Assessment

Triple Jump Examinations for Dental Student Assessment


Abstract: The triple jump examination (TJE) attempts to assess a higher level of learning with demand for analysis, critical thinking, and resolution of problems presented by written scenarios based on patient care situations. The purpose of this study was to examine the internal consistency, scale reliability, and interrater reliability of the TJE used at the Ostrow School of Dentistry, University of Southern California. On the sample of 2,227 examinations administered by seventy-seven raters across a three-year time period, the Cronbach’s coefficient alpha for internal consistency of the overall TJE was found to be good ($\alpha=0.869$). The internal consistency of the three subscales was found to be acceptable ($\alpha=0.731$), good ($\alpha=0.820$), and good ($\alpha=0.820$). Average and single measures intraclass correlation coefficients (ICC) for scale reliability were significant at $p<0.001$, indicating strong interrater reliability. There were no statistically significant differences ($p \leq 0.05$) in the mean scores assigned on the TJE between rater groups defined by rater experience level with the TJE. A very high level of agreement among rater pairs was also observed. Across the entire three-year study period, with over 19,152 ratings, the seventy-seven raters were in general agreement 99.5 percent of the time and in exact agreement 77.2 percent of the time.

Dr. Navazesh is Associate Dean, Academic Affairs and Student Life and Professor of Diagnostics Sciences, Division of Periodontology, Dental Hygiene, and Diagnostic Sciences, Ostrow School of Dentistry, University of Southern California; Dr. Rich is Associate Professor, Division of Periodontology, Dental Hygiene, and Diagnostic Sciences, Ostrow School of Dentistry, University of Southern California; Dr. Bahari Chopiuk is Clinical Assistant Professor, Division of Biomedical Sciences, Ostrow School of Dentistry, University of Southern California; and Dr. Keim is Associate Professor of Statistics and Research Methods and Faculty Fellow of the Dissertation Support Center, Rossier School of Education and Associate Professor, Division of Endodontics, Orthodontics, and Oral and Maxillofacial Surgery, Ostrow School of Dentistry, University of Southern California. Direct correspondence and requests for reprints to Dr. Mahvash Navazesh, Ostrow School of Dentistry, DEN 218, University of Southern California, 925 W. 34th Street, Los Angeles, CA 90089-0641; 213-740-1001 phone; 213-740-2376 fax; navazesh@usc.edu.

Keywords: dental education, dental students, educational methodology, assessment, triple jump exercise

Submitted for publication 9/17/12; accepted 11/1/12

The triple jump examination (TJE) was instituted in 2001 at the Ostrow School of Dentistry of the University of Southern California (OSDUSC) and has become an integral part of student outcomes assessment at the school. In 2008, a Task Force on Student Outcomes Assessment of the American Dental Education Association Commission on Change and Innovation in Dental Education (ADEA CCI) described the current status of student outcomes assessment in dental education. Its survey of U.S. dental schools revealed that traditional outcome measures of multiple-choice testing, laboratory practicals, daily grades, clinical competency exams, and procedural requirements were holding strong as assessment tools in dental education. They stated that although several newer assessment techniques, such as the TJE, have been frequently noted as best practices in the health professions literature, those tools are rarely used in dental education. With 93 percent of U.S. dental schools responding to the ADEA CCI survey, results showed that only 1 percent were using the TJE as an assessment method for measuring student performance.

In Miller’s Pyramid of Professional Competence, the TJE is an assessment categorized as a level above multiple-choice testing and student reports, which test for recall of facts or recognition of information. The TJE facilitates movement toward a higher level of learning with its demand for analysis, critical thinking, and resolution of problems presented by written and simulated problems based on patient care situations. With some urgency, the ADEA CCI Task Force advocated greater use of evaluations across all domains of competence that bear more characteristics like those of triple jump examinations.
Our curriculum at OSDUSC is learner-centered with the biomedical, behavioral, and clinical sciences integrated. A combination of learning methodologies is used, including case-based learning, service-learning, expert guidance for hands-on skills development, problem-based learning, and traditional educational methodologies. The OSDUSC TJE has previously been described in the education literature by Fincham and Shuler\(^4\) and von Bergmann et al.,\(^5\) but no outcomes data for faculty interrater reliability or student TJE examination performance were presented. After presenting an overview of the TJE, the purposes of this article are to describe the TJE examination at OSDUSC; to provide an analysis of data collected over three years from students for internal consistency, scale reliability, and interrater reliability of the TJE; and to assess suggested differences between the mean scores given by faculty raters based on differences in raters' level of experience with the TJE.

### The Triple Jump Examination

To help meet their goal of improving assessment practices in dental education, the ADEA CCI Task Force, led by Dr. Gene Kramer, director of the Department of Testing Services, American Dental Association, published a “toolbox” of available methods for assessing student performance.\(^6\) The focus was on introducing alternative competency assessment tools for measuring the knowledge, skills, and abilities necessary for entry-level dental practice. Seventeen assessment methods were described as making up the toolbox, and the strengths and limitations of each were addressed. In that toolbox, the TJE was categorized as a “multi-competency, comprehensive,” newer assessment tool.

The format of the TJE varies somewhat as used at different schools,\(^7,12\) but, as its name implies, it is consistently used as a three-step examination. Kramer et al. pointed out that the TJE may be presented in a clinical or a preclinical version.\(^6\) A clinical TJE involves a student with a live patient and has three parts: 1) interviewing and examining a patient while being observed with video or by faculty; 2) writing an assessment of the findings from the patient interview with emphasis on providing evidence from the literature to support a diagnosis and a recommended therapy; and 3) participating in an oral examination with faculty in which the student discusses the case with examiners and answers questions with regard to the pathophysiology, diagnosis, treatment, and the research evidence to support his or her ideas about the patient’s condition. As further described by Kramer et al., a preclinical TJE includes the following: 1) introducing a hypothetical patient by written scenario for which the student identifies major issues and researchable questions; 2) student investigation of the literature to find evidence related to the selected research questions; and 3) an oral report on findings, discussion of the case with examiners, answering questions, and critically analyzing evidence revealed by students’ search, as well as their own TJE performance.\(^6\) Generally, the TJE is thought of as having both subjective and objective evaluation sections with step 1 representing the more objective measurement versus the more subjective evaluation needed for steps 2 and 3.\(^8\)

A significant portion of the TJE consists of an oral presentation. The ADEA CCI Task Force description of the toolbox of assessment methods noted that while an oral presentation is well suited for evaluating critical thinking, a weakness is its vulnerability to subjectivity on the part of the raters.\(^6\) Difficulty in calibration of oral examinations has been documented in the medical education literature. Elliot and Hickam enlisted eleven faculty members to independently evaluate seventeen videotaped internal medicine student case presentations. In spite of best efforts at calibrating the faculty on four descriptors of content and six attributes of communication style, there were no significant correlations for six of ten dimensions used for testing.\(^13\)

Green et al. hypothesized that medical educators have common rather than divergent expectations of the structure and content of oral case presentations with internal medicine students.\(^14\) These researchers tested their hypothesis with a forty-two-item survey of 136 faculty members at five U.S. medical schools and concluded that the results confirmed their hypothesis. Nonetheless, Green et al. conceded that students report that educators often failed to share these expectations with them during or after the oral case presentation. Furthermore, students’ feedback was that they were frustrated with evaluators who did not consistently enforce “rules” that student presenters had been asked to follow with their case presentations. Overall, it appears that, although the study found that faculty members had common expectations of student performance, they were not consistent and well calibrated in the feedback and evaluation they provided in sessions with the stu-
Faculty Calibration

At each TJE session, all participating evaluators receive a packet with written reminder instructions including the purpose, conduct, and evaluation criteria for the TJE. New faculty evaluators are initially calibrated with close mentoring from experienced TJE faculty members and are expected to read TJE process and evaluation information available on the school intranet prior to their first sessions. They are asked to observe at examination sessions for at least six students before they serve as evaluators themselves. At their first sessions, they are paired with and observed by a senior, experienced evaluator from whom they receive feedback.

Following all TJE sessions, anonymous feedback from student evaluations, submitted by each student tested, is available to all faculty members who participate. The evaluations are e-mailed directly to faculty participants and their division chairs within two weeks of the session. The scores are taken into account at individualized, annual faculty evaluation sessions with division chairs. Additionally, students’ evaluation scores of faculty members are monitored by the associate dean of academic affairs, who tracks the performance progress of all new and experienced faculty evaluators. On rare occasions, division chairs have been requested to assign faculty members with unacceptable performance to other activities. To further assist in the calibration process, the associate dean of academic affairs, working with an assisting faculty member, assigns faculty evaluators for each session so that at least one of the two examining team members is an “expert” evaluator with a minimum of five years of TJE experience.

Steps of the TJE

The initial step of the TJE involves a one-hour, written analysis of the patient case scenario provided. The student’s analysis includes determining the patient presenting’s chief complaint and its pertinent facts, formulating mechanistic-oriented hypotheses that may explain the patient’s problem(s), and listing learning topics to be investigated with regard to each hypothesis. As a second step, the examinee conducts an overnight gathering of information to contribute to his or her learning for an oral presentation the next day. At the third step, two faculty members—one with clinical and one with biomedical expertise—assess the student’s performance and provide feedback with interactive discussion and questioning in a thirty-minute session. The students are expected to

Methods

In a preclinical version of the TJE at OSDUSC, the case and the patient are hypothetical. The emphasis is on the student’s problem-solving and critical thinking skills with regard to patient information that is provided for them in writing. The student’s abilities for searching and critically evaluating the literature and for applying investigative research findings are assessed in an oral presentation. The examination is individually administered to each student six times, at midterm and final time, in the fall, spring, and summer trimesters of the first year of dental school. This study was approved by the University of Southern California Institutional Review Board (#UP-11-00509).

dents. However, Green et al. did not provide any data from actual student evaluations that might address whether the common faculty expectations identified were associated with reliability in grading. These authors also mentioned a concern that many educators responsible for evaluations at their school were younger and less experienced than clerkship director evaluators who were also evaluators, intimating that experience level of evaluators may be a variable influencing consistency of grading/evaluation.

Smith was likewise concerned with the experience level of evaluators and other variables related to interrater reliability and validity of the TJE with medical students. Considering the difficulty, cost, and time to train examiners to uniformity in grading, he devised a method of combining scores from other independent, objective examinations such as multiple-choice question (MCQ) examinations to make a composite grade to balance against the more subjective scoring sections of the TJE. The composite grade also included Medical College Admissions Test (MCAT) scores and student self-assessment scores. In an analysis of scores for fifty-eight students, he found that TJE scores did not correlate significantly with MCAT or MCQ scores. He stated that the TJE appeared to measure different aspects of student performance than did the other examinations. Interestingly, student self-assessment scores did not correlate well with faculty assessment scores in that study; students were less inclined to score themselves as positively as did the faculty. However, Smith concluded that, overall, the “mixed examination score” was an improvement in reliability and validity for determining a student’s fitness to pass.
use their reasoning ability to explain whether they accept or reject their hypotheses from the previous day, say how they will approach a continuance of learning for the case, formulate new/revised hypotheses, and address lingering questions or further gaps in their knowledge.

Faculty evaluators engage in discussion with the students and provide immediate verbal feedback on the students’ critical analysis and reasoning. On a written TJE assessment form (Table 1), students are graded independently by the two examiners on their performance level in each of three stages of the examination: 1) initial paperwork identification of patient problems, 2) research and study approach to the problem-solving, and 3) synthesis and understanding demonstrated in the oral presentation. Following the session, students submit their anonymous evaluations of the participating faculty raters. Faculty members receive electronic copies of the evaluations, and student feedback is used for further development of faculty by the associate dean of academic affairs.

**Statistical Analyses**

Data in the form of students’ scores on the TJEs, administered from summer trimester 2009 through spring trimester 2012, were collected. For each TJE administered, two separate faculty examiners rated each student being examined. These data were then analyzed across the three-year period for scale reliability, interrater reliability, and differences in scores awarded based on rater experience level. For the purpose of internal consistency analysis, the three stages of the examination described above (see Table 1) were taken as three separate subscales: 1) Issue Identification Scale (IIS), 2) Research and Study Issues Scale (RSIS), and 3) Synthesis, Knowledge, and Understanding Scale (SKUS). All statistical analyses were conducted using Statistical Package for the Social Sciences (SPSS), version 19.

**Results**

Over the period of the study, 2,227 TJEs were administered by seventy-seven raters to individual students. Scale reliability assessments using Cronbach’s alpha coefficient\(^{15,16}\) were conducted on the TJE overall as well as for each of the three subscales. Following the recommendation of DeVellis,\(^{17,18}\) alphas between 0.7 and 0.8 were considered “acceptable,” alphas between 0.8 and 0.9 were considered “good,” and alphas over 0.9 were considered “excellent.” In our study, the Cronbach’s alpha for the overall TJE, taken across all twelve items of the TJE, was found to be good (\(a=0.869\)). Alpha values for each of the subscales were as follows: for IIS, acceptable (\(a=0.731\)); for RSIS, good (\(a=0.820\)); and for SKUS, good (\(a=0.820\)).

Interrater reliability was assessed using the intraclass correlation coefficient (ICC).\(^{19,20}\) Both average measures ICCs and single measures ICCs were calculated. The single measures ICC is an index for the reliability of the ratings for one, typical, single rater, whereas the average measures ICC is an index for the reliability of different raters averaged together.

---

**Table 1. Triple jump examination grading stages**

<table>
<thead>
<tr>
<th>Stage 1. Issue Identification (IIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Student identifies patient’s chief complaint.</td>
</tr>
<tr>
<td>2. Student selects only those facts that are relevant to the patient’s presentation.</td>
</tr>
<tr>
<td>3. Student links relevant facts to generate mechanism-oriented hypotheses.</td>
</tr>
<tr>
<td>4. Student develops focused learning objectives to validate hypotheses.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 2. Research and Study Issues (RSIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Student prioritizes ideas and related learning objectives.</td>
</tr>
<tr>
<td>6. Student critiques original hypotheses with explanations based on research.</td>
</tr>
<tr>
<td>7. Student provides evaluators with a list of resources in proper citation format.</td>
</tr>
<tr>
<td>8. Resources used are current and of appropriate quality (e.g., textbooks and peer-reviewed journal articles relevant to patient’s problem).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage 3. Synthesis, Knowledge, and Understanding (SKUS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Student demonstrates that new facts learned lead to a better understanding of the nature of the signs and symptoms.</td>
</tr>
<tr>
<td>10. Student demonstrates critical insight into linkage between different aspects of the patient’s presentation.</td>
</tr>
<tr>
<td>11. Student generates new/revised hypotheses based upon research and formulated in order to pursue advanced understanding of the patient’s problem.</td>
</tr>
<tr>
<td>12. New/revised learning objectives address specific gaps in knowledge needed to interpret the patient’s problem.</td>
</tr>
</tbody>
</table>
This ICC is always higher than the single measures ICC.²¹,²² For the entire TJE, across seventy-seven raters, the average measures ICC was 0.869, with a single measures ICC of 0.355 \((F_{2226.2446.678} = 7.615, p<0.001)\), indicating a strong interrater reliability. For the ISS, the average measures ICC was 0.731, with a single measures ICC of 0.405 \((F_{2226.6678} = 3.724, p<0.001)\), indicating acceptable interrater reliability. For the RSIS, the average measures ICC was 0.820, with a single measures ICC of 0.532 \((F_{2226.6678} = 5.546, p<0.001)\), indicating strong interrater reliability. For the SKUS, the average measures ICC was 0.820, with a single measures ICC of 0.532 \((F_{2226.6678} = 5.546, p<0.001)\), indicating strong interrater reliability.

It has been suggested that scores assigned by individual raters may differ based on the experience levels of the rater as measured by the total number of TJE scores the rater conducted. In order to assess this suggestion, the experience levels of the seventy-seven raters were divided into quartiles. Quartile 1 \((n=21)\) consisted of raters who had conducted from one to eleven previous TJE scores. Quartile 2 \((n=18)\) had conducted from twelve to twenty-four, Quartile 3 \((n=19)\) from twenty-five to forty-two, and Quartile 4 \((n=19)\) forty-three or more. When assessed by the Cronbach’s alpha statistic, qualitative ratings of scale reliability and interrater reliability did not differ between the groups based on experience levels, with the overall scale reliability rated as good, ISS as adequate, RSIS as good, and SKUS as good.

As a confirmatory measure to address the issue of potential scoring differences between raters based on experience levels, all TJE scores across the entire period of the study assigned by each individual rater were summed to create a new dependent variable representing the Total TJE Score for each rater. An analysis of variance was then conducted using experience level (by quartile) as the independent grouping variable. No significant differences were noted between the groups \((F_{4.73} = 0.925, p=0.435)\).

In evaluating students following a TJE, each of the two raters indicate the degree to which the student has met evaluation criteria for each of the twelve items on the assessment. Students are assigned a Y for Yes, they have met the criterion. They are assigned an S if they Somewhat met the criteria and an N if No, they did not meet the criterion. Raters are considered to be in general agreement if one assigned one rank and the other assigned an adjacent rank. For example, Rater 1 assigned a Y to one criterion and Rater 2 assigned an S. Raters are considered to be in exact agreement if both raters assign the student the same rank. For example, both raters assign the student an N. Across the entire study period, over 19,152 ratings, raters were in general agreement 99.5 percent of the time and in exact agreement 77.2 percent of the time.

**Discussion**

The triple jump examination at OSDUSC is intended to assess students’ critical thinking and higher-order thinking skills. The TJE has been implemented in other dental and medical schools throughout the United States and Canada but has not been as widely accepted as proponents of this type of exam had expected. Reasons for not implementing the TJE include concerns with the subjective nature of the examination, specifically with regard to scale reliability, interrater reliability, and variances in ratings awarded based on rater experience. Our study addressed each of these issues.

In the iteration of the TJE used at our institution, both the overall instrument and each of its subscales were found to be acceptable to good with respect to scale reliability, and interrater reliability was found to be good to strong overall. Additionally, no differences were found in the ratings given based on the experience levels of raters. Raters were found to be in general agreement on their assessments of student performance over 90 percent of the time. While the TJE employs a subjective grading system, that subjectivity does not result in an unreliable instrument. Quite the opposite was seen in this analysis. Each of the parameters assessed in our study, across a very large sample, indicated that concerns about this assessment technique may not be well founded. With the ADEA CCI Task Force’s advocacy of the greater use of evaluations across all competency domains, the characteristics of the TJE that allow for greater assessment of higher level of learning, skills involving analysis, critical thinking, and resolution of problems make the assessment format very desirable.

Perhaps the greatest limitation associated with this study is that the sample analyzed was taken from just one school of dentistry and that, due to this limited sample, albeit a very large one, the external validity of the findings may be somewhat limited. Generalizing the inferences made about the iteration of the TJE as administered at our institution to other iterations of the TJE as seen in other dental and medical schools may not be valid. Faculty calibration is an essential part of our TJE process and can, of course, affect reliability and validity. We recommend that
further research on the reliability of different iterations of the TJE in different schools that replicate our study, including a faculty calibration plan, may be of value in allowing overall assessment of the reliability of the TJE format.

Another possible limitation of our study is that, despite the large number of raters, most raters did not individually participate in a large number of TJEs. Given the relatively low numbers of raters who scored a high number of exams, it is difficult to ascertain characteristics associated with “good” raters. Further, more detailed studies are suggested regarding raters with high experience levels using student evaluations of faculty and student learning outcomes as criteria for defining the characteristics of “good” raters. Such studies are currently under way at OSDUSC.

Conclusion

Our study found that the iteration of the triple jump examination as used at OSDUSC possesses high internal consistency and scale reliability and that the TJE demonstrated strong interrater reliability. There were no statistically significant differences in the mean scores assigned on the TJE between groups defined by rater experience level with the TJE, and there was a very high level of agreement among rater pairs when assessing student performance.

REFERENCES