance and their predictions regarding how many correct answers they would provide for each category had a mean (and range) of $r = 0.31$ (0.14–0.51), and $r = 0.18$ (0.01–0.36) for the SAQ and MCQ conditions, respectively.

Self-monitoring: knowing when to defer (groups 1 and 2 together)

Participants chose to offer an answer the first time they were presented with a question 32.6% of the time in the SAQ condition and 60.4% of the time in the MCQ version (chi-squared = 400, $p < 0.001$), thereby indicating that participants were indeed more likely to anticipate being able to answer a question correctly when they knew response options would be presented relative to when participating in the SAQ condition.

Comparing the percent correct achieved during round 1 to that of round 2 revealed a significantly smaller difference in the MCQ version of the experiment relative to the SAQ version (for the ANOVA interaction between answer format and round $F_{1,86} = 63.7$, $p < .001$), but for both conditions, the difference between round 1 and round 2 accuracy were statistically significant ($p < 0.001$). Table 1 summarizes these findings along with those of study 1.

Self-monitoring: slowing down at the borders of competence (groups 1 and 2 together)

To determine whether participants were showing appropriate caution and slowing down at the borders of their competence, we calculated the amount of time it took to decide whether or not to answer a given question in relation to the accuracy of the eventual response. The mean decision times, illustrated in Table 2 confirmed the pattern seen in study 1 and reported by Eva and Regehr (2007).

However, again, the pattern of data was smaller for the MCQ condition. For questions answered in round 1, there was a significant interaction between accuracy and answer format ($F_{1,88} = 6.14$, $p < .05$), although there was no significant interaction for round 2 data. In fact, while the differences were in the predicted direction for the MCQ condition, they were not significant for round 1 or round 2.

Table 1 Percent correct responses provided to items respondents chose to answer in round 1 relative to those they deferred answering until round 2

	Answered in round 1 (%)	Deferred until round 2 (%)	*p*-value
Study 1	72.6	6.8	<0.001
Study 2 (short answer questions)	54.4	11.7	<0.001
Study 2 (multiple choice questions)	49.9	32.3	<0.001

Table 2 Mean time (in seconds) taken to make a decision regarding whether to answer or defer answering as a function of the decision made and the accuracy of the answer given

	Answered in round 1			Deferred until round 2		
	Correct	Incorrect	*p*-value	Correct	Incorrect	*p*-value
Study 1	6.7	11.9	<0.001	8.5	6.6	<0.001
Study 2 (short answer questions)	6.6	9.8	<0.001	7.2	6.3	0.06
Study 2 (multiple choice questions)	4.3	4.8	NS	5.0	4.9	NS

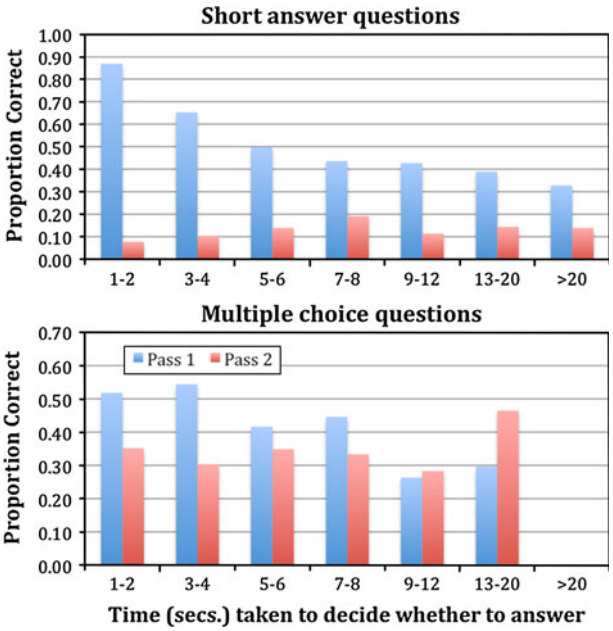

Fig. 5 Mean proportion correct as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 2

Figure 5 elaborates on the pattern for the decision time data. For the SAQ condition, the slope of the line for round 1 answers was significantly negative ($p < .001$) and the slope of the line for round 2 responses was marginally significantly positive ($p = .052$). For the MCQ condition the pattern of data trend in the appropriate directions. The slopes are not significantly different from zero, but they are also not significantly different from the corresponding slopes in the SAQ condition.

Self-monitoring: confidence ratings (group 2 data only)

Forty six respondents performed the item by item confidence ratings (28 in the SAQ condition and 18 in the MCQ condition). The correlations between confidence ratings and performance were variable across condition, but in each case the moment-by-moment assessments were more related to performance than were the general self-assessments of

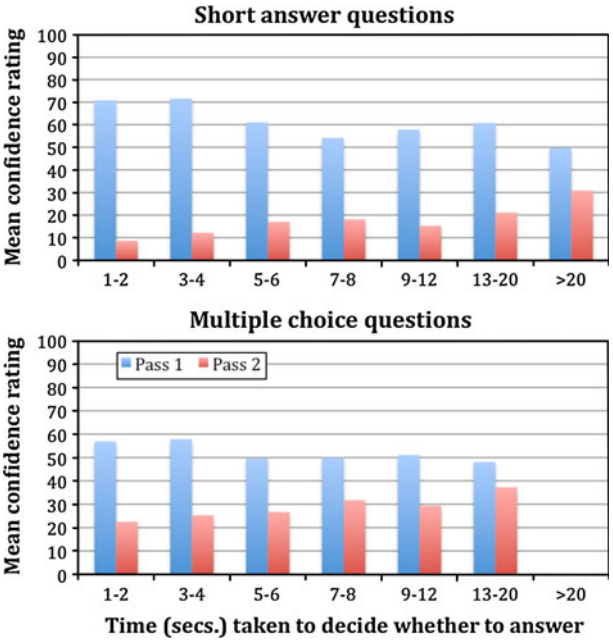

Fig. 6 Mean confidence ratings as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 2

participants' confidence in their knowledge of the domain (i.e., their predictions of performance). The mean (and range of) correlations across domain between these question-by-question confidence ratings and performance were $r = 0.47$ (0.43–0.50) in the SAQ condition and $r = 0.29$ (0.09–0.40) in the MCQ condition.

Further, confidence ratings were again found to map well onto decision times in a pattern very similar to that seen in the accuracy rates. Figure 6 replicates Fig. 5, plotting confidence ratings rather than proportion correct on the y-axis. For the SAQ condition, the slope of the line for round 1 confidence ratings was significantly negative and the slope of the line for round 2 responses was significantly positive (for both $p < .05$). For the MCQ condition the pattern of data trend in the appropriate directions. The slopes are not significantly different from zero, but they are also not significantly different from the corresponding slopes in the SAQ condition.

Learning activity: the internet search task (groups 1 and 2 together)

As seen in study 1, the proportion of questions for which participants searched revealed a consistent relationship with whether participants had earlier decided to answer or defer responding. Questions on which participants chose to defer were searched for more often (mean = 23.3%, and 14.5% for the SAQ and MCQ conditions, respectively) relative to questions participants opted to answer (mean = 6.5%, and 5.2%, respectively). These differences were statistically significant ($t > 4.5$, $p < 0.001$) in both instances. The rate of searching for question-related information was similar for both questions answered correctly and those answered incorrectly. The likelihood of searching for a response to a

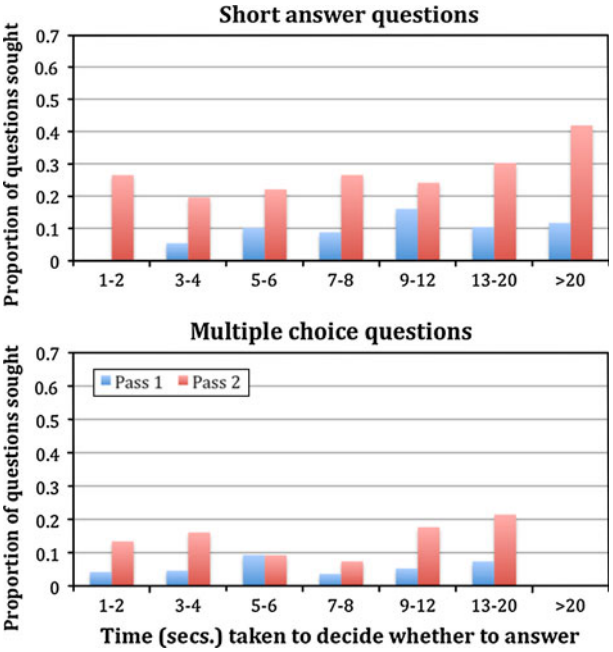

Fig. 7 Proportion of time question-relevant information was sought during a 10-min internet search task as a function of whether participant chose to answer during round 1 or to defer answering until round 2 and time taken to make the decision in study 2

question was generally greater when participants took longer to make their decision regarding whether to answer or defer as illustrated in Fig. 7.

Post-exercise self-assessments

Consistent with study 1, the self-assessment correlations post-performance were somewhat higher than those offered prior to performance. The mean (and range) of correlations across SAQ and MCQ conditions were $r = 0.39$ (0.24–0.53) and $r = 0.23$ (0.08–0.42), respectively. For ease of comparison all sets of correlations across both studies have been combined into Table 3.

Table 3 Correlations (Pearson's *r*) between overall performance score and self-assessment as a function of when and how the self-assessments were collected

Correlation between performance score and …	Study 1	Study 2 (short answer questions)	Study 2 (multiple choice questions)
…overall prediction prior to performance	0.28	0.31	0.18
…overall self-assessment collected after performance	0.66	0.39	0.23
…in the moment confidence ratings	0.81	0.47	0.29

General discussion

The intent of this pair of studies was to extend our understanding of the difference between, and the relationship between overall self-assessments of performance or ability and moment-by-moment self-monitoring of performance. In particular, we were motivated by three broad questions:

1. Was the accurate self-monitoring observed in prior work an indication of participants being consciously aware of the likelihood of responding accurately on a moment-to-moment basis?
2. Does the act of self-monitoring influence participants' self-directed information search strategies in a way that leads them to seek out information relevant to improving their performance?
3. Does the experience of self-monitoring and the opportunity to seek out information relevant to prior testing experiences alter participants' self-assessments?

This discussion will address each of these questions in turn.

1. *Was the accurate self-monitoring observed in prior work an indication of participants being consciously aware of the likelihood of responding accurately on a moment-to-moment basis?*

In prior work (Eva and Regehr 2007), two behavioural indices (deferring responding to questions on which one is less likely to be correct and taking longer to decide to make this decision for questions that were at the border of one's competence) were used as a measure of self-monitoring during performance as a mechanism to methodologically distinguish this form of self-assessment from the more traditional construction of self-assessment as an overall assessment of performance or ability. While results from that earlier work demonstrated that participants did indeed slow down when at the edges of competence and generally defer answering on items that one is objectively less likely to answer correctly, the mechanism by which this process occurred was under specified. In particular, it was not clear whether these behaviours were related to explicit confidence levels or were largely implicit and unconscious in their enactment. The current studies suggest that these phenomena are at least accessible to conscious confidence assessments, in that the confidence ratings collected on a moment-to-moment basis mimicked both deferral behaviour and response time patterns (see Figs. 2, 3, 5, 6).

That the correlations between performance and moment-by-moment confidence ratings were highly variable across study indicates that one should not assume self-monitoring to be accurate. That said, the consistency with which these correlations have been found to be more strongly related to actual performance than were participants' self-assessments of the strength of their knowledge (Table 3), further increases the likelihood that self-monitoring requires a fundamentally different cognitive process than does self-assessment. In the case of self-monitoring, the judge has many sources of information available to him that enable inferences to be drawn about the likelihood of success on a moment-to-moment basis. For example, the amount of cognitive effort engaged, as indicated by the time required to determine whether or not one knows the correct response, can indicate whether or not one's eventual response is likely to be accurate (Robinson, Johnson, and Herndon 1997). By contrast, self-assessment, defined as a more global judgment of one's strength in a given domain, requires searching one's memory (albeit perhaps unconsciously) for past events to determine not only what activities/content comprise the domain, but also what our prior rate of success has been in successfully acting within the context of those activities. In

other words, self-assessment requires accurate perception of success and also the capacity to mentally aggregate that perception over many experiences. Many areas of psychology suggest that these types of aggregate-level decisions tend to be biased by particularly memorable (i.e., available) information (Tversky and Koehler 1994; Ross and Nisbett 1991), so it is implausible to imagine that these types of decisions could yield accurate indications of ability. Further, the intuition most of us have that we can self-assess (Pronin et al. 2002) may result from our many experiences of accurate self-monitoring being translated into a belief about our ability for overall self assessment. Again, our data suggest that such a translation is inappropriate. The fact that self-monitoring may indeed be effective should not be mistaken as evidence that overall self-assessments are accurate.

Indeed, one cannot even assume that self-monitoring will always be highly accurate. While the correlations between performance and confidence ratings were consistently higher than those between performance and overall self-assessments, there was a considerable range (from a high of 0.81 in study 1 to a low of 0.29 in the MCQ version of study 2). The only difference between the SAQ version of study 2 and study 1 were the materials (i.e., questions). While preliminary in nature, finding that the effects were less pronounced in the MCQ version of study 2 relative to the SAQ questions (despite both using the same questions) suggests that it would be a mistake to presume that self-monitoring will be effective as long as judgments of the limits of one's competence are collected in the moment of the performance. By alerting respondents to the fact that response options would be presented after they decided to answer a given question we observed that respondents were almost twice as likely to choose to answer (and, therefore, half as likely to pass) relative to when they knew they would be responsible for generating their response. Further work is required to tease apart the extent to which the resulting differences in the indices of self-monitoring reported in this paper were an influence of changing the cognitive task from one of judging whether or not one could generate the response (in the free text version of the experimental design) versus adding a layer of prediction regarding the likelihood of recognizing the correct response. The latter arguably moves more towards a self-assessment from a self-monitoring judgment because it requires a prediction of the extent of one's knowledge base despite the judgment being elicited in the moment of struggle with a particular problem. Perhaps it is not the temporal nature of the judgments that enable accurate self-monitoring, but the constraint to base judgments on the readiness with which the actual problem solution can be generated rather than an abstract judgment of the likelihood of success. Indeed, the Feeling-of-Knowing literature suggests that the instruction used in the MCQ version of study 2 may be particularly effective at eliciting inaccurate Feeling-of-Knowing judgments as Widner and Smith (1996) have shown that asking participants to indicate what they believe they "know" elicits poorer accuracy in those judgments relative to asking them outright if they are likely to "recognize" the correct answer. The authors conclude that Feeling-of-Knowing judgments are better when they incorporate task-relevant information into the instructions, a phenomenon that warrants further study with respect to its implications for self-regulated learning in health professional education.

2. *Does the act of self-monitoring influence participants' self-directed information search strategies in a way that leads them to seek out information relevant to improving their performance?*

It is intriguing that participants in both of the studies in this paper, when not given any instruction regarding how to spend their time using the internet, searched for information relevant to the questions they had just been presented at a rate equal to the group of

participants who were asked explicitly to use their internet access for that purpose. Optimistically, this could provide an indication that it is a natural response to try and improve when given clear guidance about the limits of one's knowledge and the resources to do so, especially given that there was no reward associated with this activity in this context beyond the personal satisfaction of confirming one's responses or learning information that was previously unknown or misremembered. Less optimistically, the rate of questions on which participants searched was relatively low in both studies, reaching a maximum of 23% in this brief posttest situation. The simple act of prompting people to explicitly evaluate their confidence in their responses did not alter the rate at which participants sought knowledge relevant to the challenges put before them. However, opting to defer giving a response and the length of time taken to make that decision were both influential, as illustrated in Figs. 4 and 7.

Further research is required to fully explore the implications of these findings, but the results suggest that there may be motivational benefits to tailoring educational efforts to the limits of a students' ability. This notion is consistent with Bjork's theory of desirable difficulties (1999), which suggests learners must be placed in situations that elicit errors (and often make learning seem harder/less successful) for learning to be optimized. In situations in which participants could quickly decide that they knew or did not know the answer they were less inclined to spend time searching for information than they were in situations that prompted them to think for a longer period. The lack of correlation with accuracy suggests that students were not falling prey to confirmation bias, but rather, were searching for information they inferred should be within their reach even if that grasp came up short in the specific instance.

3. *Does the experience of self-monitoring and the opportunity to seek out information relevant to prior testing experiences alter participants' self-assessments?*

Above, we highlighted reasons that self-assessments, defined as global judgments of ability, come to be inaccurate. The post-test self-assessments provide further hints about why this might be the case. While the three correlations between self-assessments collected at the end of the study and performance rose relative to the pre-test self-assessments they remained lower than the correlations between performance and moment-to-moment confidence ratings even though participants had the opportunity to view the questions and search the internet for confirmation/refutation of their responses. Using Bjork's model, the presentation of specific test questions can be thought of as inducing a stimulus that provides the individual with information regarding their ability beyond the introspective judgments that arise from searching memory (explicitly or implicitly) for the answer to the question of how well have I performed in this area in the past? On one hand, the improved correlations provide further evidence that participants were aware, in the moment, of the strength of their performance as they were able to use this information to alter their perceptions. At the same time, however, that the correlations did not achieve the same magnitude as the moment-by-moment confidence ratings suggests that even within this relatively constrained context, in which memory for the experience is as active as it will ever be, the aggregation of this information was imperfect. We would hypothesize (but did not examine within these studies) that the correlation between post-performance ratings and actual performance will decline with increasing time between performance and self-assessment, a pattern that would provide further support for the notion that poor self-assessment is predominantly a reflection of the non-literal and imperfect nature of human memory.

One could alternatively argue that the pre-test correlations were poor simply because participants did not know precisely which questions would be asked, but that explanation would not account for the difference in correlations with performance observed for post-test self-assessments and self-monitoring. At the same time, that argument reinforces the more central point that self-assessment alone can not be relied upon as a mechanism for judging one's overall skill level as every situation is inevitably a narrow sample of a broader domain that cannot be fully anticipated until the individual is embedded within the experience. That participants' self-perceptions were not fully influenced by their recent experience is consistent with the arguments put forward in the social psychology literature that it is adaptive to maintain an optimistic outlook on one's ability (see Eva and Regehr 2005; Kluger and van Dijk 2010). That is, empirical findings from this literature suggest that individuals tend to perform better when they expect to perform better, thereby indicating that not only can we not create good self-assessors (thanks to the cognitive limits inherent in aggregation as (described above) and context specificity), but perhaps that we should not even try (because lowering a poor performer's self-assessment to meet their actual performance may be counter-productive, lowering the likelihood of reaching a higher level over time).

Rather, the goal of focusing upon self-assessment should be to use understanding of this construct as a means to the end of improving practice. People often express concern about poor self-assessment as a way of drawing attention to (and overcoming) issues of patient safety. Patient safety is undeniably an issue that must take on top priority, but if the findings reported in this paper generalize to clinical contexts, then we can have some confidence that individuals are self-limiting their activities more appropriately than the historical self-assessment literature would lead one to believe, even if they are generally overconfident in their abilities. It also suggests that we should take advantage of and determine if we can extend these abilities for self-monitoring by focusing our efforts on ensuring that individuals learn to recognize the cues that they should perhaps slow down (cf. Moulton et al. 2007) and look up the information they need rather than worrying about whether or not they are able to translate such experiences into broader classifications of their proficiency as a clinician (Lee 2010). Finally, we would note that whether or not there are individual differences in the capacity to self-monitor and whether or not such capacity can be taught if there are those who lack it remain to be seen. Indeed, these are the key limitations of the work presented in this report as the extent to which these materials generalize to a clinical context, to domains of greater perceived urgency or relevance relative to general knowledge of trivia, or to domains of competence beyond declarative knowledge have not yet been tested. Having used different materials across study and different response formats we are confident in the robustness of the findings (again with the caveat that the effects were generally smaller in the MCQ format), but what other contextual variables and personal factors influence that robustness remain a subject for future study. Even so, the difference in implications that are drawn from focusing on self-monitoring (i.e., the need to create situations for learners and practitioners to experience the limits of their competence in the presence of feedback and improvement strategies tailored to those experiences) rather than self-assessment (i.e., the notion that we can improve practice and education through improving the capacity of individuals to assess their own strengths and weaknesses in an effort to self-identify performance improvement strategies) emphasizes the importance of carefully distinguishing between what appear to be very different cognitive processes with very different potential given the literature that has accumulated to date.

Acknowledgments The authors would like to thank Aurora Albertina, Katie Fracalanza, Betty Howey, and Anne van Koeverden for their assistance with data collection relevant to this project.

References

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In: *Attention and performance XVII: Cognitive regulation of performance. Interaction of theory and application*. Cambridge, MA: MIT Press.

Boud, D. (1995). *Enhancing learning through self assessment*. London: Kogan Page.

Colliver, J. A., Verhulst, S. J., & Barrows, H. S. (2005). Self-assessment in medical practice: A further concern about the conventional research paradigm. *Teaching and Learning in Medicine, 17*, 200–201.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with served measures of competence: A systematic review. *JAMA, 288*, 1057–1060.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69–106.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine, 80*(10 Suppl), S46–S54.

Eva, K. W., & Regehr, G. (2007). Knowing when to look it up: A new conception of self-assessment ability. *Academic Medicine, 82*(10 Suppl.), S81–S84.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*, 762–769.

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine, 76*(10 suppl.), S87–S89.

Kluger, A., & van Dijk, D. (2010). Feedback, the various tasks of the physician, and the feedforward alternative. *Medical Education*, *44*, 1166–1174.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9*, 149–171.

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*, 1121–1134.

Lee, L. (2010). Clinical situations of uncertainty and access to resources: A study of community family physicians. Dissertation submitted to the Faculty of Graduate Studies, The University of Western Ontario, London, Ontario, Canada.

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

Nelson, T. O., & Narens, L. (1980). Norms of 300 general information questions: Accuracy of recall, latency of recall, and feeling of knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19*, 338–368.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology, 125*, 1063–1068.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*, 369–381.

Regehr, G., & Mylopoulos, M. (2008). Maintaining competence in the field: Learning about practice, through practice, in practice. *Journal of Continuing Education in the Health Professions, 28*, S19–S23.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessment of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416–425.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill, Inc.

Sargeant, J., Armson, H., Chesluk, B., Dornan, T., Eva, K., Holmboe, E., Lockyer, J., Loney, E., Mann, K., & van der Vleuten, C. (2010). Processes and dimensions of informed self-assessment. *Academic Medicine*. Epub Apr 2 2010.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.
Widner, R. L., & Smith, S. M. (1996). Feeling of knowing judgments from the subject's perspective. *American Journal of Psychology, 109*, 373–387.