Accuracy of self-monitoring during learning of radiograph interpretation

Martin V Pusic,[1] Robert Chiaramonte,[2] Sophia Gladding,[3] John S Andrews,[4] Martin R Pecaric[5] & Kathy Boutis[6]

CONTEXT Despite calls for the improvement of self-assessment as a basis for self-directed learning, instructional designs that include reflection in practice are uncommon. Using data from a screen-based simulation for learning radiograph interpretation, we present validity evidence for a simple self-monitoring measure and examine how it can complement skill assessment.

METHODS Medical students learning ankle radiograph interpretation were given an online learning set of 50 cases which they were asked to classify as 'abnormal' (fractured) or 'normal' and to indicate the degree to which they felt certain about their response (*Definitely* or *Probably*). They received immediate feedback on each case. All students subsequently completed two 20-case post-tests: an immediate post-test (IPT), and a delayed post-test (DPT) administered 2 weeks later. We determined the degree to which certainty (*Definitely* versus *Probably*) correlated with accuracy of interpretation and how this relationship changed between the tests.

RESULTS Of 988 students approached, 115 completed both tests. Mean ± SD accuracy scores decreased from 59 ± 17% at the IPT to 53 ± 16% at the DPT (95% confidence interval [CI] for the difference: −2% to −10%). Mean self-assessed certainty did not decrease (rates of *Definitely*: IPT, 17.6%; DPT, 19.5%; 95% CI for difference: +7.2% to −3.4%). Regression modelling showed that accuracy was positively associated with choosing *Definitely* over *Probably* (odds ratio [OR] 1.63, 95% CI 1.27–2.09) and indicated a statistically significant interaction between test timing and certainty (OR 0.72, 95% CI 0.52–0.99); thus, the accuracy of self-monitoring decayed over the retention interval, leaving students relatively overconfident in their abilities.

CONCLUSIONS This study shows that, in medical students learning radiograph interpretation, the development of self-monitoring skills can be measured and should not be assumed to necessarily vary in the same way as the underlying clinical skill.

Medical Education 2015: 49: 838–846
doi:10.1111/medu.12774

Discuss ideas arising from the article at www.mededuc.com discuss.

[1]Division of Education Quality and Analytics, School of Medicine, New York University, New York, NY, USA
[2]Downstate College of Medicine, State University of New York, New York, NY, USA
[3]Department of Medicine, University of Minnesota Medical School, Minneapolis, MN, USA
[4]Office of Graduate Medical Education, University of Minnesota Medical School, Minneapolis, MN, USA
[5]Contrail Consulting Services, Toronto, ON, Canada
[6]Department of Paediatrics, Hospital for Sick Children, University of Toronto, Toronto, ON, Canada

Correspondence: Martin V Pusic, Division of Education Quality and Analytics, Institute for Innovation in Medical Education, 545 First Avenue, Suite 6P, New York, NY 10016, USA.
Tel: 00 1 212 263 2053; E-mails: mpusic@gmail.com; martin.pusic@nyumc.org

INTRODUCTION

Self-assessment involves a complicated suite of cognitive and social strategies centred on identifying an individual's strengths and weaknesses. These strategies are ultimately thought to be beneficial because they lead to reflection both on and in practice.[1] The reflective practitioner internalises lessons from his or her experiences, which leads to positive adaptive behaviours such as more accurate self-direction of learning and better patient care because the reflective practitioner is more acutely aware of the limits of his or her knowledge and skill.[1,2]

Whereas self-assessment represents a judgement of one's overall ability, self-monitoring involves an 'in-the-moment' awareness of the match between personal ability and the current problem.[3]

In this paper, we focus on the accuracy of students' self-monitoring as they learn to interpret radiographs. Our goal was to characterise the co-development of expertise and self-monitoring accuracy in the setting of massed learning of radiograph interpretation. We hypothesised that self-monitoring ability would degrade over a retention interval and would be accompanied by degradation in the underlying skill. Figure 1 illustrates how, when performance is measured after a retention interval, there is generally a decay relative to the measurement of performance immediately after a learning phase.[4,5] However, self-monitoring accuracy may be based on the rate at which the material is learned (i.e. based on the early performance level: Point A in Fig. 1).

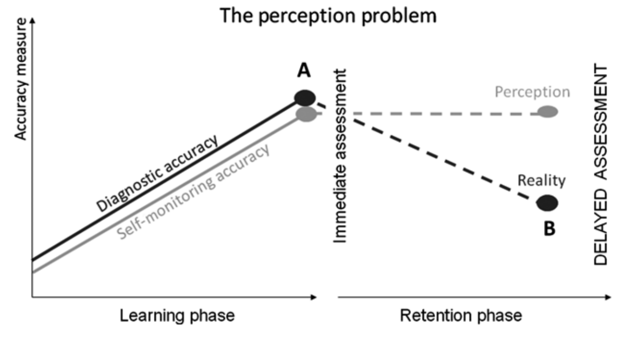

Figure 1 The perception problem. When performance is measured after a retention interval, generally there is a decay compared with performance measured immediately after a learning phase. However, self-monitoring accuracy may be based on the rate at which the material is learned and the early performance level (Point A). If a learner carries his or her self-perceptions from Point A to Point B, a significant mismatch is possible

We postulate that if a learner carries his or her self-perceptions from Point A to Point B (Fig. 1), a significant mismatch is possible.

Conceptual model

We have previously carried out a series of studies of learners developing the skill of radiograph interpretation using deliberate practice including immediate feedback.[6,7] We showed that a group of learners can reliably develop the skill in a manner that is governed by a learning curve relationship.[6,7] Like many other investigators, we noted that the skill decays considerably over a retention interval.[8]

Eva and Regehr showed that psychology students who have relatively poor *self-assessment* ability for a certain knowledge domain ('Can I predict the number of questions I will get right on this examination?') can nonetheless demonstrate significant *self-monitoring* ability ('Will I get THIS question right?').[9] Self-monitoring has also been shown to correlate with accuracy in radiograph interpretation[10,11] (Table 1).

In the current work, we measured the degree to which a measure of self-monitoring is associated with actual skill in the diagnostic classification task of radiograph interpretation.[8] We examined how the accuracy of in-the-moment self-monitoring ('Am I correct?') varies with actual diagnostic accuracy in radiograph interpretation. Additionally, we examined how any association between the two weathers a retention interval over which actual accuracy decays (Fig. 1).

METHODS

Study design

In this report, we present a second analysis, from a self-monitoring perspective, of a dataset collected during an educational experiment that has since been published.[8] In this prospective, randomised, three-arm, multicentre trial conducted in senior medical students, the experimental groups received different levels of a 'hint intervention' during deliberate practice on 50 consecutive ankle radiographs.[8] The hints consisted of a pop-up dialogue box that suggested that a participant's submitted answer may not be correct and that he or she should look further for abnormalities. The aim of this hint was to teach the students to be more persistent in searching for abnormalities. The main outcome was

Table 1 Definitions of terms

Term	Definition	Contextualised example for cognitive simulation of radiograph interpretation
Self-assessment	'Self-assessment has been defined broadly as the involvement of learners in judging whether or not learner-identified standards have been met'[1]	'My current ability to read ankle radiographs is good based on my previous test scores'
Self-monitoring	'A task-bound reflective process in which we continue to act but maintain the potential to reshape our action through cognition'[1]	'For this ankle radiograph, I am not certain if I am correct'
Self-directed learning	'A process by which individuals take the initiative, with or without the assistance of others, in diagnosing their learning needs, formulating learning goals, identify human and material resources for learning, choosing and implement appropriate learning strategies, and evaluating learning outcomes'[18]	'I am not confident of my knowledge of growth plate fractures so I will look up the definitions and practise on paediatric cases until my proficiency level is ____'

accuracy on two post-tests, one of which was carried out immediately after the practice, whereas the second was administered 2 weeks later. There proved to be no significant difference between the groups which received the hints and the control groups which did not.

During the trial, we concomitantly collected prospective data on self-monitoring by the participants. We now examine the additional self-monitoring data to establish how these data correlate with the students' accuracy on the two post-tests. Herein, we describe the original study in sufficient detail to contextualise the results of our new analysis. Full methodological details are available in the original report.[8] Because the study groups that were given hints did not differ on the main outcome, they are combined in this analysis as a single cohort.

Development of the radiograph learning cases and post-tests

Education intervention and post-intervention testing

We prospectively collected an initial pool of 234 ankle radiographs that were obtained to exclude the possibility of ankle fracture.[6] Each case included the three standard ankle radiograph views (anteroposterior, mortise, lateral), as well as the staff paediatric radiologist's report. Cases were categorised as either normal or abnormal based on the official radiology report.

From the 234 cases, we selected a 50-case learning subset based on the specific diagnosis and the radiograph's Rasch item difficulty index, which had been determined in prior research.[6,8]

The post-tests consisted of 20 cases chosen via Monte Carlo simulations from the remaining 184 cases using criteria similar to those applied to select the 50-case learning set (i.e. a 50% abnormal/normal proportion, with examples from each of the diagnostic categories).[8] The same post-test was given immediately after the first 50-case learning set (immediate post-test [IPT]) and then again 2 weeks later (delayed post-test [DPT]) to evaluate learning retention. The 20 post-test cases were presented in a fixed randomly determined order common to all participants. No feedback was provided during review of the post-test cases. The identical post-tests included 20 items each, with a range of difficulties (0.28–0.97; mean ± SD 0.59 ± 0.20).[8] Cronbach's alpha values indicated reliability was acceptable for a test with only 20 items for both the IPT (0.68) and DPT (0.67).[8] The Cohen's d effect size, comparing the first 20 items with the IPT score, for the 50-item learning set was 0.6 (95% confidence interval [CI] 0.2–0.9).[8]

Computer program to present radiographs

In our prior report, we described the computer program used to present cases to students.[8] In brief, students started with an introductory interactive tutorial on paediatric ankle radiograph interpretation consisting of 35 screens. The students then interpreted the 50 unknown cases (with feedback) and subsequently undertook the 20-case post-tests (without feedback).

For each unknown case, participants were required to review the history for the case prior to proceeding to any of the three available ankle views. Subjects classified each case as 'Probably normal', 'Definitely normal', 'Probably abnormal' or 'Definitely abnormal'. Abnormal answers also required the user to indicate the location of the suspected abnormality with a mouse click. In the cases that involved deliberate practice, once the participant had committed to his or her diagnosis, the program provided immediate feedback by highlighting the pathology on the abnormal images and providing the radiologist's report (Fig. 2). All responses were collected and stored in a MYSQL (Oracle Corp., Redwood, CA, USA) database table.

Study population and setting

The original data were derived from senior-year medical students at three universities (Site 1, Site 2, Site 3), recruited by e-mail. Students from multiple sites were selected to enhance the generalisability of the results. The research ethics boards at all the participating institutions approved this study.

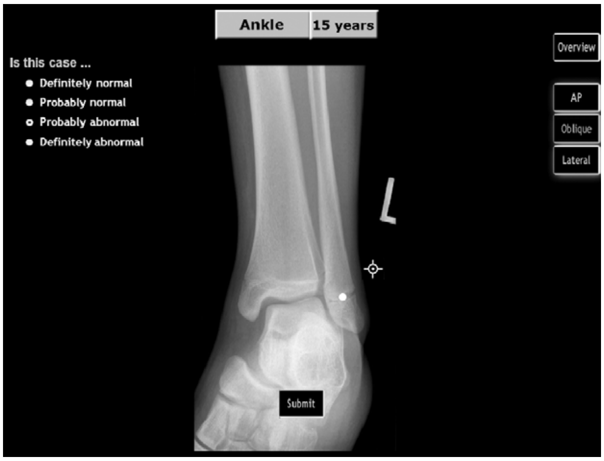

Figure 2 Screen capture of ankle radiograph case. The task is to declare the ankle radiograph either normal or abnormal (fractured). If the learner chooses one of the 'abnormal' options, the yellow marker appears allowing the student to designate the location of the fracture. Clicking 'Submit' leads to immediate feedback

Participant recruitment

Students from Site 1 and Site 3 were recruited via electronic solicitation on the listserv specific to senior-level medical students at those sites. Participants from Site 2 were approached while on a 4-week radiology selective. These students were approached at a group teaching session in the first week of their rotation. Interested students from any site replied to the research coordinator via e-mail. The coordinator obtained consent for study participation and enrolled students were provided with a link to the website and a unique username and password.

Study procedures

Students first completed the introductory tutorial and then immediately completed the 50-case learning set and the 20-case IPT. 2 weeks after the first post-test had been completed, participants were sent a reminder to log onto the system to complete the DPT. Students who completed both phases of the research were provided with a US$ 20 gift certificate and a certificate of completion.

Self-monitoring measures

The ability of participants to correctly self-monitor their responses was measured using the *Probably/Definitely* qualifier. To determine the validity of this operationalisation, we specified three *a priori* validity checks based on known self-monitoring associations. First, if the self-monitoring is valid, accuracy should be higher for cases in which *Definitely* is chosen compared with those in which *Probably* is selected. Secondly, individuals with higher ability are generally better at self-monitoring.[12] Finally, men are known to express more confidence than women at a given level of ability[13] and thus we postulated that men would choose *Definitely* more often than women.

Statistical analyses

Item coding

We considered each completed case as one item. Normal radiograph items were scored dichotomously. We defined 'accuracy' as the dependent variable in our analyses and gave it values of 0 or 1 for an individual case or, at the test level, the proportion correct. We defined 'certainty' as 0 when the qualifier chosen was *Probably* and as 1 when *Definitely* was selected.

Validity checks

We determined the proportion of questions answered accurately with *Definitely* versus *Probably*. Pearson's correlation coefficient was used to report the correlation between accuracy and an individual's number of correct self-assessments using *Definitely* (an index of correct self-monitoring). Finally, we determined the proportions of men and women who selected the *Definitely* qualifier to their responses, adjusted for accuracy.

Main analyses

Univariate analyses of continuous variables consisted of *t*-tests including 95% CIs for the difference. For the main outcome, we report a mixed-effects logistic model that accounts for the nesting of observations within individuals. Accuracy was the dependent variable. Independent variables included 'certainty', 'test timing' (immediate versus delayed), and the interaction term 'certainty*test timing'. This model allowed us to examine if accuracy was independently associated with certainty, as well as the extent to which the relationship changed between the IPT and the DPT. We report odds ratios (ORs) with respective 95% CIs.

Sample size

We had data from 115 medical students to include in this analysis. Assuming an α-value of 0.05 and a β-value of 0.8, and using the entire cohort for the present analysis, we anticipated being able to detect a univariate difference caused by certainty (*Probably* versus *Definitely*) on a post-test accuracy score as small as 5%.

RESULTS

Participant recruitment

We enrolled participants from April 2011 to February 2012. A total of 988 medical students were invited to participate in the study. Of these, 228 (23%) expressed an interest in participating and 154 (15%) consented to do so. A total of 26 students dropped out before they had completed the 50 learning cases and the IPT, and a further two participants did not complete the DPT. Thus 115 of the 143 (81%) eligible and consenting participants completed the study (Fig. 3).

Validity evidence for certainty measure

Across all post-test questions, when a student selected the *Definitely* qualifier, the probability that his or her response was correct was higher (62% versus 54%; 95% CI for the difference: 4–12%).

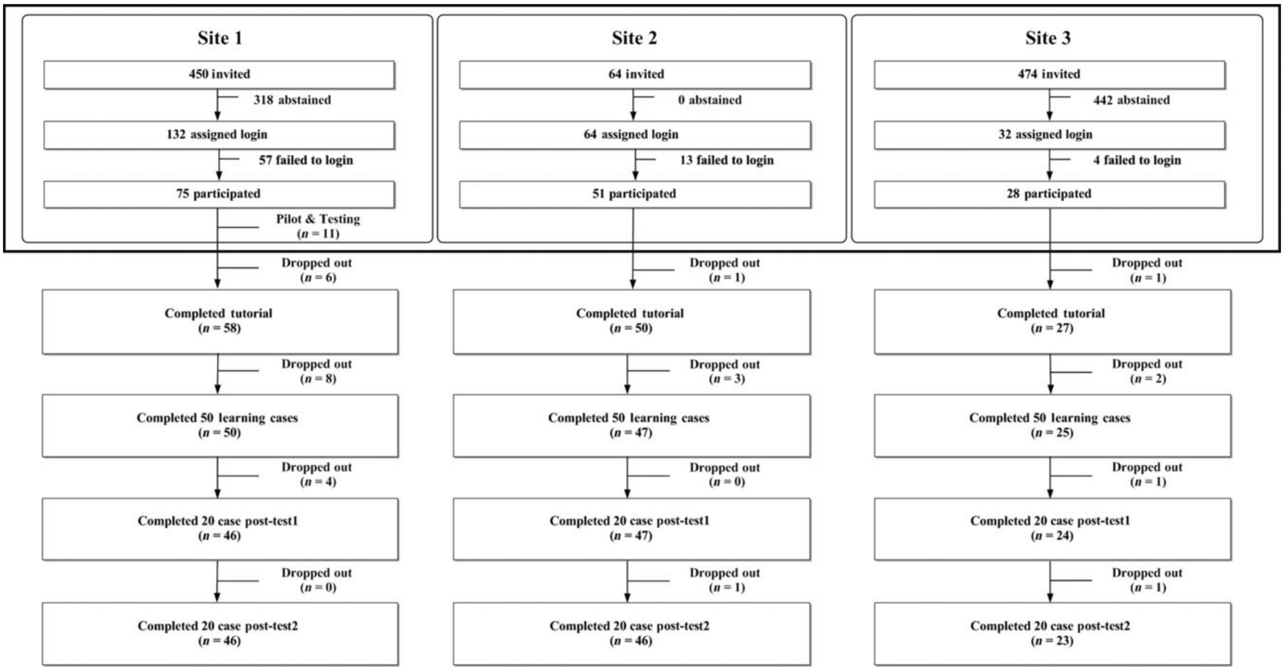

Figure 3 Study flow diagram

None of the students used the *Definitely* qualifier for every case, but 10 students used the *Probably* qualifier on all cases. The exclusion of data for these 10 students did not make a material difference to the direction or statistical significance of the subsequent analyses.

Pearson's correlation coefficient between an individual's overall accuracy and the number of correct self-assessments that person made using *Definitely* (an index of correct self-monitoring) was 0.40 ($p < 0.001$), which suggests that individuals with higher ability levels have greater self-monitoring ability (Fig. 4).

Of the 115 students who participated, 61 (53%) were male and 54 (47%) were female. In this sample, males were both more accurate overall on the 40 post-test questions (57.8% versus 53.1% accuracy, 95% CI for the difference: +1.8% to +7.6%) and more likely to use the *Definitely* qualifier (*Definitely*: 21.0% in men and 15.8% in women; 95% CI for the difference: +2.9% to +7.4%). When the choice of *Definitely* versus *Probably* is predicted using a mixed logistic regression model (accounting for within-person clustering) including both accuracy (OR 1.33, 95% CI 1.11–1.59) and male gender (OR 1.75, 95% CI 1.03–3.0), findings for both parameters are statistically significant. Thus gender, when adjusted for accuracy on the item, predicts the use of the *Definitely* qualifier.

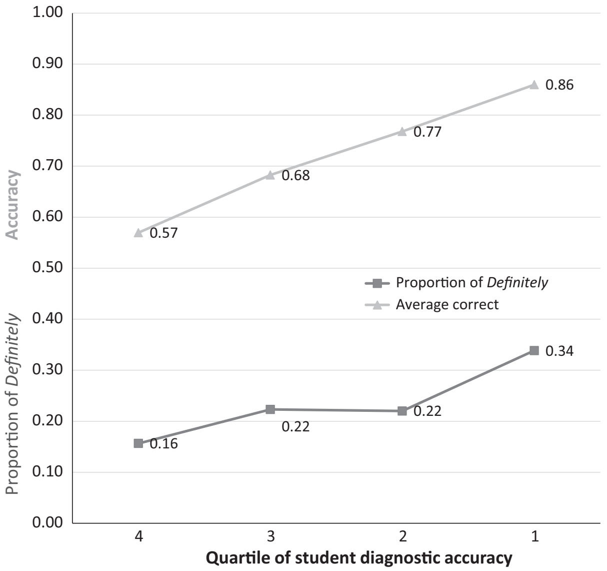

Figure 4 Certainty of interpretation by accuracy quartile. The likelihood of a student choosing to qualify his or her interpretation of a radiograph with *Definitely* (normal or abnormal) as opposed to *Probably* was associated with that student's accuracy on that radiograph (see text) and personal overall accuracy. For example, the fourth quartile (least accurate) of students responded to 57% of the cases accurately and used the *Definitely* qualifier 16% of the time

Main outcome: self-monitoring analyses

Mean ± SD accuracy scores decreased from 59 ± 17% on the IPT to 53 ± 16% on the DPT (95% CI for the difference: −2% to −10%). The percentage of cases on the two post-tests for which students used the *Definitely* qualifier for their answer did not differ statistically significantly between the IPT (17.6%) and DPT (19.5%) (95% CI for the difference: +7.2% to −3.4%). Students were more accurate on cases for which they used the *Definitely* qualifier to a statistically significant extent in comparison with those for which they used the *Probably* qualifier. This was especially so on the IPT (Table 2).

The regression model demonstrated that the odds of a student being accurate decreased from the IPT to the DPT (OR 0.81, 95% CI 0.71–0.93) and the odds of accuracy when a student selected *Definitely* (compared with *Probably*) were higher (OR 1.63, 95% CI 1.27–2.09). However, for the interaction of the variables certainty and test timing, the odds of being accurate were lower (OR 0.72, 95% CI 0.52–0.99), which suggests that accuracy of self-monitoring, as measured by reported certainty, decayed over a retention interval even more than would be accounted for by the decline in accuracy (Fig. 5).

DISCUSSION

This study of medical students learning the interpretation of radiographs under conditions of massed deliberate practice shows that medical students can demonstrate some degree of self-monitoring accuracy with deliberate practice and that this accuracy degrades over a retention interval.

A great deal has been written about the mismatch between a person's performance and his or her self-assessment of that performance.[1,3,14,15] In suggesting new approaches to the dilemma of poor self-assessment, Eva and Regehr called for new research into *why* this should be the case.[1] In particular, they called for more research into self-assessment in service of reflection in action, arguing that 'self-assessment is not a stable skill, but rather [. . .] a situationally bounded cognitive process that is con-

Table 2 Mean accuracy score by certainty selection and timing of test

	Definitely		*Probably*		
	n	**Accuracy, %, mean ± SD**	***n***	**Accuracy, %, mean ± SD**	**95% CI for difference (*Definitely–Probably*)**
Immediate post-test	405	68.8 ± 46.3	1895	56.4 ± 49.6	7–18
Delayed post-test	449	56.3 ± 49.7	1851	51.8 ± 50.0	0–10
Overall	854	62.2 ± 48.4	3746	54.1 ± 49.8	4–11

95% CI = 95% confidence interval; SD = standard deviation

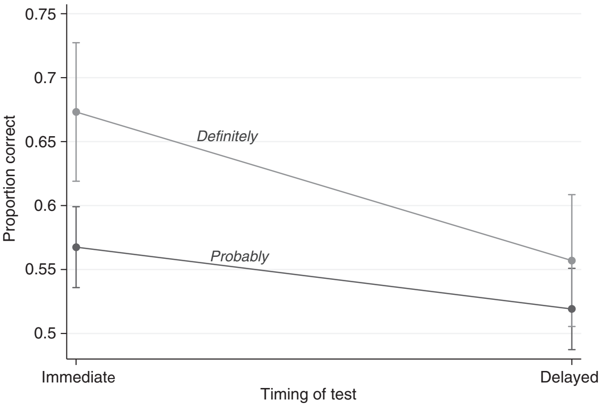

Figure 5 Interaction of certainty with test timing. Overall diagnostic accuracy decreases from the immediate post-test (IPT) to the delayed post-test (DPT). The greater the separation of the points for a given test, the greater the accuracy of self-monitoring. Separation decreases from the IPT to the DPT. The interaction term is statistically significant, indicating that self-monitoring accuracy decreases even more than would be predicted based on the decrease in accuracy alone

text specific and dependent upon expertise'.[1,12] The same investigators subsequently investigated psychology students completing examination questions to show that self-monitoring, in contradistinction to more global self-assessment, is successful in predicting 'in-the-moment' functioning.[9]

Our educational intervention was mainly geared towards improving the skill of radiograph interpretation through deliberate practice. It achieved this goal, having an effect size of 0.6 after approximately 1 hour of instruction. We demonstrated the typical findings of higher immediate test scores and then lower scores 2 weeks later; however, we did *not* find a concomitant decrease in certainty over the retention interval.

The structure of our educational experiment is common and consists of a learning phase and an immediate test followed by a retention test. What is less common is to combine the deliberate practice of a cognitive skill with deliberate practice in self-monitoring. By adding four words to the answer choices of the learners, at the cost of no additional mouse-clicks (Fig. 2), we were able to collect self-monitoring data that were sufficiently discriminatory to show the following findings: (i) students are able to predict, to some extent (a small to moderate correlation of ~0.4), the likelihood that their answer is correct; (ii) there are gender differences in certainty assessment; (iii) better students have better self-monitoring accuracy, and (iv) this self-monitoring accuracy degrades over a retention interval, potentially predisposing to overconfidence. Each of these findings could be used in the service of refining self-monitoring accuracy.

We have had difficulty finding studies that look at the relative decay of self-assessment skills over a retention interval. In our study, we found that these skills do indeed decay over a 2-week retention interval. It is important to explicitly track and develop self-monitoring ability because it facilitates self-directed learning, an important competency for health professionals.[1,16] Further, insight into one's own abilities allows more accurate clinical practice.[1,16] Different instructional strategies can result in differential accuracy of self-monitoring. For example, Brydges *et al.*[17] showed that a self-directed learning strategy for learning the skill of lumbar puncture was superior to instructor-led learning in terms of the development of accuracy in self-assessment. They recommend providing explicit opportunities for students to monitor their own progress.

In addition to this idea of more explicitly exploring the rise and fall of self-monitoring skills over time, knowledge and skill retention studies can also allow us to examine how the development of self-monitoring skills co-varies with the learning of the underlying skill. It has been well established that skill in self-assessment rises concomitantly with ability in the skill being evaluated.[12,14] Therefore, the concomitant loss in this type of self-assessment skill with loss of accuracy in radiograph diagnosis might be anticipated. However, there seem to be other factors involved because the interaction term was also significant, indicating that students were actually more overconfident relative to their accuracy on the DPT than on the IPT. We postulated at the outset that massed training, of which our educational intervention is an example, may particularly exaggerate the miscalibration of self-assessment towards overconfidence on the DPT. People confuse the speed and ease of learning with their actual competence and make the compounding error of assuming that their competence will be maintained over time.[1,14] These inherent biases may represent one explanation for the pattern of overconfidence that was exaggerated on the retention test (Fig. 1). Without a comparison group, we cannot claim this association to be proven in the context of diagnostic interpretation; however, the fact that findings in our cohort follow this pattern, which has been recognised in other fields, suggests that health professions educators would do well to incorporate self-monitoring measures into their instructional designs. Further experiments might tease out how best to inculcate these important meta-cognitive skills.

This research is subject to some limitations which should be considered. We assumed that the interpretation of each radiograph would not involve any uncertainty and that therefore our *Probably* and *Definitely* qualifiers reflected only the confidence of the participant in his or her response. However, radiographs cannot determine diagnoses with complete certainty and thus even an expert would classify a small percentage of radiographs as 'probably' indicative because of their inherent limitations. This may have introduced noise into our estimates. The participants were a non-random sample of the larger student population. The students who participated were a self-selected sample and may be likely to have a higher degree of self-regulation and other important differences compared with the general student population. However, if our study cohort did indeed consist of a more highly self-regulated group, the relatively poor self-monitoring ability found can be regarded as even more troublesome.

In summary, the results of this study show that, in medical students learning radiograph interpretation, the development of self-monitoring skills can be measured and should not be assumed to necessarily vary in the same way that the underlying clinical skill does.

Contributors: MVP provided intellectual and technical contributions to the study design and execution, contributed to the literature search and analyses, and to the preparation of all figures and the manuscript. RC contributed to the literature search and analyses, and to the preparation of all figures. SG contributed substantially to the acquisition of study data at her site. MRP oversaw all research operations in patient enrolment, data collection and entry, and was involved in the analysis and interpretation of the results. KB obtained funding, oversaw all research operations in patient enrolment, collected a significant proportion of the data, contributed to the data analysis and interpretation and provided critical revisions to the manuscript. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.
Acknowledgements: The authors wish to acknowledge the contributions of Dr. Maria Shiau of the Department of Radiology, NYU School of Medicine whose work enabled the initial study. Dr. Oleksandr Savenkov of the NYU Institute for Innovations in Medical Education performed additional statistical analyses. We are also grateful to Dr. Glenn Regehr of the Centre for Health Education Scholarship (UBC) for his input on an early version.
Funding: The original study was funded by the Educational Dean's Fund, University of Toronto.
Conflicts of interest: none.
Ethical approval: this study was approved by the institutional review boards at the three participating sites (University of Minnesota, New York University, University of Toronto).

REFERENCES

1 Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005;**80** (10 Suppl):46–54.
2 Schön DA. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions.* San Francisco, CA: Jossey Bass 1987.
3 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.
4 Rohrer D, Pashler H. Recent research on human learning challenges conventional instructional strategies. *Educ Res* 2010;**39** (5):406–12.
5 Schmidt RA, Bjork RA. New conceptualisations of practice: common principles in three paradigms suggest new concepts for training. *Psychol Sci* 1992;**3** (4):207–17.
6 Boutis K, Pecaric M, Seeto B, Pusic M. Using signal detection theory to model changes in serial learning of radiological image

interpretation. *Adv Health Sci Educ Theory Pract* 2010;**15** (5):647–58.

7 Pusic M, Pecaric M, Boutis K. How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. *Acad Med* 2011;**86** (6):731–6.

8 Boutis K, Pecaric M, Shiau M, Ridley J, Gladding SP, Andrews JS, Pusic MV. A hinting strategy for online learning of radiograph interpretation by medical students. *Med Educ* 2013;**47**:877–87.

9 Eva KW, Regehr G. Exploring the divergence between self-assessment and self-monitoring. *Adv Health Sci Educ Theory Pract* 2011;**16** (3):311–29.

10 Geller BM, Bogart A, Carney PA, Elmore JG, Monsees BS, Miglioretti DL. Is confidence of mammographic assessment a good predictor of accuracy? *AJR Am J Roentgenol* 2012;**199** (1):134–41.

11 Carney PA, Bogart TA, Geller BM *et al*. Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *AJR Am J Roentgenol* 2012;**198** (4):970–8.

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

13 Pallier G. Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles* 2003;**48** (5–6):265–76.

14 Dunning D, Heath C, Suls JM. Flawed self-assessment: implications for health, education, and the workplace. *Psychol Sci Public Interest* 2004;**5** (3):69–106.

15 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.

16 Schumacher DJ, Englander R, Carraccio C. Developing the master learner: applying learning theory to the learner, the teacher, and the learning environment. *Acad Med* 2013;**88** (11):1635–45.

17 Brydges R, Nair P, Ma I, Shanks D, Hatala R. Directed self-regulated learning versus instructor-regulated learning in simulation training. *Med Educ* 2012;**46**:648–56.

18 Knowles MS. *Self-Directed Learning: A Guide for Learners and Teachers*. Englewood Cliffs, NJ: Prentice Hall 1974.

Received 11 November 2014; editorial comments to author 23 January 2015, accepted for publication 21 April 2015