Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 4 (2018) 300–307

www.elsevier.com/locate/hpe

The Effect of Performance Standards and Medical Experience on Diagnostic Calibration Accuracy

Marloes L. Nederhand[a,*], Huib K. Tabbers[a], Ted A.W. Splinter[b], Remy M.J.P. Rikers[a,c]

[a]Department of Human Learning and Performance, Institute of Psychology, Erasmus University Rotterdam, The Netherlands
[b]Department of Internal Medicine, Erasmus Medical Center, The Netherlands
[c]The Roosevelt Center for Excellence in Education, University College Roosevelt, Utrecht University, The Netherlands

Received 21 December 2017; accepted 29 December 2017
Available online 8 January 2018

Abstract

Purpose: Medical doctors do not always calibrate accurately in terms of their diagnostic performance, which means that their evaluation of their diagnosis differs from their actual performance. Inaccurate calibration can lead to maltreatment and increased health care costs. This study was conducted to investigate whether calibration accuracy can be improved among both board certified medical specialists and medical students by providing them with a simple form of feedback (i.e., performance standards). We expected that performance standards would enhance calibration accuracy. Furthermore, we expected that medical specialists would overall be better calibrated than medical students.
Methods: Medical specialists (n = 42) and medical students (n = 43) diagnosed three clinical cases and rated their own performance, after which they did or did not receive standards (i.e., the correct diagnoses). All participants were then tested: they had to diagnose three new cases and had to rate their performance without receiving diagnostic feedback
Results: In support of our hypotheses, findings indicate that both students and specialists who received performance standards calibrated better than students and specialists who did not receive standards. Furthermore, medical specialists calibrated better than medical students.
Discussion: This study shows that providing simple forms of feedback constitute effects on calibration accuracy on new tasks.

Keywords: Calibration; Feedback; Standards; Expertise; Self-assessment

1. Introduction

Professionals often experience difficulties in adequately estimating the quality of their own diagnostic performance: they tend to overestimate themselves.[1–4] Providing an accurate estimate of one's performance is referred to as calibration accuracy.[5,6] Being able to calibrate accurately is especially important in a dynamic domain as medicine, where insights about proper treatment quickly change, and doctors have to make sure that they keep up with these developments to ensure the best treatment for their patients. However, in complex jobs, like that of a medical professional, calibration is difficult due to the little detailed and

[*]Correspondence to: Erasmus University Rotterdam, Woudestein, Mandeville Building, T13-1, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

E-mail address: m.l.nederhand@essb.eur.nl (M.L. Nederhand).

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2017.12.008

immediate performance feedback that is provided in their work.[7] Consequently, such contexts increase the susceptibility to make decisions based on incorrect or outdated knowledge or skills, resulting in inefficient or ineffective treatment, increased health care costs, and most importantly, harm to the patient.[8,9]

Previous studies have therefore argued that medical professionals can greatly benefit from becoming better calibrators to increase their performance.[10] To date, however, studies on how to improve calibration accuracy in medicine are scarce.[10] In the current study we aim to address this gap in the literature by investigating how calibration accuracy can be increased among medical professionals. We examined the effect of providing a simple form of feedback: performance standards. Studies in educational science have shown that such standards effectively help individuals become more aware of their performance.[11–13] However, whereas existing studies both within and outside medicine predominantly focused on students' calibration who bring little experience to the task at hand, the present study involved participants with high levels of experience. So besides using medical students as participants, this study also investigated calibration accuracy of board certified medical specialists.

1.1. *Improving calibration accuracy in medicine*

Miscalibration is explained by the notion that individuals generally have difficulty estimating their performance when valid cues are absent.[11,14] To improve calibration accuracy, individuals should therefore be given the opportunity to compare their own performance to a standard.[6,11,15] By comparing one's own performance to a standard, individuals gain insight in the match or mismatch of their estimated performance and actual performance.[3,16] This awareness, in turn, improves calibration accuracy.[11,14,17,18]

If providing standards indeed helps to enhance calibration accuracy, we might expect that individuals who receive standards also show better calibration accuracy on similar future tasks. That is, the capability to calibrate may transfer to new situations, where standards are not immediately available. In support of this assumption, two recent studies among psychology students showed that providing standards indeed improved calibration on a subsequent task where standards were not present.[13,19] To date, however, studies on how standards improve calibration in medicine are lacking. So based on findings outside medicine we asked the question whether providing medical professionals with standards would improve calibration on subsequent diagnostic tasks.

1.2. *Experience differences in calibration accuracy*

In studies on student calibration, top-performing students are typically compared to low-performing students.[20] These studies showed that students who perform well are generally also better calibrated; low performers, however, show poor calibration and tend to overestimate their performance. Although these studies provide insights in calibration accuracy between students who differ from each other in terms of task performance, generalization to groups that differ more substantially in experience remains unclear. For example, whereas task performance of medical students within the same cohort can vary, their general diagnostic experience is relatively similar. This raises the question whether comparing an unexperienced group of medical students to highly experienced physicians—who have received much more feedback on the accuracy of their medical diagnoses over the years—leads to the same results.

To the knowledge of the authors, there is only one study conducted in medicine on calibration in which different experience levels were included. Friedman et al. investigated the calibration accuracy of medical students, internal medicine residents, and internists. Contrary to their expectation, results showed that instead of the internists, medical students were the most accurate calibrators. However, Friedman and colleagues questioned the validity of their findings because students had much more difficulty solving the cases than the residents and internists. They argued that it would have been better if the participants were challenged with less difficult cases, making the diagnostic task equally understandable for both the students and the specialists. The question therefore remains whether medical specialists would calibrate better than medical students when both groups are provided with diagnostic cases both the specialists and students can solve.

1.3. *Present study*

The present study investigated whether providing performance standards can enhance calibration accuracy on subsequent diagnostic tasks among medical specialists and medical students, and whether medical specialists calibrate better than medical students. The specialists and students were randomly divided in two groups: The first group received performance standards (the correct diagnoses) after diagnosing three clinical case, while the second group did not receive any standards. Subsequently, all participants received three entirely new cases and then their calibration was tested again. However, this time no standards were provided to both groups. To make sure the diagnostic tasks were equally understandable for both the students and the specialists, all participants were provided with general clinical cases, instead of cases from one type of specialty (e.g., internal medicine as in the study of Friedman et al.). The cases used in our study have been shown to be suited for both students and specialists.[21,22]

We expected that providing standards on the first three cases would enhance calibration accuracy on the new cases. Furthermore, based on findings that high or experienced performers calibrated better than low or inexperienced performers, we expected medical specialists to calibrate better than the students.

Note the difference between improving diagnostic performance and calibration accuracy. Providing standards unlikely will improve diagnostic performance because this type of feedback is non-elaborate. We therefore did not expect to find any effect on diagnostic performance. Improving calibration accuracy is, however, considered a first step to eventually improve (diagnostic) performance: being aware of a mismatch between estimated performance and actual performance causes control behavior,[23] such as requesting additional tests or asking help from colleagues if poor performance is detected. So the main focus of the current study was on calibration accuracy: knowing whether the diagnosis is correct or not.

2. Method

2.1. Participants and design

Eighty-five participants were recruited, 43 second-year medical students and 42 medical specialists. The medical specialists (30 males, 12 females) were all board certified in their specialty and had a mean age of 44.73 ($SD=7.61$). They were specialized in more than 20 different medical domains, such as internal medicine, neuroscience, or cardiology. Their mean years of clinical experience was 11.83 ($SD=7.95$). The specialists were approached and tested during a professional training program by the Erasmus Medical Center and did not receive any compensation for their participation. The second-year medical students (18 males and 25 females) studied at the Erasmus University Medical School and were recruited during, and tested after, one of their lectures. Their mean age was 20.88 ($SD=2.40$). By participating, the students could join a lottery to win four small prizes. All participants provided informed consent and the study was approved by our institutional review board.

2.2. Materials and procedure

Participants were randomly assigned to a group that received standards and one group that did not receive standards. The standard group consisted of 44 participants (22 specialists and 22 students) and the no-standard group consisted of 41 participants (20 specialists and 21 students). Participants received a total of six clinical cases. Each case consisted of a short clinical scenario in which the following information was given: A patient's medical history, present complaints, physical examination, and additional investigations (e.g., lab data, ECGs). Cases were presented in a booklet in which each case was printed on a separate page with some space intentionally left blank where the participants could write down their diagnosis and provide their performance estimate (see Appendix). With the exception of receiving standards or not, the procedure for the two groups was the same. All participants received a booklet with the cases that also contained a short introduction. After reading the introduction, participants continued with diagnosing the first three cases one by one. It was stressed that the participants should be as specific as possible in their diagnoses. After providing a diagnosis, participants rated the confidence they had in their diagnosis (i.e., their performance estimate) on a 10-point Likert scale (1: ‘very unconfident’ to 10: ‘very confident’).

Depending on the group they were in, participants received a performance standard on the next page after writing down each diagnosis. The standards consisted of the confirmed diagnosis for each of the first three cases, respectively: Aneurism of the aortic artery (threatening rupture); herpes zoster; and nervous abdominal pain. The participants in the control group (i.e., no-standard) were not informed about the correct diagnosis, but had to do a filler task (i.e., answer the question how familiar they were with the case).

After the intervention, all participants were tested on three completely new cases. The confirmed diagnoses associated with the test cases were: meningitis or encephalitis as a complaint of mumps; kidney stones (colic); and epidural hematoma. All cases have been used in previous studies.[21,22] During the test phase, there were no differences between both conditions: all participants diagnosed the new cases without any standards, and had to provide a performance estimate after each diagnosis. The three test cases were used to investigate whether having received standards leads to improved calibration accuracy on new cases where these standards are missing.

2.3. Analysis

Diagnostic accuracy was scored by comparing the diagnoses of the participants with the confirmed diagnosis for each case.[21,22] Diagnoses were scored on a three-point scoring grid (0 = incorrect, 0.5 = partly correct, and 1 = fully correct) and the scoring was double checked by a medical specialist and a professor of internal medicine. [24] There were no differences between raters.

A difficulty when using Likert scales is that participants do not tend to use the extreme response options.[25] Because reluctance to use extreme response options leads to miscalibration in our experiment, we adjusted the Likert scale. Low confidence scores (1 to 3), were given the value of 0 and high confidence scores (8 to 10) were given the value of 1. The confidence scores that indicate average confidence in an answer (4 to 7), were given a value of 0.5. This converted the confidence scale into a three-point scale so it would correspond to the three-point diagnostic performance scale, enabling us to calculate calibration accuracy.

To calculate the calibration accuracy of the participants, the absolute difference between performance and confidence was calculated, and the average of the three difference scores was used as the calibration accuracy score.[11,15,26–28] Perfect calibration accuracy was thus indicated by a score of 0 (perfect match between confidence and performance scores in all three cases) and strongest miscalibration had a score of 1 (largest mismatch between confidence and performance scores in all three cases).

We analyzed our data with IBM SPSS Statistics, version 23 (IBM, New York). To compare the feedback and experience conditions, a 2×2 univariate analysis of variance (ANOVA) was conducted with Standards (Yes vs. No) and Experience level (Specialists vs. Students) as independent variables and diagnostic performance and calibration accuracy as dependent variables. A significance level of .05 was set for all analyses.

Table 1
Mean calibration scores in the test phase. The range of the calibration scores is from 0 (perfect calibration) to 1 (no calibration).

	n	Calibration scores	
		M (*SE*)	95% CI
Standards			
2nd-year students	22	0.22 (0.04)	[0.14, 0.30]
Medical specialists	22	0.14 (0.04)	[0.06, 0.22]
Total	44	0.18 (0.03)	[0.13, 0.23]
No Standards			
2nd-year students	21	0.33 (0.04)	[0.25, 0.41]
Medical specialists	20	0.18 (0.04)	[0.10, 0.27]
Total	41	0.26 (0.03)	[0.19, 0.33]

3. Results

3.1. Diagnostic performance

We first checked diagnostic performance (a score between 0 and 1) of both specialists and students, and between the two experimental conditions. We tested whether medical specialists showed better diagnostic performance than the students over the total of six cases. Results showed that diagnostic performance differed significantly between medical specialists ($M = .91$, $SD = .12$) and medical students ($M = .66$, $SD = .18$), $F(81) = 55.81$, $p < .001$, $\eta_p^2 = .408$. Medical specialists solved more cases correctly than medical students. It is important to note, however, that although students' diagnostic performance was significantly lower than that of the specialists, their level of performance was still high (66%). So, as intended the cases used in our study were solvable for both specialists and students.

The diagnostic performance over all six cases did not differ between the group that received standards on the first three cases ($M = .79$, $SD = .20$) and the group that did not receive any standards ($M = .79$, $SD = .20$), $F(81) = .03$, $p = .858$, $\eta_p^2 < .001$. There was no statistically significant interaction on diagnostic performance between standards and experience level, $F < 2$.

3.2. Calibration accuracy among specialists and students

To test whether medical students and medical specialists differed in terms of their calibration accuracy, we analyzed whether there was a main effect of experience on calibration accuracy over all six clinical cases. The main effect of experience was statistically significant, $F(83) = 46.32$, $p < .001$, $\eta_p^2 = .358$, with medical specialists having a better calibration accuracy as indicated by the lower mean score on calibration accuracy ($M = .19$, $SD = .14$) than the students ($M = .39$, $SD = .12$). This result supports our hypothesis that specialists calibrate better than students.

3.3. Enhancing calibration by receiving standards

Furthermore, we analyzed whether providing standards was associated with better calibration accuracy on the last three test cases. The main effect of standards was statistically significant, $F(1,81)=4.00$, $p=.049$, $\eta_p^2=.05$ (see Table 1 for descriptives). Participants in the standard group calibrated better on the new test cases ($M=.18$, $SD=.17$) than participants who did not receive standards ($M=.26$, $SD=.22$). Our hypothesis that providing standards can improve calibration accuracy on subsequent tasks (i.e., calibration becomes closer to zero) is therefore supported. Finally, we tested the interaction between experience level and performance standards. This interaction was not statistically significant, $F(1,81)=.69$, $p=.41$, $\eta_p^2=.01$. Standards similarly improved calibration accuracy for both specialists and students.

4. Discussion

With this study we investigated the effect of feedback (i.e., performance standards) on calibration accuracy of board certified medical specialists and medical students. We hypothesized that medical specialists would overall calibrate better than medical students because the specialists have received over the years much more feedback on the accuracy of their medical diagnoses. Furthermore, research shows that standards can be used to enhance calibration accuracy because such standards help individuals to become aware of the (mis)match between their own performance and the required performance.[11,14] We therefore predicted that standards in the form of a correct diagnosis would enhance diagnostic calibration accuracy. Our results confirm both hypotheses and hence have several educational and theoretical implications.

4.1. Calibration accuracy among specialists and students

We tested whether medical specialists are more accurate calibrators than students. In support of our expectation, medical specialists calibrated better than medical students. The specialists had a mean of eleven years of experience with treating patients. Consequently, they had many years of experience with diagnosing patients and monitoring their diagnoses. Our results show that this experience helps specialists to adequately estimate their performance on clinical cases that are not necessarily in their own domain of expertise.

Although there are many studies that argue it is important to individually differentiate in calibration accuracy,[e.g.20] little attention has been paid to differences in experience. When studies did include experience level, studies often focused on small variations. For example, students from different grades were compared.[29] The current study therefore adds to the existing literature by included two groups that substantially differ in their clinical experience.

4.2. Standards to improve calibration accuracy

Besides investigating the effect of experience on calibration accuracy, we tested the effect of standards. Findings indicate that receiving standards is associated with better calibration on new tasks (i.e., clinical cases). The results of this study are therefore in line with Nederhand et al.[13,19] and among the first to show that providing standards can help individuals to also enhance their calibration on subsequent new tasks. Our results also show that standards helped students equally well as specialists to enhance their calibration accuracy.

As intended, whereas standards affected calibration accuracy, participants did not show better diagnostic performance after receiving standards. This is because using non-elaborated forms of feedback are little effective at improving performance directly.[30] However, because standards are proven effective to enhance calibration, (diagnostic) performance can indirectly be improved.[23] For instance, when a physician knows he or she performs poor (i.e., calibration is accurate while but diagnostic performance is low), he or she can take steps to overcome this poor performance by for example asking extra help. In other words, the calibration accuracy helps them to take steps that will ultimately improve their diagnostic performance. Vice versa, better calibration accuracy among physicians that already perform well is also beneficial. For example, if a physician performs very well but is unaware of that, he or she will request too many additional tests, costing both the hospital and patient time and money. It is therefore promising that our study shows that even simple forms of feedback help to improve calibration accuracy.

4.3. Limitations and future directions

While our study provides new insight in how to improve calibration accuracy in medicine, it also has some limitations. Although there are many theoretical and empirical reasons to assume that accurate calibration also enhances diagnostic performance, we did not investigate whether our participants would indeed engage in corrective actions after they received standards. An important direction for future research is therefore to investigate the steps that are taken after a mismatch between estimated performance and actual performance is detected.

A second limitation in our study is that although we made a first attempt to measure the longitudinal effect of standards, our diagnostic test cases followed directly on the first three diagnostic cases in which participants received standards. Our design thus provides insight in whether the capability to calibrate accurately can transfer to new cases, but the effect over time still remains largely unclear. For example, when we would have asked our participants to diagnose new cases one week after our intervention, would there still have been differences between groups? Future research could investigate this question. Related to this issue, the optimal amount of feedback could further be explored. For example, it is not unlikely that, instead of our three feedback moments, more feedback is needed to constitute effects over time.

Finally, we used both board certified medical specialists and medical students as participants in our study to investigate large experience differences. As intended, the students had some degree of expertise as they were able to solve the clinical cases. In other words, although students clearly differed from medical specialists in terms of diagnostic experience, they were no full novices on the task. It is therefore important to mention here that it remains unclear whether standards would also help full novices improve their calibration and our results must be treated with caution when generalizing to a group of novices.

4.4. Conclusion and implications

Because medicine is a dynamic domain, life-long-learning of clinicians is necessary. Being able to improve oneself continuously requires self-monitoring —clinicians have to be aware of their own performance quality so they can discriminate between things that go well and things that have to be (further) developed. The current study highlights an under-studied topic in medical education: Although many studies on calibration accuracy were conducted with the aim to generalize to clinical practitioners, this group is hardly ever included as participants. We have shown that even a relatively simple feedback intervention in the form of correct diagnoses can help both medical specialists and medical students to improve their calibration accuracy on new diagnostic tasks (i.e., their awareness of their actual diagnostic performance).

Acknowledgements

The authors wish to thank all participants for their participation in our study. Many thanks to Liesbeth Dirksen-de Tombe and Lennart Rem for their valuable help and Amanda Binnendijk and John Tjen for their help during the training. Thank you Matthijs de Hoog for approving testing during the training sessions. Furthermore, the authors wish to thank Eugène Custers for providing us with the clinical cases. Finally, the authors like to dedicate their study to the memory of Prof. Dr. Splinter, one of our co-authors, who recently passed away after many years of illness. He has been a source of inspiration to all of us.

Disclosure

Ethical approval

Ethical approval has been granted from the Ethical Committee of the Department of Psychology, Education, and Child Studies (16 August 2016).

Funding

This research was funded by a Research Excellence Initiative grant from Erasmus University Rotterdam.

Other disclosures

None.

Appendix A

Description casus

01. Man, age 47, married, 3 children.

02. Occupation: storekeeper.

03. Medical history: bronchitis at age 30.

04. Had his leg broken 6 years ago, as a consequence of a car accident.

05. Four years ago: treated with medicaments for kidney stones.

06. Some of his relatives are known to have coronary disease and diabetes mellitus.

07. His wife rings up, asks the physician for an immediate visit:

08. Just like a few years ago, her husband is rolling across the room because of the pain.

09. He is also vomiting almost continuously.

10. When the physician arrives, the pain has just subsided. The patient is sitting on the sofa and recovering a bit.

11. He complains about having had a convulsive abdominal pain abreast of the navel, at the left side.

12. The pain is radiating to his groin.

13. The pain emerges very suddenly, and then gradually subsides. During an attack he almost can't stand it.

14. Earlier that day he had already seen some blood in his urine, but had no pain at the time.

15.He reports having measured 37.8° (Centigrade) temperature.

What diagnosis would you give on basis of the previous information?

How confident are you that your diagnosis is correct?

Please encircle your estimation.

Very unconfident									*Very confident*
1	2	3	4	5	6	7	8	9	10

References

1. 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 http://dx.doi.org/10.1001/jamainternmed.2013.10081.
2. 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 http://dx.doi.org/10.1111/j.1525-1497.2005.30145.x.
3. Martin D, Regehr G, Hodges B, McNaughton N. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. *Acad Med* 1998;73:1201–1206.
4. 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. *J Am Med Assoc* 2006;296(9):1094–1102 http://dx.doi.org/10.1001/jama.296.9.1094.
5. Lichtenstein S, Fischhoff B, Phillips LD. Calibration of probabilities: the state of the art to 1980. In: Kahneman D, Slovic P, Tversky A, editors. *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press; 1982. p. 306–334.
6. Dunlosky J, Thiede KW. Four cornerstones of calibration research: why understanding students' judgments can improve their achievement. *Learn Instr* 2013;24:58–61 http://dx.doi.org/10.1016/j.learninstruc.2012.05.002.
7. Ericsson KA. Acquisition and maintenance of medical expertise: a perspective from the expert-performance approach with deliberate practice. *Acad Med* 2015;90(11): 1471–1486 http://dx.doi.org/10.1097/ACM.0000000000000939.
8. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med* 2008;121(5): 2–23 http://dx.doi.org/10.1016/j.amjmed.2008.01.001.
9. Blendon RJ, DesRoches CM, Brodie, M, et al. Views of practicing physicians and the public on medical errors. *N Engl J Med* 2002;347 (24):1933–1940 http://dx.doi.org/10.1056/NEJMsa022151.
10. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005;80(10): S46–S54 http://dx.doi.org/10.1097/00001888-200510001-00015.
11. Rawson KA, Dunlosky J. Improving students' self-evaluation of learning for key concepts in textbook materials. *Eur J Cogn Psychol* 2007;19(4-5):559–579 http://dx.doi.org/10.1080/09541440701326022.
12. Dunlosky J, Hartwig MK, Rawson KA, Lipko AR. Improving college students' evaluation of text learning using idea-unit standards. *Q J Exp Psychol* 2011;64(3):467–484 http://dx.doi.org/10.1080/17470218.2010.502239.
13. Nederhand ML, Tabbers HK, Rikers RMJP *Learning to Calibrate: Providing Standards to Improve Calibration Accuracy for Different Performance Levels*. Unpublished results.
14. Koriat A. Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *J Exp Psychol Gen* 1997;126(4):349–370 http://dx.doi.org/10.1037/0096-3445.126.4.349.
15. Lipko AR, Dunlosky J, Hartwig MK, Rawson KA, Swan K, Cook D. Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *J Exp Psychol Appl* 2009;15(4):307–318 http://dx.doi.org/10.1037/a0017599.
16. 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 http://dx.doi.org/10.1097/ACM.0b013e3181d85a4e.
17. Butler DL, Winne PH. Feedback and self-regulated learning: a theoretical synthesis. *Rev Educ Res* 1995;65(3):245–281.
18. Zimmerman BJ. Attaining self-regulation: a social cognitive perspective. In: Boekaerts M, Pintrich PR, Zeidner M, editors. *Handbook of Self-Regulation*. Cambridge, MA: Academic Press; 2000. p. 13–40.
19. Nederhand ML, Tabbers HK, Abrahimi H, Rikers RMJP *Providing Standards both with and without idea-units to improve calibration accuracy*. Unpublished results.
20. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to self-assessments. *J Pers Soc Psychol* 1999;77(6):1121–1134 http://dx.doi.org/10.1037/0022-3514.77.6.1121.

21. Custers EJ, Boshuizen HP, Schmidt HG. The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Mem Cogn* 1996;24(3):384–399 http://dx.doi.org/10.3758/BF03213301.
22. Custers EJFM. *The Development and Function of Illness Scripts: Studies on the Structure of Medical Diagnostic Knowledge [dissertation].* Rotterdam: Erasmus University Rotterdam; 1995.
23. Dunlosky J, Rawson KA. Overconfidence produces underachievement: inaccurate self evaluations undermine students' learning and retention. *Learn Instr* 2012;22(4): 271–280 http://dx.doi.org/10.1016/j.learninstruc.2011.08.003.
24. Schmidt HG, Mamede S, Berge K Van Den, Gog T Van, Saase JLCM Van, Rikers RMJP. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Acad Med* 2014;89(2): 285–291 http://dx.doi.org/10.1097/ACM.0000000000000107.
25. Landy FJ, Conte JM. *Work in the 21st Century: An Introduction to Industrial and Organizational Psychology*, 3rd ed., Hoboken, New Jersey: John Wiley & Sons; 2009.
26. Pajares F, Graham L. Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemp Educ Psychol* 1999;24(2):124–139 http://dx.doi.org/10.1006/ceps.1998.0991.
27. Pajares F, Miller MD. Mathematics self-efficacy and mathematical problem solving: implications of using different forms of assessment. *J Exp Educ* 1997;65(3):213–228 http://dx.doi.org/10.1080/00220973.1997.9943455.
28. Schraw G, Potenza MT, Nebelsick-Gullet L. Constraints on the calibration of performance. *Contemp Educ Psychol* 1993;18(4): 455–463 http://dx.doi.org/10.1006/ceps.1993.1034.
29. García T, Rodríguez C, González-Castro P, González-Pienda JA, Torrance M. Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacogn Learn* 2016;11(2): 139–170 http://dx.doi.org/10.1007/s11409-015-9139-1.
30. Archer JC. State of the science in health professional education: effective feedback. *Med Educ* 2010;44(1): 101–108 http://dx.doi.org/10.1111/j.1365-2923.2009.03546.x.