Accuracy of self-monitoring: does experience, ability or case difficulty matter?

Wolf E Hautz,[1,2] Sebastian Schubert,[3,4] Stefan K Schauber,[2,5] Olga Kunina-Habenicht,[6] Stefanie C Hautz,[1] Juliane E Kämmer[3,7] & Kevin W Eva[8]

CONTEXT The ability to self-monitor one's performance in clinical settings is a critical determinant of safe and effective practice. Various studies have shown this form of self-regulation to be more trustworthy than aggregate judgements (i.e. self-assessments) of one's capacity in a given domain. However, little is known regarding what cues inform learners' self-monitoring, which limits an informed exploration of interventions that might facilitate improvements in self-monitoring capacity. The purpose of this study is to understand the influence of characteristics of the individual (e.g. ability) and characteristics of the problem (e.g. case difficulty) on the accuracy of self-monitoring by medical students.

METHODS In a cross-sectional study, 283 medical students from 5 years of study completed a computer-based clinical reasoning exercise. Confidence ratings were collected after completing each of six cases and the accuracy of self-monitoring was considered to be a function of confidence when the eventual answer was correct relative to when the eventual answer was incorrect. The magnitude of that difference was then explored as a function of year of seniority, gender, case difficulty and overall aptitude.

RESULTS Students demonstrated accurate self-monitoring by virtue of giving higher confidence ratings (57.3%) and taking a shorter time to work through cases (25.6 seconds) when their answers were correct relative to when they were wrong (41.8% and 52.0 seconds, respectively; $p< 0.001$ and $d > 0.5$ in both instances). Self-monitoring indices were related to student seniority and case difficulty, but not to overall ability or student gender.

CONCLUSIONS This study suggests that the accuracy of self-monitoring is context specific, being heavily influenced by the struggles students experience with a particular case rather than reflecting a generic ability to know when one is right or wrong. That said, the apparent capacity to self-monitor increases developmentally because increasing experience provides a greater likelihood of success with presented problems.

Medical Education 2019: 53: 735–744
doi: 10.1111/medu.13801

[1]Department of Emergency Medicine, Inselspital University Hospital Bern, Bern, Switzerland
[2]Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Oslo, Norway
[3]AG Progresstest Medizin, Charité Universitätsmedizin Berlin, Berlin, Germany
[4]Medizinische Hochschule Brandenburg, Neuruppin, Germany
[5]Centre for Health Sciences Education, Faculty of Medicine, University of Oslo, Oslo, Norway
[6]Institute of Educational Research Methods, University of Education Karlsruhe, Karlsruhe, Germany
[7]Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Berlin, Germany
[8]Centre for Health Education Scholarship, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence: Wolf E Hautz, Department of Emergency Medicine, Inselspital University Hospital Bern, Freiburgstrasse 16c, 3010 Bern, Switzerland.
Tel: 00 41 31 632 4587;
E-mail: wolf.hautz@insel.ch

INTRODUCTION

Self-assessment is of central importance to professional self-regulation. Thus, it has been incorporated into many models of lifelong learning[1,2] and become a focus of many curricular interventions in the health professions.[3,4] For decades, this has been a source of concern given the repeatedly observed disconnects between confidence and competence,[1,5] with overconfidence receiving particular attention.[6] As a result, considerable study has been conducted to refine our understanding of the way in which people judge their own abilities. Consequently, the field has moved from historical 'guess your grade' conceptions of self-assessment[7] to more differentiated models[5,8] and terminology.[9]

In contributing to this movement, Eva and Regehr proposed a distinction between summative self-assessment as a generalised, context-free estimate of one's competence in a given field and self-monitoring as an 'in-the-moment' judgement of the alignment between one's ability and the problem at hand.[10] Self-monitoring prompts surgeons 'to slow down when they should',[11] leads physicians to gather more information when they are having difficulty,[12] and helps students to 'know when to look it up'.[9] As a simple analogy for guiding learning, Eva and Regehr suggested that:

> most people are prompted to open a dictionary as a result of encountering a word for which they are uncertain of the meaning rather than out of a broader assessment that their vocabulary could be improved.[10]

Subsequent empirical studies of self-monitoring have suggested that it can be highly accurate, with moment-to-moment measures of confidence generally correlating better with performance than predictions or postdictions of one's overall ability.[8,9] Furthermore, medical and psychology students, whether directed to or not, have been observed to use self-monitoring to guide learning efforts, with particular attention paid to moments when self-monitoring and performance are misaligned.[8,13] Data collected during a national licencing examination suggested that the capacity to self-monitor was greater in high performing individuals relative to low performers, with differences in various self-monitoring indices reliably differentiating between individuals.[14] This latter finding suggests interpersonal differences in the way in which self-monitoring takes place, which offers hope that interventions can be developed to improve the effectiveness of self-monitoring. To develop such interventions requires an understanding of what cues are used to guide and what cues should be used to guide self-monitoring (i.e. which are most predictive of performance; cf.[15]).

The notion that 'cues' are used to guide self-monitoring is drawn from the cue-utilisation framework put forward by Asher Koriat.[16] His work demonstrated that humans cannot directly judge the quality of their cognitive operations, but utilise cues such as the perceived fluency of text to monitor whether or not they have understood and learned the material. Some cues, such as the speed with which a solution comes to mind, can be 'diagnostic' of actual performance (i.e. provide meaningful information regarding the likelihood of success),[14] whereas others can be misleading. In reviewing this literature, de Bruin et al. argue that we can improve self-study and clinical reasoning practice by expanding the evidence regarding the characteristics of cues that are effectively used by medical students and trainees to guide their self-monitoring. Without such information, it is impossible to know how to intervene for the sake of helping those whose efforts at regulating their learning at any given moment might be guided by the wrong cues.[15]

An important step in this regard comes from the work of Pusic et al.,[17] who used a computer-based simulation to teach medical students radiograph interpretation. On a series of 50 cases, learners were asked to classify each radiograph as normal or fractured and to note their certainty with 'definitely' or 'probably'. On subsequent post-tests, diagnostic accuracy was associated with choosing 'definitely' over 'probably' but the relationship declined after a 2-week delay. The cause of the interaction was that performance declined after 2 weeks whereas certainty in one's diagnoses did not. These results reinforce previous research suggesting that self-monitoring can be measured and extend it by suggesting that its accuracy is likely to be context dependent. If true, it suggests that the cues that are most determinant of the effectiveness of self-monitoring processes may be case (i.e. context) specific and that efforts to intervene would be best directed at considering what aspects of individual cases are most diagnostic.

Our aim, therefore, in the current study is to deepen the understanding of which factors relate to accurate self-monitoring by examining its relationship with factors related to the individual practitioner (years of training, gender, age and clinical ability) and factors related to the clinical case (its difficulty, the amount of time required and the number of tests pursued). To do so, we use confidence ratings and performance measures collected in a simulated clinical reasoning task with a cross-sectional sample of trainees to address the following research questions.

1 Are medical students more confident in their diagnostic decision making (and their clinical reasoning) when providing a correct diagnosis than when providing an incorrect diagnosis?
2 Are they slower when there is a greater risk of being incorrect?
3 Is awareness of the risk of being wrong associated with the amount of information medical students collect about a clinical case?
4 Are these tendencies associated with differences in gender, age, one's clinical ability or case difficulty?
5 How does the ability to self-monitor evolve over the course of medical school?

Our study extends the prior work described above in a variety of ways. First, this is the first study of its type to include a complex clinical reasoning task that, although simulated, better represents the scope of clinical practice than the declarative knowledge or visual interpretation tasks that have been used to examine self-monitoring accuracy previously. Second, this is the first study of self-monitoring to examine the phenomenon in individuals from a variety of training levels. Finally, this is the first study of self-monitoring for which external data regarding participants' knowledge are available. This feature allows a determination of whether the relationship between self-monitoring ability and performance observed previously[14] is task specific or related to one's ability more generally.

METHODS

Participants and setting

We recruited participants from all years of study in the reformed medical curriculum (RMC) at the Faculty of Medicine, Charité – Universitätsmedizin Berlin. The RMC offers a 5-year programme that used the assessing clinical reasoning (ASCLIRE) instrument as part of their formal assessment strategy.[18] ASCLIRE is a computer-based test for the assessment of clinical reasoning that aims to capture students' thinking during the clinical decision-making process in the manner described below. Participation in the test was mandatory and students received detailed feedback on their performance as part of their education. Consent to inclusion of their data in the study was voluntary. Students who opted to participate gave informed consent. The Curriculum Committee and the Examination Committee of the RMC, as well as the bureau of data privacy and the Institutional Review Board at Charité, granted the study their approval (EA1/170/09).

Materials

Using ASCLIRE,[18] we presented students with six patient cases of acute or subacute dyspnoea. Previous validation work performed on these cases has demonstrated the test procedure to: (i) result in scores that have good psychometric properties when administered to the population enrolled in this study; (ii) differentiate between trainees in different years of study; (iii) differentiate between experts and trainees, and (iv) have three separable latent factors: diagnostic accuracy, decision time and choice of relevant diagnostic information.[18] These factors are used here to explore the development of self-monitoring.

Procedure

We administered the test in a computer laboratory over four consecutive days with several sessions per day. Participants self-selected their testing date and time. One week prior, all participants had the opportunity to attend a plenary presentation outlining the test's purpose and providing a brief demonstration. A video-recording was made available to those who did not attend. At the beginning of each study session, we asked participants to watch the study instructions on a computer screen (Appendix Figure A1) and to complete a practice case. We did not evaluate their work on this case, but used it to familiarise participants with the computer programme.

The test cases were then presented to each participant in random sequence using Inquisit 2 (MILLISECOND SOFTWARE, Seattle, WA, USA). Participants worked through each case in the following steps (Appendix Figures A1–A3).

1 A written description of the setting and a short video of a standardised patient displaying particular symptoms was displayed.
2 Participants could select as many diagnostic procedures as they deemed important to perform,

in any order they wished, from a predefined list of 36 procedures. This presentation mimics the real world on a shortened timeline by requiring participants to determine what information they require rather than presenting them with all possible information in a fixed sequence. After clicking on a procedure, the finding was revealed using an appropriate modality: as text (e.g. pulse rate), image (e.g. electrocardiogram or chest X-ray) or audio (e.g. heart sounds). Most findings required the student to make their own interpretation. The interpretation of a radiologist, however, was provided for some imaging studies (e.g. ultrasound examinations and computerised tomography scans) that are technically difficult to present within the testing software.

3 Participants could provide their diagnosis at any time by clicking to the next screen and selecting from a list of 20 possible diagnoses. The maximum time allowed for each case was limited to 10 minutes, after which the software prompted participants for a diagnosis.

4 After completion of each case, participants were asked to record their confidence in the diagnosis they assigned and their confidence in having performed or ordered the pertinent diagnostic procedures for that case prior to moving onto the next case presentation. The exact wording (in German) read 'What is your appraisal of the likelihood (in percent) that you have made the correct diagnosis/ordered all indicated diagnostic procedures?'. We collected these ratings using a scale ranging from 0.0% to 100.0% with 10.0% increments.

Measures

Data captured by the software included: (i) the sequence in which cases were presented; (ii) the diagnosis the student provided; (iii) the procedures requested; (iv) the time each student spent on each component of the case, and (v) the confidence ratings the students assigned.

We used a dichotomous scoring key for each diagnosis (1 = correct, 0 = incorrect). The procedures requested were coded into 'relevant' and 'non-relevant' based on whether or not 50.0% of an expert sample selected that procedure for that particular case during test validation.[18] To create individualised indices of the extent to which participants were able to self-monitor their performance (i.e. the extent to which behaviour or confidence differed as a function of the likelihood of being correct), we calculated the difference in both confidence and response time between cases for which participants provided an incorrect diagnosis and those with a correct diagnosis. Creating such individualised measures avoids the problems of analysing 'self'-monitoring using group-level summary statistics[19] and replicates the approach used in preceding studies.[8,14,17]

Additional data with which the experimental measures could be matched included each participants' year of training, gender, age and performance on a progress test (i.e. a multiple-choice test that is used formatively and assesses the knowledge that a physician needs on his or her first day after graduation).[20] We used the percentage of correct answers in this test as a measure of clinical knowledge.

Analysis

For each dependent variable, an outlier check was performed by identifying scores that were 3 standard deviations (SDs) above or below the overall average. At most, 12 (0.7%) of 1698 observations within any given variable were found to be outliers and their removal had no effect on the conclusions drawn from any analyses. As a result, the data were analysed and reported with potential outliers included for the sake of completeness.

Paired samples *t*-tests and repeated measures analyses of variance (ANOVAs) were used to compare confidence and response times across cases for which participants' diagnoses were correct versus those that were incorrect to explore the effectiveness of self-monitoring. By doing so, we excluded individuals who answered all (n = 16) or no (n = 9) cases correctly, thereby making all comparisons within-subject contrasts and avoiding the risk of artificial inflation or deflation that might occur from having different individuals represented in the 'correct' versus 'incorrect' conditions.

Mean scores, 95% confidence intervals (CIs) and effect sizes (Cohen's *d*) are included. Finally, Pearson's correlations are used to examine the relationship between the self-monitoring indices created and continuous variables of age, progress test performance, and performance on the ASCLIRE.

RESULTS

A total of n = 283 students participated in our study, which constitutes 89.8% of the 315 eligible students; 283 students times six cases results in a total of 1698 possible observations. Of these, a diagnosis was not provided in 12 cases (0.7%) and diagnostic confidence ratings were not provided in 14 cases (0.8%). All remaining cases were completed within the 10 minutes given (mean = 186.8 seconds; SD = 71.9). A total of 193 (68.2%) of the participants were female, and the average age of all participants was 24.5 years (SD = 5.3). Diagnostic accuracy increased from 37.2% (SD = 21.0) in year 1 to 71.1% (SD = 18.9) in year 5. As mentioned in the methods, the data from 25 participants were excluded because they answered all questions correctly or incorrectly, thereby precluding calculation of self-monitoring indices for those individuals.

Accuracy of self-monitoring

Are medical students more confident in their diagnostic decision making (and their clinical reasoning) when providing a correct diagnosis than when providing an incorrect diagnosis?

Students were, on average, more confident in their diagnosis when it was accurate (mean confidence = 57.3%, 95% CI = 54.2–60.3%) relative to when it was not (mean = 41.8, 95% CI = 39.1–44.6; $F(1253)$ = 196, p< 0.001; d = 0.63). Similarly, their confidence in the procedures they ordered was larger when their diagnosis was accurate (mean = 46.1%, 95% CI = 43.5–48.8%) relative to when it was not (mean = 38.9, 95% CI = 36.4–41.5; $F(1253)$ = 66.2, p< 0.001, d = 0.33). A very high correlation (r = 0.90) was observed between confidence expressed in one's diagnosis and confidence expressed in one's procedure requests. Given that further analyses performed on each pair of variables yielded the same conclusions, we report only confidence in one's diagnosis below.

Are medical students slower when there is a greater risk of being incorrect?

Participants were twice as fast to decide what procedures to request when their diagnosis was accurate (mean = 25.6 seconds, 95% CI = 21.7–29.5 seconds) relative to when it was inaccurate (mean = 52.0 seconds, 95% CI = 45.1–59.0 seconds; $F(1253)$ = 31.2, p< 0.001, d = 0.57).

Overall, students completed cases more quickly when their diagnosis was accurate (mean = 179.5 seconds, 95% CI = 173.6–185.4 seconds) relative to when it was not (mean = 194.3 seconds, 95% CI = 186.9–201.7 seconds; $F(1249)$ = 20.9, p< 0.001, d = 0.27).

A moderately high correlation was observed between total time required and time spent on the procedure request screen (r = 0.66). Given that further analyses performed on each pair of variables yielded the same conclusions, we report only total time below.

Associations with self-monitoring accuracy

Is awareness of the risk of being wrong associated with the amount of information medical students collect about a clinical case?

Despite participants taking longer to choose procedures when they were less confident and less likely to be correct, their confidence was only weakly (albeit statistically) correlated with the number of procedures requested (r = 0.14, p< 0.05 for relevant procedures; r = −0.22, p< 0.01 for irrelevant procedures). Similarly, procedure requests revealed a statistically significant, but practically inconsequential, tendency to order fewer irrelevant tests when students' diagnoses were accurate relative to when they were inaccurate. No such difference was observed for the number of relevant procedures requested (Table 1).

Are self-monitoring indices associated with differences in gender, age, one's clinical knowledge or case difficulty?

Men generally gave higher confidence scores (mean = 54.9, 95% CI = 50.6–59.2) than women (mean = 44.2, 95% CI = 41.2–47.3; $F(1252)$ = 15.8, p< 0.001, d = 0.44), but self-monitoring accuracy (the difference in confidence when diagnoses were correct relative to when they were incorrect) did not vary as a function of gender ($F(1252)$ = 0.5, p> 0.45). Men completed their cases (mean = 185.7 seconds, 95% CI = 175.7–195.8) in the same time as women (mean = 186.8 seconds, 95% CI = 179.7–194.0; $F(1252)$ = 0.03, p> 0.85) and the difference in time taken when diagnoses were correct relative to when they were incorrect did not vary as a function of gender ($F(1252)$ = 0.0, p> 0.95). Men requested the same number of procedures (mean = 20.2, 95% CI = 19.1–21.4) as did women (mean = 19.7, 95% CI = 18.8–20.5; $F(1252)$ = 0.6, p> 0.4) and gender did not interact with the number of procedures ordered when diagnoses were correct relative to when they were incorrect ($F(1252)$ = 1.7, p> 0.15).

None of the self-monitoring indices created correlated substantially with age, performance on

Table 1 Number of procedures (with 95% confidence intervals) requested per case

	When diagnosis correct	**When diagnosis incorrect**	**Statistics**
Relevant procedures	12.0 (11.7–12.3)	12.1 (11.9–12.4)	$F(1253) = 0.8$, $p = 0.38$
Irrelevant procedures	6.5 (6.1–7.0)	7.0 (6.5–7.4)	$F(1253) = 5.2$, $p = 0.03$, $d = 0.14$
Total procedures	18.7 (18.1–19.3)	19.0 (18.3–19.6)	$F(1253) = 1.9$, $p = 0.16$

the progress test or overall diagnostic accuracy within the ASCLIRE test (all $r < 0.25$).

To further explore the extent to which participants' capacity to monitor their likelihood of success is influenced by the specifics of the situation encountered, we examined the relationship between each of the outcomes described above (confidence ratings, time taken and procedures ordered) in relation to case difficulty (defined based on the overall diagnostic accuracy aggregated across all participants). Table 2 reveals that average accuracy correlated with the difference in confidence ratings between cases for which participants were correct and cases for which their diagnosis was incorrect (delta) ($r = 0.57$). The trend largely arose because the most difficult case showed little evidence of accurate self-monitoring. Similarly, the correlation between accuracy and the difference in time taken was $r = -0.54$, largely because of the most difficult case. Finally, the correlation between accuracy and the difference in total number of procedures ordered was small ($r = -0.23$).

How does the ability to self-monitor evolve over the course of medical school?

Mixed-design ANOVA performed on the confidence ratings with accurate diagnosis versus inaccurate diagnosis as a repeated measure and year of training as a between-subjects measure revealed that absolute levels of confidence increased with seniority ($F(4249) = 16.5$, $p< 0.001$) and that the gap between confidence when accurate and confidence when inaccurate differed with year of training ($F(4249) = 7.4$, $p< 0.001$). The self-monitoring indices (i.e. delta) illustrated in Table 3 demonstrate that self-monitoring accuracy tended to increase as students became more senior. It is noteworthy, however, that confidence increased with seniority regardless of diagnostic accuracy, with fifth-year students being more confident when they were inaccurate than were first-year students when they were accurate.

Mixed-design ANOVA with time spent per case as a dependent variable, accurate diagnosis versus inaccurate diagnosis as a repeated measure and year of training as a between-subjects measure revealed that the absolute amount of time taken did not significantly change with seniority ($F(4249) = 0.9$, $p> 0.45$), nor did the time difference between accurately and inaccurately diagnosed cases interact with seniority ($F(4249) = 0.3$, $p> 0.8$) (Table 3).

Finally, mixed-design ANOVA performed on the number of procedures ordered with accurate versus inaccurate diagnosis as a repeated measure and year of training as a between-subjects measure revealed that the total number of procedures ordered did not significantly change with seniority ($F(4249) = 0.5$, $p> 0.75$) nor did seniority interact with diagnostic accuracy ($F(4249) = 0.7$, $p> 0.6$) (Table 3).

DISCUSSION

The capacity to monitor the likelihood that one's impressions of a clinical case are correct is a critical competence required of all health care professionals. With respect to clinical reasoning, one's confidence in one's diagnostic suppositions is likely to have considerable impact on the choice of management options and one's willingness to take action when required. Generating an absolute conclusion is not always the goal of clinical reasoning,[21] but the more confident one is in one's hypotheses, the less likely it is that other important possibilities will be given due consideration. Also, the less confident one is, the greater the risk of becoming paralysed by indecision. As such, it is no wonder that de Bruin et al.[15] have emphasised the importance of determining how to help trainees and practitioners self-monitor accurately. In framing the challenge, they review the psychology and educational literatures and present a cue utilisation framework that emphasises the need to develop interventions that increase the use of cues that are predictive of achievement or performance rather than those that are likely to be misleading. For any such intervention to be effective, it will have to help practitioners overcome the use of non-predictive cues that are used spontaneously (i.e. without intervention). This study sought to shed light on what cues are related to accurate self-monitoring when learners are left to their own devices to evaluate their likelihood of success with a clinical reasoning task.

We observed a strong relationship between confidence ratings, collected in the moment after the experience of working on each case, and diagnostic accuracy (Table 2). This finding reinforces claims that individuals have a greater sense of their likelihood of success with particular problems than the general literature on summative self-assessment as an aggregate indication of one's ability has led us to believe. It is consistent with the literature on self-monitoring by replicating results previously reported in a variety of articles.[8,11,14,22] However, most of the research concerning factors that affect self-monitoring ability has been conducted using tests of declarative[8,23] or procedural[24,25] knowledge, leading Eva and Regehr to note 'The extent to which these [findings] generalize to a clinical context, to domains of greater perceived urgency or relevance [... has] not yet been tested'.[8] In testing whether or not such generalisation occurs, we were also able to observe that the feeling of uncertainty that arises when

Table 2 *Confidence, time taken (in seconds) and number of procedures ordered by decreasing case difficulty (average accuracy)*

		Diagnostic confidence				Time taken				Procedures ordered			
Case	Average accuracy (%)	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value
1	36.1	42.5	47.9	5.4	0.09	182.4	189.3	6.9	0.44	18.5	18.6	0.1	0.92
6	41.3	35.6	53.6	18.0	<0.001	181.7	165.9	−15.7	0.08	22.2	21.2	−1.1	0.19
5	45.0	34.0	66.8	32.8	<0.001	201.5	213.3	11.9	0.13	18.3	18.5	0.2	0.72
3	64.9	37.6	58.6	21.0	<0.001	169.5	177.5	8.0	0.31	19.0	20.1	1.1	0.15
2	69.3	38.4	60.5	22.1	<0.001	205.3	184.0	−21.2	0.03	19.2	18.3	−0.9	0.24
4	77.4	29.2	62.5	33.3	<0.001	198.4	176.7	−21.7	0.04	22.8	21.6	−1.2	0.24

* Correct minus incorrect of the two preceding columns. See text for details.

Table 3 *Confidence, time taken (in seconds) and number of procedures ordered by year of study*

		Diagnostic confidence				Time taken				Procedures ordered			
Year	Overall accuracy	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value	When incorrect	When correct	Delta*	Post-hoc *t*-test p-value
1	37.2	29.3	37.5	8.2	<0.001	186.1	174.3	−11.8	0.06	20.6	20.7	0.1	0.89
2	48.6	37.1	55.3	18.2	<0.001	185.8	174.6	−11.2	0.09	19.7	19.1	−0.6	0.20
3	56.7	41.6	52.0	10.4	<0.001	196.6	183.1	−13.4	0.06	19.6	19.2	−0.4	0.42
4	69.4	42.6	68.6	26.0	<0.001	202.1	179.9	−22.2	0.02	20.3	19.4	−0.9	0.06
5	71.0	51.7	70.1	18.4	<0.001	201.0	185.6	−15.5	0.04	20.3	19.6	−0.7	0.14

* Correct minus incorrect of the two preceding columns. See text for details.

participants had a greater likelihood of being wrong was associated with taking more time to work on the case, but not with the amount of information collected (Tables 1 and 2). This suggests that the time required to think about a case is a better cue (i.e. has higher diagnosticity) regarding clinicians' likelihood of success than is their efforts to deliberately sample as many sources of information as possible.

Theoretically, the relationship observed between accuracy and time has implications for the cue-utilisation framework developed by Koriat. When exploring judgements of learning (i.e. assessments of the success of one's efforts to learn particular material), Koriat and colleagues have noted an important distinction between being 'data driven' (i.e. guided in judgements by cues related to the experience of the task, such as ease of memorisation) and being 'goal driven' (i.e. guided by the relative importance of learning particular material).[16] In the former case, judgements of learning decrease with study time: the longer it takes to study material the less confident we are that it has been learned. In the latter, judgements of learning increase with study time: the longer one has attended to important information the more confident people feel that it has been learned. The relationship observed here suggests that confidence derived when self-monitoring one's performance increases the more rapidly one can generate a diagnosis. That could suggest that clinical trainees are using 'data driven cues' to self-monitor their performance or it could suggest that the 'data driven' versus 'goal driven' distinction is less applicable to judgements of performance than it is to judgements of learning. Whether regulation of effort in clinical reasoning tasks draws upon cues differently to regulation of effort in learning how to reason clinically remains an open question that might be explored by varying the stakes or incentive schemes presented in the context of performance-based tasks.

Another unique and particularly interesting finding emerged when self-monitoring was examined in relation to year of study. Although the gap between confidence ratings reported when students were correct and those reported when students were incorrect increased with year of training, seniority appears to have a more substantial effect on the confidence one feels than it does on one's capacity to self-monitor. More senior students were more likely to be accurate in their diagnostic reasoning, but even so, Year-5 students were as confident in their inaccurate diagnoses as were Year-1 students in their accurate diagnoses (Table 3). Various studies have suggested that experience with a type of problem is a particularly salient cue to learners regarding their capacity to solve that problem.[4,26] This might help to explain Pusic et al.'s[17] discovery that the relationship between confidence and accuracy declines with the passage of time after learning if it indicates that the cues one draws upon to monitor performance during learning (e.g. the amount of time required to struggle with the material) differ from those that are used when separated in time from learning (e.g. the knowledge of having experienced similar problems before). In this regard, it is also compelling to note that the improvement of self-monitoring that came with seniority was unrelated to the number of procedures one felt compelled to order (Table 3), whereas time on task was a better indicator of self-monitoring accuracy.

Participants' gender was predictive of confidence, but neither that nor overall academic ability was related to the 'accuracy of self-monitoring' indices created. In other words, being male corresponded with giving higher confidence ratings on average, but both male and female participants were equally likely to give higher confidence ratings when they were correct relative to when they were not. This finding reinforces the importance of within-person comparisons in research on self-monitoring, because confidence ratings are highly variable between persons and appear to systematically vary by gender. As an implication for practice, however, the lack of difference in self-monitoring indices suggests that we do need to consider interventions that focus more upon the cues present within any given case than on the characteristics of the individual or the demographic groups to which they belong.

That self-monitoring ability was related to the difficulty of particular cases is reminiscent of Kruger and Dunning's seminal work indicating that self-assessment is particularly poor in domains in which we don't have ability, but extends it further by suggesting that 'domain' might need to be more narrowly defined than it has been in the past.[27] That is, self-monitoring accuracy does not appear to simply be a function of increased skill overall (as evidenced by the lack of correlation with progress test performance), but rather, is driven more specifically by the capacity to work effectively through a specific case.

Limitations

Several limitations to this study must be considered. First, we report a cross-sectional study that did not longitudinally follow individual students, thereby allowing the potential for some of the 'year of training' differences to be illustrative of cohort effects rather than developmental differences. Second, although we took great care to mimic the process of a patient presenting symptoms to a physician who, after data collection, makes a conclusion regarding possible diagnosis, our study still employed an artificial laboratory setting, neglecting aspects such as collegial advice, a nurse's triage or interaction with other professionals, all of which might have additional effects on the alignment between confidence and diagnostic accuracy in the clinical context.[28,29] Finally, being induced to provide explicit confidence ratings may have led people to deviate from their normal thought processes about the cases presented, leading them to be more deliberate and analytic. Our prior work, which manipulated that very factor, suggested that requesting confidence ratings is not detrimental to the valid measurement of self-monitoring,[8] but that work was conducted with less complex materials, creating the risk that such findings do not generalise to this context.

CONCLUSIONS

This study's findings add to the field's understanding of self-monitoring by suggesting that its accuracy is context specific. Overall ability was not associated with the self-monitoring indices generated, but the difficulty of particular cases was. As students gained more experience, they showed a greater gap between confidence ratings reported for cases on which their conclusions were correct relative to cases on which their conclusions were incorrect, but they were more likely to be confident when they were wrong relative to more junior students who were right. So, although the capacity to self-monitor increases developmentally as a result of students' experience, that might be a by-product of experience increasing the likelihood of success on clinical reasoning problems.

Contributors: WEH and OK-H designed the study and collected the data; WEH, SS, SKS, SCH, JEK and KWE analysed the data. WEH and SS wrote the first version of the manuscript; all authors (WEH, SS, SKS, OK-H, SCH, JEK and KWE) revised the manuscript for important intellectual content and approved the final version for submission.
Acknowledgements: none.
Funding: none.
Conflicts of interest: WEH has received research funding from Mundipharma Medical Basel, Switzerland; research support in kind from Prytime Medical Boerne, USA; support for a conference he chaired from Mundipharma Medical Basel, Switzerland, Isabel Healthcare, UK, EBSCO, Germany, and VisualDx, USA; and speakers honorarium from AO Foundation Zurich, Switzerland.
Ethical approval: the Curriculum Committee and the Examination Committee of the RMC, as well as the bureau of data privacy and the Institutional Review Board at Charité granted the study their approval (EA1/170/09).

REFERENCES

1 Davis DA, Mazmanian PE, Fordis M, van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA* 2006;**296** (9):1094–102.
2 Arnold L, Willoughby TL, Calkins EV. Self-evaluation in undergraduate medical education: a longitudinal perspective. *J Med Educ* 1985;**60** (1):21–8.
3 Stroben F, Schröder T, Dannenberg KA, Thomas A, Exadaktylos A, Hautz WE. A simulated night shift in the emergency room increases students' self-efficacy independent of role taking over during simulation. *BMC Med Educ* 2016;**16**:177.
4 Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten CPM. The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med* 2010;**85** (7):1212–20.
5 Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005;**80** (10 Suppl):S46–54.
6 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;**121** (5 Suppl):S2–23.
7 Colliver JA, Verhulst SJ, Barrows HS. Self-assessment in medical practice: a further concern about the conventional research paradigm. *Teach Learn Med* 2005;**17** (3):200–1.
8 Eva KW, Regehr G. Exploring the divergence between self-assessment and self-monitoring. *Adv Health Sci Educ Theory Pract* 2011;**16** (3):311–29.
9 Eva KW, Regehr G. Knowing when to look it up: a new conception of self-assessment ability. *Acad Med* 2007;**82** (10 Suppl):S81–4.
10 Eva KW, Regehr G. "I'll never play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof* 2008;**28** (1):14–9.
11 Moulton C, Regehr G, Lingard L, Merritt C, MacRae H. 'Slowing Down When You Should': initiators and influences of the transition from the routine to the effortful. *J Gastrointest Surg* 2010;**14** (6):1019–26.
12 Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Intern Med* 2013;**173** (21):1952.
13 Agrawal S, Norman GR, Eva KW. Influences on medical students' self-regulated learning after test completion. *Med Educ* 2012;**46** (3):326–35.

14 McConnell MM, Regehr G, Wood TJ, Eva KW. Self-monitoring and its relationship to medical knowledge. *Adv Health Sci Educ Theory Pract* 2012;**17** (3):311–23.

15 De Bruin ABH, Dunlosky J, Cavalcanti RB. Monitoring and regulation of learning in medical education: the need for predictive cues. *Med Educ* 2017;**51** (6):575–84.

16 Koriat A, Nussinson R, Ackerman R. Judgments of learning depend on how learners interpret study effort. *J Exp Psychol Learn Mem Cogn* 2014;**40** (6):1624–37.

17 Pusic MV, Chiaramonte R, Gladding S, Andrews JS, Pecaric MR, Boutis K. Accuracy of self-monitoring during learning of radiograph interpretation. *Med Educ* 2015;**49** (8):838–46.

18 Kunina-Habenicht O, Hautz WE, Knigge M, Spies C, Ahlers O. Assessing clinical reasoning (ASCLIRE): instrument development and validation. *Adv Health Sci Educ* 2015;**20** (5):1205–24.

19 Ward M, Gruppen L, Regehr G. Measuring self-assessment: current state of the art. *Adv Health Sci Educ Theory Pract* 2002;**7** (1):63–80.

20 Freeman A, van der Vleuten CPM, Nouns Z, Ricketts C. Progress testing internationally. *Med Teach* 2010;**32**:451–5.

21 Ilgen JS, Eva KW, Regehr G. What's in a label? Is diagnosis the start or the end of clinical reasoning? *J Gen Intern Med* 2016;**31** (4):435–7.

22 Tweed M, Purdie G, Wilkinson T. Low performing students have insightfulness when they reflect-in-action. *Med Educ* 2017;**51** (3):316–23.

23 Lynn DJ, Holzer C, O'Neill P. Relationships between self-assessment skills, test performance, and demographic variables in psychiatry residents. *Adv Health Sci Educ Theory Pract* 2006;**11** (1):51–60.

24 Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, Fine PL, Miller TM, Elstein AS. Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *J Gen Intern Med* 2005;**20** (4):334–9.

25 Cavalcanti RB, Sibbald M. Am I right when I am sure? Data consistency influences the relationship between diagnostic accuracy and certainty. *Acad Med* 2014;**89** (1):107–13.

26 Jowett N, LeBlanc V, Xeroulis G, MacRae H, Dubrowski A. Surgical skill acquisition with self-directed practice using computer-based video training. *Am J Surg* 2007;**193** (2):237–42.

27 Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol* 1999;**77** (6):1121–34.

28 Hautz WE, Kämmer JE, Exadaktylos A, Hautz SC. How thinking about groups is different from groupthink. *Med Educ* 2017;**51** (2):229.

29 Hautz WE, Kammer JE, Schauber SK, Spies CD, Gaissmaier W. Diagnostic performance by medical students working individually or in teams. *JAMA* 2015;**313** (3):303–4.

Received 25 September 2018; editorial comments to authors 13 November 2018; accepted for publication 5 December 2018

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Figure A1. Study design.
Figure A2. Design of the experimental task.
Figure A3. Screenshot of the data collection screen.