Received: 19 July 2019 | Revised: 20 December 2019 | Accepted: 2 January 2020

DOI: 10.1111/medu.14057

ORIGINAL RESEARCH

Self-monitoring accuracy does not increase throughout undergraduate medical education

Juliane E. Kämmer[1,2] | Wolf E. Hautz[3] | Maren März[4]

[1]Inistitute of Health and Nursing Science, Charité – Universitätsmedizin Berlin, Berlin, Germany

[2]Centre for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[3]Department of Emergency Medicine, Inselspital University Hospital, University of Bern, Bern, Switzerland

[4]AG Progress Test Medicine, Charité – Universitätsmedizin Berlin, Berlin, Germany

Correspondence
Juliane E. Kämmer, Max Planck Institute for Human Development, Centre for Adaptive Rationality, Lentzeallee 94, 14195 Berlin, Germany.
Email: kaemmer@mpib-berlin.mpg.de

Abstract

Context: Accurate self-assessment of one's performace on a moment-by-moment basis (ie, accurate self-monitoring) is vital for the self-regulation of practising physicians and indeed for the effective regulation of self-directed learning during medical education. However, little is currently known about the functioning of self-monitoring and its co-development with medical knowledge across medical education. This study is the first to simultaneously investigate a number of relevant aspects and measures that have so far been studied separately: different measures of self-monitoring for a broad area of medical knowledge across 10 different performance levels.

Methods: This study assessed the self-monitoring accuracy of medical students (n = 3145) across 10 semesters. Data collected during the administration of the formative Berlin Progress Test Medicine (PTM) were analysed. The PTM comprises 200 multiple-choice questions covering all major medical disciplines and organ systems. A self-report indicator (ie, confidence) and two behavioural indicators of self-monitoring accuracy (ie, response time and the likelihood of changing an initial answer to a correct rather than an incorrect item) were examined for their development over semesters.

Results: Analyses of more than 390 000 observations (of approximately 250 students per semester) showed that confidence was higher for correctly than for incorrectly answered items and that 86% of items answered with high confidence were indeed correct. Response time and the likelihood of the initial answer being changed were higher when the initial answer was incorrect than when it was correct. Contrary to expectations, no differences in self-monitoring accuracy were observed across semesters.

Conclusions: Convergent evidence from different measures of self-monitoring suggests that medical students self-monitor their knowledge on a question-by-question basis well, although not perfectly, and to the same degree as has been found in studies outside medicine. Despite large differences in performance, no variations in self-monitoring across semesters (with the exception of the first semester) were observed.

1 | INTRODUCTION

Adequate self-assessment is essential for physicians' self-regulation, which is why it has been integrated into many models of lifelong learning[1,2] and is a focus of medical education.[3,4] For decades, this has been a cause for concern because numerous studies have found little to no connection between self-assessment and actual performance.[1,5] More recent conceptual advances, however, distinguish between summative *self-assessment* as a general, context-free assessment of one's own competence in a particular area and *self-monitoring* as a momentary judgement on the correspondence between one's own ability and the current problem.[6,7] For example, consider a junior physician diagnosing an incoming patient with respiratory problems in the emergency room. Obviously, this situation requires the application of a variety of professional skills and knowledge; less obviously, it also requires an adequate evaluation of whether the available information, knowledge and capabilities suffice for effectively and efficiently treating the patient. In the event of evaluating these factors as insufficient, the physician needs to recognise whether he or she should slow down[8] or consult a colleague,[9] for example. This process of assessing one's performance in the very moment is termed 'self-monitoring.'[6] Accurate self-monitoring is imperative for the provision of safe, efficient and effective health care,[5,9-12] and is an important guide for self-regulated learning during and after medical education.[5,13] However, our understanding of the functioning of self-monitoring and its development across medical education is limited. Addressing this research gap was the main goal of this study.

In research on self-monitoring, study participants usually encounter a knowledge task (ranging from very specific medical knowledge [eg, Pusic et al[14]] to general knowledge questions [eg, Eva and Regehr[9]]) they are required to complete. Self-monitoring is then assessed using either measures of conscious awareness of one's performance, such as confidence ratings, or behavioural indicators as measures of, instead, unconscious processing, such as response time and response changing. Reflecting a relatively good level of self-monitoring, these studies found that participants were more confident in their correct than in their incorrect answers,[10,11,14,17-19] took more time to answer an initially incorrectly answered item, flagged incorrect items for future consideration or deferred from answering them altogether.[6,8,9,18,19] Despite this convergent evidence from different studies using very different materials, measures and study populations, confidence and behavioural indicators have never been simultaneously assessed in a clinical context. The current study fills this gap in a low-stakes context.

According to the Dunning-Kruger framework, the expertise needed to perform well in a given situation is the same as that required to judge performance; thus, poor performers lack the necessary expertise not only to perform well, but also to recognise the gaps in their knowledge or ability.[20] Likewise, on the intra-individual level, low competence in a given domain leads to lower self-monitoring accuracy in that domain.[5,21] Given that students' medical knowledge increases tremendously during medical education,[22] we expect that levels of self-monitoring accuracy will increase each semester. Indeed, the three existing studies of self-monitoring that involved students from more than one semester showed that self-monitoring accuracy (operationalised as the relationship between confidence and performance) increases over time in training.[11,16,23] These studies are, however, limited to the evaluation of self-monitoring accuracy in limited areas, such as dyspnoea or pharmacology, or to a single measure of self-monitoring. Both limitations are addressed in the study presented here. Furthermore, a study of diagnostic decision making by Friedman and colleagues found that, somewhat counter-intuitively, the calibration of confidence to performance took a U-shaped form in a comparison of final-year students, residents and consultants in internal medicine.[15] Thus, more training does not necessarily imply better self-monitoring. Consequently, a cross-sectional study of self-monitoring across the whole spectrum of undergraduate education will be likely to inform such education.

Going beyond previous studies,[6,10,11,14,16,18,23] this study is the first to simultaneously investigate a number of relevant aspects that have so far been studied separately: specifically, it combines the assessment of a self-reported measure of self-monitoring (ie, confidence) with the assessment of two behavioural indicators of self-monitoring (ie, response time and response changing) in medical students of *all* semesters (and thus 10 different performance levels) concerning their knowledge of *all* major medical disciplines and organs. By doing this, we aim to gain a more comprehensive picture of the co-development of medical knowledge and self-monitoring.

2 | METHODS

2.1 | Setting

The medical degree course at Charité University Hospital Berlin comprises 10 semesters plus two semesters of electives. At the beginning of each of the 10 semesters, all students are requested to sit the Berlin Progress Test Medicine (PTM).[22,24] The test is administered online by an adjunct service provider, which is not involved in regular student grading. Although it is compulsory to take part in the test once per semester (no participation is sanctioned by no admission to further courses), it is sufficient to log in only; there is no need to answer all or any of the questions. The test is formative, with no pass/fail decision. Students can return to any item and change their answers at any time as they take the test.

The test comprises the same 200 multiple-choice items for students of all semesters. Each item comprises an item stem and 3-6 response options with one single best answer. Items are drawn from a database of approximately 5000 questions according to a fixed blueprint that ensures that each test covers all major medical disciplines and organ systems.[24] Students have 3 hours to complete the test within a 3-week time window.

For each answer given, students additionally indicate their confidence level (as: *I am very sure*, *I am fairly sure*, or *I am guessing*). In the

feedback students receive, items rated with *fairly sure* or *very sure* score 1 point if correct and −1 point if incorrect, whereas any items rated as *guessed* score 0 points.

2.2 | Participants

All medical students (n = 4644; 62.4% women; mean age: 25.1 years) enrolled at the Charité University Hospital Berlin were eligible for the present study.

2.3 | Measures

Self-monitoring accuracy was assessed using the following measures. First, we established mean confidence in correct and incorrect answers. For this, confidence ratings were interval-scaled numerically, following previous studies,[16,17] so that a rating of *I am guessing* = 0, a rating of *I am fairly sure* = 0.66, and a rating of *I am very sure* = 1. By subtracting the mean confidence for correct answers from the mean confidence for incorrect answers, we obtained self-monitoring indices (delta) within persons.[6,14,18,23] Second, we examined the proportion of correct answers amongst all answers at each confidence level. We compared the respective proportions with the overall chance level, which was 22.5%. Third, we calculated mean response times (in seconds) for initially correct versus incorrect answers, and fourth, we compared the mean percentage of response changes for initially correct versus incorrect answers. For the last two indicators of self-monitoring, we based our calculations on the *initial* responses because they best capture whether participants were aware of whether their (initial) answer was correct or not (cf. McConnell et al[18]).

2.4 | Data analyses

For each of the four dependent variables (mean confidence for accurate and inaccurate responses, proportion of correct responses per confidence level, response time and percentage of responses changed), we ran mixed-design analyses of variance (ANOVAs) with semester as a between-subjects factor. Additionally, in the first ANOVA, the accuracy of the *final* answer served as a within-subject variable and confidence as a dependent variable (mean confidence score for correctly answered questions versus mean confidence score for incorrectly answered questions). In the second ANOVA, confidence level (ie, *I am guessing*, *I am fairly sure*, *I am very sure*) served as a within-subject variable and the final percentage correct per confidence level as a dependent variable. Planned comparisons between confidence levels were conducted. In the third and fourth ANOVAs, *initial* accuracy served as a within-subject variable, and response time and percentage of responses changed as dependent variables, respectively. We conducted two separate ANOVAs with response time and percentage of responses changed as dependent variables instead of conducting one multivariate ANOVA (MANOVA) in order to obtain results that could be compared with findings in previous research in which these variables had been studied in a univariate context[18,25] and because further explorations of the data revealed that these two measures were poorly correlated (median intra-individual r_{pb} = −0.004).

3 | RESULTS

3.1 | Participants

A total of 3145 students from semesters 1-10 sat the PTM in April 2018 (Figure 1); data for 174 students (5.6%) were excluded as a

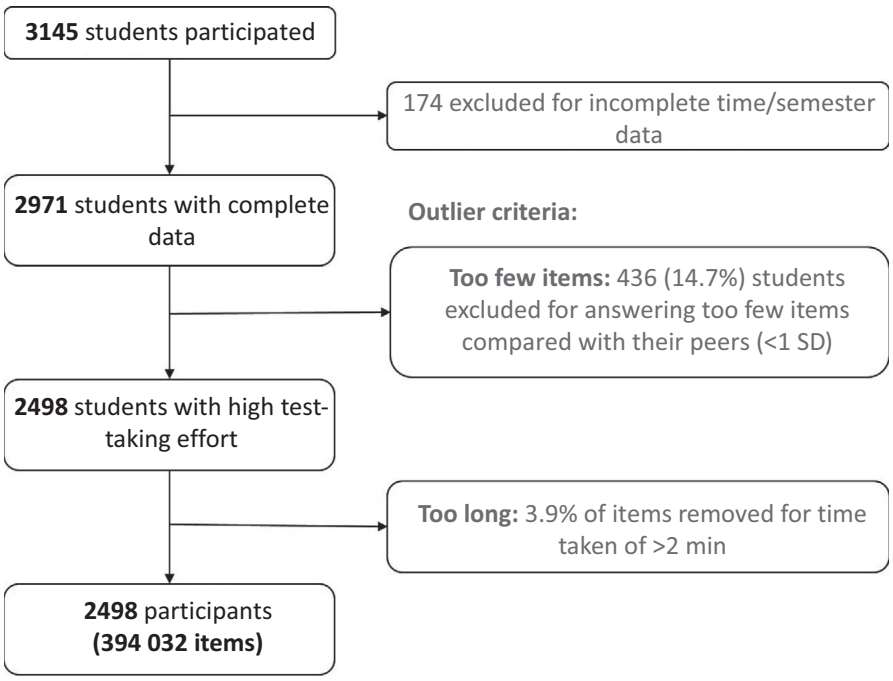

FIGURE 1 Flowchart of the study data-cleaning procedure. Abbreviation: SD, standard deviation

TABLE 1 Descriptive statistics by semester

Semester	Students, n	Items answered (including guesses), n Mean ± SD	Correct answers (including guesses), % Mean ± SD	*Very sure* answers, % Mean ± SD	*Fairly sure* answers, % Mean ± SD	*Guessed* answers, % Mean ± SD	Confidence score Mean ± SD
1	262	162.62 ± 60.07	41.89 ± 10.32	11.58 ± 10.73	28.39 ± 16.82	60.03 ± 23.79	0.28 ± 0.38
2	275	114.58 ± 72.34	57.29 ± 16.47	23.28 ± 16.82	31.64 ± 18.04	45.08 ± 28.32	0.33 ± 0.41
3	256	129.63 ± 67.14	56.35 ± 15.70	25.95 ± 17.28	34.28 ± 17.42	39.77 ± 24.72	0.41 ± 0.42
4	253	134.22 ± 60.78	59.37 ± 13.95	30.02 ± 18.31	33.75 ± 16.32	36.22 ± 24.14	0.47 ± 0.42
5	250	154.16 ± 51.95	62.64 ± 12.25	31.45 ± 17.95	34.21 ± 12.86	34.33 ± 19.83	0.51 ± 0.42
6	232	161.39 ± 46.83	65.14 ± 10.53	37.62 ± 19.78	35.52 ± 15.82	26.86 ± 17.26	0.59 ± 0.41
7	220	167.39 ± 44.66	64.54 ± 11.39	34.88 ± 18.86	36.48 ± 15.15	28.63 ± 17.32	0.57 ± 0.41
8	268	175.20 ± 37.29	67.45 ± 10.39	40.25 ± 20.74	34.88 ± 15.15	24.87 ± 16.25	0.62 ± 0.40
9	271	177.75 ± 36.54	67.50 ± 9.49	40.89 ± 20.12	36.24 ± 15.79	22.87 ± 15.51	0.64 ± 0.39
10	248	180.81 ± 32.09	71.25 ± 9.04	45.75 ± 20.60	36.39 ± 16.34	17.86 ± 12.83	0.70 ± 0.36

Abbreviation: SD, standard deviation.

result of missing details (no semester indicated or time not recorded for technical reasons).

We imposed two outlier criteria on the remaining 2971 students to filter out participants with low test-taking effort because high variations in test-taking effort can impair the evaluation of formative assessments.[26-28] First, we excluded 436 students (14.7%) who answered substantially fewer questions than their peers with medium or high levels of confidence (ie, for whom the number of questions answered [excluding guesses] was more than 1 standard deviation [SD] below the respective mean for students of that semester). Second, because students undertook the test without supervision, they theoretically had opportunities to consult others, look things up or take breaks. As the average time available for each question was 54 seconds (180 minutes for 200 items), we excluded all items with a response time longer than 2 minutes (3.9% of all items) in an attempt to filter out such behaviours. Our final sample consisted of 2498 students, with approximately 250 students per semester (Table 1, for raw data see reference [29]), and a total of 394 032 observations. It should be noted that the pattern of our results did not change when all data were included.

3.2 | Test performance

With the exception of semester 1, during which participants answered an exceptionally high number of items (and reported many guesses), the number of questions answered in total, answered correctly, and answered with medium/high confidence increased by semester, as did average confidence (Table 1).

3.3 | Accuracy of self-monitoring: Confidence

We ran a first mixed-design ANOVA with mean confidence for correct versus incorrect answers (including guesses) as a repeated measure withi-subjects and semester as a between-subjects measure. Students were, on average, more confident in correct (mean ± SD: 0.64 ± 0.19) than in incorrect answers (mean ± SD: 0.39 ± 0.22) ($F_{(1, 2524)}$ = 13 378.260, P < .001, η_p^2 = 0.841) (Table 2). We also found that the absolute level of confidence increased with semester, regardless of performance ($F_{(9, 2524)}$ = 77.943, P < .001, η_p^2 = 0.217), primarily as a result of an increase from semester 1 to semester 2 (0.10, 95% confidence interval [CI] 0.05-0.15) as was revealed by a Bonferroni-adjusted post hoc analysis (P < .001). Moreover, there was a small interaction effect between self-monitoring and semester ($F_{(9, 2524)}$ = 15.439, P < .001, η_p^2 = 0.052) (Table 2).

We ran a second mixed-design ANOVA with the proportion of correct answers (Figure 2A) out of all answered items per confidence level (Figure 2B) as a repeated measure and semester as a between-subjects measure. It revealed a main effect of the confidence level ($F_{(2, 4804)}$ = 9865.396, P < .001, η_p^2 = 0.804). On average, responses rated as *very sure* were more likely to be correct (mean ± SD: 85.72 ± 12.06) than responses rated as *fairly sure* (mean ± SD: 61.05 ± 15.37), which were more likely to be correct than responses rated as *guessed* (mean ± SD: 40.52 ± 16.28) (Bonferroni-adjusted P < .001). The ANOVA also revealed a medium effect of semester ($F_{(9, 2402)}$ = 20.971, P < .001, η_p^2 = 0.073), primarily due to an increase in the proportion of correct answers from semester 1 to semester 2 as revealed by Bonferroni-adjusted post hoc analysis (difference: 9.32%, 95% CI 6.12-12.52%; P < .001), without any further changes in later semesters. No interaction was revealed ($F_{(18, 4804)}$ = 1.551, P = .064).

3.4 | Accuracy of self-monitoring: Response time and response changing

In a third mixed-design ANOVA with response time as a dependent variable and a fourth ANOVA with the percentage of changed responses as a dependent variable, we entered *initial* accuracy as a repeated measure and semester as a between-subjects measure. The third ANOVA revealed a main effect of accuracy ($F_{(1, 2519)}$ = 501.558,

TABLE 2 Self-monitoring accuracy by semester

	Mean ± SD confidence if *final* answer was …			Mean ± SD time (in s) if *initial* answer was …			Mean ± SD response changes, % if *initial* answer was …		
Semester	Correct	Incorrect	Delta[a]	Correct	Incorrect	Delta[a]	Correct	Incorrect	Delta[a]
1	0.41 ± 0.19	0.22 ± 0.18	−0.18 ± 0.09	32.67 ± 12.68	32.13 ± 14.73	−0.53 ± 5.90	1.15 ± 1.19	3.86 ± 2.97	2.70 ± 2.54
2	0.54 ± 0.21	0.30 ± 0.23	−0.24 ± 0.12	39.85 ± 14.90	42.25 ± 18.91	2.72 ± 9.87	0.90 ± 1.24	2.93 ± 2.84	2.03 ± 2.61
3	0.60 ± 0.18	0.35 ± 0.22	−0.25 ± 0.11	37.50 ± 14.03	40.20 ± 18.65	2.79 ± 8.39	0.80 ± 1.10	2.85 ± 2.66	2.05 ± 2.42
4	0.63 ± 0.18	0.37 ± 0.23	−0.26 ± 0.12	36.53 ± 13.24	38.88 ± 17.03	2.34 ± 7.83	0.91 ± 1.16	2.98 ± 2.56	2.08 ± 2.33
5	0.64 ± 0.16	0.37 ± 0.20	−0.27 ± 0.10	35.88 ± 11.34	38.11 ± 13.92	2.30 ± 6.87	0.91 ± 0.88	3.33 ± 2.50	2.42 ± 2.22
6	0.71 ± 0.14	0.43 ± 0.19	−0.27 ± 0.10	34.87 ± 10.52	39.49 ± 13.25	4.63 ± 6.08	1.03 ± 1.03	3.15 ± 2.30	2.11 ± 2.09
7	0.69 ± 0.14	0.42 ± 0.19	−0.27 ± 0.10	34.38 ± 9.21	38.31 ± 11.99	3.93 ± 6.09	0.97 ± 1.00	3.26 ± 2.42	2.29 ± 2.16
8	0.72 ± 0.14	0.45 ± 0.20	−0.27 ± 0.11	33.92 ± 9.40	37.63 ± 11.51	3.71 ± 5.48	0.98 ± 0.97	3.02 ± 2.23	2.04 ± 1.92
9	0.73 ± 0.13	0.47 ± 0.20	−0.26 ± 0.11	33.06 ± 8.94	37.03 ± 11.04	3.96 ± 5.55	0.97 ± 0.95	3.26 ± 2.58	2.29 ± 2.32
10	0.77 ± 0.12	0.52 ± 0.19	−0.25 ± 0.11	33.31 ± 8.29	38.37 ± 10.51	5.06 ± 5.48	1.05 ± 0.91	3.12 ± 2.55	2.07 ± 2.30
Mean	0.64 ± 0.19	0.39 ± 0.22	−0.25 ± 0.11	35.17 ± 11.66	38.23 ± 14.70	3.06 ± 7.08	1.00 ± 1.05	3.17 ± 2.59	2.21 ± 2.31

Abbreviation: SD, standard deviation.

[a]Calculated by subtracting the mean value for correct answers from the mean value for incorrect answers.

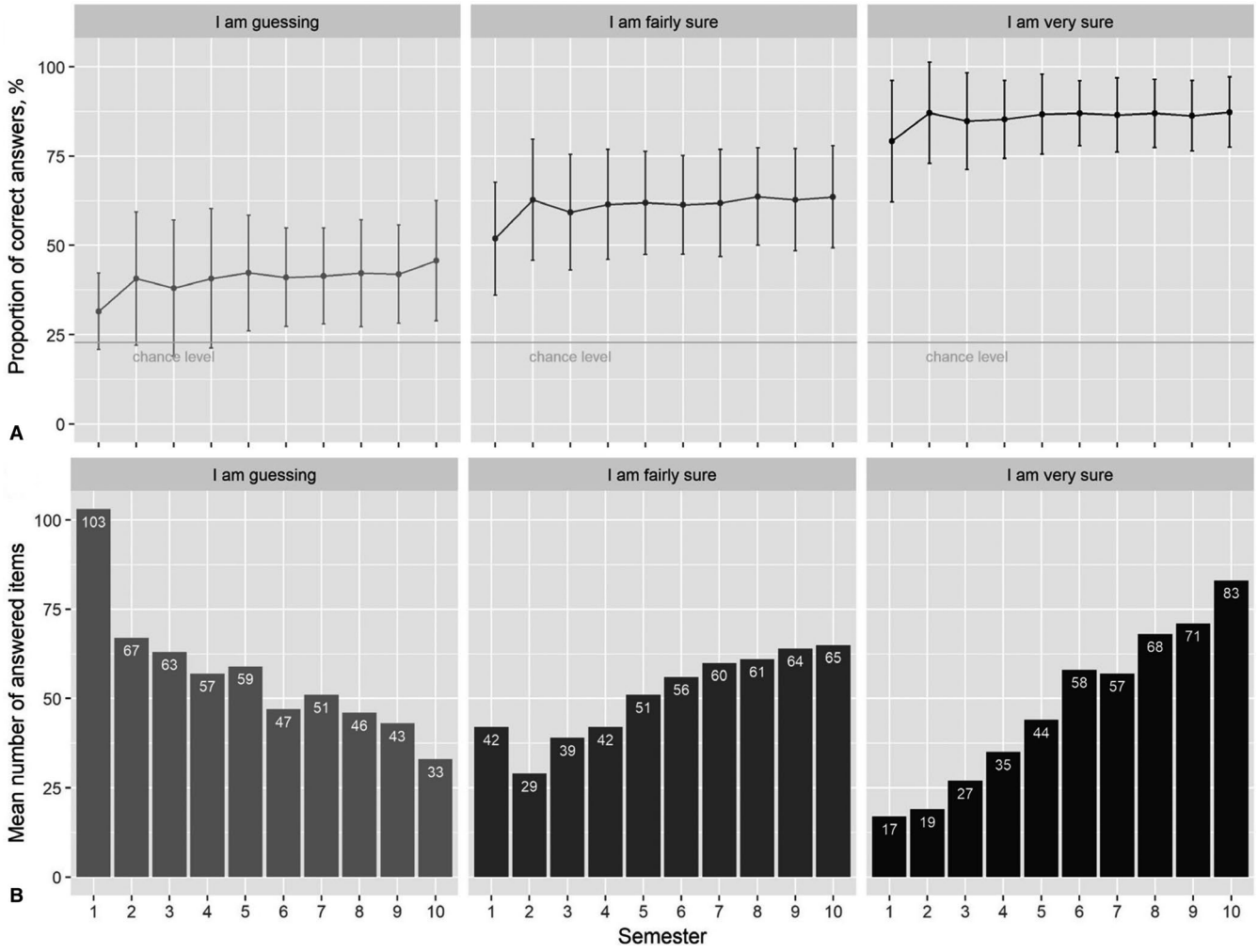

FIGURE 2 Self-monitoring accuracy by semester. A, proportion of correct answers by confidence level and semester (±1 standard deviation), relative to the chance level of 22.5%. B, mean number of items answered by confidence level and semester

$P < .001$, $\eta_p^2 = 0.166$), with students taking longer to respond when their initial answer was incorrect (mean ± SD: 38.23 ± 14.70 seconds) than when it was correct (mean ± SD: 35.23 ± 11.71) (Table 2). It further revealed a small effect of semester ($F_{(9, 2519)} = 8.549$, $P > .001$, $\eta_p^2 = 0.030$), due to an increase in response times from semester 1 to 2 (8.49, 95% CI 4.92-12.06) as revealed by a Bonferroni-adjusted post hoc analysis ($P < .001$). Lastly, the ANOVA revealed a small interaction effect of semester and accuracy on response time ($F_{(9, 2519)} = 13.362$, $P < .001$, $\eta_p^2 = 0.046$).

The fourth ANOVA revealed that students were more likely to change their response when the initial answer was incorrect (mean ± SD: 3.17 ± 2.59%) than when it was correct (mean ± SD: 1.00 ± 1.05%) ($F_{(1, 2524)} = 2551.615$, $P < .001$, $\eta_p^2 = 0.503$) (Table 2). In half of the cases in which students changed their response, they changed it to the correct option (mean ± SD: 2.04 ± 1.81%). In the other half, they were equally likely to switch from the correct to an incorrect option or from one incorrect option to another. There was again a small effect of semester ($F_{(9, 2524)} = 6.3237$, $P < .001$, $\eta_p^2 = 0.022$) and a small interaction effect of semester and accuracy on change of response ($F_{(9, 2524)} = 17.443$, $P < .001$, $\eta_p^2 = 0.059$).

4 | DISCUSSION

Do medical students finish their degrees with an adequate level of self-monitoring accuracy? Do different measures of self-monitoring reflect the same underlying construct? How do final-year students compare with novice students? Extending previous findings[11,18,23] to a broader area of medical knowledge and encompassing more performance levels across the whole spectrum of undergraduate medical education, our analyses of low-stakes test data provide evidence for relatively good self-monitoring accuracy from semester 2 onwards. In fact, contrary to our expectations, self-monitoring accuracy did not vary across the 10 semesters of medical education except that semester 1 stood out in all analyses (ie, students in semester 1 answered an exceptionally high number of items, reported the most guesses, took least time per item and changed their initial answers most often), probably because first-semester students may not have known what to expect or how to use a progress test.

With regards to self-monitoring accuracy, we found that medical students were more confident in their correct answers than in their incorrect answers, replicating previous findings.[23]. Second, the percentage of correct answers per confidence level, which offers a more fine-grained evaluation of self-monitoring accuracy than average confidence levels, showed that only approximately 86% of items answered with a high level of confidence were indeed correct. This degree of overconfidence is comparable with that reported in numerous previous studies (for reviews, see Berner and Graber[12] and Fischhoff et al[30]) and thus does not seem to be specific to the medical domain. Accordingly, 14% of answers for which respondents rated themselves as being *very sure* were incorrect. Such errors may prove dangerous in practice because highly confident physicians request fewer diagnostic tests, which potentially results in more diagnostic errors.[31] Hence, these are the items that most urgently require students' attention. Within the context of a progress test, feedback could be provided on these items in the form of the correct answers accompanied by further explanations. Moreover, students with high proportions of such items could be flagged and approached individually within the framework of a mentoring system.

Approximately 61% of answers for which respondents rated themselves as being *fairly sure* were indeed correct, in line with previous quantifications of that rating.[16,17] Informing students about their performance levels here might boost their confidence and self-esteem, and reduce the tendency to request additional and unnecessary tests.

The accuracy level of *guesses* was much better than would be expected by chance. This may be a side-effect of the confidence scale used. As there was no rating category for *I am fairly unsure*, the gap between the categories *I am guessing* and *I am fairly sure* was quite large.[16,17] Respondents who were neither confident nor unconfident in their answer may thus have chosen the *I am guessing* category, thereby increasing the overall level of accuracy in this category.

Concerning the behavioural indicators response time and response changing, we found that, in line with previous studies,[6,18] students took more time to think about a question and changed their response more often when their initial answer was incorrect (although the total proportion of changes was very small), thus providing convergent evidence for self-monitoring accuracy within one study. These findings further strengthen the validity of the assumption that the latent variable 'self-monitoring' can be operationalised through both behavioural (eg, response time) and self-reported (eg, confidence) measures.

This study can be regarded as a baseline study of a sample of students who had no experience in the explicit assessment of their self-monitoring accuracy. Although the students who participated in this study had experience in filling in the progress test (except for those in the first semester), they had never been prompted to indicate their level of confidence and had not received any feedback on their self-monitoring accuracy before. Future longitudinal studies need to test whether the provision of early detailed feedback not only on performance but also on self-monitoring accuracy helps students choose appropriate learning goals and activities, and ultimately leads to improvements in performance and self-monitoring accuracy.[32,33] We expect that providing students with external guidance in the form of feedback[14,21,33,34] or cue prompts[32] will improve their self-monitoring accuracy and thus contribute to their achieving expert performance in the long run.[35] In particular, the cue utilisation framework proposed by de Bruin and colleagues[32,36] may be used in developing research and education that identify the cues students can use to adequately guide their self-monitoring. A formative progress test (preferably in combination with a mentoring system) offers the potential to provide specific, timely feedback on the full range of knowledge tested and cues relevant for accurate self-monitoring.

4.1 | Limitations

This study has several limitations. First, as a cross-sectional study, it does not allow for longitudinal conclusions. Second, it is not advisable to generalise the present findings to clinical settings and domains of competence beyond declarative knowledge.[18,21] Third, the possibility of 'opting out' of the progress test (ie, by deferring an exceptionally high number of answers, used by approximately 17% of students) may mean that our sample is a non-random sample of the larger student population, namely, that it includes students who put greater effort into test taking. The relationship between test-taking motivation and self-monitoring remains a subject for future study.

5 | CONCLUSIONS

This study found that medical students across all semesters of study have relatively good levels of self-monitoring accuracy but that there is still room for improvement. We hope that an easy and early intervention, such as one that provides students with feedback on their self-monitoring accuracy, will promote the development of self-monitoring accuracy during undergraduate medical training, identify potentially dangerous knowledge gaps, help students to set appropriate learning goals and ultimately improve medical care.

ACKNOWLEDGEMENTS

The authors would like to thank Stefan Schauber (Centre for Health Sciences Education, Faculty of Medicine, University of Oslo, Oslo, Norway) for his helpful comments on an earlier version of the paper and Susannah Goss for editing the manuscript.

AUTHOR CONTRIBUTIONS

JEK and MM designed the study and collected the data. JEK analysed the data and wrote the first version of the manuscript. All authors (JEK, WEH and MM) interpreted the data, revised the manuscript for important intellectual content, approved the final version for submission, and have agreed to be accountable for all aspects of the work.

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 a speaker's honorarium from AO Foundation Zurich, Switzerland.

ETHICAL APPROVAL

This is an analysis of Berlin PTM data, which are collected routinely. The data were anonymised for the purposes of this study (retaining only information about the respondent's semester). The study was approved by the Institutional Review Board at the Charité Medical University Hospital of Berlin (EA1/101/19).

ORCID

Juliane E. Kämmer https://orcid.org/0000-0001-6042-8453
Wolf E. Hautz https://orcid.org/0000-0002-2445-984X

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):1094-1102.
2. Arnold L, Willoughby TL, Calkins E. Self-evaluation in undergraduate medical education: a longitudinal perspective. *J Med Educ*. 1985;60(1):21-28.
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(1):177-184.
4. Sargeant J, Armson H, Chesluk B, et al. The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med*. 2010;85(7):1212-1220.
5. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*. 2005;80(10 Suppl):S46-S54.
6. Eva KW, Regehr G. Exploring the divergence between self-assessment and self-monitoring. *Adv Health Sci Educ Theory Pract*. 2011;16(3):311-329.
7. Eva KW, Regehr G. 'I'll never play professional football' and other fallacies of self-assessment. *J Contin Educ Health Prof*. 2008;28(1):14-19.
8. 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):1019-1026.
9. Eva KW, Regehr G. Knowing when to look it up: a new conception of self-assessment ability. *Acad Med*. 2007;82(Suppl):S81-S84.
10. Rangel RH, Möller L, Sitter H, Stibane T, Strzelczyk A. Sure, or unsure? Measuring students' confidence and the potential impact on patient safety in multiple-choice questions. *Med Teach*. 2017;39(11):1189-1194.
11. Tweed M, Purdie G, Wilkinson T. Low performing students have insightfulness when they reflect-in-action. *Med Educ*. 2017;51(3):316-323.
12. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med*. 2008;121(5):S2-23.
13. Royal College of Physicians and Surgeons of Canada. CanMEDS 2000: extract from the CanMEDS 2000 Project Societal Needs Working Group Report. *Med Teach*. 2000;22(6):549-554.
14. 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):838-846.
15. Friedman CP, Gatti GG, Franz TM, et al. Do physicians know when their diagnoses are correct? Implications for decision support and error reduction. *J Gen Intern Med*. 2005;20(4):334-339.
16. Kampmeyer D, Matthes J, Herzig S. Lucky guess or knowledge: a cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th year medical students. *Adv Health Sci Educ Theory Pract*. 2015;20(2):431-440.
17. Kolbitsch J, Ebner M, Nagler W, Scerbakov N. Can confidence assessment enhance traditional multiple-choice testing? Conference on Interactive Computer Aided Learning, ICL2008, 23-26 September 2008, Villach.
18. 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):311-323.

19. Moulton C, Regehr G, Lingard L, Merritt C, MacRae H. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med*. 2010;85(10):1571-1577.
20. Dunning D. The Dunning-Kruger effect: on being ignorant of one's own ignorance. *Adv Exp Soc Psychol*. 2011;44:247-296.
21. Eva KW. On the generality of specificity. *Med Educ*. 2003;37(7):587-588.
22. Osterberg K, Kölbel S, Brauns K. The progress test Medizin. *GMS J Med Educ*. 2006;23(3):Doc46.
23. Hautz WE, Schubert S, Schauber SK, et al. Accuracy of self-monitoring: does experience, ability or case difficulty matter? *Med Educ*. 2019;53(7):735-744.
24. Nouns ZM, Georg W. Progress testing in German speaking countries. *Med Teach*. 2010;32(6):467-470.
25. Huberty CJ, Morris JD. Multivariate analysis versus multiple univariate analyses. *Psychol Bull*. 1992;105(2):302-308.
26. Schüttpelz-Brauns K, Kadmon M, Kiessling C, Karay Y, Gestmann M, Kämmer JE. Identifying low test-taking effort during low-stakes tests with the new Test-taking Effort Short Scale (TESS) – development and psychometrics. *BMC Med Educ*. 2018;18(1):101.
27. Schüttpelz-Brauns K, Hecht M, Hardt K, Karay Y, Zupanic M, Kämmer JE. Institutional strategies related to test-taking behavior in low stakes assessment. *Adv Health Sci Educ Theory Pract*. 2019;doi: 10.1007/s10459-019-09928-y. [Epub ahead of print.]
28. Wise SL, DeMars CE. Examinee non-effort and the validity of program assessment results. *Educ Assess*. 2010;15(1):27-41.
29. Kämmer, JE, Hautz, WE, & März, M. Self-monitoring accuracy does not increase throughout undergraduate medical education. 2020. Retrieved from osf.io/uhjad. 10.17605/OSF.IO/UHJAD.
30. Fischhoff B, Slovic P, Lichtenstein S. Knowing with certainty: the appropriateness of extreme confidence. *J Exp Psychol Hum Percept Perform*. 1977;3(4):552-564.
31. 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):1952-1958.
32. 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):575-584.
33. Ryan A, McColl GJ, O'Brien R, et al. Tensions in post-examination feedback: information for learning versus potential for harm. *Med Educ*. 2017;51(9):963-973.
34. Brydges R, Nair P, Ma I, Shanks D, Hatala R. Directed self-regulated learning versus instructor-regulated learning in simulation training: self-regulated learning on a simulator. *Med Educ*. 2012;46(7):648-656.
35. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004;79(10 Suppl):S70-S81.
36. de Bruin ABH, van Gog T. Improving self-monitoring and self-regulation: from cognitive psychology to the classroom. *Learn Instr*. 2012;22(4):245-252.

How to cite this article: Kämmer JE, Hautz WE, März M. Self-monitoring accuracy does not increase throughout undergraduate medical education. *Med Educ*. 2020;54:320–327. https://doi.org/10.1111/medu.14057