Adv in Health Sci Educ (2015) 20:431–440
DOI 10.1007/s10459-014-9537-1

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

Daniela Kampmeyer · Jan Matthes · Stefan Herzig

Received: 5 March 2014 / Accepted: 24 July 2014 / Published online: 8 August 2014

Abstract Multiple-choice-questions are common in medical examinations, but guessing biases assessment results. Confidence-based-testing (CBT) integrates indicated confidence levels. It has been suggested that correctness of and confidence in an answer together indicate knowledge levels thus determining the quality of a resulting decision. We used a CBT approach to investigate whether decision quality improves during undergraduate medical education. 3rd- and 5th-year students attended formative multiple-choice exams on pharmacological issues. Students were asked to indicate their confidence in a given answer. Correctness of answers was scored binary (1-correct; 0-wrong) and confidence levels were transformed to an ordinal scale (guess: 0; rather unsure: 0.33; rather sure: 0.66; very sure: 1). 5th-year students gave more correct answers (73 ± 16 vs. 49 ± 13 %, $p < 0.05$) and were on average more confident regarding the correctness of their answers (0.61 ± 0.18 vs. 0.46 ± 0.13, $p < 0.05$). Correlation of these parameters was stronger for 5th-year students ($r = 0.81$ vs. $r = 0.52$), but agreement of confidence and correctness ('centration') was lower. By combining the Bland-and-Altman approach with categories of decision-quality we found that 5th-year students were more likely to be 'well-informed' (41 vs. 5 %), while more 3rd-students were 'uninformed' (24 vs. 76 %). Despite a good correlation of exam results and confidence in given answers increased knowledge might be accompanied by a more critical view at the own abilities. Combining the statistical Bland-and-Altman analysis with a theoretical approach to decision-quality, more advanced students are expected to apply correct beliefs, while their younger fellows are rather at risk to hesitate or to act amiss.

D. Kampmeyer · J. Matthes (✉) · S. Herzig
Department of Pharmacology, University of Cologne, Priv.-Doz. Dr. med. Jan Matthes, Gleueler Strasse 24, 50931 Cologne, Germany
e-mail: jan.matthes@uni-koeln.de

D. Kampmeyer
e-mail: daniela.kampmeyer@gmail.com

S. Herzig
e-mail: stefan.herzig@uni-koeln.de

Keywords Bland and Altman-analysis · Confidence-based testing · Multiple-choice-question

Background

According to Sveiby (1997) 'knowledge is the capacity to act'. This capacity is crucial in a doctor's daily life, as she or he has to make many decisions a day, some of them simple, others far-reaching, therapy related and essential for the patients' well-being. So is for instance drug therapy, which is present in almost all fields of medicine and thus are medication errors. Adverse drug reactions are often based on dosing errors or drug interactions (Egger et al. 2003; Keers et al. 2013). Tobaiqy et al. reported that 30 % of foundation year 1 doctors in the UK rated their knowledge of clinical pharmacology as poor or even worse (Tobaiqy et al. 2007). Similarly Cologne medical students being closest to their final (practical) year judged their expertise significantly lower than their fellows in earlier years (Matthes et al. 2013). Since students' knowledge increases during their medical studies (Albano et al. 1996; Osterberg et al. 2006) these findings support the idea that expertise and performance are not only a matter of knowledge but also of the confidence in it. This assumption dates back to the Chinese philosopher Confucius describing knowledge already around 500 b.c. to be both 'knowing a thing and the recognition of knowing it' (Confucius and Legge 1971). Only if recognized, knowledge can be transferred to an intentional action. Thus, recognition is one important metacognitive jigsaw piece of medical competence. It is tempting to add another level to Miller's pyramid (Miller 1990): knows that she or he knows—respectively knows that she or he does not know. In medical examinations however we tend to oversimplify the complex concept of knowledge considering simply the correctness of a given answer. In Multiple-Choice-Questions (MCQs) we even concede points to students, who luckily guessed the right answer, leading to 'an artificial inflation of marks' (Bush 2006) and a decrease of reliability (Zimmerman and Williams 2003). According to Hunt not only the correctness but also the confidence of an assumption determines the quality of an answer and thus the quality of the subsequent decision (Hunt 2003). Accordingly four categories of decision quality can be defined (Fig. 1): well informed, misinformed, uninformed, and partially informed. The crucial difference between these categories is that firstly—depending on the level of confidence—a decision will lead to an action or not and secondly—depending on its correctness—this may be beneficial or harmful. The significant role of confidence is considered in the approach of confidence-based testing, an advancement of multiple-choice questions, where students not only choose an answer, but additionally state how sure they are that their choice is the right one. Though this approach dates back to the early twentieth century (Henmon 1911) it took almost a century until it was introduced into summative exams within medical curricula (Gardner-Medwin and Gahan 2003). Among the different ways of how to score and interpret the confidence, correlation of correctness and confidence has predominantly been used (Kruger and Dunning 1999; Rippey and Voytovich 1985). Though for a given cohort a high correlation coefficient suggests a good agreement of the two parameters, no conclusion can be drawn to an individual student. Another way of comparing two different measures is by using the statistical approach by Bland and Altman, who stated that tools measuring the same variable should not only show a perfect correlation but also a perfect agreement of values (Bland and Altman 1986, 2012). As an

index for the agreement of correctness and confidence Leclercq proposed the so-called 'centration' (Leclercq 1983). The centration is calculated as the difference of confidence and correctness and allows a feedback on the degree of agreement, i.e. on over- or underestimation of knowledge. We here for the first time use the Bland–Altman approach to analyze the relation of correctness and confidence and to subsequently identify students' affiliation to Hunt's categories of being informed. By doing so we address the question whether more advanced (5th year) medical students reach a higher level of being informed compared to their younger (3rd year) fellows.

Methods

At the University of Cologne medical students have to attend two major courses on pharmacology, namely that on basic pharmacology (3rd year) and that on clinical pharmacology (5th year). In addition there are several interdisciplinary lectures as well as small-group sessions, spreading from year 1 to 6, dealing with pharmacological issues. During winter term 2011/2012, we provided voluntary, formative pharmacology tests to participants from either pharmacology course. We chose a computer based multiple-choice approach, executed 1 week before the respective final (summative) exam. The opportunity to take the formative test was announced verbally during lectures as well as by e-mail. Students were informed that the test was formative and part of an educational study.

Similar to the respective final exams the 3rd year test consisted of 40 MCQs and the one for 5th year students of 29. A single best answer had to be chosen from the five provided options per MCQ. Taking advantage of some topics covered in both courses, six questions were identical in both tests. Immediately after choosing an answer, students had to indicate their confidence in their choice. For each multiple-choice question answered correctly students received one point, while none for a wrong answer was rewarded. The confidence was indicated by choosing statements based upon Kolbitsch that for analysis were (semi-) quantified as follows (Kolbitsch et al. 2008):

I am guessing	0
I am rather uncertain	0.33
I am rather sure	0.66
I am very sure	1

We first analyzed an apparent relationship of correctness and confidence using Spearman's correlation coefficient. Applying Confucius' definition we assumed that ideally correctness and confidence of an answer could be regarded as two different measures of knowledge. Thus a student with a low level of correctness is expected to display a low level of confidence (i.e. rather towards 'I am guessing' = 0), while a student with a high level of correctness is expected to indicate higher confidence values (i.e. rather towards 'I am very sure' = 1). As one measure we thus defined the 'knowledge value' calculated as arithmetic mean of a student's correctness and confidence. According to the idea of Bland and Altman we as a measure of agreement used Leclercq's 'centration', i.e. the difference between confidence (ranging from 0 to 1) and correctness (ranging from 0 to 100 %). We then applied the Bland-and-Altman analysis using the two parameters 'knowledge value'

wrong answer		correct answer	
sure of correctness	unsure		sure of correctness
misinformed	uninformed	partially informed	well informed

Fig. 1 Categories of decision quality adapted from those proposed by Hunt (2003)

and 'centration'. Furthermore the values were used to estimate the quality of decision-making according to Hunt (2003), with 60 % of correct answers and a 66 % confidence defining the threshold between incorrect and correct as well as unconfident and confident, respectively.

Results

131 of 160 (82 %) 3rd year students attending the course on basic pharmacology and 78 of 145 (54 %) 5th year students attending the course on clinical pharmacology took part. Three students did not fulfill the inclusion criterion of answering at least 75 % of the questions and thus were excluded from our analysis.

The proportion of questions answered correctly was lower in the 3rd year cohort compared to 5th year students (0.49 ± 0.13 vs. 0.73 ± 0.16, $p < 0.05$). Assuming that wrong answers are more often associated with guessing than with being convinced of their correctness, we expected a lower proportion of correct answers to correspond with a lower level of confidence and vice versa. Indeed, using the score according to Kolbitsch (0, 0.33, 0.66, 1) the corresponding mean confidence values were calculated to be 0.46 ± 0.13 for 3rd year and 0.61 ± 0.18 for 5th year students ($p < 0.05$; Fig. 2a). Of note the correlation coefficient of correctness versus confidence was closer to unity for 5th year students (0.77; Fig. 2b).

Bearing in mind that a higher correlation does not necessarily imply a higher agreement of two parameters, we used the Bland and Altman approach for further analysis (Bland and Altman 1986, 2012). Therefore we used the two parameters 'centration' and 'knowledge value' that both consider correctness and confidence (cp. methods). Knowledge values of 5th year students were significantly higher compared to their 3rd year fellows (0.67 ± 0.16 vs. 0.47 ± 0.11; $p < 0.05$). 3rd year students showed a slight but statistically significant deviation of their mean centration towards negative values (-0.03 ± 0.13; Fig. 3a), indicating a weak discrepancy between actual knowledge and confidence. This deviation was more pronounced in the 5th year cohort (-0.12 ± 0.11; Fig. 3b). Thus, compared to 3rd year students agreement of correctness and confidence appeared to be lower in 5th year students suggesting that the latter underestimate their knowledge. This apparent underestimation can be explained by the fact that compared to 3rd-year students the portion of questions where 5th year students were convinced of being right but actually were wrong was decreased while the fraction of questions answered correctly but where students were unsure did not change (19.2 vs. 7.8 and 16 vs. 16.1 %, respectively) (Fig. 4). We now interpreted the data processed according to Bland and Altman by applying Hunt's

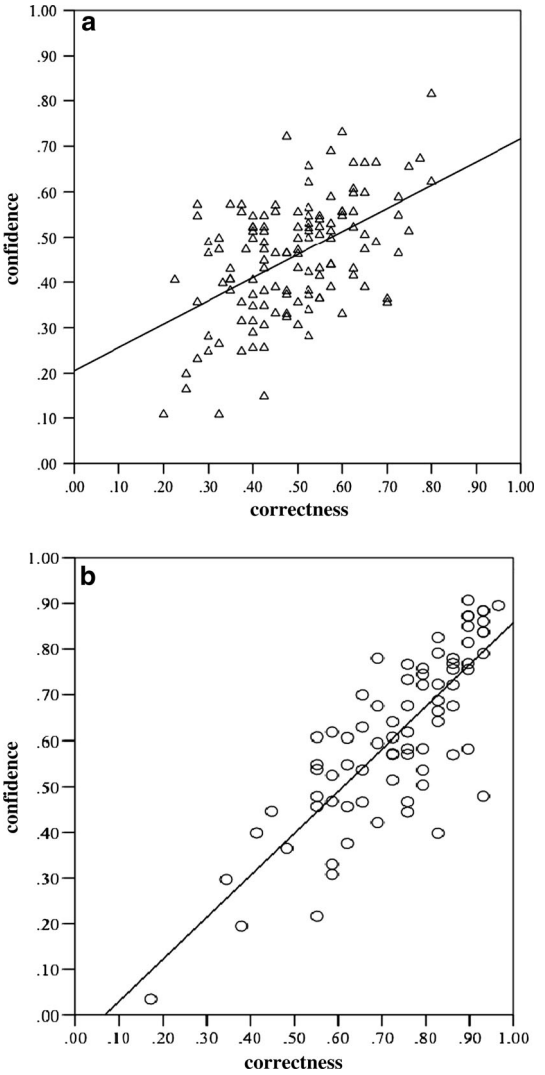

Fig. 2 Correlation of correctness (i.e. fraction of correctly answered multiple-choice questions) and the mean level of confidence as stated by individual 3rd **(a)** and 5th **(b)** year medical students

categories of decision quality. We differentiated between 1.) students who answered with sufficient correctness—considering a limit of 60 % corrects answers, which accords to the pass-fail limit in our summative exams—and those, who failed to reach this requirement, as well as between 2.) students who achieved a confidence level of ≥66 %—since according to Hunt then a subsequent application of knowledge is likely—and those, who answered rather unsure. By doing so, we found the vast majority (76 %) of the 3rd year students to be classified as 'uninformed' while this was the case for only 24 % of the 5th year students. Furthermore 41 % of students in the more advanced semester were 'well informed' compared to only 5 % in the 3rd year cohort (Fig. 5a, b).

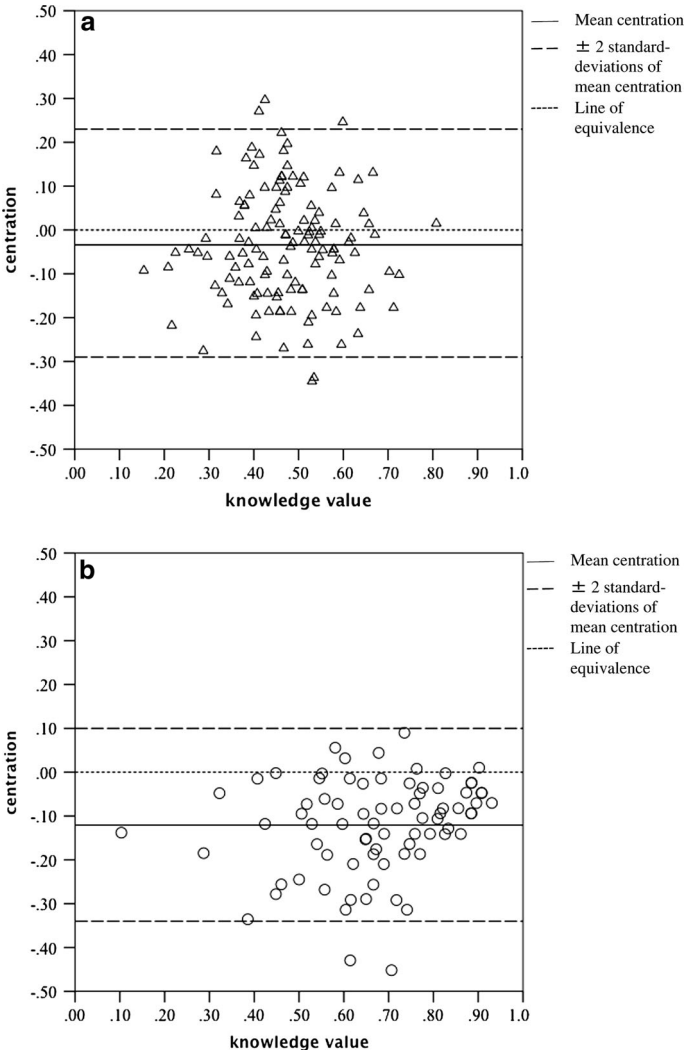

Fig. 3 Adapting an approach proposed by Bland and Altman (1986) we plotted knowledge levels [=(correctness plus confidence)/2] of individual 3rd **(a)** and 5th **(b)** medical students against the referring centration values [=(confidence minus correctness)]

Discussion

We found that compared to their younger fellows more advanced medical students gained higher levels of correctness of and confidence in answers within a formative test on pharmacological issues. Both cohorts (3rd and 5th year) showed a similar significant correlation of correctness and confidence. In contrast, our analysis according to Bland and Altman revealed that 5th year students seemed to underestimate their knowledge more clearly. In fact applying the criteria of Hunt's levels of decision-making, most of the 3rd year students turned out to be 'uninformed' while clearly more 5th year students were

Fig. 4 Distribution of answered questions according to the correctness of the answer and the confidence in that answer on the level of individual 3rd **(a)** and 5th **(b)** year students

a 3rd year students

	correct answer given	wrong answer given
rather sure or even convinced	30%	19%
rather unsure or even guessing	16%	35%

b 5th year students

	correct answer given	wrong answer given
rather sure or even convinced	57%	8%
rather unsure or even guessing	16%	19%

'well-informed'. Translating these categories into putative probability of knowledge application we conclude that most 3rd year students would hesitate to apply knowledge due to uncertainty or even act amiss due to being convinced of wrong answers, while significantly more 5th year students would apply correct content.

Simple correlation of correctness and confidence is likely to be biased. For example the Kruger-Dunning effect describes the phenomenon of weaker students overestimating their performance and better students underestimating themselves (Kruger and Dunning 1999). Furthermore a weaker correlation of correctness and confidence was found at higher educational levels (Friedman et al. 2001). To overcome these shortcomings of simple correlation we applied an approach of Bland and Altman originally developed to compare diagnostic tools. To our knowledge, the Bland and Altman method has not yet been used to analyze the results of confidence-based testing. Though correctness and confidence are different parameters, we assume on the one hand that confidence in someone's knowledge should increase with the level of correctness, i.e. the gain of knowledge, thus being the prerequisite for correct knowledge to be recognized and applied. On the other hand, low levels of knowledge should be accompanied by uncertainty, making someone to hesitate rather than to apply wrong knowledge and eventually—in the worst case—to cause harm.

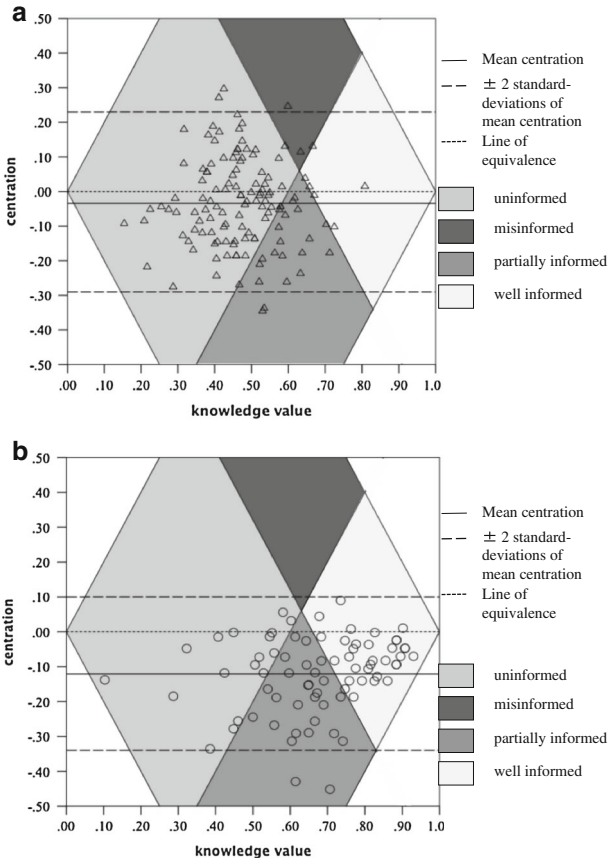

Fig. 5 Combination of the Bland–Altman plot (knowledge levels versus centration values) and the categories of decision quality proposed by Hunt (2003) regarding data of individual 3rd **(a)** and 5th **(b)** year medical students

Thus congruence of correctness and confidence as considered by the Bland and Altman approach should help to identify either those, who are not aware of their wrong knowledge and thus are likely to harm or those, who are not aware of being right and thus are likely to 'withhold' beneficial knowledge and/or actions.

There are some limitations of our study. Voluntary participation in our tests may have led to some selection bias and may have impaired comparability considering the different levels of participation (82 % of 3rd but 54 % of 5th year students). The formative character of our assessment may have led to a different attitude in stating the confidence of an assumption compared to summative conditions. On the other hand distress as a potential confounder in summative knowledge tests should have been lower in our setting. Since the applied questions were limited to pharmacological topics, transferability to other medical fields or even the clinical-practical context is limited.

Though benefits of multiple-choice-questions are well known, there are limitations discussed extensively since decades (Anderson 2004). To overcome the lack of information due to MCQ tests being widely limited to assessing mainly declarative knowledge we in our study combined the testing procedure of confidence-based testing with the statistical Bland and Altman approach to identify different levels of decision-quality of medical students. Our findings indicate that indeed the relationship of knowledge and confidence

changes during medical studies suggesting a movement towards 'convinced doing the right thing'. Procedures like the Bland and Altman analysis may lead to a more differential, and perhaps more meaningful interpretation of assessment results in medical education.

Acknowledgments The authors would like to thank Dr. med. Dipl.-Math. Hartmut Stützer for his advice on the statistical analysis of the data.

Conflict of interest Daniela Kampmeyer is working since January 2013 for the editorial staff of AMBOSS, an e-learning platform dealing with medical multiple choice questions, MIAMED enterprise, Cologne, Germany. The other authors declare that they have no competing interests.

References

Albano, M. G., Cavallo, F., Hoogenboom, R., Magni, F., Majoor, G., Manenti, F., et al. (1996). An international comparison of knowledge levels of medical students: The Maastricht Progress Test. *Medical Education, 30*(4), 239–245.

Anderson, J. (2004). Medical teacher 25th anniversary series multiple-choice questions revisited. *Medical Teacher, 26*, 110–113.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*(8476), 307–310.

Bland, J. M., & Altman, D. G. (2012). Agreed statistics: Measurement method comparison. *Anesthesiology, 116*, 182–185.

Bruno, J. (2005). Method and system for knowledge assessment and learning incorporating feedbacks. United States Patent: 6921268.

Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education, 14*, 398–404.

Confucius & Legge, J. (1971). *Confucian analects, the great learning, and the doctrine of the mean.* New York: Dover Publications.

Egger, S. S., Drewe, J., & Schlienger, R. G. (2003). Potential drug–drug interactions in the medication of medical patients at hospital discharge. *European Journal of Clinical Pharmacology, 58*, 773–778.

Friedman, C., Gatti, G., Elstein, A., Franz, T., Murphy, G., & Wolf, F. (2001). Are clinicians correct when they believe they are correct? Implications for medical decision support. *Studies in Health Technology and Informatics, 1*, 454–458.

Gardner-Medwin, A.R., Gahan, M. (2003). Formative and summative confidence-based assessment. Presented at the 7th International Computer-Aided Assessment Conference, Loughborough.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review, 18*, 186–201.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital, 4*, 100–113.

Keers, R. N., Williams, S. D., Cooke, J., & Ashcroft, D. M. (2013). Prevalence and nature of medication administration errors in health care settings: A systematic review of direct observational evidence. *Annals of Pharmacotherapy, 47*, 237–256.

Kolbitsch, J., Ebner, M., Nagler, W. & Scerbakov, N. (2008). Can confidence assessment enhance traditional multiple choice testing? (Paper presented at the Interactive Computer aided Learning (ICL) International Conference, Villach).

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121.

Leclercq, D. (1983). Confidence marking, its use in testing. In Postlethwaite, Choppin (Eds.), *Evaluation in education* (pp. 161–287). Oxford: Pergamon.

Matthes, J., Johannsen, W., & Herzig, S. (2013). Change of medical students' self-appraisal of pharmacological knowledge and skills over time. *Naunyn-Schmiedebergs. Archives of Pharmacology, 386*(Suppl 1), S52.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), 63–67.

Osterberg, K., Kölbel, S. & Brauns, K. (2006). Der Progress Test Medizin: Erfahrungen an der Charité Berlin. *GMS Zeitschrift für Medizinische Ausbildung 23(3)*, Doc46.

Rippey, R. M., & Voytovich, A. E. (1985). Anomalous responses on confidence-scored tests. *Evaluation and the Health Professions, 8*, 109–119.

Sveiby, K. E. (1997). *The new organizational wealth: Managing and measuring knowledge-based assets*. California, SA: Berrett-Koehler Publishers.

Tobaiqy, M., McLay, J., & Ross, S. (2007). Foundation year 1 doctors and clinical pharmacology and therapeutics teaching. A retrospective view in light of experience. *British Journal of Clinical Pharmacology, 64*, 363–372.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education, 7*, 63–80.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*, 357–371.