Development and Evaluation of High-Fidelity Simulation Case Scenarios for Pediatric Resident Education

Mark D. Adler, MD; Jennifer L. Trainor, MD; Viva Jo Siddall, MS; William C. McGaghie, PhD

Objective.—Pediatric residency programs need objective methods of trainee assessment. Patient simulation can contribute to objective evaluation of acute care event management skills. We describe the development and validation of 4 simulation case scenarios for pediatric resident evaluation.

Methods.—We created 4 pediatric simulation cases: apnea, asthma, supraventricular tachycardia, and sepsis. Each case contains a scenario and an unweighted checklist. Case and checklist development began by reaching expert consensus about case content followed by 92 pilot simulation sessions used for content revision and rater training. After development, 54 first- and second-year pediatric residents participated in 108 simulation test cases to assess the validity of data from these tools for our population. We report outcomes for interrater reliability, discriminant validity, and the impact of potential confounding factors on validity estimates.

Results.—Interrater reliability (κ) ranged from 0.75 to 0.87. There were statistically and educationally significant differences in summary scores between first- and second-year residents for 3 of the 4 cases. Neither previous simulation exposure nor the order in which the cases were performed were found to be significant factors by multivariate analysis.

Conclusions.—Simulation can be used to reliably measure and discriminate resident competencies in acute care management. Rigorous measurement development work is difficult and time-consuming. Done correctly, measurement development yields tangible and lasting benefits for trainees, faculty, and residency programs. Development studies that use systematic procedures and large trainee samples at multiple sites are the best approach to creating measurement tools that yield valid data.

KEY WORDS: medical education; simulation

Ambulatory Pediatrics 2007;7:182–186

Pediatric residency programs have an obligation to assess the competency of residents to care for seriously or critically ill children. To fulfill this need, educators first must consider the substance of the assessment. It is important to conduct both knowledge-based assessment and ones based on assessing clinical skills. George Miller used a pyramid model to describe his hierarchy of assessment, with knowledge ("knows" and "knows how") at the base and performance measures ("show how" and "does") at the apex. To assess competency in acute care skills, educators seek to directly observe learner performance.

The low frequency of acute care events in childhood is a barrier to assessment in the clinical setting. It is extremely difficult to observe pediatric trainees in emergent care situations with a sufficient number and variety of patients to allow for a valid assessment of acute care knowledge and skills. Nadel and colleagues reported data from a large tertiary pediatric residency where 44% of senior residents reported never having led a resuscitation. It would be extremely difficult to arrange for independent evaluators to be present for these unscheduled encounters, even in settings in which more training opportunities arise.

If direct clinical observation is limited, what other evaluation tools are available? Videotape recording of emergent events in the emergency department or intensive care unit is also limited by the infrequency of these events and the rare instances in which residents play a leadership role. Standardized patients have been used to replicate provider-patient interactions including high-stakes assessment (eg, the United States Medical Licensing Examination Step 2 Clinical Skills), but these actors cannot simulate critical illness in children.

High-fidelity human patient (mannequin) simulation offers a practical assessment method that addresses many evaluation problems. A variety of simulated critical events can be developed and standardized for content (disease, symptoms) and setting (emergency room, clinic, number and type of assistants). Simulation offers the opportunity to have residents “show how.” Most importantly, simulation-based assessment occurs without risk or inconvenience to patients and families. In addition, residents who do not meet mastery standards can be offered further training until competence is achieved.

Useful assessment tools yield reliable and valid data. Validity is a construct that addresses the meaning of scores derived from an educational assessment and how the
scores may be interpreted. Reliability refers to "reproducibility of assessment data or scores, over time or occasions." In performance ratings, interrater variation is the main threat to reliability.

This report describes the development of 4 pediatric simulation cases and linked rating checklists. We also assess the reliability and validity of this checklist-based assessment for a sample of pediatric residents. In addition, we describe the efforts required to achieve this goal, the barriers encountered, and the limitations of this type of educational endeavor.

**METHODS**

**Study Design**

This study had a single-group, posttest-only quasi-experimental design.

**Setting/Study Participants**

We enrolled first and second year pediatric residents from Children's Memorial Hospital's (Chicago, Ill) pediatric residency program. During the development phase (July 2003 to June 2004), 51 residents were enrolled. In the evaluation phase (July 2004 to June 2005), 54 residents were enrolled, which represented 100% of the residents. First-year residents starting in 2003 participated in both years of the study. Participation was voluntary. All residents gave written consent to participate.

Pediatric residents did not have any exposure to other mannequin-based simulation at Children's Memorial Hospital during the study period. The institutional review boards of Northwestern University and Children's Memorial Hospital reviewed and approved this study.

**Simulator**

This study was conducted with a high-fidelity human patient simulator, the PediaSIM mannequin (METI, Sarasota, Fla). The simulator is located in the Patient Safety Simulation Center at Northwestern Memorial Hospital, Chicago, Ill. The center provides a lifelike and highly controlled environment for medical education and evaluation. The model is a representation of a 5-year-old child. The mannequin simulates pupil and eyelid functions, respiration, heart sounds, and peripheral pulses. The mannequin can be endotracheally intubated and ventilated as well as electrically cardioverted or defibrillated. A patient monitor displays physiological parameters. The expected steps for each case were preprogrammed. If a resident did not follow the expected path, the facilitator could use a wireless laptop to modify the behavior of the simulator.

**Case and Checklist Development**

We began by creating and iteratively revising 4 pediatric simulation-based evaluation cases (Figure). A standard pediatric mock code reference served as a starting point for case topics. By using pediatric emergency medicine sub-board-certified colleagues as expert resources, we chose the following case topics: (a) apnea (ingestion), (b) asthma, (c) supraventricular tachycardia, and (d) sepsis (oncology patient). We sought to match an a priori list of key skills we intended to assess with the types of cases that work best with the simulator. Cases were constructed with a uniform framework: initial assessment, diagnostic tests, management, and disposition. Most cases required both basic and advanced airway management.

The cases were outlined and designed to last approximately 10 to 15 minutes. Checklist rating instruments were developed concurrently. The case flow was translated into software settings on the simulator. The METI simulator has a sophisticated physiology model with many possible settings. Each case was evaluated to determine whether the programming achieved the desired simulator behavior, particularly when a subject followed an unexpected path. Each case took approximately 20 to 25 work hours to develop before pilot testing.

Fifty-one residents participated in 92 simulation cases during the 2003–2004 academic year (development phase). Each resident completed one or two cases. Residents participated in pairs, with one leading the case and the second serving as a nonverbal assistant. Two study personnel were in the room, one providing the history (the proctor) and the other assisting with care (the nurse confederate) for most cases. The development team revised the cases daily. Obvious problems were addressed in the first few sessions. For example, residents almost never put on gloves despite a prompt from the proctor. Although it is possible that residents simply do not wear gloves, we could not eliminate the possibility that the simulator did not prompt this behavior in the same way that a patient might.

Checklists were developed following Stufflebeam's approach, which used an unweighted dichotomous scale: completed versus not completed. Each checklist had 20 to 30 items at initial development, a number chosen to balance rater effort with the need to have representative items. Each case checklist was revised repeatedly during the development year. The 2 pediatric emergency medicine study faculty (MA and JT) suggested revisions with the help of peer consultation. Checklist items were action
based, reflecting the assessment goal. The final 4 checklists ranged from 22 to 32 items.

In the final 3 months of the development process, the 3 study raters rated 13 cases (35 resident sessions). We reviewed each checklist and removed or clarified items that had poor interrater agreement. For example, we included an item that read “omits action x.” However, we could not assess whether the resident actively avoided action x or did so through ignorance. This type of checklist item has no assessment use unless the case design forces residents to unpack their reasoning through standardized prompting.

**Resident Evaluation Procedure**

Fifty-four residents participated in 2 scenarios during the validation phase. Case assignments were equally distributed by training year. We held validation case sessions once a week over the academic year in 1- to 3-hour blocks as scheduling permitted. Cases lasted about 15 minutes. In contrast to the development phase, residents were evaluated individually to prevent them from observing other residents’ performance. We made no changes to the scenarios or checklists during the evaluation phase. Second-year residents from July 2004 to June 2005 participated in both phases of the study because we had no other trainee source and we could not wait another year to have 2 new groups who were unfamiliar with the checklists. Eight residents (30% of the second-year group) participated in the same case in the development and validation years with a mean time lapse of 16 months (range, 10–20 months).

Rating was accomplished by videotape review. Our raters served in other roles during the case scenarios, so they could not rate during the scenario itself. One of the physician raters had clinical experience with the residents and was not blinded. The other physician and the third rater had no previous experience with the residents but were not blind to the residents’ training year. The video showed the mannequin and resident through 1 of 3 cameras. A technician changed camera views to maintain the best perspective. The simulated patient-monitor screen appeared on the video image as an overlay. One session was lost when the tape was erased accidentally. The resident completed a different case on a later date. Rating required viewing each complete session. Thus, each rater reviewed 16 to 20 hours of video.

**Data Analysis**

All analyses refer to the validation phase of the study. For each of the 54 subjects, we had achievement data for each checklist item from 3 raters. We first analyzed for interrater reliability by using a modified form of Cohen’s k coefficient, described by Brennan and Prediger. To obtain 3-rater k coefficients, we calculated the mean k coefficients (k) for each rater dyad for each checklist item (MA-VS, JT-VS, and MA-JT). A composite k score for each checklist was obtained by averaging the k value for each checklist item.

To analyze whether our cases could discriminate between first- and second-year residents, we used independent t-tests to evaluate for differences in summary scores between training levels. Summary scores were calculated as a percentage of completed checklist items (possible range 0%–100%).

Finally, we analyzed whether resident (prior simulation experience) or case factors were associated with resident performance. Resident demographic data (gender and training year) and exposure to simulation prior to this study were collected. The effect of previous simulation experience was evaluated with an independent t-test. Repeated-measures analysis of variance was used to assess the effect of case order (training effect of having completed a simulation case earlier) and the relationship between study months 1 through 12 on summary scores. Mixed general linear modeling was used to assess the relative impact on summary scores of all independent factors. We also assessed the effect of having participated in the case in both the development and validation years by an independent t-test. Statistics were calculated by Stata version 7.0 (Stata, College Station, Tex) and SAS version 9.1 (SAS, Cary, NC).

**RESULTS**

All residents who were approached agreed to participate. Technical problems caused one session to be rescheduled. Seventy-two percent of subjects were female (78% and 67% of first- and second-year residents, respectively). Fifty-six percent reported previous simulation exposure in medical school (41% and 70% of first- and second-year residents, respectively). Reliability data are listed in the Table. The mean adjusted k coefficients for each case range from 0.75 to 0.87, consistent with acceptable levels of interrater agreement for clinical evaluation.

The Table presents residents’ summary scores for each checklist item.

**Table.** Checklist Descriptive Statistics and Reliabilities

<table>
<thead>
<tr>
<th>Case Scenario</th>
<th>No. of Items</th>
<th>Summary Score (95% Confidence Interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First Year</td>
<td>Second Year</td>
</tr>
<tr>
<td>Asthma kₕ = 0.87</td>
<td>32</td>
<td>61.7 (56.7–66.7)</td>
</tr>
<tr>
<td>Apnea kₕ = 0.83</td>
<td>27</td>
<td>62.0 (57.6–66.4)</td>
</tr>
<tr>
<td>SVS kₕ = 0.75</td>
<td>22</td>
<td>61.0 (54.6–67.4)</td>
</tr>
<tr>
<td>Sepsis kₕ = 0.79</td>
<td>30</td>
<td>65.5 (59.4–71.6)</td>
</tr>
</tbody>
</table>

Bold indicates difference significant at P ≤ 0.05 level. kₕ = Mean Kappa coefficient.
mary scores for each case scenario by training year. Mean scores for all residents ranged from 66.5% to 68.7%. Second-year residents had statistically higher mean summary scores than first-year residents for the asthma, apnea, and supraventricular tachycardia cases. Residents with a previous exposure to simulation before this study had a significantly higher summary score, 69.9% (95% confidence interval [95% CI], 67.1–72.7) than those without exposure 64.4% (95% CI, 60.6–68.1) when all cases are considered together (P = .02). Within training years, the effect of previous simulation experience was not statistically significant. For all subjects, the mean summary score for the case that was performed second was higher than the case performed first (69.0% vs 65.6%, P = .03). The month the case was performed was not related to summary score (F = 1.96, P = .06). Having participated in the same case in both years of the study did not affect scores.

Multivariate analysis was done to assess the relative contribution of postgraduate year, previous simulation experience, and case order. Month performed and case repetition were not included in this model because these were not found to be statistically significant predictors by univariate analysis. In a mixed effects linear regression, postgraduate year was the only significant predictor of summary score (F = 16.25, P = .0002).

**DISCUSSION**

In this study, we demonstrated that assessing residents with high-fidelity simulation cases can be accomplished with high reliability. Our case content was developed and reviewed by content experts in pediatric emergency medicine and medical education. The residents’ response processes produced by the cases using the simulator in the controlled environment closely approximated the behaviors needed to provide acute pediatric care in emergent situations. For 3 of the 4 cases, we demonstrated that higher-level residents performed significantly better than their junior counterparts (discrimination). This difference is preserved after accounting for possible confounding factors. Finally, this approach evaluated performance of key skills, an important goal for resident acute care assessment (consequences for residents and future patients).

For the sepsis case, we did not find a significant difference between first- and second-year residents. The higher mean score for first years (65.5%) versus the other 3 cases (61%–62%) could be explained by the significant exposure in our tertiary-care residency to a large oncology practice, which provided residents more opportunity to manage septic shock than respiratory or cardiac critical events.

Anticipating the major time commitment to develop and pilot test cases, to conduct assessments, and to review and rate performance is particularly important in planning this type of evaluation. We committed more than 100 person-hours in direct work plus an equal amount of preparation and logistical work in each year of this study. The simulation laboratory space was provided by another department’s simulation laboratory without cost, which was a benefit. However, the laboratory was located 10 miles from our campus and required shuttling residents to the study location. Coordination of resident off-campus time was a hurdle and resulted in a lengthy study time span. We believe that the assessment process could have been accomplished over a much shorter time span (but not with less time commitment) if a local simulation laboratory had been available.

We chose a checklist-based assessment because it is easier to implement than other methods. However, checklists are only as good as the quality of the assessment items. Care must be taken to neither omit key factors nor overrepresent unnecessary steps that might be skipped by an expert. Alternative assessment approaches also have limitations. Global assessments can be less reliable unless substantial effort is made to train and calibrate raters and expert raters are expensive and hard to schedule. Time-to-action metrics (eg, time until chest compressions start) are useful alternatives but require carefully constructed starting points and must avoid settings in which more than one alternative approach would be acceptable.

There are limitations to this effort. We cannot generalize these results from a single location without replication of this process at other programs. The case repetition between the development and validation phases did not result in a significant difference in scores. The long time span between repetitions is likely responsible for this effect.

We did not have the resources to have independent raters who could be blinded to resident identity and training level. The lack of blinding to training level might produce a bias that increases the difference among groups, a possibility we cannot exclude. The use of structured checklists with clearly defined, action-oriented items and rigorous rater training removes some of the subjectivity in assessment, which may minimize the effect of bias.

We did not randomize participants but rather allocated residents so that an equal number of residents per training year would participate in each case. The uncertainty of resident availability until 1 month before their session limited our ability to use a random allocation. The lack of randomization may have led to allocation of residents in a biased manner, although we allocated subjects by choosing the case with the fewest participants to date, without consideration of the subject’s identity.

The power to detect significant differences (or effects of confounders) is limited by the sample size of residents, the number of cases, and the number of raters. We enrolled 2 of the 3 training years at our program, which was the limit of our resources and time. Studies of this type are often limited to single programs by logistical and financial constraints.

Resident acute care skills cannot be generalized from a 4-case simulation-based assessment alone. Multiple evaluations of both knowledge and skill from a variety of sources are needed to assess general clinical competence. In this study, we describe the initial process of case development and validation. We plan a series of simulation cases, covering a number of competency domains, which will have sufficient areas of overlap to achieve generalizability. We intend to use these cases as part of a formative evaluation process for our residency training program.
Nonetheless, this study showed that simulation can be used to measure and discriminate resident competency in acute care management. Systematic development and validation of simulation-based assessment tools can improve the quality of assessment data. Rigorous measurement development work is difficult and time-consuming. Done correctly, measurement development yields tangible and lasting benefits for trainees, faculty, and residency programs. Development studies that use systematic procedures and large trainee samples at multiple sites are the best approach to create valid measurement tools.

**ACKNOWLEDGMENTS**

We are grateful for the provision of simulation laboratory time and personnel from the Patient Safety Simulation Center and the Department of Anesthesiology (M. Christine Stock, MD, Chair), Feinberg School of Medicine, Northwestern University. We also thank Leonard D. Wade for his help as simulation laboratory technician.

**REFERENCES**