Fostering Dental Student Self-Assessment of Knowledge by Confidence Scoring of Multiple-Choice Examinations

C. Alex McMahan, Ph.D.; R. Neal Pinckard, Ph.D.; Anne Cale Jones, D.D.S.; William D. Hendricson, M.A., M.S.

Abstract: Creating a learning environment that fosters student acquisition of self-assessment behaviors and skills is critically important in the education and training of health professionals. Self-assessment is a vital component of competent practice and lifelong learning. This article proposes applying a version of confidence scoring of multiple-choice questions as one avenue to address this crucial educational objective for students to be able to recognize and admit what they do not know. The confidence scoring algorithm assigns one point for a correct answer, deducts fractional points for an incorrect answer, but rewards students fractional points for leaving the question unanswered in admission that they are unsure of the correct answer. The magnitude of the reward relative to the deduction is selected such that the expected gain due to random guessing, even after elimination of all but one distractor, is never greater than the reward. Curricular implementation of this confidence scoring algorithm should motivate health professions students to develop self-assessment behaviors and enable them to acquire the skills necessary to critically evaluate the extent of their current knowledge throughout their professional careers. This is a professional development competency that is emphasized in the educational standards of the Commission on Dental Accreditation (CODA).

Dr. McMahan and Dr. Pinckard contributed equally to this study. Dr. McMahan is Professor, Department of Pathology, The University of Texas Health Science Center at San Antonio; Dr. Pinckard is Distinguished Teaching Professor, Department of Pathology, The University of Texas Health Science Center at San Antonio; Dr. Jones is Distinguished Teaching Professor, Department of Pathology, The University of Texas Health Science Center at San Antonio; and Mr. Hendricson is Assistant Dean for Education and Faculty Development, School of Dentistry, The University of Texas Health Science Center at San Antonio. Direct correspondence and requests for reprints to Dr. C. Alex McMahan, Department of Pathology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900; 210-567-4026; mcmahan@uthscsa.edu.

Keywords: dental education, assessment, self-assessment, multiple-choice questions, formula scoring, lifelong learning

Submitted for publication 1/7/14; accepted 5/15/14

In most learning environments, the principal objective for giving examinations is to estimate the extent of information that students have learned and retained, i.e., "what they know." However, a second crucial objective for examinations should be to enable students to become aware of what they have not actually learned, i.e., "what they do not know." Attaining this latter objective necessitates students' development of self-assessment behaviors and skills to allow them to better evaluate the extent of their current knowledge. The value of the development of self-assessment skills has been recognized and acknowledged as an important educational objective.[1]

Creating a learning environment that fosters students' development of self-assessment behaviors and skills is a particularly important objective in the education of health professionals. Thus, as health professions educators, we must strongly encourage students to develop lifelong learning and self-assessment skills to allow them to be cognizant of the current state of their knowledge, including recognizing and admitting what they do not know. This essential element in health professions learning environments has been emphasized by the American Dental Education Association (ADEA) in its Statement on Professionalism in Dental Education.[2] In the discussion of competence, this statement notes that students must "Develop the habits and practices of lifelong learning, including self-assessment skills. Accept and respond to fair negative feedback about your performance (recognize when you need to learn)." It further states that students must "Know the limits of your knowledge and skills and practice within them; learn when and how to refer."

Standard 2-10 of the 2013 predoctoral education standards of the Commission on Dental Accreditation (CODA) states that "Graduates must demonstrate the ability to self-assess, including the

development of professional competencies and the demonstration of professional values and capacities associated with self-directed lifelong learning."[3] The accompanying intent statement for CODA Standard 2-10 is the following: "Educational programs should prepare students to assume responsibility for their own learning. The educational program should teach students how to learn and apply evolving and new knowledge over a complete career as a health care professional. Lifelong learning skills include student assessment of learning needs."

In many educational institutions, including our health science center, most examinations are comprised of multiple-choice questions that are scored by computation of the percent of questions answered correctly. We believe this type of examination question format and scoring is not particularly conducive for fostering self-assessment skills. It is widely acknowledged that, because of cueing and guessing, scores on multiple-choice examinations with number-correct scoring overestimate the extent of actual student knowledge and conceptual understanding, particularly among lower performing students. Indeed, many students become very adept at choosing the correct answer, using a variety of clues unrelated to subject matter knowledge. Moreover, inflation of examination scores is exacerbated when students are able to use their "partial knowledge" to eliminate one or more of the incorrect distractors prior to selecting their final answer from remaining options. This score inflation leads faculty and students to overestimate the extent of students' knowledge and reduces the necessity for students to consciously identify what they do not know. Compounding these problems, there are few, if any, consequences levied upon students even for random, no knowledge guessing; to the contrary, students often are encouraged to guess and are rewarded for doing so. Although many students may be aware when they are guessing, they most likely do not fully appreciate and consciously distinguish between whether they are guessing randomly, making partial knowledge guesses, or truly believe they know the correct answer. Moreover, even if students were able to recognize the shortcomings in their knowledge, they are not encouraged to admit these limitations either to themselves or others. A willingness to openly admit deficiencies is an essential part of the "assessment of learning needs" required for "lifelong learning skills."[3]

In comparison to cued multiple-choice questions, the primary assessment advantage of un-cued short-answer and essay questions is that students must generate their own thoughts, integrate them into a coherent whole, and communicate an original response "in their own words." This, according to the modified Bloom's cognitive hierarchy,[4] is a more sophisticated intellectual challenge than recognizing a correct, or an incorrect, statement, depending on question format, from an array of options. Based on our experience, use of un-cued questions as an assessment format encourages students to prepare differently and in more depth as compared to a multiple-choice examination. Reviews of the pros and cons of various assessment methods to measure knowledge acquisition and capacity to apply this knowledge typically identify the ability of essays to measure deeper levels of knowledge than multiple-choice questions.[5-8] Although beyond the scope of this article, reliability of essay-type examinations can be influenced by the clarity of rating criteria, number of criteria to be judged, number of qualitative rating points on the judges' assessment form, presence or absence of evaluator training (calibration exercises), and the number of evaluators, which may reduce the reliability of an un-cued assessment.[9]

However, in spite of the educational benefits of un-cued examination questions, at least for the foreseeable future, use of multiple-choice questions will continue to prevail as the examination question format of choice for the following reasons: 1) the ease of scoring multiple-choice questions as opposed to the time and effort required to evaluate answers to essay or short-answer questions; 2) a perceived better objectivity of scoring multiple-choice questions whereas scoring essay or short-answer questions is often considered to be subjective; and 3) multiple choice being by far the prevailing question format for major assessments used during the admission process and in national board examinations in dentistry, medicine, and other health professions. Taking the likelihood of continued utilization of multiple-choice questions into account, we propose in this article the introduction of a modified confidence scoring algorithm[10-13] (deduction/reward) for multiple-choice questions. This modified multiple-choice question scoring algorithm not only could allow better estimation of student academic achievement but also potentially would foster more effective student learning behaviors and acquisition of self-assessment skills. The use of such a confidence scoring algorithm would be one proactive method dental schools could employ to strongly encourage students to develop skills to better self-assess their current knowledge and provide evidence that students are indeed actively engaged in self-assessment activities.[3,8,14]

Review of Our Previous Studies

We previously examined the effects of two examination question formats (multiple-choice and short-answer) and two scoring algorithms for multiple-choice questions (number-correct and formula scoring) on the academic performance of four consecutive classes of second-year dental students taking an oral and maxillofacial pathology course.[15-19] In these studies, we documented that examination question format and scoring algorithm had significant but separate effects on student academic performance (course scores).

Short-Answer Questions

Relative to all multiple-choice question examinations, utilization of all short-answer format questions significantly improved overall student performance; a comparable increase in performance also was achieved when the examinations were comprised of one-half short-answer questions and one-half multiple-choice questions.[17,18] These studies strongly supported the concept that use of un-cued format questions for at least some portion of examination questions increases student performance, presumably by modifying their behavior in preparing for the more challenging un-cued questions. However, scoring of short-answer questions required a major time commitment by the faculty. Thus, for practical considerations, we acknowledge that the use of cued multiple-choice questions will remain a fact of academic life.

Formula Scoring of Multiple-Choice Examinations

In number-correct scoring of multiple-choice examinations, +1 point (full credit) is assigned for a correct answer to a question, and 0 is assigned for an incorrect answer. Formula scoring,[20] or the standard correction for guessing, is a modification of number-correct scoring. For the five-option multiple-choice questions we have investigated to date,[15-18] the standard formula for correction for guessing consisted of assigning +1 point for a correct answer, -1/4 point for an incorrect answer, and 0 points for a question not answered. Assuming a random (no knowledge) selection among one of the five options, the probability of guessing the correct answer is 1/5 (0.20), and the probability of choosing an incorrect answer is 4/5 (0.80). Thus, using the standard correction for guessing, the expected value of the number of points gained due to random guessing is zero [(1/5)(1)+(4/5)(-1/4) = 0]. In general, for κ possible response options per question, $-1/(\kappa-1)$ is assigned for an incorrect answer.

We previously reported that after retrospectively applying the preceding formula scoring algorithm, the corrected multiple-choice scores for individual students agreed significantly better with their respective scores on the short-answer portions of the examinations testing the same body of material.[15] We concluded, therefore, that utilization of formula scoring increased the validity of the scores obtained on the multiple-choice portions of the examinations, thereby providing a better estimate of actual student achievement.[15]

Subsequently, in a class in which students were informed that the multiple-choice portions of examinations would be scored by formula scoring (prospective implementation), we unexpectedly found significant increases in the course scores, particularly by the lower performing students.[16,18] Of note, this enhanced performance was observed not only on the multiple-choice questions but on the short-answer portions of the examinations as well.[16,18] Thus, implementation of formula scoring not only increased the validity of the multiple-choice examination scores, but also fostered improved overall student performance, presumably by motivating positive modifications in student learning behaviors.

As usually applied, the parameter κ in the formula scoring algorithm is the number of options presented in a multiple-choice question. Students often eliminate one or more of the incorrect options (distractors) before making their final selection from among the remaining options. One possible modification of the formula scoring algorithm would be to base the adjustment on the estimated number of functioning distracters—that is, those incorrect options that are used by students.[19] We previously identified functioning distractors as those distractors used by at least 1 percent of students.[19] This definition resulted in the number of functioning distractors being related to question difficulty and discrimination.

The formula scoring algorithm based on functioning distractors should assign +1 for a correct answer, $-1/\delta$ points for an incorrect answer, and 0 points for a question not answered.[19] We estimated the numbers of functioning distractors used by three academic classes in our oral and maxillofacial

pathology course, in which the examinations were comprised either of all or one-half multiple-choice questions with five options.[19] These analyses indicated that the number of functioning distractors had a distribution ranging from zero to four. Thus, our previous estimates of score inflation based on a model of random (no knowledge) guessing among five items on multiple-choice examinations have been underestimates.[15,16,18] Therefore, we advocated that a formula scoring algorithm be applied only after determining the number of functioning distractors (labeled as δ).[19] In that study, our analyses showed that the harmonic mean of the number of functioning distractors for questions missed by an individual student represented the appropriate average adjustment (appropriate value of δ) for that student and that the harmonic mean of the number of functioning distractors for all students in a class represented an appropriate adjustment for the overall class average.[19] Additionally, our findings supported the validity of using the class harmonic mean as a constant adjustment value instead of requiring a different adjustment for each multiple-choice question.[19] Estimation of the number of functioning distractors is particularly difficult in small classes.[19] Furthermore, we do not anticipate that determining the number of functioning distractors would be done for each examination but could be developed for each course over a period of time.

The arithmetic mean and the harmonic mean of the number of functioning distractors varied significantly among the three classes of the oral and maxillofacial pathology course having different examination environments.[19] The difference in number of functioning distractors between the class given all multiple-choice questions and the two classes given one-half short-answer questions and one-half multiple-choice questions was in accord with the expectation based on academic performance (course score). Specifically, the better performing classes given one-half short-answer questions used fewer distractors than the poorer performing class given all multiple-choice questions. However, in the two classes given one-half short-answer questions and one-half multiple-choice questions, the better performing class in which formula scoring was prospectively applied did not use a lower number of functioning distractors. This outcome was in contrast to results anticipated because of overall course performance and suggests that the prospective application of formula scoring had an effect on students that was not indicated by the very low number of questions left unanswered.[16] Finally, our analyses indicated an overall average of three functioning distractors in the multiple-choice examinations for the three classes of our oral and maxillofacial pathology course;[19] thus, a value of δ=3 was deemed an appropriate parameter to be used in the formula scoring algorithm. We stress, however, that an appropriate parameter (δ) for other courses and learning environments must be determined separately.

Confidence Scoring Algorithm

Our previous studies showed that use of formula scoring increased the validity of multiple-choice scores[15,16] and also improved student academic achievement.[16,18] To build upon these positive findings, we now propose the application of a version of confidence scoring to multiple-choice examinations to help students learn to better self-assess their current knowledge.[10-13,21-23] This scoring algorithm assigns +1 point for a question answered correctly, a deduction of $-p_D$ points for a question answered incorrectly, and a reward of p_R points for a question left unanswered. Under these circumstances, we believe that when students decide to answer a question, they are expressing high confidence that they know the correct answer; and when students decide to leave a question unanswered, they are openly admitting low confidence that they know the correct answer. Thus, our scoring algorithm requires that students make a realistic assessment about their actual knowledge or lack of knowledge of the correct answer (students must decide whether they have high or low confidence as opposed to requiring a student to specify a numerical statement of confidence as others have done[21]).

The values of the deduction ($-p_D$) and reward (p_R) relative to the point value assigned for a correct answer (+1) must be selected by the faculty member. One criterion for selecting the relative levels of the deduction ($-p_D$) and reward value (p_R) is to select values such that the maximum expected value of points gained due to random guessing, even after elimination of some distractors, is never greater than the reward for leaving a question unanswered. Thus, after selection of a value for either the deduction or reward, the foregoing criterion will determine the value of the other. The greatest likelihood for guessing a correct answer occurs when a student can eliminate all but one of the distractors, thereby reducing the problem

to a true/false question in which the probability of randomly guessing the correct answer is 1/2. In this situation, the expected value of the points gained due to random guessing is $+1(1/2) - p_D(1/2)$; our proposed criterion for the confidence scoring algorithm requires that this quantity be equal to or less than the reward, that is, $1/2 - p_D/2 \leq p_R$. Deductions of -1/4, -1/3, -1/2, and -1 correspond to formula scoring for correction due to random guessing among 5, 4, 3, and 2 options, respectively; the corresponding rewards under our proposed criterion would be 3/8, 1/3, 1/4, and 0. Table 1 shows the expected outcomes of each of these deductions and corresponding rewards when a student randomly guesses after distractor elimination. Thus, if a student were to eliminate three distractors and had a 1/2 probability of guessing the correct answer, the maximum expected points gained due to random guessing would only equal the reward for leaving the question unanswered. Although the reward and expected gain may be identical, students must realize that the reward is guaranteed, while randomly selecting an answer exposes them to a possible deduction.

Although the confidence scoring algorithm should discourage random guessing, knowledgeable guessing could be of benefit to students. Table 2 shows the expected gain in score for three groups of students having differing levels of knowledge. Knowledge level is described by an odds ratio relative to the random guessing situation (see Appendix); an odds ratio of 1.00 corresponds to random guessing, and increasing odds ratios greater than 1.00 indicate increasing levels of knowledge and therefore ability to recognize the correct answer. In these examples, students with the illustrated knowledge levels are expected to gain, relative to the reward, if they can reduce the problem to a true/false question. With the highest odds ratio presented, students have to eliminate only one distractor to expect to achieve an increase in score greater than the reward. Lower knowledge levels require elimination of more distractors to achieve gains bigger than the reward. Thus, to make the best choice, students not only have to carefully evaluate their ability to eliminate distractors, but also their confidence in their knowledge that they know the correct answer among the remaining options. We anticipate that even very capable students could be motivated toward self-assessment under a confidence scoring approach.

Table 3 shows how the expected scores obtained by students would be affected by their choices in realistically assessing what they do not know. The calculations demonstrate the situations of using deductions of -1/4, -1/3, -1/2, and -1 along with corresponding rewards for an examination in which the number of functioning distractors is three ($\delta=3$).[19] The lowest scores occur if students guess randomly when the scoring system imposes a deduction for incorrect answers; if no questions are left unanswered, the magnitude of the reward is irrelevant. Not answering some fraction of the questions and accepting the reward would help some students to improve their examination scores.

The confidence scoring algorithm includes a reward component to motivate students to recognize and admit what they do not know. If students were to effectively dichotomize their knowledge into what they know and what they do not know and then leave all questions for which they do not know the correct response unanswered, their examination score

Table 1. Expected gain due to random guessing after eliminating incorrect distractors in a multiple-choice question with five possible answers

Guessing Status	Question Category[a]	Probability		Expected Gain Due to Guessing			
		Correct	Incorrect	$-p_D=-1/4$ $p_R=3/8$	$-p_D=-1/3$ $p_R=1/3$	$-p_D=-1/2$ $p_R=1/4$	$-p_D=-1$ $p_R=0$
Random	$\delta=4$	1/5	4/5	0	-0.067	-0.2	-0.6
	$\delta=3$	1/4	3/4	0.063	0	-0.125	-0.5
	$\delta=2$	1/3	2/3	0.167	0.111	0	-0.333
	$\delta=1$	1/2	1/2	0.375[b]	0.333[b]	0.25[b]	0[b]

[a]δ represents the number of functioning distractors; thus, the student makes final selection of response to question from among the functioning distractors and the correct answer.
[b]Equal to reward.

Note: A student is assigned a score of +1 for a correct answer, $-p_D$ deduction for an incorrect answer, and p_R if unanswered (i.e., the student recognizes and admits that he or she does not know).

Table 2. Examples of expected gain due to knowledgeable guessing after eliminating incorrect distractors in a multiple-choice question with five possible answers

Knowledge of Correct Answer (Odds Ratio[a])	Question Category[b]	Probability		Expected Confidence Score Due to Guessing			
		Correct	Incorrect	$-p_D$=–1/4 p_R=3/8	$-p_D$=–1/3 p_R=1/3	$-p_D$=–1/2 p_R=1/4	$-p_D$=–1 p_R=0
2	δ=4	1/3 = 0.333	2/3	0.167	0.111	0.000	-0.333
	δ=3	2/5 = 0.400	3/5	0.250	0.200	0.100	-0.200
	δ=2	1/2 = 0.500	1/2	0.375[c]	0.333[c]	0.250[c]	0.000[c]
	δ=1	2/3 = 0.667	1/3	0.583[d]	0.556[d]	0.500[d]	0.333[d]
3	δ=4	3/7 = 0.429	4/7	0.286	0.238	0.143	-0.143
	δ=3	1/2 = 0.500	1/2	0.375[c]	0.333[c]	0.250[c]	0.000[c]
	δ=2	3/5 = 0.600	2/5	0.500[d]	0.467[d]	0.400[d]	0.200[c]
	δ=1	3/4 = 0.750	1/4	0.688[d]	0.667[d]	0.625[d]	0.500[d]
4	δ=4	1/2 = 0.500	1/2	0.375[c]	0.333[c]	0.250[c]	0.000[c]
	δ=3	4/7 = 0.571	3/7	0.464[d]	0.429[d]	0.357[d]	0.143[d]
	δ=2	2/3 = 0.667	1/3	0.583[d]	0.556[d]	0.500[d]	0.333[d]
	δ=1	4/5 = 0.800	1/5	0.750[d]	0.733[d]	0.700[d]	0.600[d]

[a]Odds ratio describes student knowledge relative to random guessing (Table 1); see Appendix.
[b]δ represents the number of functioning distractors; thus, the student makes final selection of response to question from among the functioning distractors and the correct answer.
[c]Equal to reward.
[d]Greater than reward.

Note: A student is assigned a score of +1 for a correct answer, $-p_D$ deduction for an incorrect answer, and p_R if unanswered (i.e., the student recognizes and admits that he or she does not know).

would increase substantially. In fact, if all unknown questions were left unanswered and the deduction was -1/4 or -1/3 and the reward +3/8 or +1/3, then the scores would be higher than the expected scores even if students could guess without consequence for a wrong answer (number-correct scoring).

In number-correct scoring, students are rewarded when they are able to use their partial knowledge to eliminate one or more of the incorrect distractors prior to selecting their final answer from the remaining options. In a confidence scoring algorithm, there is no reward for differences in partial knowledge between two students who can eliminate different numbers of distractors but still choose the option to leave the question unanswered. It is only when they choose to answer a question that their partial knowledge comes into play and is rewarded. While with confidence scoring we would give up some rewarding of a student's partial knowledge, our principal aim is to promote better self-assessment, which we believe will ultimately be of more benefit to students in their professional careers.

The results of the calculations in Table 3 show that the rewards we investigated are small enough that it would be hard for a poorly prepared student to achieve a passing score based on leaving questions unanswered. Thus, we see no requirement to impose a limit on the number of questions left unanswered that would be rewarded. Moreover, a limit might defeat the purpose of encouraging students to self-assess their knowledge.

By strongly encouraging students to admit what they do not know, confidence scoring could make the examinations a more effective learning experience. When students leave a question unanswered, they have recognized and admitted that they do not fully understand the material; thus, they receive immediate feedback as to the extent of their knowledge of the material as opposed to waiting until after the examination is scored. Nonetheless, for those questions on which they marked an answer, students must wait until the exam is scored before they can completely assess their knowledge. If students recognized that they were not highly confident of the correct answer but marked an answer, they have realized their lack of knowledge. On the other hand, the questions on which students marked an incorrect answer but were highly confident that they knew the correct answer

Table 3. Illustration of consequences of students' decisions regarding how they deal with a question for which they do not know the correct answer

		Knowledge of Correct Answer (%)			
Student Decision	Confidence Scoring Formula	60%	70%	80%	90%
Random guessing δ=3[a]	$-p_D=0$ $p_R=0$ [b]	70	77.5	85	92.5
Random guessing δ=3,[a]	$-p_D=-1/4$ $p_R=3/8$	62.5	71.875	81.25	90.625
no questions unanswered	$-p_D=-1/3$ $p_R=1/3$	60	70	80	90
	$-p_D=-1/2$ $p_R=1/4$	55	62.25	77.5	88.75
	$-p_D=-1$ $p_R=0$	40	55	70	85
Random guessing δ=3,[a] one-quarter	$-p_D=-1/4$ $p_R=3/8$	65.623	74.219	82.813	91.406
of questions for which student does	$-p_D=-1/3$ $p_R=1/3$	63.333	72.5	81.667	90.833
not know correct answer are unanswered	$-p_D=-1/2$ $p_R=1/4$	58.75	69.063	79.375	89.688
	$-p_D=-1$ $p_R=0$	45	58.75	72.5	86.25
Random guessing δ=3,[a] one-half of	$-p_D=-1/4$ $p_R=3/8$	68.75	76.563	84.275	92.188
questions for which student does not	$-p_D=-1/3$ $p_R=1/3$	66.667	75	83.3330	91.667
know correct answer are unanswered	$-p_D=-1/2$ $p_R=1/4$	62.5	71.875	81.25	90.625
	$-p_D=-1$ $p_R=0$	50	62.5	75	87.5
All questions for which student does not	$-p_D=-1/4$ $p_R=3/8$	75	81.25	87.5	93.75
know correct answer are unanswered	$-p_D=-1/3$ $p_R=1/3$	73.333	80	86.667	93.333
	$-p_D=-1/2$ $p_R=1/4$	70	77.5	85	92.5
	$-p_D=-1$ $p_R=0$	60	70	80	90

[a]δ represents the number of functioning distractors; thus, the student makes final selection of his or her response to question from among the functioning distractors and the correct answer.
[b]Corresponds to number-correct scoring.

notifies them that their understanding of the material was extremely poor or completely lacking.

The faculty also could receive more feedback from the proposed confidence scoring algorithm about actual or perceived student knowledge. The authors, based on our experiences as faculty members, interpret questions that were answered incorrectly as indicating far less knowledge and greater misunderstanding by a student than a question left unanswered. Thus, with the availability of data describing different degrees of lack of knowledge, faculty members can better identify areas that are poorly understood by students and take corrective action. Faculty members also would receive information on the development of self-assessment behaviors and skills based on the questions left unanswered.

Our previous prospective imposition of formula scoring in one class of the oral and maxillofacial pathology course resulted in extremely negative reactions by the students.[16] We interpreted this as being due to the fact that the students viewed use of formula scoring as punitive because we took away their reward for guessing. By using confidence scoring, we have changed the reward aspect. Under formula scoring, the differential between the deduction and "reward" of 0 for leaving a question unanswered was small (only 1/4 point). The reward included in our proposed confidence scoring formula increases the differential between a wrong answer deduction and the reward for leaving a question unanswered; this should encourage students to recognize and admit when they do not know the correct answer.

Traub et al. used a reward of 1/5 point for omitting an answer rather than penalizing for incorrect answers in a vocabulary test for ninth-grade students.[24] This magnitude of reward is precisely the expected gain due to random guessing from five options. The reward is a certainty, and students are not at risk of guessing themselves into a lower score. These authors found that the promise of reward was more effective in getting examinees to omit answers than was a deduction for guessing. These results suggest that the reward component may reduce the poor reaction by the students,[16] even though the Traub et

al. study was conducted with students vastly different from our dental students.

We do not expect students to like a confidence scoring system, but we do expect them to learn to function within it, to learn that self-assessment is important for their careers, and to acquire some self-assessment skills. We anticipate that our proposed confidence scoring algorithm will be applicable primarily in the basic science portion of the dental school curriculum and serve as a foundational learning experience for students that will facilitate their transition into the clinical environment. In the clinical portion of the curriculum, students are continually questioned by the faculty about treatment plans and patient care decisions. Students must make self-assessments as they answer questions and explain their treatment plans. Self-assessment is a substantial part of the clinical competency evaluation process at our school; all clinical skill assessments and other graded clinical events include a student self-assessment component. Students may not like this ongoing question and answer component of clinical education, but they accept it as part of their professional development.

Calibration

The choice of deduction and reward will affect student attitudes and behaviors. Faculty members also are required to evaluate the level of a student's knowledge. We previously have shown that formula scoring improved agreement between multiple-choice scores and short-answer scores, that is, the validity of the multiple-choice scores was improved.[15,16] To investigate this when confidence scoring is used, Figure 1 illustrates application of the confidence scoring algorithm with two choices of deduction and reward (deduction -1/4, reward 3/8; and deduction -1/3, reward 1/3) to the multiple-choice

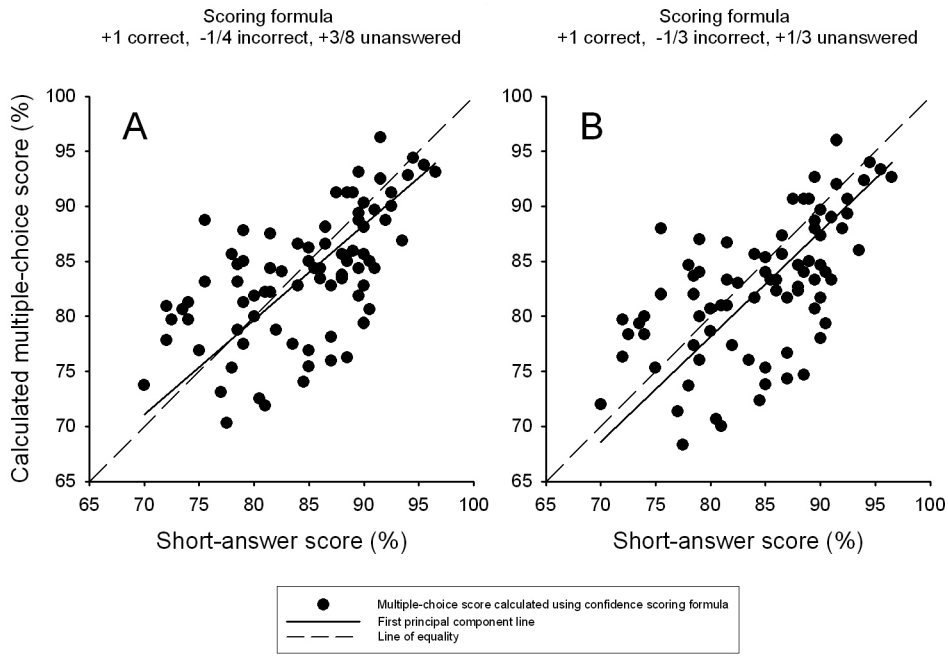

Figure 1. Scatter diagrams and principal component line relating scores calculated using confidence scoring formulas on multiple-choice portion of examinations to scores on short-answer portion

Note: These examinations were given in an oral and maxillofacial pathology course. Multiple-choice (MC) questions made up one-half of the examinations, with short-answer questions making up the other half. The MC component was originally scored using formula scoring (+1 correct answer, -1/4 incorrect answer, 0 unanswered) with complete prospective knowledge by the students; that is, students could elect to leave a question unanswered and not expose themselves to the deduction imposed for an incorrect answer. In the left panel, filled circles represent MC scores calculated using the confidence scoring formula +1 for a correct answer, -1/4 for an incorrect answer, and +3/8 for a question left unanswered. The solid line represents the first principal component line; the dashed line represents equality. In the right panel, filled circles represent MC scores calculated using the confidence scoring formula +1 for a correct answer, -1/3 for an incorrect answer, and +1/3 for a question left unanswered. The solid line represents the first principal component line; the dashed line represents equality.

scores in the oral and maxillofacial pathology course in which prospective formula scoring was used for the half of the examination comprised of multiple-choice questions. Multiple-choice examinations in this class were scored originally with formula scoring (no reward), and students had the opportunity to leave questions unanswered, although very few actually made this choice;[16] thus, in this example, few scores show the benefit of the reward. The deduction of -1/4 would be appropriate for formula scoring assuming random guessing among five options, and the deduction of -1/3 would be appropriate for formula scoring assuming random guessing among four options. The deduction for an incorrect answer for this class should have been -1/3 based on the estimated harmonic mean number of functioning distractors for this class.[19]

The relationships between the confidence scores and the short-answer scores were examined by using principal component lines estimated from the variance-covariance matrix.[15-18] The first principal component line more accurately estimates the linear relationship between these X and Y variables than would ordinary linear regression analysis. The improved accuracy from the use of principal component analysis was because both the X and Y variables were random variables; in ordinary linear regression analysis, only the Y variable is considered to be a random variable, and the estimator of the line is biased if X also is a random variable.

The principal component line relating multiple-choice scores to short-answer scores for deduction -1/4 and reward +3/8 is visually close to the line of equality. The principal component line for deduction -1/3 and reward +1/3 is approximately parallel to the line of equality; this indicates that the scale of the confidence scores and short-answer scores are similar, that is, a 1 point change in score in one examination question format corresponds to a 1 point change in score in the other examination question format. Of note, the principal component line is below the line of equality; this also has been noted previously in unadjusted scores in which the better students (students having higher scores on the short-answer components) had scores on the multiple-choice component that were lower than their short-answer scores.[15] We attributed this to the fact that their better understanding of the material may have enabled them to identify deficiencies in the multiple-choice questions resulting in confusion; alternatively, these lower multiple-choice scores may represent negative cueing. Deductions such as -1/2 result in the principal component line being farther below the line of equality and having a slope greater than 1 (not shown).

We believe that there is a considerable benefit to including questions using an un-cued format as part of the examination if only for the reason of calibration. As shown in Figure 1, the multiple-choice scores computed with a confidence scoring algorithm compared to short-answer scores clearly could be used to help select the pass/fail cutoff value as well as other cutoff values. Use of other question formats for calibration is especially important until experience with the confidence scoring formula is obtained.

Implementation

To be most effective in developing self-assessment behaviors and skills, use of the confidence scoring algorithm should be the policy throughout the curriculum for a health professions education program. In this regard, the decision to implement confidence scoring must be strongly supported by the teaching faculty. In addition, it is imperative that the intent and application of the confidence scoring algorithm be explained in detail to students immediately upon their enrollment. This will make it clear to students that developing self-assessment behaviors and skills to allow them to recognize what they do not know represents a crucial constituent of their professional growth and a major goal of the institution.

A particular selection of the magnitude of the deduction and reward likely will not be uniformly applicable throughout the curriculum, and the magnitudes of the deduction and reward would vary. Nonetheless, it must be kept in mind that any variation in the algorithm might affect student acceptance. A crucial point for sustaining student acceptance is to explain the educational reasons underlying any variation in the scoring algorithm. For instance, beginning dental students likely will have no prior experience in clearly identifying and admitting what they do not know; this would support the case for a smaller deduction (to reduce anxiety about the effect on grades) and a larger reward. Thus, the choice of deduction and reward may need to be modified as students progress through the curriculum from basic science, to preclinical courses, to clinical courses.

We previously implemented the use of the formula scoring algorithm with a -1/4 deduction and no reward in one class of the oral and maxillofacial pathology course.[16] Unexpectedly, we observed significantly improved academic performance by this class, particularly among the lower performing

students.[16,18] A subsequent investigation showed that a deduction of -1/3 better estimated the appropriate formula scoring correction based on the harmonic mean number of functioning distractors used in a class assessed using all multiple-choice questions.[19] We would anticipate improved student performance by using the slightly larger deduction of -1/3 because of the increased challenge.[16,18]

One concern is that use of confidence scoring might lead students to spend more time answering questions because they made a sincere effort to better assess their level of knowledge, and their slower progress through the examination could lead to a shortage of time. As they do currently, faculty members must ensure an appropriate balance between the number of questions and the time available for the examination. Changing the number of questions potentially affects coverage of material and examination reliability, while increasing time for the examination represents a practical scheduling problem in the curriculum.

The knowledge of probability necessary to understand our proposed confidence scoring algorithm could potentially favor, at least initially, students with a better background in probability. Thus, it is incumbent upon faculty members to ensure that students have an adequate understanding of the underlying probability concepts. Students should become more familiar with the reasoning required by the confidence scoring algorithm if, as we recommend, this scoring system is used throughout the curriculum. We believe one of the advantages of our confidence scoring algorithm over an algorithm that requires a quantitative statement of confidence is that it requires only a general understanding of probabilities. While numbers such as we have presented are relevant to student decisions, students would not be presented with such extensive information and choices of deduction/reward; students would be presented with only one deduction and reward. Moreover, we would anticipate presenting example probabilities to students and not the odds ratios, which are more likely familiar to dental faculty members from their reading of the scientific literature.

An important aspect of explaining this system to students is to emphasize self-assessment. Self-assessment is absolutely required as students move into clinical training, and hopefully the clinical instructors will continue to reinforce this. The goal of the procedure we are advocating for multiple-choice examinations is to start the students thinking this way very early in their professional education.

We believe that use of multiple-choice format questions will prevail as the question format of choice at least for the foreseeable future. The confidence scoring algorithm is only one approach to improving student performance and fostering acquisition of self-assessment behaviors and skills. Other approaches to this end could include interactive classroom questioning of randomly selected students and use of non-cued questions in examinations; we note that confidence scoring also could be applied to un-cued question. These procedures could increase challenges to students to more accurately assess their knowledge. However, non-cued questions and questioning during class would require changes and significant time commitments by the faculty, whereas the proposed confidence scoring algorithm could be implemented electronically.

The use of a confidence scoring approach, even if implemented across the curriculum, is an important step but will not be sufficient to promote self-assessment and development of lifelong learning skills. These must be encouraged by the faculty throughout a student's preclinical and clinical training.

Conclusion

In this article, we propose implementation of a confidence scoring algorithm for multiple-choice examinations as one approach to help promote the development of students' self-assessment behaviors and skills to recognize and to admit what they do not know. Our proposed confidence scoring algorithm is thoroughly grounded in our empirical research on student academic performance in an oral and maxillofacial pathology course. The deduction for incorrect answers contained in the confidence scoring formula should promote better examination preparation and academic achievement by students because it increases the demands placed upon them. The reward component of the confidence scoring formula should motivate students to learn to recognize and admit what they do not know. Careful consideration of the confidence students have in their knowledge is required to maximally achieve under this algorithm. Institution-wide implementation of confidence scoring will help provide a learning environment for students to develop the self-assessment behaviors and skills that will increase their awareness of the limits of their knowledge and thereby encourage lifelong learning.

REFERENCES

1. Colthart I, Bagnall G, Evans A, et al. The effectiveness of self-assessment of learner needs, learner activity, and impact on clinical practice. BEME guide no. 10. Med Teach 2008;30:124-45.
2. American Dental Education Association. ADEA statement on professionalism in dental education. J Dent Educ 2014;78(7):1071-6.
3. Commission on Dental Accreditation. Self-study guide for dental education programs. Chicago: American Dental Association, 2012:Section 2-10.
4. Anderson LW, Krathwohl DR, eds. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Addison Wesley Longman, 2001.
5. LaDuca A, Downing SM, Henzel TR. Systematic item writing and test construction. In: Impara JC, ed. Licensure testing: purposes, procedures, and practices. Lincoln, NE: Buros Institute of Mental Measurements, 1995:117-48.
6. Haladyna TM. Developing and validating multiple-choice test items. 3rd ed. Mahwah, NJ: Lawrence Erlbaum, 2004.
7. Downing ST, Haladyna TM, eds. Handbook of test development. Mahwah, NJ: Lawrence Erlbaum, 2006.
8. Kramer GA, Albino JEN, Andrieu SC, et al. Dental student assessment toolbox. J Dent Educ 2009;73(1):12-35.
9. Linn RL, Gronlund NE. Measurement and assessment in teaching. 7th ed. Upper Saddle River, NJ: Prentice-Hall, 1995.
10. Kohler RA. A comparison of the validities of conventional choice testing and various confidence marking procedures. J Educ Meas 1971;8:297-303.
11. Echternacht GJ. The use of confidence testing in objective tests. Rev Educ Res 1972;42:217-36.
12. Rippy RM, Smith S. Improving the reliability and validity of confidence-scored test by adjusting for realism. Eval Health Professions 1979;2:100-9.
13. Davidoff F. Confidence testing: how to answer a meta-question. Am Coll Phys Observer, February 1995.
14. Albino JEN, Young SK, Neumann LM, et al. Assessing dental students' competence: best practice recommendations in the performance assessment literature and investigation of current practices in predoctoral dental education. J Dent Educ 2008;72(12):1405-35.
15. Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. J Dent Educ 2006;70(4):378-86.
16. Prihoda TJ, Pinckard RN, McMahan CA, et al. Prospective implementation of correction for guessing in oral and maxillofacial pathology multiple-choice examinations: did student performance improve? J Dent Educ 2008;72(10):1149-59.
17. Pinckard RN, McMahan CA, Prihoda TJ, et al. Short-answer examinations improve student performance in an oral and maxillofacial pathology course. J Dent Educ 2009;73(8):950-61.
18. Pinckard RN, McMahan CA, Prihoda TJ, et al. Short-answer questions and formula scoring separately enhance dental student academic performance. J Dent Educ 2012;76(5):620-34.
19. McMahan CA, Pinckard RN, Prihoda TJ, et al. Improving multiple-choice questions to better assess dental student knowledge: distractor utilization in oral and maxillofacial pathology course examinations. J Dent Educ 2013;77(12):1593-609.
20. Lord FM. Formula scoring and number-right scoring. J Educ Meas 1975;12:7-12.
21. Farrell G, Leung Y. Convergence of validity for the results of a summative assessment with confidence measurement and traditional assessment. In: Khandia F, ed. 12th CAA International Assisted Assessment Conference: proceedings of the conference on 8th and 9th July 2008 at Loughborough University. Loughborough, UK: Loughborough University, 2008:123-36.
22. Dinsmore DL, Parkinson MM. What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. Learn Instruct 2013;24:4-14.
23. Hattie J. Calibration and confidence: where to next? Learn Instruct 2013;24:62-6.
24. Traub RE, Hambleton K, Singr B. Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. Educ Psychol Meas 1969;29:847-61.

APPENDIX

Students can use their knowledge to eliminate incorrect distractors and thus reduce their problem of answering a multiple-choice question of κ items to answering a question comprised of the correct answer and δ functioning distractors, that is, the options that the student cannot definitely identify as incorrect. If reducing the problem to that level represents the limit of their knowledge, then students' choice is either to leave the question unanswered or to guess randomly among the δ+1 remaining items. The probability of randomly guessing the correct answer for such a multiple-choice question is π_{Random}=1/δ+1.

Students may have knowledge that would enable them to go beyond simply identifying the set of items containing the correct answer and functioning distractors, that is, they have some knowledge as to which of the remaining options is the correct answer. Thus, they may be able to use their partial knowledge to improve their chance of guessing the correct answer among this identified set of options. The probability describing the students' true knowledge of the correct answer is specified by π. This probability, which is used by an individual student to make a decision about how to answer the question, is a personal probability or degree of belief. Personal probabilities will vary among students and may overestimate or underestimate the student's true knowledge. To make use of the numbers that we calculate, we assume that the personal probabilities of the student are calibrated and thus the probability π correctly estimates the student's knowledge in a relative frequency interpretation. The relative frequency approach requires that, in a large number of questions for which the student expresses this same degree of certainty that he or she knows the correct answer, the proportion (relative frequency) of this large number of questions for which the student's answer is correct is π.

The level of knowledge beyond the knowledge that enabled the student to simply identify the set of items containing the correct answer and functioning distractors can be expressed by this odds ratio:

$$\phi = \frac{\pi(1-\pi_{Random})}{(1-\pi)\pi_{Random}}$$

Alternatively, the knowledge of the student expressed in terms of the foregoing odds ratio and the random guessing probability is expressed with this ratio:

$$\pi = \frac{\phi\pi_{Random}}{1-\pi_{Random}+\phi\pi_{Random}}$$

If $\phi = 1$, the student has no additional knowledge ($\pi = \pi_R$). The odds ratio $\phi \to \infty$ represents the student's knowing the correct answer ($\pi = 1$). The odds ratio $\phi > 1$ represents $\pi > \pi_{Random}$ and describes some knowledge in addition to defining the set of items containing the correct answer. The odds ratio $\phi < 1$ represents $\pi < \pi_{Random}$ and describes a situation in which the student has some knowledge but is applying that knowledge incorrectly and thus is not expected to achieve even what he or she could accomplish by guessing randomly.