Assessment

Comparison of Three Evidence-Based Practice Learning Assessment Methods in Dental Curricula

Asim A. Al-Ansari, DScD; Maha M.A. El Tantawi, PhD

Abstract: Incorporating evidence-based practice (EBP) training in dental curricula is now an accreditation requirement for dental schools, but questions remain about the most effective ways to assess learning outcomes. The purpose of this study was to evaluate and compare three assessment methods for EBP training and to assess their relation to students’ overall course grades. Participants in the study were dental students from two classes who received training in appraising randomized controlled trials (RCTs) and systematic reviews in 2013 at the University of Dammam, Saudi Arabia. Repeated measures analysis of variance was used to compare students’ scores on appraisal assignments, scores on multiple-choice question (MCQ) exams in which EBP concepts were applied to clinical scenarios, and scores for self-reported efficacy in appraisal. Regression analysis was used to assess the relationship among the three assessment methods, gender, program level, and overall grade. The instructors had acceptable reliability in scoring the assignments (overall intraclass correlation coefficient=0.60). The MCQ exams had acceptable discrimination indices although their reliability was less satisfactory (Cronbach’s alpha=0.46). Statistically significant differences were observed among the three methods with MCQ exams having the lowest overall scores. Variation in the overall course grades was explained by scores on the appraisal assignment and MCQ exams (partial eta-squared=0.52 and 0.24, respectively), whereas score on the self-efficacy questionnaire was not significantly associated with overall grade. The results suggest that self-reported efficacy is not a valid method to assess dental students’ RCT appraisal skills, whereas instructor-graded appraisal assignments explained a greater portion of variation in grade and had inherent validity and acceptable consistency and MCQ exams had good construct validity but low internal consistency.

Dr. Al-Ansari is Assistant Professor, Department of Preventive Dental Sciences and Vice Dean for Academic Affairs, College of Dentistry, University of Dammam, Saudi Arabia; Dr. El Tantawi is Associate Professor, Department of Preventive Dental Sciences, College of Dentistry, University of Dammam, Saudi Arabia. Direct correspondence to Dr. Maha M.A. El Tantawi, University of Dammam, College of Dentistry, P.O. Box 1982, Dammam 31411, Saudi Arabia; 00966138574928; maha_tantawy@hotmail.com.

Keywords: dental education, evidence-based practice, evidence-based dentistry, assessment, educational measurement, Saudi Arabia

Submitted for publication 5/2/14; accepted 7/16/14
Attitudes, Access, and Confidence of Evaluation (KACE). Developers of that tool suggested it can be used in classroom settings but acknowledged that the tool was not intended for high-stakes (mostly summative) assessment and that their main purpose was to develop a tool to assess several dimensions of EBP training in a short time (20 minutes). As its name implies, KACE does not include a component assessing acquisition of EBP skills.

In 2011, the Fifth International Conference of Evidence-Based Health Care Teachers and Developers proposed a classification for EBP learning assessment tools within what it called the CREATE framework (Classification Rubric for EBP Assessment Tools in Education). This conference proposed classifying assessments according to targeted change regarding reaction to educational experience, attitude, self-efficacy, knowledge, skills, behaviors, or benefits to patients. The other dimension of the framework was the steps of EBP (Ask, Acquire, Appraise, Apply, and Assess). This conference also pointed to the need to indicate the characteristics of the audience (student, clinician, etc.) as well as assessment aim (formative or summative) and applied the framework to classify already validated tools in existence at the time. Only the Fresno test and the Berlin questionnaire were available to assess performance of skills, but neither is applicable to dental settings. A survey of existing assessment instruments clearly showed a limited number that can be applied in dental settings and particularly among predoctoral students. In addition, among the validated instruments, none assesses EBP skills.

In the College of Dentistry, University of Dammam, Saudi Arabia, EBP training has been incorporated into the curriculum, which was restructured to correspond to international standards and prepare competent dentists with defined critical thinking skills. A hybrid approach was used with stand-alone basic courses and several integrated components in other courses. In view of the absence of a standardized validated tool to assess the acquisition of EBP skills in a dental education setting, multiple assessment methods were used during and at the end of the courses to evaluate learning outcomes. The aims of this study were to evaluate and compare the various methods of summative assessment in EBP courses and to assess the relation between students’ performance in each assessment method and their overall performance in the course.

Materials and Methods

We obtained the approval of the Research Ethics Committee at the College of Dentistry, University of Dammam to conduct the study. Students gave their implied consent to participate by responding to the study questionnaire.

The College of Dentistry, University of Dammam is a new college that began accepting students in 2001. Students join the program after successfully passing one year of preparatory studies directly after high school. The program is ten semesters in length with a class size of 30-50 students. Ten semesters are offered in fall and spring extending over five years with no summer semesters. Students start clinical courses in the fifth semester (third year). Limited experience with research begins usually around the fifth semester (third year), in which students may be required to read scientific articles to prepare presentations in some clinical courses. This involvement with the dental literature is entirely up to the course directors and varies by course and academic year. On average, a student reads one or two articles per semester from the fifth semester till the last (tenth) semester.

As part of curricular reform, two stand-alone basic EBP courses (in the third and seventh semesters) were planned to cover the first three steps of EBP (Ask, Acquire, and Appraise). These courses addressed a prior deficiency in which no courses/components were delivered to train students in EBP so they were graduating with no exposure to this form of practice. The third-semester course was planned to focus on developing skills for appraising observational studies and randomized controlled trials (RCTs), whereas the seventh-semester course was intended to focus on developing these skills for diagnostic studies and systematic reviews. The initial design of the two courses was that they complement each other. The last two steps (Apply and Assess) were to be handled in components integrated in later/simultaneous clinical courses. At the time of implementation, developers faced a choice either to focus exclusively on younger cohorts and provide EBP courses gradually in the program as students progressed or to also include senior students to give them exposure to EBP before graduation. The latter option was selected, and a modified version of the two EBP courses was delivered to senior students in the seventh semester. The courses thus delivered in academic year 2013-14 to third- and seventh-semester students were similar,
with the exception that the seventh-semester course included appraisal of systematic reviews that was not part of the third-semester course. Table 1 shows the features of the courses.

Two instructors (the authors) developed the courses and delivered the parts addressed in this study regarding the RCT appraisal component. One author (AA) had formal training in EBP and is responsible for similar training in several Saudi organizations. The other (MT) had previous experience with EBP courses at another institution.

Two methods of instruction were used: 1) lectures explaining principles of the three EBP steps, and 2) class discussion of nongraded assignments in which students were divided into groups of five or six and appraised articles away from class using an appraisal worksheet. The students were given the article two to three days before the discussion session and were allowed to consult together to appraise it. Each of the instructors independently appraised the article using the worksheet, and consensus was reached between them through discussion. Discussions in class included instructors’ feedback on performance and explanation of correct/alternative answers. During the discussions, the instructor who delivered the lecture conducted the session with the other instructor present to expose the students to various points of view. Including both instructors also allowed them to gain perspective into the students’ grasp of the subject matter to direct any needed modification and revision of the new courses. Course time was divided between lectures and discussions approximately in the ratio 1:1.

Learning outcomes were evaluated using assignments about the groups’ appraisal of the articles and written multiple-choice question (MCQ) exams. The total course grade consisted of scores on the appraisal assignments and MCQ exams in addition to

<table>
<thead>
<tr>
<th>Learning Outcome</th>
<th>Teaching Methods</th>
<th>Contact Hours Allocated</th>
<th>Assessment Methods</th>
<th>Included in Study</th>
<th>Included in Course 3rd Semester</th>
<th>Included in Course 7th Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify major types of epidemiologic studies, and compare and contrast their strengths and limitations.</td>
<td>• Lectures</td>
<td>2</td>
<td>• Exams</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Differentiate primary and secondary sources of evidence, and list evidence according to strength.</td>
<td>• Lectures</td>
<td>1</td>
<td>• Exams</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Formulate a PICO question (ASK).</td>
<td>• Lectures, • Class discussions</td>
<td>3</td>
<td>• Exams</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Develop a strategy to search for evidence, and use information technology to retrieve evidence (ACQUIRE).</td>
<td>• Lectures, • Class discussions, • Computer lab session</td>
<td>4</td>
<td>• Group assignments</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Assess the presence of threats to validity in epidemiological studies, and evaluate study validity.</td>
<td>• Class discussions</td>
<td>1</td>
<td>• Exams, • Group assignments</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Appraise evidence retrieved from articles reporting on observational studies (APPRAISE).</td>
<td>• Class discussions</td>
<td>6</td>
<td>• Exams, • Group assignments</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Appraise evidence retrieved from scientific publications based on clinical trials (APPRAISE).</td>
<td>• Class discussions</td>
<td>6</td>
<td>• Exams, • Group assignments</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Appraise evidence retrieved from scientific publications based on systematic reviews (APPRAISE).</td>
<td>• Class discussions</td>
<td>3</td>
<td>• Exams, • Group assignments</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: Lectures were delivered by instructors. There were two exams during the semester and one at the end.
assignments and exams for other parts of the courses not included in the study (Table 1). The worksheet used for the nongraded and graded assignments was that adopted by the Jeddah evidence-based medicine working group (http://capps.ngha.med.sa/ebm/SGW/Saudi-Grou/Jeddah/EBM-Jeddah/jed-2_gp1.doc_cvt.htm). Table 2 shows the items included in the worksheet. The worksheet guided the appraisal by requiring students to respond to or comment on the items listed. The responses were in the form of yes/no in addition to free text that provided further explanation.

Two articles were appraised during discussions of the ungraded assignments in each of the third- and seventh-semester classes. Students then appraised two other articles and were graded on them using the same worksheet and grading scheme. The same article was appraised by all groups in each class per assignment. Among the four articles in the ungraded and graded assignments, two were the same for the two classes. The two classes were held on different days of the week and they had different sequences of sessions, so students were not appraising the same article at the same time. In addition, since students had two to three days to read the article and appraise it, chances to exchange information between the two classes were reduced. Students were warned that any form of academic dishonesty would not be tolerated and would result in cancellation of the assignment score for the guilty groups. They were also told that class discussion would probe their understanding of the appraisal. No evidence of similarity of responses between groups in the same class or between the two classes was observed.

One MCQ exam assessed application of the skills of RCT appraisal in two scenarios in which students were required to analyze information, assess validity, and calculate or interpret measures of treatment effectiveness. The two scenarios had a total of 14 MCQs. Other MCQs on the exam assessed factual recall of knowledge related to EBP but were not included in the analysis for this study. The same questions were used in the exams for the two classes.

| Table 2. Items in randomized controlled trial appraisal worksheet used for assignment |
|-----------------------------------|---------------------------------|
| Item                              | Details                           |
| Section 1: Are the study results valid? |
| 1.1 Randomization                  | Were groups randomly allocated?  |
| 1.2 All patients accounted for at end of study | How was randomization performed? |
| 1.3 Blinding                       | Was allocation concealed?        |
| 1.4 Were groups similar at baseline? | Did most subjects complete the study? |
| 1.5 Were groups treated equally aside from the intervention? | Were patients analyzed using intention to treat analysis? |
| 1.6 Overall, are the results valid? | Were baseline factors balanced?  |
| Section 2: What are the results?   | Were there single, double, or triple blinding? |
| 2.1 What is the size of treatment effect? | Was there unblinding during the study? |
| 2.2 How precise was the estimate of the size of treatment effect? | Were there co-intervention? |
| Section 3: Will the results help in caring for my patients? |
| 3.1 Do the results apply to my patient care? | Were there contamination? |
| 3.2 Were all clinically important outcomes considered? | Were there compliance? |
| 3.3 Are the likely treatment benefits worth the potential harm and cost? | Were groups treated equally aside from the intervention? |

February 2015 ■ Journal of Dental Education
These exams were conducted at the same time to preclude the classes’ sharing of information.

At the end of the courses, the students were given a questionnaire designed for the study to assess their perceived self-efficacy in performing appraisal skills on a scale from 1 to 10. Table 3 shows the statements indicating self-efficacy in RCT appraisal. The overall self-efficacy score was not included in the course grades.

The reliability of the instructors in scoring the assignments was assessed using intraclass correlation coefficient (ICC) for consistency and absolute agreement in a two-way random model.10 Difficulty and discrimination indices of the MCQ exams for the two classes were calculated.11,12 Some MCQs were combined to produce scores for appraisal aspects similar to those in the assignment worksheet, and their internal consistency was assessed using Cronbach’s alpha. Similarly, some self-efficacy statements were combined for analysis to represent the same aspects of RCT appraisal, and their internal consistency was assessed using ICC. Each instructor independently aligned the items in the assignment worksheet with respective MCQs (when available) and self-efficacy items, and differences between the instructors were resolved by discussion until consensus was reached.

Repeated measures analysis of variance was used to assess differences in scores among the three methods using pairwise comparison with Bonferroni adjustment for multiple comparisons. Regression analysis was used to assess the relation between the appraisal assignments, MCQ exams, and self-efficacy scores and the students’ overall performance in the courses measured by their course grade. Year and gender were included in the regression analysis. Regression coefficients and confidence intervals were calculated as well as partial eta-squared to assess the contribution of each variable to students’ overall course grades. SPSS version 17.0 was used for statistical analysis. Table 4 shows the statistical methods used to evaluate each of the three assessment methods and to compare them.

Table 3. Statements in self-efficacy questionnaire to assess randomized controlled trial appraisal skills

Please rate your ability to determine the following on a scale from 1 to 10:

1. Subjects were randomly allocated.
2. Study duration was adequate.
3. Percentage of subjects followed up to study end was sufficient.
4. Intention to treat analysis was performed.
5. Blinding of subjects/any others was performed.
6. Comparability of groups at baseline.
7. Need for adjustment to ensure groups’ comparability at baseline.
8. Equality of groups as regards any treatment other than the intervention.
11. Interpretation of measures of treatment effect.
13. Assessing applicability of study findings to patients.

Table 4. Characteristics of three assessment methods and statistical methods used to evaluate each and compare them

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Appraisal Assignment</th>
<th>MCQ Exam†</th>
<th>Self-Efficacy Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max score</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>N assessment events</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>N points or questions per event</td>
<td>23</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>N events included in course grade</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>% of course grade</td>
<td>20%</td>
<td>20%†</td>
<td>0</td>
</tr>
<tr>
<td>Assessment performed in class</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Allowed to use external resources</td>
<td>Yes</td>
<td>No</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Group or individual</td>
<td>Group</td>
<td>Individual</td>
<td>Individual</td>
</tr>
<tr>
<td>Closed-ended responses</td>
<td>Partly</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Assessment of validity</td>
<td>Face and content</td>
<td>Construct validity: discrimination index</td>
<td>Face and content</td>
</tr>
<tr>
<td>Differences among three methods</td>
<td>ICC</td>
<td>Cronbach’s alpha</td>
<td>ICC</td>
</tr>
<tr>
<td>Effect on overall course grade</td>
<td>Regression analysis of variance</td>
<td>Regression analysis (with gender and year included)</td>
<td></td>
</tr>
</tbody>
</table>

†Only questions related to randomized controlled trial (RCT) appraisal are included. Other RCT questions and other non-RCT questions on the exam were not included in the analysis.
MCQ=multiple-choice question; ICC=intraclass correlation coefficient
Results

This study included a total of 96 students in the third- and seventh-semester classes (n=68 and 28, respectively). There were 33 female students in the third-semester class (34.4% of all students in the study); all students in the seventh-semester class were male.

Table 5 shows the reliability of the two instructors’ scoring of the assignment. Three areas (judging the comparability of groups aside from the intervention, overall validity of the study, and applicability of findings to patients) had almost no reliability of scoring (ICC≤0.15). Two other areas (consideration of all clinically important outcomes and judging if benefits of intervention outweigh harms) had somewhat higher reliability values (ICC≥0.44). The differences between the two instructors’ scores ranged from 0.06 (randomization) to as large as 2.96 (comparability of groups at baseline), and they had no direction so that one instructor sometimes scored higher or lower than the other.

The mean difficulty indexes of the MCQ exams for the third- and seventh-semester classes were 68.7% and 59.9%, respectively. No questions had negative discrimination, and the mean discrimination indices for the two exams were 0.36 and 0.50. The exams consisted of one MCQ assessing students’ ability to judge the presence of randomization, four MCQs to evaluate their ability to assess whether patients enrolled in the study were accounted for at its conclusion (Cronbach’s alpha=0.10), two MCQs to evaluate their ability to identify the presence of blinding (Cronbach’s alpha=0.38), two MCQs to assess overall validity (Cronbach’s alpha=0.10), and five MCQs to evaluate their ability to calculate and interpret effect size values (Cronbach’s alpha=0.46). The overall internal consistency for all MCQs was 0.46.

For the statements assessing self-efficacy, the ICC for three statements expressing ability to judge whether all patients were accounted for at the end of the study was 0.75; the ICC for two statements indicating ability to assess comparability of groups at baseline was 0.62; and the ICC for two statements assessing self-efficacy to estimate and interpret the size of treatment effect was 0.81. The overall internal consistency of all statements assessing self-efficacy was 0.92.

Table 6 shows the comparison of scores obtained with the three assessment methods in each appraisal aspect. MCQs were used to evaluate students’ skills in five aspects. In three of these, MCQ scores were significantly lower than scores on the appraisal assignments and reported self-efficacy questionnaire. The same was observed regarding the overall score. The appraisal assignments and reported self-efficacy scores were not significantly different in two out of nine aspects of appraisal as well as in the overall score. In the remaining seven aspects, the greatest difference between scores was observed in the aspect related to judging the comparability of groups at baseline of the study.

Table 5. Reliability of instructors’ scores (out of 10) in various areas of randomized controlled trial (RCT) appraisal in assignment

<table>
<thead>
<tr>
<th>Item</th>
<th>Consistency</th>
<th>Absolute Agreement</th>
<th>Difference in Instructors’ Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICC (CI)</td>
<td>ICC (CI)</td>
<td>Mean (CI)</td>
</tr>
<tr>
<td>Randomization</td>
<td>0.83 (0.69, 0.91)</td>
<td>0.84 (0.70, 0.91)</td>
<td>0.06 (-0.33, 0.46)</td>
</tr>
<tr>
<td>Enrolled patients accounted for at trial end</td>
<td>0.67 (0.45, 0.82)</td>
<td>0.68 (0.45, 0.82)</td>
<td>0.10 (-0.65, 0.84)</td>
</tr>
<tr>
<td>Blinding</td>
<td>0.73 (0.54, 0.86)</td>
<td>0.70 (0.44, 0.84)</td>
<td>-0.35 (-2.30, -0.06)</td>
</tr>
<tr>
<td>Comparability of groups at baseline</td>
<td>0.75 (0.56, 0.86)</td>
<td>0.71 (0.45, 0.85)</td>
<td>-2.96 (-4.15, -1.78)</td>
</tr>
<tr>
<td>Comparability of groups aside from intervention</td>
<td>-0.15 (-0.45, 0.19)</td>
<td>-0.13 (-0.41, 0.18)</td>
<td>1.72 (0.06, 3.39)</td>
</tr>
<tr>
<td>Overall validity of study</td>
<td>-0.03 (-0.35, 0.30)</td>
<td>-0.02 (-0.30, 0.28)</td>
<td>-1.94 (-3.60, -0.30)</td>
</tr>
<tr>
<td>Size of treatment effect</td>
<td>0.69 (0.47, 0.83)</td>
<td>0.64 (0.36, 0.81)</td>
<td>-1.76 (-2.95, -0.57)</td>
</tr>
<tr>
<td>Precision of estimate</td>
<td>0.69 (0.47, 0.83)</td>
<td>0.63 (0.31, 0.81)</td>
<td>-1.99 (-3.16, -0.81)</td>
</tr>
<tr>
<td>Applicability to patient population</td>
<td>-0.06 (-0.38, 0.27)</td>
<td>-0.06 (-0.36, 0.27)</td>
<td>0.58 (-0.24, 1.40)</td>
</tr>
<tr>
<td>Consideration of all clinically important outcomes</td>
<td>0.44 (0.14, 0.67)</td>
<td>0.39 (0.07, 0.63)</td>
<td>1.00 (0.37, 1.64)</td>
</tr>
<tr>
<td>Benefits outweigh harms</td>
<td>0.28 (-0.05, 0.57)</td>
<td>0.24 (-0.05, 0.51)</td>
<td>1.28 (0.40, 2.15)</td>
</tr>
<tr>
<td>RCT appraisal assignment (including items with acceptable internal consistency only)</td>
<td>0.76 (0.58, 0.87)</td>
<td>0.70 (0.36, 0.86)</td>
<td>0.76 (0.36, 1.16)</td>
</tr>
<tr>
<td>RCT appraisal assignment (including all items)</td>
<td>0.60 (0.35, 0.78)</td>
<td>0.58 (0.31, 0.76)</td>
<td>-0.48 (-0.90, -0.06)</td>
</tr>
</tbody>
</table>

ICC=intraclass correlation coefficient; CI=confidence interval
baseline, with students’ overestimating their ability to perform this part of the appraisal.

Table 7 shows the effect of various factors on overall course performance. The greatest effect (partial eta-squared) was for the appraisal assignment scores followed by MCQ exam scores. Students who scored higher by one mark in the appraisal assignments eventually were about five marks higher than others in the overall score, whereas an increase of one mark in the MCQ exam corresponded to 1.5 higher overall score. Students’ self-efficacy scores were not significantly related to their overall course performance. Females and senior students did significantly better than the other students, scoring on average about five and four marks, respectively, higher than males and third-semester students.

Discussion

To our knowledge, this study is the first to report on an objective assessment of EBP skills in a dental education setting. It included instructors with diverse yet relevant EBP experience that is likely to be found in most dental schools. The study included a number of students at different levels of the program. None of these students had previous experience with EBP, so their response to the training provided in the study reflected their ability to learn and apply the content. A concern, however, is that the clinical experience of the senior students could have helped them to some extent with the course and may be considered a limitation of this study; therefore, the difference between juniors and seniors should be interpreted with caution. Whereas the size of the study sample is not particularly large, it reflects the actual class size in the school. Larger sample sizes could be obtained in the same setting by including findings from several classes in more than one semester, which would potentially introduce the problem of variation across time. Alternatively, the study could be replicated in other schools to confirm our findings through larger sample sizes and different learning conditions. Including other EBP skills

Table 6. Comparison of scores obtained in three assessment methods for various aspects of randomized controlled trial (RCT) appraisal

<table>
<thead>
<tr>
<th>RCT Appraisal Aspect</th>
<th>Appraisal Assignment</th>
<th>MCQ Exam</th>
<th>Self-Efficacy Questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomization</td>
<td>8.40 (8.06, 8.73)*</td>
<td>6.85 (5.88, 7.82)*</td>
<td>7.59 (7.16, 8.01)*</td>
</tr>
<tr>
<td>Enrolled subjects accounted for</td>
<td>6.70 (6.31, 7.09)*</td>
<td>6.25 (5.83, 6.67)*</td>
<td>7.02 (6.63, 7.41)*</td>
</tr>
<tr>
<td>Blinding</td>
<td>6.71 (6.18, 7.23)*</td>
<td>7.94 (7.30, 8.57)*</td>
<td>7.42 (6.99, 7.85)*</td>
</tr>
<tr>
<td>Comparability at baseline</td>
<td>2.99 (2.42, 3.57)*</td>
<td>–</td>
<td>6.46 (6.05, 6.86)*</td>
</tr>
<tr>
<td>Equality of treatment other than intervention</td>
<td>7.89 (7.58, 8.19)*</td>
<td>–</td>
<td>7.08 (6.59, 7.57)*</td>
</tr>
<tr>
<td>Overall validity</td>
<td>8.50 (8.13, 8.86)*</td>
<td>3.32 (2.64, 3.99)*</td>
<td>7.00 (6.57, 7.43)*</td>
</tr>
<tr>
<td>Treatment effect size</td>
<td>5.58 (4.87, 6.28)*</td>
<td>2.92 (2.67, 3.18)*</td>
<td>7.56 (7.11, 8.01)*</td>
</tr>
<tr>
<td>Precision of effect size estimate</td>
<td>5.03 (4.36, 5.71)*</td>
<td>–</td>
<td>6.62 (6.06, 7.18)*</td>
</tr>
<tr>
<td>Usefulness of results to patients</td>
<td>9.36 (9.19, 9.53)*</td>
<td>–</td>
<td>7.26 (6.80, 7.72)*</td>
</tr>
<tr>
<td>Average overall score</td>
<td>7.17 (6.90, 7.46)*</td>
<td>6.04 (5.67, 6.41)*</td>
<td>7.27 (6.94, 7.61)*</td>
</tr>
</tbody>
</table>

CI=confidence interval; MCQ=multiple-choice question
Note: Different letters (a, b, c) denote statistical difference.

Table 7. Factors affecting overall performance in courses (adjusted R²=0.70)

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>95% CI</th>
<th>Partial Eta-Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appraisal assignment score</td>
<td>4.95*</td>
<td>3.93, 5.97</td>
<td>0.52</td>
</tr>
<tr>
<td>MCQ exam score</td>
<td>1.51*</td>
<td>0.94, 2.09</td>
<td>0.24</td>
</tr>
<tr>
<td>Self-efficacy questionnaire score</td>
<td>0.18</td>
<td>-0.48, 0.84</td>
<td>0.003</td>
</tr>
<tr>
<td>Males vs. females</td>
<td>-4.71*</td>
<td>-7.38, -2.03</td>
<td>0.13</td>
</tr>
<tr>
<td>3rd semester vs. 7th semester</td>
<td>-4.12*</td>
<td>-6.74, -1.50</td>
<td>0.10</td>
</tr>
</tbody>
</table>

B=regression coefficient; CI=confidence interval; Partial Eta-Squared: measure of effect size; MCQ=multiple-choice question
*Statistically significant at 5% level
such as appraisal of various study types in addition to question formulation and searching for evidence will also allow generalizability of our findings to be based on firmer grounds.

The main focus of the study was the evaluation of appraisal skills assessment in an educational setting using methods readily utilized by instructors in everyday situations. These methods can be repeatedly used across successive classes by modifying individual components (articles used in the appraisal assignments and questions on the exams). The courses’ intended outcomes were the development of skills in line with the restructuring of the program to be competency-based. In addition, for dental students, appraisal of RCTs is a valuable skill in view of the limited number of secondary sources of evidence available in dentistry compared to medicine. Developing a validated assessment tool was not the focus of the study because the use of standardized tools for students’ assessment would allow them to predict the questions with their correct answers. Hence, they cannot be used in high-stakes assessment. This may explain why summative assessment of students in educational settings does not usually rely on standardized questionnaires although their psychometric properties are superior. A 2008 survey of assessment practices in 53 out of 56 U.S. dental schools reported a number of methods used to assess performance in different domains, none of which was related to or used standardized questionnaires. For the domain in which students are supposed to apply EBP skills, the most common methods of assessment reported in that study were MCQ exams (45%), and the least common were triple jump (3%) and critically appraised topics (2%).

A limitation of our study is that traditional approaches to evaluate validity and reliability were not followed. This is partly because there is no standardized tool assessing appraisal skills in dental settings that can be used for validation. In addition, the study describes an actual learning experience including courses taught to students without previous experience of EBP skills, so that assessing test-retest reliability, discriminative validity, and responsiveness cannot be performed.

Considerable variation existed in the reliability of grading the assignment between the two instructors. The least reliability was observed in areas in which the instructors made subjective judgments, such as whether there was equal treatment of groups aside from the intervention, overall study validity, applicability of findings to patients, consideration of clinically important outcomes, and whether benefits outweighed harms. In other areas, acceptable reliability was observed, indicating that the instructors agreed about the scores given to students. High values of absolute agreement indicated that scores could be given by either one of the two instructors interchangeably. Whereas it is preferred that no single instructor scores assignments individually, situations such as the temporary absence of an instructor may call for this to happen. In these cases, ensuring absolute agreement in scoring between instructors is an advantage.

The overall scores on the appraisal assignments had acceptable consistency and absolute agreement, which reduces the concern for areas of the assignments in which scoring reliability was low. Better phrasing and structuring of these areas of the worksheet may close the gap between the scores to achieve better reliability. It is interesting to notice that the level of reliability between the instructors did not always translate into similarity or difference of scores. For example, an area with low reliability such as applicability to patient population had almost 24% of the score difference between the two instructors as an area with high reliability such as comparability of groups at baseline. Furthermore, if all items are included regardless of the reliability of their scoring, the overall reliability (consistency and absolute agreement) would still be acceptable. The absolute difference in the overall assignment scores between the two instructors was about 0.5 out of 10. Since scores on these assignments constituted 20% of the overall course grade, one instructor could potentially have over- or underestimated scores for a student by one mark out of 100, which is very minimal. Further research is needed to assess the relation between reliability of scoring and type of article assessed. For example, it would be interesting to know if the instructors would have better reliability in scoring when the entire class was assigned the same article or if each group has a different article. The potential disadvantage of the approach used in this study, however, is the higher likelihood of sharing appraisal results among groups with potential impact on scores and needs for measures against academic dishonesty.

The difficulty indices of the MCQ exams for the third- and seventh-semester classes indicate that the exams had reasonable difficulty and their discrimination indices were also acceptable in ensuring construct validity, although they had low internal reliability. This may partly be attributed to the comparatively small number of questions in each
indicating that students overestimated their abilities on the two other assessment methods in some aspects, call for action to improve the internal reliability of weaknesses in student performance. tailor instruction to emerging needs, and address instructors to progressively support student learning, performance is important to student and instructor; it helps was done for the MCQ exams. Feedback on perfor assignments helped explain vague points and stress MCQs. Discussion of the training and non-graded training to manage these higher cognitive level may explain the lower scores is the need for student appraisal assignments. Another factor that considered the questions them individually as compared to the group work area although the items in the self-efficacy question-naire, which had similar or fewer items, did not suffer from this problem. Another explanation for the low reliability is that the MCQs did not cover all areas of the content. For example, the MCQs did not assess students’ skills in appraising the articles as regards groups’ comparability at baseline, equality of groups’ treatment aside from intervention, or judging the applicability of study findings to patients. The overall internal consistency of the 14 MCQs included in the study was 0.46, which is lower than the reliability of 0.59 that Lake and Chambers reported for a test of 50 MCQs. McComas et al. reported the reliability of a number of objective structured clinical examinations (OSCEs) that were modified to include MCQs. Their values of alpha for reliability ranged from -0.20 for a domain with two questions to 0.77 for a domain with ten questions, with an overall alpha=0.66 for all 25 questions. Those authors considered the questions to lack acceptable reliability to be used to assess critical thinking skills. Their study and ours cast doubt on the idea that increasing the number of questions may increase reliability to an acceptable level. Similarly, William proposed a formula for calculating how long a test has to be to reach a specific level of reliability. Based on his formula, there would have to be around 66 questions for the exam to reach a reliability of 0.80. This is too large a number of questions to assess the skills of appraising RCT, especially when considering there are other course outcomes to be assessed in the same exam.

Overall, the MCQ exams had the lowest scores of the three assessment methods. A possible explanation for this may be that the students answered them individually as compared to the group work in the appraisal assignments. Another factor that may explain the lower scores is the need for student training to manage these higher cognitive level MCQs. Discussion of the training and non-graded assignments helped explain vague points and stress concepts to students. No such training or discussion was done for the MCQ exams. Feedback on performance is important to student and instructor; it helps instructors to progressively support student learning, tailor instruction to emerging needs, and address weaknesses in student performance. These factors call for action to improve the internal reliability of MCQs assessing various aspects of RCT appraisal.

Self-efficacy scores were higher than the scores on the two other assessment methods in some aspects, indicating that students overestimated their abilities to appraise RCTs. This was particularly noticed in the case of judging whether patients enrolled in the study were accounted for, whether groups were comparable at baseline, estimating the effect size, and the precision of that estimate in addition to the overall score. These appraisal areas depended to some extent on statistical concepts that the students may erroneously conclude they grasp but later fail to apply. These same skills had generally the lowest scores in the three assessment methods compared to other appraisal areas. Other studies have reported a low correlation between medical students’ and physicians’ self-reported skills and objective assessment of these skills.

The greatest effect on overall course grade in our study was the appraisal assignment scores. These scores are based on an objective method of assessment, numerous assessment events, and proven consistency of scoring. The MCQ exam scores had almost half the effect on the overall course score. Together, these two objective methods of assessment had an effect that far outweighed all other variables in the regression model. Self-efficacy scores had a very minimal and non-statistically significant effect on overall grade in the course. This indicates that students’ expression of ability to perform some appraisal skills should be dealt with cautiously. Their impression of how good or bad they are is not actually reflected in their grades in the EBP courses. This may be attributed to students’ unconsciously ascribing the change in their performance to factors other than the learning experience.

Female students in our study outperformed their male colleagues. This is especially notable since all female students were in the third-semester class, which was shown by the model to have modest performance compared to senior students in the seventh-semester class. These results are based on a single class with female students and should therefore be interpreted cautiously and need to be supplemented by further observation over the years. Other studies found no gender difference in EBP knowledge score or reported implementation of EBP.

Third-semester students in our study had significantly lower overall course grades than the seventh-semester students. This should be taken into consideration when planning the placement of the courses in the program in future. Early introduction of EBP training is an advantage as long as it has a satisfactory effect on students’ learning outcomes as reflected in the course score. If the students perform less than expected and/or less than their senior peers, it may be desirable to move the course to a later part.
of the program. In this way, students can be more comfortable with the course, which is understandably demanding since it addresses higher cognitive skills. The experience of some schools in introducing EBP courses early in the program may not be directly comparable to students in our study. Students in U.S. dental schools spend an average of four years in college after high school before they enroll in postdoctoral dental programs. This education is expected to develop their cognitive and critical thinking skills to some extent. In contrast, in Arab countries in general and Saudi Arabia in particular, students spend one preparatory year in education after they leave high school and before they can be accepted into dental programs. It is only during this preparatory year that they have experience with scientific English through studying basic science courses (biology, physics, chemistry, etc.). These differences between the two systems, therefore, make direct transfer of courses’ content, teaching, and assessment methods or placement rather impractical.

Using several methods of assessment supports differences in students’ learning styles and increases the validity of the assessment by offering students various ways to demonstrate their competence. The validity of high-stakes assessment is increased as multiple types are used on a regular basis. Hence, the use of multiple assessment methods as done in the current study is conceptually sound. Appraisal assignments have inherent validity as a method of assessing appraisal skills since they involve doing the same task that is supposed to be done later in professional life. When consistency of scoring by different examiners is ensured, they offer a straightforward, reliable and valid assessment of appraisal skills. On the other hand, MCQ exams allow assessment of individual students’ achievement of learning outcomes in an efficient way in terms of conserving instructors’ time and effort compared to assignments. Those exams are also less subject to potential problems of reliability of scoring compared to assignments and can be used to assess higher cognitive skills such as interpretation, application, and synthesis. In the area of EBP learning assessment, MCQs have been used as part of the Berlin test and KACE, as well as in assessment of EBP knowledge using non-standardized tools.

A combination of two methods in which appraisal of assigned articles is done through MCQs has the potential to make use of the strengths of both methods. In future EBP courses in our college, we intend to implement question formats in which students appraise RCTs by selecting from a set of multiple choices citing areas of the article and drawing conclusions on validity, importance of results, and applicability to patients. The validity and reliability of this assessment method will have to be evaluated through research. Another useful direction of future research is the long-term follow-up of students’ achievement of learning outcomes in subsequent courses targeting the last two steps of EBP as a function of their performance in various assessment methods.

**Conclusion**

This study focused on methods to assess EBP training, an important issue for dental educators. Existing validated assessment tools for EBP are suitable for medical settings, and the only validated tool designed for dental settings does not assess skills. As more educators and students become involved in EBP, the need for assessment methods becomes more acute. The value of this study is to help dental educators decide on the usefulness of assessment methods with which they are familiar. It does not undermine the importance of standardized questionnaires that have superior psychometric properties but are less likely to be used in semester to semester course assessment. In this study, EBP training was provided in a dental didactic course to develop RCT appraisal skills. The appraisal assignments had inherent construct validity, and their scoring by the instructors was consistent. MCQs, on the other hand, had acceptable construct validity demonstrated by discrimination indices, although internal consistency was low. Self-efficacy scores were neither related to overall course grades nor to the other methods of assessment and should therefore be interpreted with caution. Based on the study findings, appraisal assignments seem to be the most valid and reliable method to assess students’ skills in appraising RCTs.

**REFERENCES**