American Journal of Pharmaceutical Education **2020; 84 (7) Article 7782.**

RESEARCH

Perception of Learning Versus Performance as Outcome Measures of Educational Research

Adam M. Persky, PhD,[a,c] Edward Lee, BS,[a] Lauren S. Schlesselman, MA, Ed Psych, PharmD[b,c]

[a] University of North Carolina at Chapel Hill, Eshelman School of Pharmacy, Chapel Hill, North Carolina
[b] University of Connecticut, Center for Excellence in Teaching and Learning, Storrs, Connecticut
[c] Associate Editor, *American Journal of Pharmaceutical Education*, Arlington, Virginia

Submitted July 25, 2019; accepted December 4, 2019; published July 2020.

Objective. To determine whether perception of student learning equates to learning gains.
Methods. Two-hundred seventy-seven college-aged students and student pharmacists participated in the study. Participants were assessed before and after completing a reading intervention and reported their perceptions of learning by responding to various Likert-scale questions. Relationships between perception and performance were assessed by correlation analysis, trend analysis, and using measures of metacognitive accuracy.
Results. There was a lack of correlation between measures of the perception of learning and actual gains in knowledge. There were weak correlations between the perception of learning and post-reading scores. Comparing student-pharmacists to college-aged individuals, both had similar metacognitive accuracy and there were little differences after the intervention.
Conclusion. Perceptions of learning may not reflect knowledge gains, and perception data should be used cautiously as a surrogate for evidence of actual learning.

Keywords: learning assessment, metacognition, perception of learning

INTRODUCTION

Much of pharmacy education research is aimed at evaluating the impact of an educational intervention, whether that be a new technology, new technique, unique setting, or unique content area. Some studies that have been published, including randomized controlled trials, only report participants' perceptions and confidence related to achieving the intended learning objective. In contrast, a randomized controlled study evaluating the same educational intervention may measure students' actual achievement of the knowledge, skill, or attitudes taught.

Suppose that a professor conducts a randomized controlled study comparing a standard educational practice to the use of a new technology to teach communication skills. As a result of the intervention, participants report higher confidence in their communication skills. A second professor conducts a study on this same technology's influence on communication skills, using the first professor's study design; however, rather than measuring students' confidence in their ability, she measures students' demonstration of the desired communication skill. After the intervention, participants demonstrated better communication skills. Which study would and should be more influential in convincing the academic community about the effectiveness of the new technology? To explore this debate, this study aims to document whether student perception of their learning is equivalent to their actual performance on criterial tasks.

John Flavell coined the term *metacognition* in referring to "cognition about cognitive phenomena," or in more common terms, "thinking about thinking."[1] Metacognition consists of cognitive knowledge and cognitive regulation. Metacognitive knowledge involves knowing oneself as a learner, knowing strategies to improve learning, and knowing when to use these strategies. Metacognitive regulation involves monitoring one's cognition, such as awareness of comprehension and performance and efficacy of strategies. Both the knowledge and regulation aspects of metacognition can limit or enhance learning depending on the quality of students' knowledge, monitoring, and control processes. For example, when a student perceives her ability as being higher than it actual is (ie, overconfident) the individual may pre-maturely terminate studying and therefore not reach the desired level of competency.[2]

In general, people tend to be overconfident when evaluating their performance. In agreement with Darwin's "Ignorance more frequently begets confidence than does

Corresponding Author: Adam M. Persky, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, 325 Beard Hall, Chapel Hill, NC 27599. Tel: 919-966-9104. Email: apersky@unc.edu

knowledge,"[3] Dunning and Kruger made the observation that less competent individuals are unable to recognize their own incompetence and therefore tend to overrate their ability.[4] Rather than recognizing their cognitive bias, those without knowledge and metacognition lack the ability to recognize that insufficiency. This lack of metacognition, known as *illusory superiority*, is when people overestimate their positive qualities and abilities and underestimate their negative qualities relative to others. Studies have found that a certain degree of competency is necessary for a person to recognize their incompetence.[4,5] Many reports have corroborated the Dunning-Kruger effect, including a 2017 study by Pennycook and colleagues that found participants with the greatest number of errors on a cognitive reflection test overestimated their performance by a factor of more than three and that this overestimation actually decreased as their performance increased.[6] The Dunning-Kruger effect is not restricted to students. In 1977, Cross demonstrated the "better than average" effect among teachers, with 94% of faculty members rating themselves as above average, which would be impossible given the definition of average.[7] A study by Motta and colleagues found that more than a third of respondents, who were members of the general public, thought they knew as much or more than doctors and scientists about the causes of autism, with overconfidence highest among those with low levels of knowledge about the causes of autism and high levels of endorsement of misinformation.[8]

Despite documentation of the Dunning-Kruger effect in multiple settings, many educational research projects claim the effectiveness of an educational intervention based solely on student perceptions of learning. In 1959, Donald Kirkpatrick first published his model of evaluation of training based on four levels. Historically, the Kirkpatrick classification puts perception of training as the lowest form of learning, while changes in behavior indicate a higher level of learning.[9] Each successive level of Kirkpatrick's evaluation adds precision to the measure of effectiveness of training. At the first and lowest level, participant reactions are measured, such as "did they like it," "did they consider it relevant," "did they learn anything." Such perceptions are subjective in nature. In Kirkpatrick's second level, the learning level, objective measures of an increase in knowledge or intellectual capability from before to after the learning experience must be provided. Such evaluations, based on the learning objectives of the training, often involve assessments administered before and after the training to determine achievement. In the third level, Kirkpatrick describes evaluating the sustainability of behavioral change associated with the learned knowledge. Typically, these evaluations are performed months after the training. In the final level, evaluation is intended to measure the improved performance of the trainees, including their achievement of key performance indicators.

Because so many educational research projects employ students' perceptions of learning as the only measure of success and do not measure students' actual learning, the primary objective of this study, which involved pharmacy and non-pharmacy students, was to examine the relationship between students' perceptions of their learning and the actual knowledge gained about four non-pharmacy topics. Because of the intensity of pharmacy education and the frequently changing topics covered in the pharmacy curriculum, whether they realize it or not, pharmacy students often rely on metacognition to identify the best learning strategies for each topic and to evaluate how much they know or do not know. Thus, the secondary objective of this study was to evaluate the metacognitive accuracy of these students.

METHODS

A total of 277 college-aged students (18-25 years old) and student pharmacists completed this study. Two different groups were recruited to allow for greater generalizability of the results and to have a sufficient sample size. College-aged students were recruited through Mechanical Turk (Amazon, Seattle, WA) and paid $4 for completing the study (n=150); student pharmacists were recruited via email from a convenience sample of four schools of pharmacy (three public and one private). Three hundred thirty-nine individuals started the study, however, seven were excluded because they failed the attention checks embedded in the survey. Additionally, 55 student pharmacists were excluded because they did not complete all of the survey items.

As an overview to the study design, first, participants made judgements on their level of understanding of several topics (ie, a judgement of learning or JOL). They were asked to indicate their perception of knowledge of four topics (bats, prions, ozone, and "sweet tracks") (eg, I am knowledgeable about this topic) and if they were to receive a quiz on the respective topics, they were asked to estimate their score. Next, participants' baseline knowledge was established by having them complete a pre-quiz on the four topics. Students then completed an instructional intervention in which they read a short passage related to one of the four topics. After each passage, participants answered questions about their perceptions of learning. After completing this process for all four topics, participants completed a filler activity designed to distract them prior to testing their retention of knowledge. Participants then completed the pre-test again to assess

knowledge gained. Attention checks were incorporated into the survey to ensure survey compliance and to prevent arbitrary responses.[10,11] The sequence of the four topics was randomizes to reduce any order effects, and backward navigation was not allowed. The study was conducted using Qualtrics (Qualtrics, Provo, UT).

The instructional intervention involved participants completing a simple reading comprehension task involving reading a short passage (~500 word) adopted from a standardized examination preparation book (ie, SAT or Test of English as a Foreign Language [TOEFL]). Learning was measured using a pre- and post-intervention quiz of basic knowledge about four non-pharmacy topics. The pre- and post-intervention quizzes were identical and consisted of 16 multiple-choice questions (four per topic), each with six answer choices (one correct answer, five distractors). Both the reading material and questions were taken from various standardized practice examinations and have been used in prior research.[12,13]

Perception of learning was measured in several ways, all using a five-point Likert scale (Appendix 1). These measures were adopted from a variety of measures used within the literature. At baseline, participants were asked to rate their knowledge of each of the four topics (ie, I am knowledgeable about [topic]). Then, after completing the reading, students were asked how familiar they were with the topic, (eg, "How familiar was the topic to you?"), how much they learned (eg, "I learned a lot while studying the above passage."), did the reading enhance their understanding (eg, "After reading and studying this passage, the passage enhanced my understanding of the material."), confidence (ie, "After reading and studying this passage, the passage improved my confidence in the material.") and their knowledge (ie, "I am knowledgeable about this topic."). These perception measures were made on a five-point Likert scale.

The primary outcome was the relationship between the various measures of the perception of learning (ordinal scale) and learning performance (continuous scale) using a Spearman rank correlation (SPSS Statistics for Windows, Version 24.0, IBM Corp, Armonk, NY). Cutoffs were established for a weak relationship ($r_s < .3$), a relationship of moderate strength ($.3 \leq r_s \leq .5$), and a strong relationship ($r_s > .5$) for behavioral data.[14] A secondary analysis was conducted using a one-way ANOVA assessing a linear trend in the relationship between the five-point Likert scale and performance. We used cutoffs on *f* for a small effect ($f < .25$), moderate effect ($.25 \leq f \leq .40$), and strong effect ($f > .40$).[14] We examined these relationships using change in scores (ie, post-quiz score minus the pre-quiz score) and the post-quiz score only. A Tukey post hoc analysis was used to detect differences between scale points and their respective performance data. In addition, we assessed metacognitive accuracy by examining the relationship between predicted quiz scores and actual quiz performance data. We calculated a bias (ie, prediction minus actual performance) and absolute bias (ie, absolute value of prediction minus actual performance).[15] Bias and absolute bias are examples of absolute accuracy or the magnitude of the effect because it reflects the absolute match between judgment magnitude and target performance. Bias is a measure of overconfidence (confidence greater than actual score) or underconfidence (confidence less than actual score). Negative bias values indicated the participant was underconfident while positive bias values indicated the participant was overconfident. Absolute bias indicates the magnitude of the difference between the confidence judgement and performance regardless of directionality (ie, accuracy). We also assessed participants relative accuracy (intraindividual accuracy) with a Goodman Kruskal gamma.[16] Relative accuracy is the degree to which the judgments discriminate between different levels of performance across items (eg, prions, bats, "sweet tracks," ozone) on an individual level. We calculated an effect size, when appropriate, using Cohen d with $< .5$ as a small effect, $.5 \leq d \leq .8$ as a medium effect, and $> .8$ as a large effect.[14] This study was deemed exempt by the University of North Carolina at Chapel Hill institutional review board.

RESULTS

At baseline, participants were asked to indicate using a Likert scale how knowledgeable they were about the four topics. Based on the combined percent of participants marking either strongly agree or somewhat agree for each topic, this was the rank order of the topics: ozone (46%), prions (35%), bats (31%), and "sweet tracks" (8%). Because we had college-aged students and student pharmacists, we wanted to assess their similarity with respect to metacognitive accuracy. We measured metacognitive accuracy by examining participants' absolute bias score, ie, the absolute value of the difference between their predictive score and actual score. No difference was found when examining the means across topics; thus, we concluded that there was no difference in accuracy between the two groups. Therefore, we combined the data from student pharmacists and that from the college-aged cohort for the remaining analysis (Table 1).

After completing the educational intervention, participants' performance on the post quiz did improve (mean score of 48% vs 78%, $p < .001$, $d = 1.3$). Participants' knowledge of each topic significantly improved

Table 1. Summary of Metacognitive Judgments (Bias and Absolute Bias) by Cohort

		Pre		Post		Total	
Subject		College (n=150)	Pharmacy (n=127)	College (n=150)	Pharmacy (n=127)	Pre (n=277)	Post (n=277)
Prions	Predicted Score (%)	23 (24)	42 (27)[a]	55 (25)	70 (20)[a]	31 (27)	62 (24)
	Actual Score (%)	42 (29)	62 (30)[a]	73 (30)	83 (29)[a]	51 (31)	78 (30)
	Bias	-19 (33)	-20 (32)	-18 (31)	-13 (25)	-19 (32)	-16 (28)
	Absolute Bias	30 (22)	30 (23)	28 (22)	22 (17)[a]	30 (22)	26 (20)
	Correlation	.22[b]	.22[b]	.37[b]	.54[b]	.38[b]	.48[b]
Bats	Predicted Score (%)	40 (26)	35 (23)	61 (21)	59 (25)	38 (25)	60 (23)
	Actual Score (%)	49 (27)	53 (27)	74 (28)	78 (30)	51 (27)	76 (29)
	Bias	-9.4 (32)	-18 (31)[a]	-14 (29)	-17 (29)	-13 (32)	-17 (31)
	Absolute Bias	27 (19)	30 (20)	26 (19)	29 (18)	28 (19)	30 (19)
	Correlation	.25[b]	.25[b]	.24[b]	.48[b]	.46[b]	.39[b]
Ozone	Predicted Score (%)	41 (28)	42 (22)	64 (25)	68 (22)	42 (25)	66 (24)
	Actual Score (%)	59 (26)	63 (22)[a]	81 (25)	81 (21)	61 (24)	81 (23)
	Bias	-18 (29)	-22 (30)	-17 (28)	-13 (22)	-19 (30)	-15 (26)
	Absolute Bias	27 (21)	31 (21)	26 (20)	20 (16)[a]	39 (21)	23 (18)
	Correlation	.39b	.39b	.38b	.43b	.30b	.40b
"Sweet Tracks"	Predicted Score (%)	19 (24)	19 (18)	60 (25)	59 (21)	18 (21)	59 (22)
	Actual Score (%)	36 (28)	25 (23)[a]	76 (29)	75 (32)	31 (27)	75 (30))
	Bias	-18 (34)	-6.7 (29)[a]	-16 (30)	-14 (33)	-13 (33)	-15 (31)
	Absolute Bias	27 (21)	31 (21)	30 (19)	28 (21)	28 (21)	28 (21)
	Correlation	.14	.14	.38[b]	.30[b]	.19[b]	.34[b]
Total	Predicted Score (%)	31 (21)	34 (18)[a]	59 (22)	65 (18)	32 (20)	62 (20)
	Actual Score (%)	46 (20)	51 (16)	76 (24)	79 (24)	48 (19)	77 (24)
	Bias	-16 (24)	-17 (21)	-16 (24)	-15 (21)	-16 (23)	-16 (22)
	Absolute Bias	24 (16)	22 (15)	23 (17)	22 (14)	23 (16)	23 (15)
	Correlation	.32[b]	.27[b]	.50[b]	.57[b]	.32[b]	.53[b]

Data presented as mean (standard deviation). Bias is calculated from predicted score – actual score; absolute bias is the absolute value of bias. Correlation is the correlation between predicted score and actual score using Spearman's Rank

[a] $p<.05$, compared to college-aged students

[b] $p<.05$

because of the reading material provided (range d=.85 – 1.5), with scores on the topic with which participants were most familiar (ozone) showing the smallest effect size (d=.85) and scores on the topic with which they were the least familiar ("sweet tracks") showing the largest effect size (d=1.5). Thus, the educational intervention resulted in learning.

The primary outcome examined was the relationship between students' perception of learning and the actual knowledge gained. Table 2 shows the correlations between the five-point Likert scale and quiz performance with quiz performance measured as a change score (post minus pre) and the quiz score after reading (post-intervention score). Most of the correlations between change in performance (post score minus the pre score) and perception were not significant. When examining the post-intervention quiz score and perceptions of learning, we noted some weak correlations ($r_s<.4$). We also conducted a correlation between the predicted score after reading and the actual score and found weak correlation (Table 2).

We conducted an additional analysis using a one-way ANOVA for all perceptions of learning and the change in score (ie, post minus pre) and post-reading score. For change in performance and measures of perception of learning, only the "sweet tracks" topic had significant linear trends on four of the five scales, ie, confidence ($p=.002$), familiarity ($p<.001$), learned a lot ($p<.001$), and understanding ($p<.001$), but not for knowledge ($p=.82$). No other topics demonstrated a linear trend. For post-score and perception, some linear trends were found (Figure 1). Interestingly, when asked how familiar the topic was, there was negative trend found with the highest scores being associated with "less familiarity." To note, while trends exist, there were inconsistent differences in each Likert point relative to the adjacent points; that is, a lower level scale (eg, strongly disagree) was not necessarily statistically lower than the adjacent disagree, or neither

Table 2. Spearman Correlations for Examining Perception of Learning Measures and Measures of Learning

		Perception Scales				
Subject	**Knowledge Scales**	**"I learned a lot"**	**"Enhanced Understanding"**	**"Improved confidence"**	**"I am knowledgeable"**	**"How familiar"**
Prions	Gain	.16[a]	.08	-.18[a]	.20[a]	.29[a]
	Score	.22[a]	.34[a]	.34[a]	.37[a]	.27[a]
Bats	Gain	.15[a]	.022	-.048	.10	.20[a]
	Score	.29[a]	.27[a]	.21[a]	.36[a]	-.031
Ozone	Gain	.11	.21[a]	-.081	-10	-.098
	Score	.27[a]	.28[a]	.26[a]	.28[a]	.14[a]
"Sweet Tracks"	Gain	.24[a]	.27[a]	.26[a]	-.029	-.43[a]
	Score	.36[a]	.32[a]	.28[a]	.05	-.37[a]

Gain is the difference between the post-reading score and pre-reading score. Score represents the post-score. Correlations below .29 are considered weak relationships, .3 to .5 moderate strength relationships, and over .5 strong relationships (n=277)

[a] $p<.05$

disagree or agree (ie, the neutral position). This suggests, despite a linear trend, adjacent scales may not discriminate differences in learning.

The relative accuracy of each participants' self-assessment of knowledge was also examined, ie, how well did a student's judgments discriminate between concepts they understood well versus those they understood less well. We tested each measure of the perception of learning and the respective change score in performance. All relationships were poor ($-.39<\gamma<.21$). This suggests the participants were poor at discriminating their performance gains across topics. Subjects were better discriminators when the perception of learning was compared to post-intervention score ($.045<\gamma<.39$) but this was variable between the various perception measures.

DISCUSSION

This study examined the relationship between students' actual learning and their perception of learning. This study is important because within the scholarship of education, researchers often use data on students' perception of learning as a surrogate for proof of learning (ie, documented increase in knowledge or skills).

Overall, students in this study associated perception of learning more strongly with post-reading scores than with change scores, suggesting they ignored their baseline knowledge. Regardless of the comparator, these associations were weak, suggesting the students were relatively poor judges of their own learning; for these relationships, topic was an influential factor. As such, perceptions of learning are subjective measures of perceived learning, not actual learning. They are probably better measures of self-efficacy than ability.

Historically, the Kirkpatrick classification puts perception of learning as the lowest form of learning, with change in attitude as the second lowest. Furthermore, this classification system puts changes in knowledge/skill and changes in behavior as higher forms of learning.[9] Despite educators' recognition of student perception of learning as a poor measure of training effectiveness because of its subjective nature, it continues to be used in educational research without triangulation with other measures of actual learning. The results of this study are consistent with Kirkpatrick's classification of learning and with the metacognitive literature. Zell reviewed the metacognitive accuracy of individuals relative to various professional skills (eg, academics, sports, etc).[17] They found a weak correlation between an individual's perceived ability and their actual ability ($r<.3$).[17] This study agrees with the prior literature because we also found a lack of correlation for change scores. Part of this poor correlation is because the participants ignored their baseline performance. This is consistent with the *base rate fallacy* or the bias that individuals tend to ignore an earlier premise (eg, base rate probability, baseline knowledge).[18,19]

Metacognition is not static and can be influenced in different situations or individuals. For example, acute stress is associated with altered cognitive function, in particular less flexible cognitive processing.[20-22] Metacognition is impacted by stress, with acute stress correlating with less accurate assessment of performance.[23] Pharmacy and other college students are often under huge amounts of stress from heavy course loads and non-academic-related pressures. While this study was conducted in a low-stakes fashion, the same is not true of college courses. When educational interventions are conducted as part of a course, students are often under stress and therefore exhibit less metacognitive accuracy.

Previous studies have also found differences in metacognitive accuracy based on gender. A study by

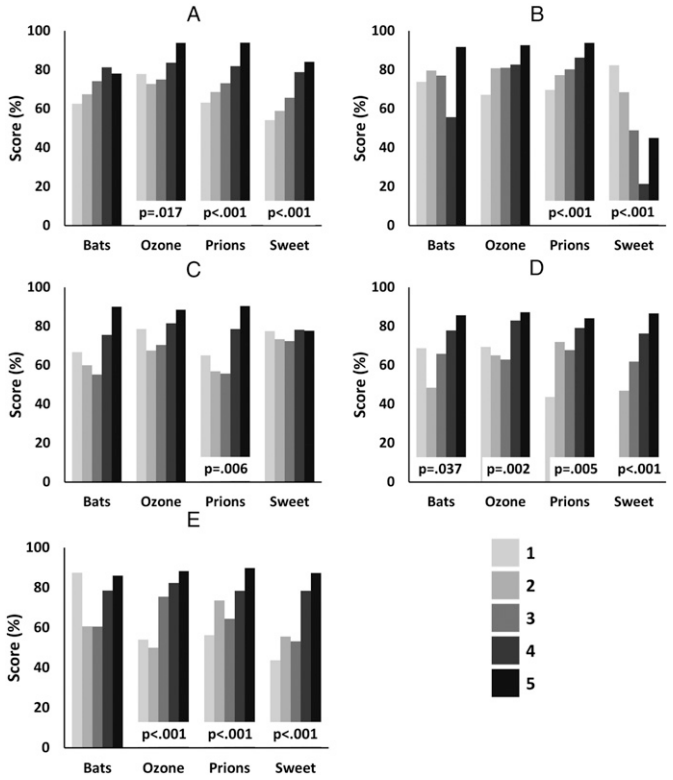

Figure 1. One-way ANOVA for linear trend for examining perception of learning measures and post-score. (n=277). Where the difference was significant, a *p* value is given. A) "After reading and studying the passage, the passage improved my confidence in the material"; B) "How familiar was the topic to you"; C) "I am knowledgeable about the topic"; D) "I learned a lot while studying the passage"; E) "After reading and studying the passage, the passage enhanced my understanding of the material". Shades of grey indicate the response category: category 1=strongly disagree or not familiar at all; category 2 =somewhat disagree or slightly familiar; category 3=neither agree or disagree or moderately familiar; category 4=somewhat agree or very familiar; category 5=strongly agree or extremely familiar

Ariel and colleagues evaluating gender differences in confidence and accuracy found that female students displayed lower confidence in their performance, even when no actual accuracy difference existed.[24] In educational research based on metacognitive accuracy, the gender of the participants may also play a factor in the findings. Although not evaluated as part of this study, it is important to consider the impact of gender on metacognitive accuracy.

The strength of this study is the sample size and diversity of the subject pool, which included both college-aged students and student pharmacists. The material used for the study was intentionally not pharmacy related. However, it was on topics with which most students would be unfamiliar or only somewhat familiar, much like the material first-year student-pharmacists encounter within the Doctor of Pharmacy curriculum. Notably, domain knowledge, ease of processing of material to be learned, incentives, or interest in a topic can impact a person's perception of learning. However, these factors play a minimal role in the accuracy of a person's judgements, ie, they do not make a person's judgement of their learning more accurate compared to their actual performance.[25-27] In this study, we used a variety of measures of perception, which helps generalize the findings that perception data does not reflect learning. We also attempted to emulate the pre-post design of other studies; however, we did not use the pre-post format for perception scales; we only used it for the knowledge domain. We expected that participants' perceptions of learning would also change after completing the learning intervention but would still not accurately reflect gains in learning.

Given that students are poor judges of their own learning, educational research studies should be designed using more objective measures of learning. Determining the intended student learning objectives and establishing measures to directly evaluate these objectives results in a stronger, more accurate evaluation of student knowledge and abilities. If measures of student perceptions of knowledge are included in a study, they should only be used to triangulate other measures of success.

CONCLUSION

Perception of learning data should not be the primary measure of learning for educational research. The most appropriate use of perception data would be in assessing student metacognition or self-efficacy. Even in this domain, actual knowledge or skill acquisition is important to calibrate the measurements.[15,28,29]

REFERENCES

1. Flavell JH. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *Am Psychol*. 1979;34(10):906-911.
2. Dunlosky J, Rawson KA. Overconfidence produces underachievement: inaccurate self evaluations undermine students' learning and retention. *Learn Instruct*. 2012;22(4):271-280.
3. Darwin C. *The Descent of Man, and Selection in Relation to Sex*. New York, NY: D. Appleton and Company; 1871.
4. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Personal Soc Psychol*. 1999;77(6):1121-1134.
5. Dunning D, Johnson K, Ehrlinger J, Kruger J. Why people fail to recognize their own incompetence. *Curr Dir Psychol Sci*. 2003; 12(3):83-87.
6. Pennycook G, Ross RM, Koehler DJ, Fugelsang JA. Dunning–Kruger effects in reasoning: theoretical implications of the failure to recognize incompetence. *Psychon Bullet Rev*. 2017;24(6):1774-1784.
7. Cross KP. Not can, but will college teaching be improved? *New Dir Higher Educ*. 1977(17):1.

8. Motta M, Callaghan T, Sylvester S. Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Soc Sci Med*. 2018;211:274-281.
9. Kirkpatrick D. Techniques for evaluating training programs. *Training Develop*. 1996;50(1):54-59.
10. Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Meth*. 2016;48(1):400-407.
11. Meade AW, Craig SB. Identifying careless responses in survey data. *Psychol Meth*. 2012;17(3):437-455.
12. Fazio LK, Agarwal PK, Marsh EJ, Roediger HL. Memorial consequences of multiple-choice testing on immediate and delayed tests. *Mem Cogn*. 2010;38(4):407-418.
13. Roediger HL, Marsh EJ. The positive and negative consequences of multiple-choice testing. *J Exp Psychol Learn Mem Cogn*. 2005; 31(5):1155-1159.
14. Cohen J. A power primer. *Psychol Bullet*. 1992;112(1):155-159.
15. Dunlosky J, Thiede KW. Four cornerstones of calibration research: why understanding students' judgments can improve their achievement. *Learn Instruct*. 2013;24(1):58-61.
16. Nelson TO, Narens L. A new technique for investigating the feeling of knowing. *ACta Psychol*. 1980;46(1):69-80.
17. Zell E, Krizan Z. Do people have insight into their abilities? a metasynthesis. *Persp Psychol Sci*. 2014;9(2):111-125.
18. Bar-Hillel M. The base-rate fallacy in probability judgments. *Acta Psychologica*.1980;44(3):211-233.
19. Kahneman D, Tversky A. On the Psychology of Prediction. *Psychol Rev*. 1973;80(4):237.
20. Lupien SJ, Maheu F, Tu M, Fiocco A, Schramek TE. Elsevier Scientific Publishing Co NYNY. The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition. *Brain Cogn*. 2007;65(3):209-237.
21. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. *Proc Nat Acad Sci*. 2013;110(52):20941-20946.
22. Starcke K, Brand M. Decision making under stress: a selective review. *Neurosci Biobehav Rev*. 2012;36(4):1228-1248.
23. Reyes G, Silva JR, Jaramillo K, Rehbein L, Sackur J. Self-knowledge dim-out: stress impairs metacognitive accuracy. *PLoS ONE*. 2015;10(8):e0132320-e.
24. Ariel R, Lembeck NA, Moffat S, Hertzog C. Are there sex differences in confidence and metacognitive monitoring accuracy for everyday, academic, and psychometrically measured spatial ability? *Intelligence*. 2018;70:42-51.
25. Begg I, Duft S, Lalonde P, Melnick R, Sanvito J. Memory predictions are based on ease of processing. *J Mem Languag*. 1989;28(5):610-632.
26. Saenz GD, Geraci L, Tirso R. Improving metacognition: a comparison of interventions. *Appl Cogn Psychol*. 2019;33(5):918-929.
27. Shanks LL, Serra MJ. Domain familiarity as a cue for judgments of learning. *Psychonom Bull Rev*. 2014;21(2):445-453.
28. Keren G. Calibration and probability judgments - conceptual and methodological issues. *Acta Psychol*. 1991;77(3):217-273.
29. Schraw G, Kuch F, Gutierrez AP. Measure for measure: calibrating ten commonly used calibration scores. *Learn Instruct*. 2013;24(1):48-57.

Appendix 1. Perception of learning questions and their scales

Abbreviation	Question	Scale	Measured
Knowledgeable	I am knowledgeable about the topic	Strongly agree Somewhat agree Neither agree or disagree Somewhat disagree Strongly disagree	Pre and Post reading
Familiarity	How familiar was the topic to you?	Extremely familiar Very familiar Moderately familiar Slightly familiar Not familiar at all	Post reading
Learned A Lot	I learned a lot while studying the passage	Strongly agree Somewhat agree Neither agree or disagree Somewhat disagree Strongly disagree	Post reading
Understanding	After reading and studying the passage, the passage enhanced my understanding of the material	Strongly agree Somewhat agree Neither agree or disagree Somewhat disagree Strongly disagree	Post reading
Confidence	After reading and studying the passage, the passage improved my confidence in the material	Strongly agree Somewhat agree Neither agree or disagree Somewhat disagree Strongly disagree	Post reading
Predicted Score	If I was to be given a test on this material, I would get a	Percent correct (0-100%)	Pre and Post reading