Assessing the Reliability, Validity, and Use of the Lasater Clinical Judgment Rubric: Three Approaches

Article in Journal of Nursing Education - November 2011
DOI: 10.3928/01484834-20111130-03 - Source: PubMed

CITATIONS
39

READS
2,097

4 authors:

Katie Adamson
University of Washington Tacoma
44 PUBLICATIONS 410 CITATIONS
SEE PROFILE

Paula Gubrud-Howe
Oregon Health and Science University
17 PUBLICATIONS 191 CITATIONS
SEE PROFILE

Stephanie Sideras
Oregon Health and Science University
24 PUBLICATIONS 331 CITATIONS
SEE PROFILE

Kathie Lasater
Oregon Health and Science University
50 PUBLICATIONS 1,058 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

- Developing performance appraisal instrument designed for use in undergraduate nursing simulation. View project
- Population Health and Nurse Education View project
Assessing the Reliability, Validity, and Use of the Lasater Clinical Judgment Rubric: Three Approaches

Katie Anne Adamson, PhD, RN; Paula Gubrud, EdD, RN, FAAN; Stephanie Sideras, PhD, RN; and Kathie Lasater, EdD, RN, ANEF

ABSTRACT
The purpose of this article is to summarize the methods and findings from three different approaches examining the reliability and validity of data from the Lasater Clinical Judgment Rubric (LCJR) using human patient simulation. The first study, by Adamson, assessed the interrater reliability of data produced using the LCJR using intraclass correlation (2,1). Interrater reliability was calculated to be 0.889. The second study, by Gubrud-Howe, used the percent agreement strategy for assessing interrater reliability. Results ranged from 92% to 96%. The third study, by Sideras, used level of agreement for reliability analyses. Results ranged from 57% to 100%. Findings from each of these studies provided evidence supporting the validity of the LCJR for assessing clinical judgment during simulated patient care scenarios. This article provides extensive information about psychometrics and appropriate use of the LCJR and concludes with recommendations for further psychometric assessment and use of the LCJR.

C onsistent, reliable evaluation of students’ and new graduates’ clinical performance has long been a challenge for educators. The expanding use of simulation, which provides learners with opportunities to demonstrate clinical abilities, has intensified the challenge. In particular, educators from both schools of nursing and practice agencies recognize that new graduates often lack the clinical thinking required to meet the needs of acutely ill patients (del Bueno, 2005; Gillespie & Paterson, 2009; Newton & McKenna, 2007). Furthermore, safety initiatives are being implemented in response to reports and quality improvement programs of preventable deaths in acute care settings (Cronenwett et al., 2007; Institute of Medicine, 1999; Joint Commission, 2010). Although critical, these initiatives have added to the complexity of nurses’ work, requiring superior clinical judgment (Ebright, 2004; Ebright, Patterson, Chalko, & Render, 2003). In response to the need for students and graduate nurses to be competent in clinical judgment, methods to evaluate progress in this area are of great interest.

Tanner (2006) defined clinical judgment as “an interpretation or conclusion about a patient’s needs, concerns, or health problems, and/or the decision to take action (or not), use or modify standard approaches, or improvise new ones as deemed appropriate by the patient’s response” (p. 204). Clinical judgment needs to be flexible, not linear, using a variety of ways of knowing, including theoretical knowledge and practical experience (Benner, Tanner, & Chesla, 2009). The purpose of this article is to summarize the psychometric findings from three studies that examined the application of the Lasater Clinical Judgment Rubric (LCJR) in the setting of simulation as a method of identifying performance ability.

BACKGROUND
Development of the Lasater Clinical Judgment Rubric
scribes the development of noticing, interpreting, responding, and reflecting through 11 clinical indicators. Through leveling of the clinical indicators, the LCJR offers language to form a trajectory for development of clinical judgment, opportunity for self-assessment, and facilitating nurse educators’ evaluation of clinical thinking (Cato, Lasater, & Peeples, 2009; Lasater, 2011). The common language provides the potential for use of the LCJR as a research instrument or evaluation tool. Table 1 provides a summary of the aspects and clinical indicators described in the LCJR.

The rubric has been used extensively for educational and research purposes (Adamson, 2011; Blum, Borglund, & Parcells, 2010; Dillard et al., 2009; Gubrud-Howe, 2008; Lasater, 2007a; Lasater & Nielsen, 2009; Mann, 2010; Sideras, 2007). Like any evaluation instrument, the reliability and validity of data produced using the LCJR is of key importance (Kardong-Edgren, Adamson, & Fitzgerald, 2010). The remainder of the background section describes reliability in classical test theory and the relationship between reliability and validity.

Reliability in Classical Test Theory

Reliability is a measure of consistency. In classical test theory, the observed score is a result of a combination of the true score, along with any error in measurement (Nunnally & Bernstein, 1994). A limitation of classical test theory is that error measurement is viewed as a single entity. However, when the goal of the performance appraisal is to evaluate the ability of the learner to respond to a clinical problem that is presented in the highly realistic setting of simulation, identifying extraneous sources of variability becomes important. In performance-based evaluations, there are several sources of variability, including the rater, the simulation case, and the learner’s performance. Rater variability can come from within the raters as a bias they bring to the evaluation setting, such as a belief that older students with more life experience are more capable than younger students, or it can emerge from expected differences between raters, with some being more stringent or lenient (Williams, Klamen, & McGaghie, 2003). Case variability occurs in simulation as a result of how consistently the clinical problem is presented, which can vary with the nature and type of questions that the learner brings to the situation. To obtain a true evaluation of performance, variation by rater and by case needs to be minimal. A reliable and valid performance appraisal requires the use of an instrument that accurately reflects the learner’s ability over the influence of the raters or the specific case.

The Relationship Between Reliability and Validity

Although reliability and validity have traditionally been viewed as two distinct concepts, the most current Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) identify validity as a unitary concept. Validity, according to Messick (1989), is “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). This description is echoed in the Standards (AERA, APA, & NCME, 1999), regarding different types of validity evidence, rather than types of validity. Reliability of data from an instrument provides evidence based on internal structure that supports or refutes the validity argument. Approaching validity and reliability from this angle is particularly important in the educational evaluation of performance. The complexities inherent in the domain-specific knowledge and the behaviors and communication skills required for effective clinical judgment necessitate evaluation using multiple sources of evidence (Downing, 2003).

LITERATURE REVIEW

Case Specificity and Clinical Judgment

In health care, it is a consensus opinion that those who make clinical judgments use multiple processes, including analytic thinking, narrative reasoning, and intuition (Banning, 2007; Norman, 2005; Simmons, 2010; Tanner, 2006). The primary difference between expert judgments and those of novices is the ability to bring domain-specific knowledge to the patient encounter (Norman, 2005; Tanner, 2006). Novices lack the ability to differentiate salient features of a situation, which slows interpretation and decision making regarding interventions (Dreyfus, 2004). For example, the Tanner Model of Clinical Judgment (2006) identified the components of situational context, background knowledge, and relationship with the patient as characteristics that set up nurses’ initial expectations and frame their ability to gain an initial grasp of the situation. The acquisition of such domain-specific contextual knowledge is developmental and visible in the nurses’ ability to fluidly respond to ongoing situational change (Benner, 2004).

### Table 1

<table>
<thead>
<tr>
<th>Clinical Judgment Aspects and Performance Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aspects of Clinical Judgment</strong></td>
</tr>
<tr>
<td>Effective noticing</td>
</tr>
<tr>
<td>Recognizing deviations from expected patterns</td>
</tr>
<tr>
<td>Information seeking</td>
</tr>
<tr>
<td>Effective interpreting</td>
</tr>
<tr>
<td>Making sense of the data</td>
</tr>
<tr>
<td>Prioritizing</td>
</tr>
<tr>
<td>Effective responding</td>
</tr>
<tr>
<td>Calm, confident manner</td>
</tr>
<tr>
<td>Clear communication</td>
</tr>
<tr>
<td>Well-planned intervention and flexibility</td>
</tr>
<tr>
<td>Being skillful</td>
</tr>
<tr>
<td>Effective reflecting</td>
</tr>
<tr>
<td>Evaluation and self-analysis</td>
</tr>
<tr>
<td>Commitment to improvement</td>
</tr>
</tbody>
</table>

Copyright 2007, Kathie Lasater, EdD, RN, ANEF. Reprinted with permission. All rights reserved.
As clinical judgment ability varies more by level of domain-specific knowledge than by application of a particular problem-solving method, it is important to note that the clinical indicators within each of the four aspects of the LCJR (Lasater, 2007a) add further definition to each element. For example, the aspect of noticing comprises the dimensions or clinical indicators of focused observation, recognizing deviations from expected patterns, and information seeking. Thus, the LCJR provides a means to measure demonstration of domain-specific knowledge. The absence of case specificity of the LCJR focuses on the construct of clinical judgment. However, the score obtained by the learner is reflective of ability on only a performance evaluation on the construct of clinical judgment. However, the score obtained by the learner is reflective of ability on only a specific case. Therefore, case variability becomes an important aspect when examining reliability evidence.

Rater Training

One of the greatest threats to the reliability of data produced from observation-based performance evaluation instruments is perception, or human judgment; one rater may perceive a performance differently than another rater and subsequently rate it differently (Shrout & Fleiss, 1979). Redder (2003) conducted research to explore the effect of rater training when using rubrics to assess student performance. Her research affirmed the importance of training raters as a means to establishing interrater reliability. She found that training had a positive effect on interrater reliability, primarily because trained raters (1) construct a mental image of the rubric text and scoring guide, (2) take a more iterative approach to scoring, and (3) tend to make multiple evaluative decisions. Conversely, untrained raters tended to use a more linear approach to scoring student work when using rubrics and are more likely to base their scores on personal experience and their individual understanding of constructs guiding the rubric. Raters review the student performance and then study the anchors to become acquainted with the scoring criterion differences between levels. This type of training is known as performance dimension training and has the goal of helping raters make dimension-relevant decisions (Woehr & Huffcutt, 1994). Raters are encouraged to refer to the anchor performances throughout the scoring process. Wiggins and McTighe (1998) reinforced the notion of anchor performance and further suggested that rubrics should always be accompanied by exemplars of student work to assist raters in developing a mental schema of the knowledge and concepts that the rubric aims to assess. The second activity proposed by Moskal and Leydens (2000) involves practicing scoring sessions and follow-up discussion among raters regarding discrepancies of scores. Differences in interpretation are discussed, and appropriate adjustments to the rubric are negotiated. This process supports the development of a frame of reference within the rater group and is effective both in increasing accuracy of observation and in decreasing error of leniency/stringency and halo (Williams et al., 2003).

Several strategies have been identified for establishing interrater reliability of data produced using observation-based evaluation tools. Moskal and Leydens (2000) asserted that establishing interrater reliability when using rubrics to assess student performance begins by posing the following questions regarding the clarity of the rubric: “1) Are the scoring categories well defined? 2) Are the differences between the score categories clear? and 3) Will two independent raters arrive at the same score for a given response based on the scoring rubric?” (p. 8). In answer to the first two of the three questions, the LCJR includes well-defined scoring categories, and the differences between categories are clear (Gubrud-Howe, 2008). Therefore, to answer the third question about two independent raters arriving at the same score, it is necessary to systematically assess the consistency of multiple raters’ scores. The three studies discussed in this article are specifically concerned with answering this third question.

<table>
<thead>
<tr>
<th>Study</th>
<th>Raters</th>
<th>Subject of Ratings</th>
<th>No. of Scenarios and Ratings</th>
<th>Rater Inclusion Criteria</th>
<th>Rater Training</th>
<th>Reliability Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adamson (2011)</td>
<td>N = 29</td>
<td>3 videoarchived vignettes portraying students performing at various levels of proficiency</td>
<td>3 videoarchived scenarios; 174 ratings</td>
<td>Current teaching experience using simulation; nursing practice experience; no conflict of interest with instrument authors; access to necessary technologies</td>
<td>1-hour telephone or videoconference training</td>
<td>Intraclass correlation (2,1)</td>
</tr>
<tr>
<td>Gubrud-Howe (2008)</td>
<td>N = 2</td>
<td>36 second-year associate degree students in the final quarter of the nursing program</td>
<td>8 scenarios; 72 ratings</td>
<td>Full-time educator; both theory and practical knowledge of simulation; working knowledge of clinical judgment model</td>
<td>7-hour live training using videorecorded scenarios</td>
<td>Percent agreement</td>
</tr>
<tr>
<td>Sideras (2007)</td>
<td>N = 4</td>
<td>22 junior students; 25 senior students</td>
<td>3 scenarios; 141 ratings</td>
<td>Master's degree; full-time educator; both theory and practical knowledge of simulation; working knowledge of clinical judgment model</td>
<td>6-hour seminar</td>
<td>Percent agreement</td>
</tr>
</tbody>
</table>
Findings

The following sections describe three independent studies that assessed the reliability and validity of data produced using the LCJR. Although each of the studies endeavored to answer similar questions, the study designs varied significantly. The Adamson (2011) study examined reliability when individual case variation was minimized but raters had the opportunity to see a broad range of cases (from below expectations to above expectations). The Gubrud-Howe (2008) study examined reliability when the cases and the raters varied. The specific aspects of each study that will be discussed include rater selection, rater training, data collection, and data analyses. Table 2 summarizes the study designs, including characteristics of the raters and ratees and analytic strategies used. Table 3 displays the interrater reliability results and validity evidence from each study. Each of the studies was reviewed by and received exemption certificates or approval from the appropriate institutional review boards.

The Adamson Study

The primary focus of this study (Adamson, 2011) was to pilot a new method for assessing the reliability of simulation evaluation instruments using technology to allow a large number of raters to view the same students in the same simulation and then evaluate the performance using the same simulation evaluation instrument. This method allowed the researcher to minimize individual case variation, thus isolating potential variation caused by the raters. In the past, recruiting a large number of raters to view and score the same scenario has been challenging. To overcome this logistical challenge, this study used videoarchived vignettes that portrayed students in simulated patient care scenarios. The investigators produced three vignettes scripted to depict student nurses performing in simulated patient care scenarios at three levels of proficiency: below, at, or above expectations for a senior baccalaureate nursing student. Twenty-nine nurse educators scattered around the United States, who were masked from the intended level of the scenarios, viewed and scored the students in the vignettes using the LCJR. Intraclass correlations were used to assess the interrater and intrarater reliability of the scores.

Rater Selection. Investigators of this study contacted potential participants via e-mail using a simulation interest electronic mailing list and professional contacts. Potential participants, by self-report, were required to meet the following inclusion and exclusion criteria: currently teach in an accredited, prelicensure, U.S. baccalaureate nursing (BSN) program; have at least 1 year of experience using human patient simulation in prelicensure, BSN education; have clinical teaching or practice experience in an acute care setting as an RN during the past 10 years; not be a primary contributor to the original development of the instrument; have a U.S. Postal Service address, e-mail, and Internet access; and consent to participate. Strict adherence to these criteria provided a relatively homogenous sample of raters, which is ideal for establishing reliability.

Rater Training. Interested, qualified potential raters were sent packets that included additional information about the study and an invitation to attend a video or telephone conference training. As part of the training, the investigator provided background information about the LCJR and the study procedures. Then the rater was asked to view a sample scenario that provided a demonstration of how to score a simulation using the LCJR. Raters were also provided with the investigators’ contact information in case they had any questions or concerns. The one-on-one standardized video and telephone conference trainings were designed to ensure consistency of raters’ training and preparation and lasted approximately 45 minutes each.

Data Collection. Upon completion of the training, raters began the 6-week data collection procedures. Each week, for 6 weeks, participants received e-mails inviting them to score a randomly selected, videoarchived scenario. The three scenarios, each depicting a different level, were coded with symbols (circle, triangle, and square) to mask the participants from the intended level of the scenario they were viewing. A schematic of a sample sequence of study participation is presented in Table 4.

Interrater Reliability Results. Interrater reliability was assessed using intraclass correlation (2,1) agreement. This selection was based on three specifications: two-way ANOVA design; raters were considered random effects—that is, they were intended to represent a random sample from a larger popula-

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Validity Evidence Based on Relationships to Other Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>Interrater Reliability</td>
</tr>
<tr>
<td>Adamson (2011)</td>
<td>ICC (2,1) = 0.889</td>
</tr>
<tr>
<td>Gubrud-Howe (2008)</td>
<td>Agreement = 92% to 96%</td>
</tr>
<tr>
<td>Sideras (2007)</td>
<td>Agreement = 57% to 100%</td>
</tr>
</tbody>
</table>

Note. LCJR = Lasater Clinical Judgment Rubric.
tion; and the unit of analysis was the individual rating (Shrout & Fleiss, 1979). According to Everitt (1996), “The intraclass correlation coefficient can be directly interpreted as the proportion of the variance of an observation due to the between-subjects variability in the true scores” (p. 293). As noted in Table 2, ICC (2,1) = 0.889.

Validity Results. In addition to providing reliability evidence, the results from this study provided validity evidence based on relationships with measures of other variables: the intended levels of the scenarios. The Figure displays the scores assigned to the below expectations, at level of expectations, and above expectations scenarios using the LCJR. These scores were consistent with the intended levels of the scenarios.

The Gubrud-Howe Study

The primary focus of this exploratory study (Gubrud-Howe, 2008) was to better understand the development of clinical judgment in nursing students using the How People Learn (Bransford, Brown, & Cocking, 2000) framework to design instructional strategies in high-fidelity simulation environments. To assess the likelihood that students’ scores would vary between raters, the study design required that the inter-rater reliability of the LCJR be established before the tool was used as an instrument for data collection. This article will focus on the interrater reliability assessment portion of the study.

The interrater reliability assessment portion of this study took place in two phases: first, as part of the initial rater training prior to the initiation of data collection for the larger study, and second, using data collected during the course of the larger study. To assess the interrater reliability of scores assigned using the LCJR prior to the initiation of data collection for the larger study, the researcher identified five previously recorded simulations to serve as anchor performances. The simulations came from a library of recorded scenarios used previously in nursing courses. The five scenarios that were chosen included varying levels of students. Two recordings, one featuring beginning students and another featuring advanced students, were selected as scenarios to use as anchors. The researcher viewed and scored the recordings using the LCJR and developed written comments and instructions regarding the rationale for each score assigned.

Rater Selection. The two raters who were selected to assess interrater reliability were both nursing faculty and had attended a half-day workshop on a Research-Based Model of Clinical Judgment in Nursing (Tanner, 2006). These raters also functioned as instructors during the simulated learning experiences, had extensive, recent experience as nurse educators, and had been using simulation for the previous 18 months prior to the study. However, neither rater had completed any formal training related to the evaluation of simulation activities.

Rater Training. The investigator developed a summary document describing the study, including the study’s conceptual framework. The summary document provided an overview of the study procedures to orient the raters. This orientation was congruent with Redder’s (2003) claim that tactics are needed to assist scorers in developing a mental map or picture of the constructs and criteria that the rubric aims to assess. Once the two participating faculty verbalized that they understood the study and their roles, they, along with the investigator, viewed the previously recorded simulations that served as anchor performances. Results from these ratings were shared and compared.
The investigator facilitated dialogue that promoted a think-aloud format to encourage the raters to describe the reasoning related to each assigned score. A total of five anchor simulation scenarios were assessed in this way.

Comparisons of rater scores after scoring each recorded scenario indicated that the ratings were almost always identical on all items, and the raters verbalized similar rationales for scores given. This process lasted approximately 3 hours; after the fifth scoring, the investigator was satisfied that adequate interrater reliability had been achieved and the study could proceed. Statistical analysis using SPSS software confirmed this assessment, as the alpha coefficient was 0.87.

Data Collection. The raters found that they were best able to complete the rubric when they were close to the scenario action, so they were each situated at opposite sides of the patient room during the simulations. During debriefing, they sat in opposite corners of the room. The raters functioned as spectator observers (Patton, 2002) and did not participate in the scenarios or debriefings while collecting data. Students were accustomed to being evaluated by faculty in a similar manner in both the laboratory and the clinical setting and did not seem to be affected by the raters’ presence.

In addition to observing the scenarios and debriefings to complete the rubric, the raters viewed the digital recordings before finalizing their evaluations. Immediately after each simulation session, the technician replayed each scenario for the raters in the control room. The raters watched each scenario individually and affirmed or adjusted their ratings accordingly. The raters did not confer with each other during the rating process.

This study used a pretest–posttest design, so the Lasater instrument was completed twice for each student enrolled in the study. The first set of ratings occurred during week 2 of a 10-week quarter. The second set was completed at week 9. A total of eight different simulation scenarios were used to collect the data. Four scenarios were used during the first phase of data collection. The second phase of data collection used another four scenarios. All scenarios were designed for a pair of students to participate in the role as a registered nurse. The raters scored two students at a time, and each simulation session with students to participate in the role as a registered nurse. The raters scored two students at a time, and each simulation session with students to participate in the role as a registered nurse. The raters functioned as spectator observers (Patton, 2002) and did not participate in the scenarios or debriefings while collecting data. Students were accustomed to being evaluated by faculty in a similar manner in both the laboratory and the clinical setting and did not seem to be affected by the raters’ presence.

In addition to observing the scenarios and debriefings to complete the rubric, the raters viewed the digital recordings before finalizing their evaluations. Immediately after each simulation session, the technician replayed each scenario for the raters in the control room. The raters watched each scenario individually and affirmed or adjusted their ratings accordingly. The raters did not confer with each other during the rating process.

This study used a pretest–posttest design, so the Lasater instrument was completed twice for each student enrolled in the study. The first set of ratings occurred during week 2 of a 10-week quarter. The second set was completed at week 9. A total of eight different simulation scenarios were used to collect the data. Four scenarios were used during the first phase of data collection. The second phase of data collection used another four scenarios. All scenarios were designed for a pair of students to participate in the role as a registered nurse. The raters scored two students at a time, and each simulation session with the debriefing lasted 50 minutes. A total of 72 ratings were completed and were used to calculate the interrater reliability findings.

Reliability Results. The interrater reliability of data produced during the rater training conducted prior to the initiation of the larger study indicated there was mean score of 92% agreement between raters when examining the 11 clinical indicators of the LCJR. Interrater reliability improved, as data produced as part of the larger study indicated 96% agreement between raters when combining pretest with posttest scores. Findings from one-way ANOVA were also completed to assess for significant differences between raters on each of the 11 clinical indicators. The F ratios for each clinical indicator were all less than 4.84, and all p values were greater than 0.05. These findings confirmed that acceptable interrater reliability was established and that the LCJR was a reliable instrument to use for meeting the study’s aim.

The Sideras Study

The primary focus of this study (Sideras, 2007) was the assessment of the construct validity of the LCJR. The study hypothesis was that graduating senior nursing students would demonstrate a significantly higher level of clinical judgment, as measured by the LCJR, than end-of-year junior nursing students as a result of their increased domain-specific nursing knowledge and amount of clinical experience. The study design compared the clinical judgment performance of the two groups of students using three simulation case scenarios of increasing complexity.

Rater Selection. Four raters were recruited using the following criteria: possession of a master’s degree in nursing, full-time status as a nurse educator, experience with both the theoretical and practice aspects of simulation, and working knowledge of the Tanner (2006) Model of Clinical Judgment. Faculty who had any knowledge of the educational level of the student participants were excluded from the study. The final group of four raters was geographically dispersed and had no prior joint teaching experiences.

Rater Training. Initial rater training consisted of a 6-hour seminar that provided raters with a baseline understanding of clinical judgment theory, the Tanner (2006) Model of Clinical Judgment, and training regarding sources of rater error. The training also provided an opportunity to engage in the active practice of the application of the LCJR to develop rater understanding of the clinical indicators of clinical judgment and begin to establish a joint frame of reference. The goal of this initial seminar was to achieve an interrater percentage of agreement greater than 90%. This goal was initially not met. As a result, supplemental, follow-up modules were developed, and faculty were asked to continue to practice independently, communicating their scoring via e-mail. The flexibility of the modular method was effective in moving faculty raters forward in their application of the rubric. The number of supplemental cycles of training varied across all raters from one to four cycles necessary to attain a greater than 90% level of agreement.

Data Collection. Students from each of the two groups, junior or senior level, participated individually in three simulation cases and the data were recorded to DVDs. Interrater percent agreement was assessed throughout the course of the study to determine if the initial consensus interpretation of the LCJR would be maintained overtime. Overlap DVDs between pairs of raters were scheduled at the fourth, eighth, and 13th rounds of faculty rating. To compare pairs of raters at the overlap points, percent agreement across the three simulations were averaged.

Reliability Results. Calculating interrater reliability using percent agreement is founded on the assumption that each indicator is reasonably independent (Downing, 2004). The 11 clinical performance indicators in the LCJR are highly intercorrelated. To compensate for this intercorrelation, the definition of level of agreement was expanded by one point, so ratings of performance that differed by one level or less were considered equal, and those that differed by two levels or more were considered unequal. Percent agreement varied between pairs and over time. At round four, percent agreement ranged from \( r = 0.75 \) to 1.0; at round eight, \( r = 0.91 \) to 1.0; and at round 13, \( r = 0.85 \) to 0.57. Although these levels of reliability are insufficient for making definitive decisions (Downing, 2004), the limitation of only comparing pairs of raters must be acknowledged.

Validity Results. The validity argument proposed in this study was whether performance measured using the LCJR would find...
the known differences between the two groups of students. This study found that faculty could accurately differentiate performance between junior and senior nursing students. Statistically significant differences were found across all four aspects. Effect size and $z$ score were calculated to provide a gauge of the magnitude of the differences between the two groups (Table 5). Using Cohen’s (1988) guidelines for interpretation, these effect sizes are large, indicating sizable differences between the two groups.

### CONCLUSION

Nurse educators need robust instruments to evaluate all aspects of students’ abilities, including the ability to make clinical judgments. Valid and reliable evaluation instruments allow educators to provide feedback to students that is both specific and accurate. In turn, specific and accurate feedback about student performance and progress helps educators to identify deficits and modify teaching methods. Psychometric assessment of performance-based evaluation instruments is challenging, especially due to the limitations of classical test theory. However, by summarizing the findings of three studies that used diverse methods to assess the reliability of the LCJR, some of the limitations of classical test theory can be mitigated. In addition, several conclusions can be made regarding the validity, appropriate use, and psychometric properties of the LCJR.

### Validity and Appropriate Use of the LCJR

First, the three studies described herein affirm that although student demonstration of clinical judgment is case specific, clinical judgment ability and development are visible in the setting of high-fidelity simulation and measurable using the LCJR. Each of the three studies provided validity evidence supporting the ability of raters to evaluate this construct using the LCJR. The four aspects with their associated clinical indicators provide effective descriptors of clinical judgment that are helpful for the evaluation of this construct. The Adamson (2011) study provided evidence that nursing faculty raters could accurately and consistently identify the “true” or intended level of student performance using the LCJR. The Sideras (2007) study found that faculty could apply the LCJR and accurately differentiate between known levels of student ability, and results from the Gubrud-Howe (2008) study supported the validity of the LCJR from a more theoretical perspective by finding that students who worked to increase their domain-specific nursing knowledge demonstrated improved clinical judgment, as evaluated using the LCJR.

### Reliability

Second, the data from the three studies provided evidence that rater selection, rater training, data collection and analytic strategies affect reliability results. When the raters or the cases used to establish reliability are held stable, data from the LCJR are reliable. The Adamson (2011) study held the cases stable by using three videoarchived vignettes that all 29 raters scored and found the intrarater reliability to be 0.889 using intraclass correlation (2,1). The Gubrud-Howe (2008) study used only two raters who scored a variety of scenarios and found the intrarater reliability to be $r = 0.92$ to 0.96. However, the Sideras (2007) study, in which there was variability in both raters and cases, resulted in a range of reliability results ranging from $r = 0.57$ to 1.0.

### RECOMMENDATIONS

These results prompt several recommendations for future research of the LCJR and other instruments used to evaluate student performance. Specific to the LCJR, a generalizability study must be done to identify the location of the variability in reliability under different conditions to determine whether there is an interaction effect between the raters and the case. Second, researchers and educators need to think carefully about developing simulation scenarios that reveal the true range of students’ clinical judgment abilities. Cases must be appropriately complex to avoid a floor or ceiling effect. Similarly, raters need to view and score a wide range of simulation performances to adequately assess the reliability of a simulation evaluation instrument. Finally, standards for rater training need to be established to decrease rater variability and isolate alternative sources of error.

Given these short-term and long-term suggestions for future research, the authors wish to offer several recommendations about the immediate use of the LCJR. First, no single instrument can provide a comprehensive evaluation of student

### Table 5

Comparison of Groups on Clinical Judgment Aspect, as Rated by Faculty

<table>
<thead>
<tr>
<th>Lasater Clinical Judgment Rubric – Aspect</th>
<th>Juniors</th>
<th></th>
<th>Seniors</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Noticing</td>
<td>2.03</td>
<td>0.68</td>
<td>2.58</td>
<td>0.76</td>
<td>-2.54</td>
<td>0.015$^*$</td>
<td>0.76</td>
</tr>
<tr>
<td>Interpreting</td>
<td>1.94</td>
<td>0.68</td>
<td>2.63</td>
<td>0.76</td>
<td>-3.15</td>
<td>0.003$^{**}$</td>
<td>0.94</td>
</tr>
<tr>
<td>Responding</td>
<td>2.13</td>
<td>0.69</td>
<td>2.72</td>
<td>0.73</td>
<td>-2.77</td>
<td>0.008$^{**}$</td>
<td>0.83</td>
</tr>
<tr>
<td>Reflecting</td>
<td>2.19</td>
<td>0.67</td>
<td>2.82</td>
<td>0.69</td>
<td>-3.14</td>
<td>0.003$^{**}$</td>
<td>0.93</td>
</tr>
</tbody>
</table>

$p < 0.05.$  
$p < 0.01.$
performance or of clinical judgment skill. Likewise, clinical judgment cannot be evaluated in a single episode or summative demonstration (Norman, 2005). Many factors enter into making clinical judgments that cannot be measured or represented in a rubric (Lasater, 2011; Tanner, 2006); therefore, evaluation data from the LCJR should be considered as one component, or a snapshot in time, of a broader evaluation picture. Second, as evidenced by the results from the studies described in this article, reliability results are affected by characteristics of both the raters and the scenarios. Consequently, these results provide evidence supporting the immediate use of the LCJR, along with a caution about the generalizability of any reliability results.

REFERENCES


