Title

Identifying Student Misconceptions in Biomedical Course Assessments in Dental Education

Permalink

https://escholarship.org/uc/item/8cq7j1fx

Journal

Journal of Dental Education, 76(9)

ISSN

0022-0337

Authors

Curtis, Donald A
Lind, Samuel L
Dellinges, Mark
et al.

Publication Date

2012-09-01

DOI

10.1002/j.0022-0337.2012.76.9.tb05373.x

Peer reviewed

Identifying Student Misconceptions in Biomedical Course Assessments in Dental Education

Donald A. Curtis, D.M.D.; Samuel L. Lind, Ph.D.; Mark Dellinges, D.D.S.; Kurt Schroeder, D.D.S.

Abstract: Dental student performance on examinations has traditionally been estimated by calculating the percentage of correct responses rather than by identifying student misconceptions. Although misconceptions can impede student learning and are refractory to change, they are seldom measured in biomedical courses in dental schools. Our purpose was to determine if scaling student confidence and the clinical impact of incorrect answers could be used on multiple-choice questions (MCQs) to identify potential student misconceptions. To provide a measure of student misconception, faculty members indicated the correct answer on twenty clinically relevant MCQs and noted whether the three distracters represented potentially benign, inappropriate, or harmful application of student knowledge to patient treatment. A group of 105 third-year dental students selected what they believed was the most appropriate answer and their level of sureness (1 to 4 representing very unsure, unsure, sure, and very sure) about their answer. Misconceptions were defined as sure or very sure incorrect responses that could result in inappropriate or harmful clinical treatment. In the results, 5.2 percent of the answers represented student misconceptions, and 74 percent of the misconceptions were from four case-based interpretation questions. The mean student sureness was 3.6 on a 4.0 scale. The students' sureness was higher with correct than with incorrect answers ($p<0.001$), yet there was no difference in sureness levels among their incorrect (benign, inappropriate, or harmful) responses ($p>0.05$). This study found that scaling student confidence and clinical impact of incorrect answers provided helpful insights into student thinking in multiple-choice assessment.

Dr. Curtis is Professor, Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco; Dr. Lind is Associate Professor, School of Economics and Business Administration, Saint Mary's College of California; Dr. Dellinges is Health Sciences Professor, Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco; and Dr. Schroeder is Health Sciences Assistant Clinical Professor, School of Dentistry, University of California, San Francisco. Direct correspondence and requests for reprints to Dr. Donald A. Curtis, Department of Preventive and Restorative Dental Sciences, School of Dentistry, University of California, San Francisco, 707 Parnassus Avenue, San Francisco, CA 94143; 415-476-5827; don.curtis@ucsf.edu.

Keywords: dental education, dental students, assessment, educational methodology, multiple-choice questions

Submitted for publication 4/18/11; accepted 11/27/11

Assessments provide a measure of student progress, motivate students to excel, provide feedback for curriculum refinement, and can be a valuable means to improve students' conceptual understanding.[1-8] Assessments are also an important tool in certifying and credentialing processes.[8,9] The collective use of assessments in the development and certification of the dental professional is intended to provide a measure of confidence for both the graduating student and the general public that the health care provider is professionally competent.

One of the most widely used instruments in health professions education is the multiple-choice assessment. Multiple-choice questions (MCQs) offer an efficient way to assess a large number of students over wide subject domains with reliability.[3,4,7,10,11] While MCQs are most often written to assess knowledge and understanding, higher levels of learning such as application and synthesis can also be evaluated.[12] MCQs can effectively discriminate high and low achievers, and both students and faculty members are familiar with the format.[12]

Despite the many efficiencies of MCQs, criticisms of this type of assessment are numerous, varied, and often justified.[2,13,14] Many of the criticisms relate to the difficulty and skill necessary to develop valid context-based MCQs and the unfortunate situation that many educators have neither the time nor the training necessary to write the type of MCQs that test higher-order learning.[13] Not surprisingly, a frequent criticism of MCQs is that they foster study habits consistent with superficial or strategic learning in which the student is interested in recognizing isolated facts rather than relationships or higher levels of learning.[15] Another concern is that MCQs are not a pure measure of knowledge because reading comprehension and test-taking ability are also relevant to the final score.[13,16] MCQs are thought to overestimate learning compared to short-answer questions because all the students need to do is recognize the answer from the options in front of them whereas on a short-answer question they need to think of the response. Finally, MCQs provide little feedback to the instructor about what students do not know or the confidence students have in their responses. Therefore, misconceptions by students who are confident in their incorrect answers are often not identified by conventional MCQs.[18-20]

A misconception is defined as an erroneous thought, idea, or notion that results in a student's misunderstanding or misinterpretation of information. Student misconceptions generally originate with previous student learning, are difficult to detect, and are resistant to change, all of which can interfere with learning.[17,21,22] Misconceptions can also occur when students believe an incorrect choice on a multiple-choice examination is actually correct, a situation termed a negative testing effect.[2,23,24] Thus, not only can refractory student misconceptions occur because of previous student mislearning, but they also can be introduced through testing.[2]

The importance of misconceptions to student learning has been recognized and studied in academic areas such as physics.[25-28] Kohnle et al. argued that pre-existing knowledge and beliefs can strongly influence how new concepts in physics are understood and used a free-response pretest to identify misconceptions and a multiple-choice assessment to better understand student reasoning and misconceptions.[27] Prather used MCQs, written justifications, and interviews to help identify misconceptions on the subject of radioactive decay.[26] Unfortunately, in the biomedical sciences, misconceptions are seldom measured, and their implications seldom evaluated.[29]

Students with high confidence in their incorrect responses on MCQs often have misconceptions that are highly resistant to modification.[2,24,30] Conversely, studies have suggested that students with low-confidence incorrect responses are comparatively responsive to learning. In a study of 167 adults who completed a pretest, course of instruction, and posttest with measures of correctness and confidence, learners confident of an incorrect response on the pretest had only 21 percent correct on the posttest, while learners unsure and incorrect on the pretest had 72 percent correct responses on the posttest.[30] This study found that the unsure incorrect learners were at least three times more likely to learn from a course of study than confidently incorrect learners. Similarly, Butler et al. found in a study of thirty students with immediate and delayed feedback on consecutive tests that high-confidence errors were shown to be refractory.[24] Not only are misconceptions difficult to change, but students may also distort or ignore new information that conflicts with their existing perceptions.[30,31] Additionally, for health care professionals, misconceptions are not always benign and may lead to poor clinical decisions that are inappropriate or potentially harmful to the patient. Therefore, developing assessments that identify student misconceptions is an important goal, and these assessments can provide important feedback to students as well as to faculty.

MCQs are poor identifiers of student misconceptions for at least three reasons: 1) if an uninformed student guesses correctly, there is no feedback to prompt any improvement in his or her understanding; 2) in traditional MCQs, all distracters are considered equal, when, in fact, some distracters may identify different degrees of misconception; and 3) most importantly, there is no feedback as to the student's confidence in incorrect answers. For example, if a student is incorrect but unsure, the student is uninformed, while if a student is incorrect but sure, he or she is laboring under a misconception.[24]

One way to identify misconceptions on MCQs is to ask students to indicate their confidence about test responses.[2,17,18,32] When students know they will be asked to consider their confidence level in an answer, they are more likely to attempt to resolve uncertainties.[33] Testing student confidence in MCQs helps identify students' misconceptions, encourages their preparation for assessments, and increases their interest in additional learning.[34] Klymkowsky et al. proposed a scale to measure sureness and correctness and found that measuring student confidence provided instructors with valuable feedback about student confusion and misinformation.[17]

Traditionally, assessments have been used to provide a quantitative measure of student understanding rather than a measure of student misconceptions, so that we generally have no idea how sure students were of their incorrect answers. Additionally, by counting only correct responses, we assume all distracters are equally consequential—which is not likely. Some distracters may represent student thinking that would result in benign application to clinical outcomes, while others could result in unintended or potentially harmful clinical results. Understanding both the student confidence in and clinical impact of incorrect answers on a multiple-choice examination would allow multidimensional scaling of information not possible with conventional MCQs. This scaling would provide a measure of misconception, allowing insights into student thinking not measured by the traditional univariate method of percent correct responses.

Although investigators have previously used measures of student confidence, combining a multiple-choice assessment with a measure of confidence and clinical impact of incorrect answers has not been attempted to our knowledge. Our purpose was to determine if scaling student confidence and the clinical impact of incorrect answers could be completed on MCQs to identify potential student misconceptions.

Methods

One hundred and five third-year University of California, San Francisco (UCSF) dental students (fifty-nine male, forty-six female), consisting of eighty-two traditional predoctoral students and twenty-three international students from the graduating class of 2011, were considered for inclusion in this study. All students from each group participated. As part of their routine requirements, they completed an endodontic assessment that included twenty multiple-choice questions and a measure of how sure they were about each response.

Clinical faculty members developed twenty multiple-choice, context-based questions to evaluate students' diagnostic ability, clinical assessment, and clinical judgment. For each question, there were four possible responses: one most appropriate answer and three incorrect distracters. Two faculty members independently scaled the distracters as benign, inappropriate, or harmful (not all questions contained benign, inappropriate, and harmful responses), and a percent agreement between the two faculty members was completed.

Assessment Instrument

Selected questions from the assessment instrument used in the study are shown in the Appendix. In the questions shown there, the correct answer is indicated by an asterisk, and the faculty members' designation of the clinical impact of each distracter is noted.

Each distracter was designated by faculty members as benign, inappropriate, or harmful in order to provide a relative measure of danger should the action be completed on a patient. The categories were defined as follows. A "benign" answer results in an inconsequential or harmless patient outcome. An example can be found in question 23 (see Appendix), in which students were asked about the most important characteristic of a restorative material on the pulp. "The speed the material set" is a benign response as its application to patient well-being would be negligible. An "inappropriate" response means that the answer was either unsuitable or would delay appropriate patient care. For example, on question 13, students were asked about the clinical approach for a patient with a fractured tooth. The option "antibiotics and monitoring" is an inappropriate answer because it does not address the clinical problem, delays appropriate care, and is an inappropriate use of antibiotics. A "potentially harmful" answer would result in direct and irreversible detrimental treatment to the patient. For example, on question 13, "extraction of the tooth" would be an unnecessary, detrimental patient outcome.

For each question, the students were asked to state their level of sureness or confidence on a rating scale of 1, 2, 3, and 4 representing very unsure, unsure, sure, and very sure, respectively. The mean level of sureness was identified for each level of clinical impact (benign, inappropriate, and harmful) to determine if students were less confident when they selected answers that were potentially more dangerous. Students were told their sureness responses were important to document, but were not told how their sureness responses would affect grades.

For purposes of this study, a misconception was defined as a student response of sure or very sure to an incorrect answer that, if applied to clinical care, would result in potentially inappropriate or harmful care. We realize that all strongly held incorrect beliefs, even those with benign outcomes, represent misconceptions; however, our interest was in identifying those misconceptions more likely to adversely affect patient care.

One examination was administered to all students with results evaluated in two ways. First, the data collected using the assessment as a traditional MCQ examination resulted in percentage of correct answers. Second, the data combining student sureness with the designation of clinical impact provided a measure of student misconception. Test results were collected with code numbers replacing names so that confidentiality was maintained. The protocol for the study was reviewed and approved by the Committee of Human Research at UCSF.

Data Analysis

Descriptive statistical measures included the percentage of correct and incorrect answers and the range of scores by individual student and by each question. Descriptive measures for misconception included a determination of questions resulting in the most misconceptions, the average number of student misconceptions, and the students with the most misconceptions. The sureness score for each student on each question was recorded as 1, 2, 3, or 4 for very unsure, unsure, sure, and very sure, respectively. Additionally, the mean sureness score was calculated at each level of clinical impact (benign, inappropriate, or harmful) to determine if student confidence changed with more dangerous responses.

We wanted to know the relationship between clinical impact and sureness. For example, are students who select a harmful response less confident than when they select a benign response? Practically, we would hope for at least some hesitation when students select a potentially dangerous alternative. Insight into this important question was evaluated two ways. First, we compared the mean level of student sureness at each level of clinical impact (correct answers, incorrect benign, incorrect inappropriate, and incorrect harmful answers) for statistical significance ($p<0.05$) by Analysis of Variance (ANOVA) with Tukey-Kramer Multiple Comparisons Test. Second, we completed a nonparametric Spearman's correlation between sureness and clinical impact for all 2,100 student responses and a second correlation for the 204 incorrect answers. We wanted to know if student confidence was lower when a student selected an incorrect answer that was more dangerous. The percent agreement was also calculated for the two faculty members' concurrence of their designation of benign, inappropriate, and harmful responses for the sixty distracters on the twenty questions.

Results

The exam consisted of twenty questions paired with sureness responses from 105 students for a total of 2,100 student responses to evaluate. The mean exam score was 90 percent correct (range of 55 percent to 100 percent). Of the 204 total incorrect answers, 110 or 53.9 percent were student misconceptions: eight (3.9 percent) were potentially harmful application of student knowledge to clinical care in which the student was very sure; five (2.5 percent) were potentially harmful application of student knowledge to clinical care in which the student was sure; forty-eight (23.5 percent) were potentially inappropriate application of student knowledge to clinical care in which the student was very sure; and forty-nine (24.0 percent) were potentially inappropriate application of student knowledge to patient care in which the student was sure.

Of the correct responses (no danger), the mean student sureness was 3.70±0.55, benign 3.24±0.80, inappropriate 3.07±0.86, and harmful 3.11±0.97 (Table 1). Mean sureness varied by level of clinical impact. ANOVA showed sureness was significantly higher ($p<0.001$) in correct as opposed to all incorrect

Table 1. Numbers of student responses in each category and mean sureness scores

	Correct		Incorrect	
Sureness Responses	No Error	Benign	Inappropriate	Harmful
Very sure	1,423	25	48	8
Sure	391	17	49	5
Unsure	77	12	29	5
Very unsure	5	0	5	1
Mean sureness score	3.70*	3.24	3.07	3.11

*statistical significance at $p<0.001$

Note: Sureness responses are the number of responses in each category; there were 2,100 total responses. Clinical impact ranged from no impact (correct) to incorrect harmful. Student sureness was based on a four-point scale on which 1, 2, 3, and 4 represented very unsure, unsure, sure, and very sure, respectively. The mean sureness levels were higher on correct than incorrect answers ($p<0.001$), but there was no significant difference among incorrect answers ($p>0.05$). Therefore, students were just as confident of harmful incorrect answers as they were of benign answers.

answers, but that there was no difference in sureness among incorrect answers ($p>0.05$) (Table 1). In other words, student confidence was similar when students selected harmful, inappropriate, or benign incorrect answers. The Spearman's correlation between sureness and clinical impact was consistent with the ANOVA finding. We found a statistically significant, although weak, correlation between sureness and clinical impact for all 2,100 student responses (rho=0.26; $p<0.001$), but insignificant among the 204 incorrect responses (rho=0.076; p=0.281). Four questions (#9, #11, #13, and #15) accounted for 74 percent (81/109) of the misconceptions. Thus, 20 percent of the questions accounted for 74 percent of the student misconceptions.

There was no statistically significant difference between male and female students with respect to their mean correct answers (90.6 percent correct for male, 89.9 percent for female), mean composite sureness score (3.6 for both male and female), or misconceptions as a percentage of total answers (2.7 percent, male; 2.5 percent, female). There was little difference in the sureness/clinical impact correlation of the 2,100 responses when comparing males versus females (males, rho=0.26, $p<0.001$; females, rho=0.25, $p<0.001$). Seventy students had either one or more misconceptions, with seven students (four male and three female) having more than three responses identified as misconceptions.

The agreement between the two faculty members on the designation of benign, inappropriate, and harmful outcomes was 52/60 or 87 percent agreement. In all the disagreements, one faculty member identified the response as benign, while the other one identified it as inappropriate; the benign designation was used for all eight.

Discussion

This study found that additional information was obtained by identifying the clinical impact of incorrect answers (benign, inappropriate, or harmful) and the students' identifying the sureness of their answers. We had four specific findings. First, the group of students with the lowest percent scores was similar, but not identical, to the group of students with the most misconceptions. Second, the students were more sure about correct versus incorrect answers ($p<0.001$). Third, when evaluating the sureness of incorrect answers across different levels of clinical impact (benign, inappropriate, or harmful), we found no difference in students' level of sureness ($p>0.05$). Fourth, there was no difference between male and female students with respect to performance on the traditional or misconception measures.

Of the total 2,100 student responses, there were 110 misconceptions on which students were sure or very sure of an incorrect answer that could result in inappropriate or harmful clinical care. Seventy students had one or more misconceptions. Of those seventy, seven (four male and three female) had three or more misconception responses: 25 percent of the total.

Students with numerous misconceptions may require careful review. These students were relatively uninformed, often incorrect, and yet confident in decisions that could result in inappropriate or harmful care to patients. These students would likely need remediation above and beyond simply becoming informed about what they had not learned. This is because reshaping student misconception is much more difficult than informing the uninformed student.[2,24,30]

Being uninformed is different from holding a misconception, and faculty feedback is significantly different in each situation.[2,24,30,35] Students are uninformed when they select an incorrect answer and admit they are unsure. This combination of being incorrect and unsure is considered a very appropriate "teaching moment," during which students are especially responsive to faculty feedback and to learning.[30] Similarly, low-confidence correct answers are also an opportunity during which early feedback increases retention and improves metacognitive monitoring.[24] In a series of experiments on thirty students, prompt feedback was found to significantly improve retention of low confidence incorrect responses.[24] In contrast, students have a misconception when they select an incorrect answer but state that they are sure or very sure of their response, i.e., what they believe is wrong, which is very different from just being uninformed. Ecker et al. found that even with strong retractions, faculty members often fail to eliminate continued effects associated with relatively weak encoding of student misconceptions.[35] The students identified by the measure of student misconception but not the traditional assessment may need reshaping of misinformation rather than additional factual knowledge. Rather than measuring a knowledge domain, identifying misconception is providing a measure of student misunderstanding.

The students in our study did poorly on the four questions that used radiographs or a clinical picture to ask about diagnosis and/or treatment decisions.

These four questions (#9, #11, #13, and #15) required clinical judgment to answer correctly and are consistent with case-based reasoning (CBR) theory, which requires higher-order cognition.[36,37] Case-based questions offer the advantages of testing knowledge at an operational level, allowing candidates to propose solutions in domains that are not completely understood, and more closely tie clinical findings to actions.[36,37] This makes CBR effective in testing for problem-solving, a critical and difficult skill in establishing clinical competence.[36,37] The low levels of student correctness and high levels of student misconception we found on these four questions will result in curriculum modifications to increase student exposure to case-based questions.

The students' sureness levels did not vary significantly with increasing danger of incorrect responses. The students were just as sure of harmful incorrect responses as they were of benign incorrect answers (Table 1). Ideally, we would hope that students would be less confident in responses that have a potentially harmful impact on patient care, but we did not find that. With the average sureness score at 3.6/4.0, the students' average level of confidence in their answers was relatively high. The low level of calibration between confidence/correctness overall and confidence/clinical impact of incorrect answers (Spearman's rho=0.076, p=0.281) is difficult to explain, but is obviously important in our efforts to develop clinical judgment and self-awareness in future dentists. Recent research has found that judgment and decision making can be distorted by many cognitive and motivational biases, and often students will see bias in others but not in their own performance.[31] Much of human judgment is driven by nonconscious processes, and Pronin has concluded from the current literature that self-enhancement biases (a person's inclination to see him- or herself in a positive light, even when evidence suggests otherwise) can compromise the quality of human judgment and decision making.[31] Our findings of student difficulty in the confidence/correctness and confidence/clinical impact calibration may relate to the issue of self-enhancement biases discussed by Pronin.

We may have underreported the percentage of student misconceptions for two reasons. First, there was disagreement between the two faculty members who rated the distracters as benign, inappropriate, or harmful. Where faculty members disagreed on the designation, we used benign, resulting in 5.2 percent misconceptions; using the alternate rater's designation of inappropriate would have increased the percentage of student misconceptions to 6.0 percent. Second, we defined misconception as a confident (sure or very sure) incorrect answer in which clinical application of that answer could result in inappropriate or harmful care. We used this very narrow definition to provide a measure of consequential validity. It could be argued that strongly held incorrect beliefs of any sort are misconceptions. If we defined misconception more broadly to include all incorrect strongly held (sure or very sure) answers and benign as well as inappropriate and harmful responses, then the total misconceptions would be 7.2 percent rather than our reported 5.2 percent.

Limitations of this study include having results from just one assessment. Our results may or may not generalize to other subject areas. Future investigations will be to evaluate student confidence and correctness over time and in multiple subject areas.

Conclusions

The primary findings of our study were the following. First, 5.2 percent of the student responses were misconceptions in which students were sure or very sure of an inappropriate or potentially harmful response. Identifying misconceptions is relevant in that remediation strategies are different for misinformed, as opposed to uninformed, students. Second, sureness levels did not differ significantly among incorrect student answers. Students were just as confident of incorrect harmful answers as they were of incorrect benign answers; therefore, feedback to students about the clinical impact of their decision making becomes important. Third, the students had the highest number of misconceptions with diagnosis and treatment planning questions that required clinical judgment interpretations from radiographs and/or clinical pictures.

Educators and students are familiar with MCQs, which may invite complacency regarding this means of assessment from both groups. Students see MCQs at every level of education, from grade school to graduate studies. The content changes, but the basic underlying premise remains unchanged: MCQs are employed to provide a quantitative measure of knowledge and understanding because we assume our primary objective is to inform the uninformed. We use correct answers as the learner outcome to measure. We do not measure student misconception; we do not measure student confidence; we do not look at the interaction between misunderstanding and student confidence. We just look at correct responses. Students learn to satisfy our stated requirements by passing multiple-choice exams, and we gather objective measures that tell us those students have become informed. We do a pretty good job at this and so do the students. Missing from this process is consequential validity because we are only evaluating what a student knows and not his or her misconceptions.

A primary goal in health professions education has not changed much since the fifth century BCE, when Hippocrates famously advised, "At least do no harm." Doing no harm is a basic tenet of providing clinical care and is an integral part of quality assurance programs in hospitals and clinics. Unfortunately, we seldom measure the potential for doing harm in biomedical education for dental students. This is a problem. The primary outcome measured in the didactic portion of these students' education is the correctness of student responses. Assessments are generally focused on dichotomous outcomes—whether the student knows something or not—rather than on what the implications of incorrect responses might be. Identifying misconceptions should be emphasized more in biomedical courses in dental education. Identifying and reshaping what students *think* they know may be as important as measuring what students *do* know.

Acknowledgments

The authors would like to recognize and thank Dr. David Chambers for his contributions to this article.

REFERENCES

1. Roediger H, Karpicke J. Test-enhanced learning: taking memory tests improves long-term retention. Psychol Sci 2006;17(3):249–55.
2. Chang C, Yeh T, Barufaldi JB. The positive and negative effects of science concepts tests on student conceptual understanding. Int J Sci Educ 2010;32(2):265–82.
3. Ghiabi E, Taylor KL. Teaching methods and surgical training in North American graduate periodontics programs: exploring the landscape. J Dent Educ 2010;74(6):618–27.
4. Albino JE, Young SK, Neumann LM, Kramer GA, Andrieu SC, Henson L, 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.
5. Curtis DA, Lind SL, Brear S, Finzen FC. The correlation of student performance in preclinical and clinical prosthodontic assessments. J Dent Educ 2007;71(3):365–72.
6. Curtis DA, Lind SL, Dellinges M, Setia G, Finzen FC. Dental students' self-assessment of preclinical examinations. J Dent Educ 2008;72(3):265–77.
7. Norcini J, McKinley D. Assessment methods in medical education. Teach Teacher Educ 2007;23:239–50.
8. Buckley S, Coleman J, Davison I. The educational effects of portfolios on undergraduate student learning: a best evidence medical education (BEME) systematic review. BEME guide no. 11. Med Teach 2009;31:282–98.
9. DeChamplain A. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010;44:109–17.
10. Tasdemir T. A comparison of multiple-choice tests and true-false tests used in evaluating student progress. J Instructional Psychol 2010;37(3):258–66.
11. Tarrant M, Ware J, Mohammed A. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ 2009;9(40):1–8.
12. Schuwirth L, Vleuten C. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;26(8):974–9.
13. Boland R, Lester N, Williams E. Writing multiple-choice questions. Acad Psychiatry 2010;34(4):310–6.
14. Koivula N, Hassmen P, Hunt D. Performance on the Swedish scholastic aptitude test: effects of self-assessment and gender. Sex Roles 2001;44(11/12):629–45.
15. Entwistle N, Skinner D, Entwistle D, Orr S. Conceptions and beliefs about "good teaching": an integration of contrasting research areas. Higher Educ Res Development 2000;19:5–26.
16. Park J. Constructive multiple-choice testing system. Br J Educ Tech 2010;41(6):1054–64.
17. Klymkowsky M, Taylor L, Spindler S, Garvin-Doxas R. Two-dimensional, implicit confidence tests as a tool for recognizing student misconceptions. J Coll Sci Teach 2006:44–8.
18. Walker D, Thompson J. A note on multiple choice exams, with respect to students' risk preference and confidence. Assessment Eval Higher Educ 2010;26(3):261–7.
19. Pelaez N, Boyd D, Rojas J, Hoover M. Prevalence of blood circulation misconceptions among prospective elementary teachers. Adv Physiol Educ 2005;29:172–81.
20. McCourbie P. Improving the fairness of multiple-choice questions: a literature review. Med Teacher 2004;26(8):709–12.
21. Liu T, Lin Y, Tsai C. Identifying senior high school students' misconceptions about statistical correlation and their possible causes: an exploratory study using concept mapping with interviews. Int J Sci Math Educ 2008;7:791–820.
22. Draper S. Catalytic assessment: understanding how MCQs and EVS can foster deep learning. Br J Educ Technol 2008;40(2):285–93.
23. Marsh E, Agarwal P, Roediger H. Memorial consequences of answering SAT II questions. J Exp Psychol Appl 2009;15(1):1–11.
24. Butler A, Karpicke J, Roediger H. Correcting a metacognitive error: feedback increases retention of low confidence correct responses. J Exp Psychol Learn Mem Cogn 2008;34(4):918–28.

25. Burgoon J, Heddle M, Duran E. Re-examining the similarities between teacher and student conceptions about physical science. J Sci Teacher Educ 2011;22:101–14.
26. Prather E. Students' beliefs about the role of atoms in radioactive decay and half-life. J Geosci Educ 2005;53(4):345–54.
27. Kohnle A, Mclean S, Aliotta M. Towards a conceptual diagnostic survey in nuclear physics. Eur J Phys 2011;32:55–62.
28. Gonzalez-Espada W, Birriel J, Birriel I. Discrepant events: a challenge to students' intuition. The Physics Teacher 2010;48:508–11.
29. Loftus E. Planting misinformation in the human mind: a thirty-year investigation of the malleability of memory. Learn Mem 2005;12(4):361–6.
30. Wakabayshi T, Guskin K. The effect of an "unsure" option on early childhood professionals' pre- and posttraining knowledge assessments. Am J Eval 2010;31(4):486–98.
31. Pronin E. Perception and misperception of bias in human judgment. Trends Cogn Sci 2007;11(1):37–43.
32. Hunt D. Effects of human self-assessment responding on learning. J Appl Psychol 1982;67:75–82.
33. Gardner-Medwin A. Confidence-based marking: towards deeper learning and better exams. London: Routledge, 2006.
34. Prihoda TJ, Pinckard RN, McMahan CA, Littlefield JH, Jones AC. 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.
35. Ecker U, Lewandowsky S, Swire B, Chang D. Correcting false information in memory: manipulationg the strength of misinformation encoding and its retraction. Psychon Bull Rev 2011;18:570–8.
36. Eshach H, Bitterman H. From case-based reasoning to problem-based learning. Acad Med 2003;78(5):491–6.
37. Kim H, Hannafin M. Developing situated knowledge about teaching with technology via web-enhanced case-based activity. Computers Educ 2011;57:1378–88.

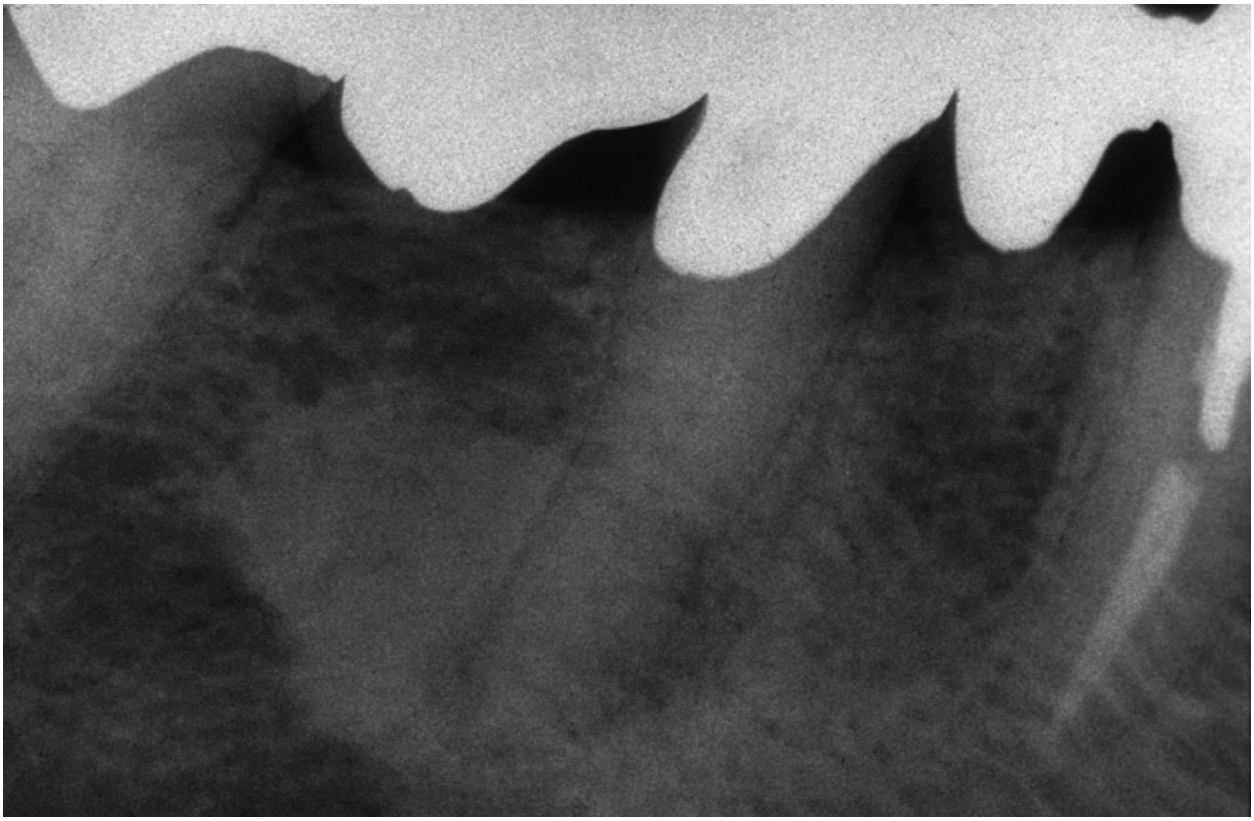
APPENDIX

Selected Questions on Assessment Instrument Used in Study

Note: The instrument was an endodontic examination using context-based questions designed to assess a student's diagnostic and treatment planning skills. The correct answer is indicated on each question with an asterisk. For each distracter, the faculty members' designation of benign, inappropriate, or harmful has been added. For a copy of the entire instrument, contact the corresponding author.

3. A patient has a sensitive tooth #4, a non-endodontically treated tooth that had a crown placed two years previously. An apical radiolucency is now evident on the radiograph. This patient most likely requires:
    a. New crown with less occlusal contact — Inappropriate
    b. Additional diagnostic information before treatment*
    c. Extraction if the pain has persisted for weeks — Harmful
    d. Apicoectomy if the patient would rather not have a root canal — Harmful

4. How sure are you of your answer in question #3?
    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure



APPENDIX *(continued)*

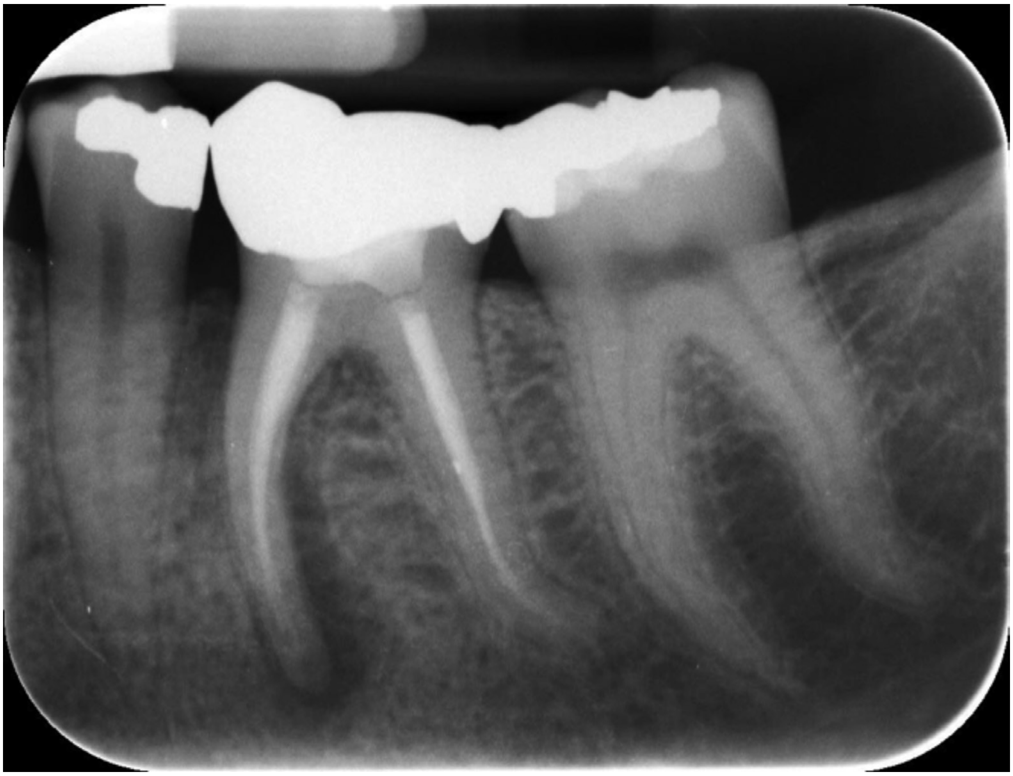
9. Based on the radiograph shown above and the patient report of pain from tooth #29, the most likely treatment would be:

    a. Endodontic treatment with monitoring*
    b. Extraction, biopsy, and monitoring — Harmful
    c. Apicoectomy and monitoring — Harmful
    d. Antibiotics and monitoring — Inappropriate

10. How sure are you of your answer in question #9?

    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure

11. A patient presents with an isolated periodontal pocket with slight swelling adjacent to the periodontal pocket (picture with j shaped lesion). The most likely diagnosis is:

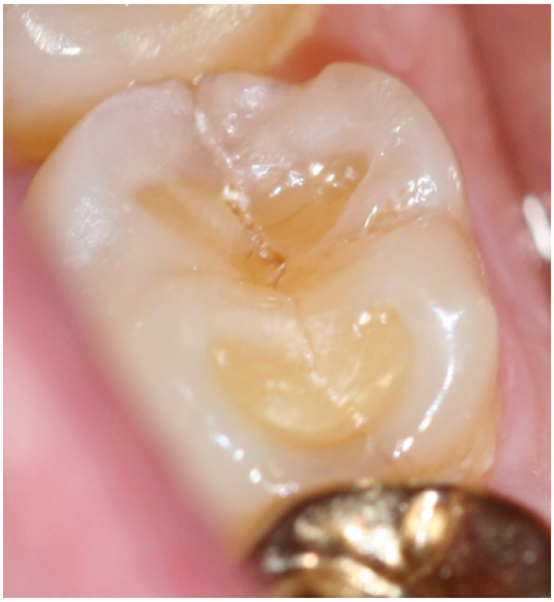


    a. Abscess of periodontal origin — Inappropriate
    b. Vertical root fracture*
    c. Untreated apical inflammation — Benign
    d. Generalized periodontal condition — Inappropriate

12. How sure are you of your answer in question #11?

    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure

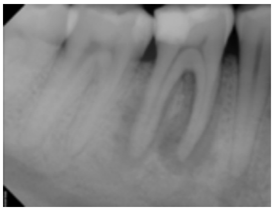
13. Based on the clinical photograph shown below, tooth #31 needs evaluation. The tooth is percussion-sensitive; there is no periodontal pocket, no spontaneous pain, and the tooth tests vital. The most appropriate treatment sequence would be the following:



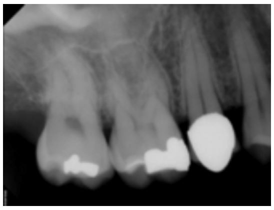
a.	Extraction of the tooth because the crack is extensive	Harmful
b.	Prepare tooth for a crown to determine extent of crack*	
c.	Referral to a periodontist to evaluate periodontal status	Inappropriate
d.	Antibiotics and monitoring as a conservative approach	Inappropriate

14. How sure are you of your answer in question #13?
    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure

15. Which of the teeth labeled in the following pictures would be most prone to ledge formation when filing?

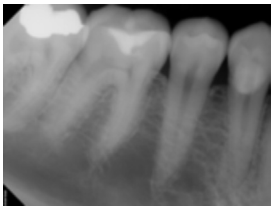


A. Tooth #30*

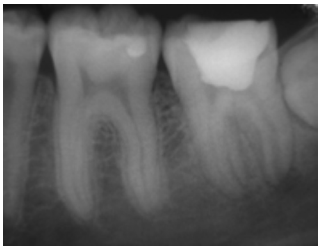


B. Tooth #4 Inappropriate

APPENDIX *(continued)*



C. Tooth #28 Inappropriate



D. Tooth #18 Inappropriate

16. How sure are you of your answer in question #15?
    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure

23. What is the most important characteristic of any restorative material in determining its effect on pulp tissue?
    a. Speed the material sets Benign
    b. Ability to form a marginal seal*
    c. Amount of fluoride released by the material Benign
    d. The type of material is not important Inappropriate

24. How sure are you of your answer in question #23?
    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure

37. A 46-year-old patient presents to your office with a swelling in the lower right posterior quadrant for a one-day duration. Patient is febrile with tender lymph nodes. Clinically, there is a PFM crown on tooth #30. Fluctuant swelling is evident on the buccal vestibule adjacent to tooth #30. Periodontal pockets are 3-4 mm, and teeth #29, 30, and 31 are tender to percussion and palpation. #30 has no response to cold stimuli; however, #29 and 31 do have a response that does not linger. Radiographically, the PDL on tooth #30 is slightly widened. Based on the pulp testing results, what is the endodontic diagnosis and how would you treat this patient?
    a. Refer to a periodontist immediately Inappropriate
    b. Provide an occlusal adjustment of the opposing dentition Inappropriate
    c. Pulpal debridement, incision and drainage, and antibiotics*
    d. A vertical root fracture is likely and extraction recommended Harmful

38. How sure are you of your answer in question #37?
    a. Very sure
    b. Sure
    c. Unsure
    d. Very unsure