A Multirater Instrument for the Assessment of Simulated Pediatric Crises

Aaron W. Calhoun, MD
Megan Boone, RN, MSN, CCRN
Karen H. Miller, PhD
Rebecca L. Taulbee, MD
Vicki L. Montgomery, MD
Kimberly Boland, MD

Abstract

Background Few validated instruments exist to measure pediatric code team skills. The goal of this study was to develop an instrument for the assessment of resuscitation competency and self-appraisal using multirater and gap analysis methodologies.

Methods Multirater assessment with gap analysis is a robust methodology that enables the measurement of self-appraisal as well as competency, offering faculty the ability to provide enhanced feedback. The Team Performance during Simulated Crises Instrument (TPDSCI) was grounded in the Accreditation Council for Graduate Medical Education competencies. The instrument contains 5 competencies, each assessed by a series of descriptive rubrics. It was piloted during a series of simulation-based interdisciplinary pediatric crisis resource management education sessions. Course faculty assessed participants, who also did self-assessments. Internal consistency and interrater reliability were analyzed using Cronbach α and intraclass correlation (ICC) statistics. Gap analysis results were examined descriptively.

Results Cronbach α for the instrument was between 0.72 and 0.69. The overall ICC was 0.82. ICC values for the medical knowledge, clinical skills, communication skills, and systems-based practice were between 0.87 and 0.72. The ICC for the professionalism domain was 0.22. Further examination of the professionalism competency revealed a positive skew, 43 simulated sessions (98%) had significant gaps for at least one of the competencies, 38 sessions (86%) had gaps indicating self-overappraisal, and 15 sessions (34%) had gaps indicating self-underappraisal.

Conclusions The TPDSCI possesses good measures of internal consistency and interrater reliability with respect to medical knowledge, clinical skills, communication skills, systems-based practice, and overall competence in the context of simulated interdisciplinary pediatric medical crises. Professionalism remains difficult to assess. These results provide an encouraging first step toward instrument validation. Gap analysis reveals disparities between faculty and self-assessments that indicate inadequate participant self-reflection.

Aaron W. Calhoun, MD, is Director of SPARC Program and Assistant Professor of Pediatrics at University of Louisville; **Megan Boone, RN, MSN, CCRN,** is Nursing Director of SPARC Program and Clinical Nurse Specialist at Just for Kids Critical Care Center at Kosair Children's Hospital; **Karen H. Miller, PhD,** is Adjunct Graduate Faculty of College of Education and Human Development and Assistant Professor of Graduate Medical Education at University of Louisville; **Rebecca L. Taulbee, MD,** is Assistant Professor of Pediatrics at University of Louisville; **Vicki L. Montgomery, MD,** is Chief of Division of Pediatric Critical Care and Professor of Pediatrics at University of Louisville; and **Kimberly Boland, MD,** is Director of Pediatric Residency Program and Associate Professor of Pediatrics at University of Louisville.

The authors wish to acknowledge the assistance of the Paris Simulation Center in the completion of this study.

Program funding was provided by the Children's Hospital Foundation. This entity was not involved in data collection, analysis, or manuscript preparation.

Corresponding author: Aaron W. Calhoun, MD, Division of Pediatric Critical Care Medicine, University of Louisville, 571 South Floyd Street Suite 332, Louisville, KY 40202, 502.852.3720, aaron.calhoun@lousville.edu

Received March 25, 2010; revision received June 17, 2010; accepted October 7, 2010.

DOI: 10.4300/JGME-D-10-00052.1

Editor's Note: The online version of this study contains the Team Performance during Simulated Crises Instrument (TPDSCI) assessment tool used in this study.

Background

During the past decade, simulation has been increasingly used to teach the knowledge and skills needed to manage pediatric medical crises.[1–4] In a typical pediatric crisis simulation, participants receive a clinical story and are then asked to manage the patient as they would in real life, using a high-fidelity mannequin as a patient proxy. Participants are then debriefed regarding their experience. This debriefing encourages reflection upon clinical performance and gives the opportunity to modify future behavior.[5,6] Traditionally, written feedback is not a part of debriefing, even though it could enhance learning by giving participants materials to reflect on after the session's conclusion. Written

TABLE 1 **Comparison of 5 Currently Existing Assessment Tools Focusing on Pediatric Resuscitation in the Simulated Environment**[a]

Name of Assessment Tool	Overall Psychometric Validity	No. of Questions	Question Style	Anchored to ACGME Core Competencies	Validation Environment
Ottawa Crisis Resource Management Global Rating Scale (OCRMGRS)[13]	ICC = 0.59–0.61	5	7-Point descriptively anchored scale	No	Adult medical and surgical ICU
Ottawa Crisis Resource Management Checklist (OCRMC)[17]	ICC = 0.55–0.63	12	3-Point behavior checklist	No	Adult medical and surgical ICU
Standardized Direct Observation Tool (SDOT)[15]	CA = 0.95 ICC = 0.81	26	4-Point behavioral checklist	Yes	Adult emergency department
Neonatal Resuscitation Program (NRP) Megacode Checklist[14]	CA = 0.70	20	3-Point behavioral checklist	No	Neonatal intensive care unit
Tool For Resuscitation Assessment Using Computerized Simulation (TRACS)[16]	ICC = 0.8	72	2-Point (yes-no) behavioral checklist	No	Pediatric intensive care unit

Abbreviations: ACGME, Accreditation Council for Graduate Medical Education; CA, Cronbach α; ICC, intraclass correlation; ICU, intensive care unit.

[a] CA and ICC are statistics commonly used to measure an instrument's internal consistency and interrater reliability. For both, values of 0.7 or above are typically considered as ideal.

feedback, however, is contingent on the existence of valid assessment instruments. To date, several tools have been constructed to meet this need, and their characteristics are presented in TABLE 1.[7–11] Although each has its strengths, none take advantage of recent developments in multirater feedback methodology.

Multirater feedback is a technique derived from the business domain that has been successfully adapted to medical education.[12–17] Its advantage lies in the synthesis of multiple perspectives to achieve a more stable, global rating.[18–20] An additional property is the ability to incorporate learner self-assessment in what is known as gap analysis.[16,18] Gap analysis examines the difference between the combined scores of "expert" faculty raters and the learner's self-score to obtain a measure of that individual's self-appraisal. "Positive" gaps indicate areas where learners underappraise their abilities, while "negative" gaps indicate areas where those abilities are overappraised, a concerning phenomenon suggesting an inability to accurately reflect on performance. By quantifying self-appraisal, gap analysis enables faculty to give focused feedback tailored to alter inaccurate learner self-perception.[16] This technique has been applied to communication skills assessment[16] but has not been used to date in the assessment of code team skills.

Given the enhanced feedback made possible by these techniques, we sought to develop a multirater tool with gap analysis for use in simulation-based pediatric crisis resource management (CRM) courses. An explicit goal of our development process was to adhere as closely to the Accreditation Council for Graduate Medical Education (ACGME) core competencies as possible. In July 2008 we introduced this tool, the Team Performance during Simulated Crises Instrument (TPDSCI), into our pediatric CRM course at the University of Louisville and Kosair Children's Hospital. The goal of this article is to report the psychometric and gap analysis data derived from this pilot.

Methods

This study was approved by the University of Louisville Institutional Review Board.

Tool Development

During the initial development phase, we discovered that competency scores based on individual performance have limited utility due to the team-oriented nature of medical crises. Therefore, we chose to use the entire code team, rather than the team member, as the unit of assessment. The code management literature, which suggests that team cohesiveness affects outcome more significantly than individual performance,[21,22] supported this decision.

During tool development, we assessed the ACGME core competencies (patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice[23]) as they related to code team skill. We sought construct validity by adhering to this framework. Practice-

based learning did not appear measurable in the context of a time-limited training session, as this competency reflects long-term patterns of growth and development, and was therefore not included.[24] We chose to focus the patient care competency on procedural aspects of patient care, as other elements of this competency do not pertain to medical crises,[24] and further limited the procedures involved to key airway, breathing, and circulatory interventions such as bag-mask ventilation, intubation, and chest compressions.

We used a short, rubric-based format to allow for faster, less educationally disruptive assessment. Rubrics were developed by listing key code team behaviors that might fit within each competency and ranking them in order of importance for a successful outcome. This list was transformed into descriptive statements that were further compiled into a series of 5 ranked paragraphs per competency. Each ranked paragraph was given a descriptor based on content and assigned a numerical value (1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent) to enable score averaging. Paragraphs containing multiple behaviors were phrased such that all must be present for a particular score to be merited. The completed tool was reviewed by 2 additional pediatric intensivists for accuracy and transparency of language. Suggested alterations were incorporated into the final version of the tool (online supplemental material).

Tool Implementation

The tool was piloted for 1 year during sessions of the Simulation for Pediatric Assessment, Resuscitation, and Communication Program's CRM training program. This program consists of 1-hour simulation-based CRM sessions conducted in the University of Louisville's Paris Simulation Center, Kosair Children's Hospital Pediatric Critical Care Center, and Kosair Children's Hospital Emergency Department. Participants include pediatrics and medicine-pediatrics residents, pediatric critical care and emergency nurses, respiratory therapists (RTs), and pharmacists. All sessions included participants with varying levels of clinical experience as defined by postgraduate year for physicians and years spent in critical or emergency care settings for nurses, RTs, and pharmacists. Each session begins with an introduction to the simulator followed by a discussion of CRM principles and a brief case history. Session participants then engage in a simulated pediatric crisis, organizing into an interdisciplinary code team to address the situation. Cases included primary cardiac, respiratory, and neurologic derangements. After each simulation the team is debriefed regarding teamwork, procedural skills, and medical knowledge.

During the pilot, the TPDSCI was completed by course faculty. Up to 3 raters were used, depending on faculty availability. A total of 15 faculty raters contributed during the course of the study. Participant teams reached consensus and completed self-evaluations using the tool. Raters received no instruction prior to using the TPDSCI, as our intent was to create a tool transparent enough to be usable ''off-the shelf'' without need for training sessions. Pilot data were then analyzed. Sessions in which intensive care unit faculty participated were excluded from analysis to minimize bias.

Scoring

Competency-specific scores were calculated by averaging individual faculty scores for each competency. Overall score was determined by first averaging the individual competency scores for each faculty rater and then averaging this score between all faculty raters. Average scores of 3 (good) or greater were defined as meeting or exceeding expectations, and scores of less than 3 (poor or fair) were defined as needing improvement. Gap analyses were calculated by subtracting the self-score for each competency and, for the overall resuscitation, from the corresponding average faculty score. An absolute gap of 0.5 has been suggested in the literature as a cutoff for significance.[25] Based on this, we defined a gap of 0.5 or greater as signifying self-underappraisal, and a gap of −0.5 or less as signifying self-overappraisal.

Statistical Analysis

We examined the TPDSCI for internal consistency using Cronbach α, calculated using aggregate data from all included sessions. To address the issue of potential intrasubject correlation, a second α value was calculated by randomly sampling 1 rater from each simulated session, determining an α for that subset, and then repeating this process until all raters were accounted for. The resulting numbers were then averaged to provide a composite Cronbach α.

To determine interrater reliability, intraclass correlation (ICC) coefficients were calculated for each competency as well as for the overall score. Data from sessions rated by less than 3 faculty members were excluded from this analysis as a stable number of raters are needed to produce the statistic. A statistical assessment of construct validity could not be performed as each team was a heterogeneous mix of clinicians with differing experience levels, and we possessed no other information that could serve as a ''gold standard'' for team crisis management ability.[26] We instead chose to present the range of scores descriptively (FIGURE 1).

Gap analysis results were examined for frequency of significant positive (self-underappraisal) and negative (self-overappraisal) gaps. Gap data were correlated with faculty rater scores for domains in which the groups scored less than 3 (defined as needing improvement) or achieved a score of 3 or greater (defined as meeting or exceeding expectations). These results are presented descriptively in TABLE 2.

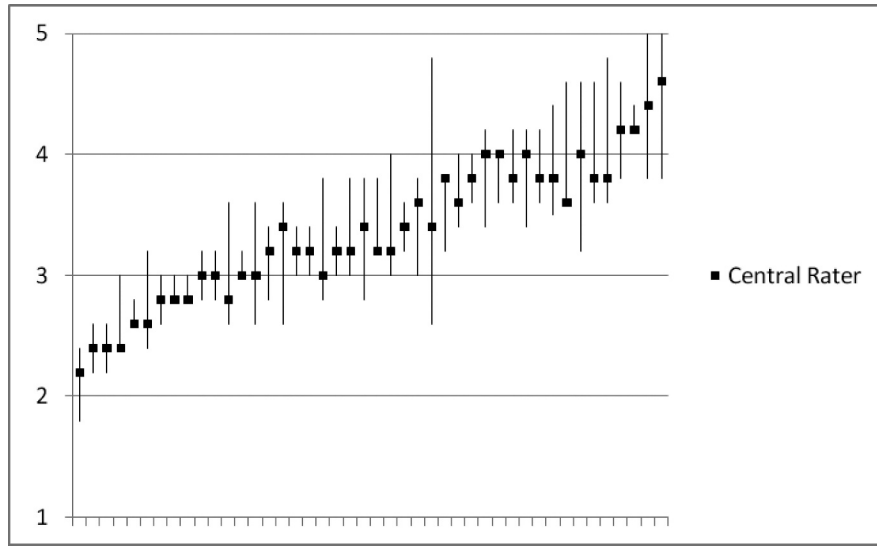

FIGURE 1 **Range of Scores for Each Session as Sorted by Overall Average Score**

Each line corresponds to one session, with the top point of the line corresponding to the highest score given for that session and the bottom point of the line corresponding to the lowest score given for that session. The position of the central rater score is also marked.

Results

Learner Characteristics

After initial exclusions, 54 teams, composed of a total of 79 residents, 128 nurses, 9 RTs, and 8 pharmacists, were found suitable for analysis by Cronbach α. Ten of those teams were subsequently excluded from ICC analysis. FIGURE 2 gives a schematic of the exclusion process. Individual teams consisted of, on average, 4 residents (range, 1–7) and 4 nurses (range 0–8). Experience levels varied between participants, with teams typically containing both experienced and inexperienced members. Pharmacists and RTs attended 19% (10/54) and 13% (7/54) of sessions, respectively. Due to logistical issues, only residents were able to attend Paris Simulation Center sessions.

Instrument Internal Consistency and Reliability

Overall Cronbach α for the TPDSCI was 0.72. Using the averaging process described, we found the Cronbach α to be 0.69. ICC values for the medical knowledge, clinical skill, communication skill, and systems-based practice competencies ranged between 0.72 and 0.87. ICC for the professionalism domain was 0.22. The overall ICC for the TPDSCI was 0.82. Details regarding these statistics are displayed in TABLE 3. FIGURE 1 depicts the range of

TABLE 2 **Correlations Between Objective Faculty Ratings and Gap Analysis Values for All Competencies in Which a Significant Gap Was Noted**[a]

	Competencies Showing Self-Overappraisal (Gap ≤−1)	Competencies Showing Self-Underappraisal (Gap ≥−1)
Competencies scored as meeting or exceeding expectations (scores ≥3)	Percent of all competencies with significant gap = 51%	Percent of all competencies with significant gap = 15%
	Interpretation: These areas represent known strengths of the participants not subject to active self-reflection. Only brief attention should be focused here during debriefing.	Interpretation: These areas represent unrecognized strengths. Attention can be drawn to these areas during debriefing to improve participant's sense of their own abilities.
Competencies scored as needing improvement (scores <3)	Percent of all competencies with significant gap = 31%	Percent of all competencies with significant gap = 3%
	Interpretation: These are concerning areas representing unrecognized weaknesses. These areas should receive special focus during debriefing as they are unlikely to change without external intervention.	Interpretation: These areas represent known weakness subject to active self-reflection. Although these areas should be mentioned during debriefing, attention should not be focused here.

[a] This table examines all 136 domains in which a significant gap was noted, categorizing them by both the type of gap (self-overappraisal versus self-underappraisal) and the overall objective faculty rating for that domain. Interpretations were derived from Calhoun et al.[16]

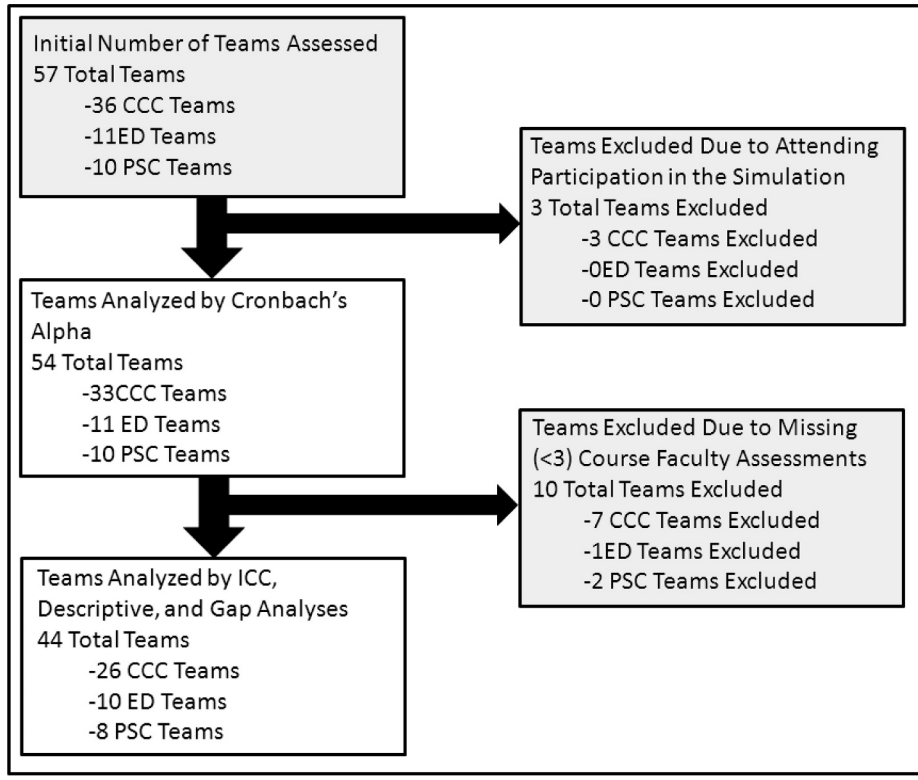

FIGURE 2 **Flow of the Study, Including Initial Teams Assessed, Subsequent Exclusions, and Number and Type of Team Participants Remaining at Each Step**

Analyses performed at each step are detailed above, as well as whether individual simulations occurred at the Kosair Children's Hospital Pediatric Critical Care Center (CCC), Kosair Children's Hospital Pediatric Emergency Department (ED), or the University of Louisville Paris Simulation Center (PSC).

scores present within each individual sessions arranged from lowest to highest average score.

Gap Analysis

Significant gaps were found among all 5 competencies assessed by the tool. Forty-three (97%) sessions assessed were found to have significant gaps between faculty and self-assessment in at least 1 competency; 38 (86%) sessions had at least one significant gap indicating participant team self-overappraisal, and 15 (34%) sessions had at least one significant gap indicating self-underappraisal. Significant gaps were found between overall faculty and team scores for 23 (52%) sessions. Individual domain scores for each team were further examined to determine the prevalence of

TABLE 3 **Global and Domain-Specific Intraclass Correlation (ICC) Coefficients for the Team Performance During Simulated Crises Instrument (TPDSCI)[a]**

Domain	ICC Value	*P* Value	95% Confidence Interval
Medical knowledge	0.76	<.001	0.61–0.86
Clinical skill	0.72	<.001	0.54–0.84
Communication skills	0.87	<.001	0.79–0.93
Professionalism	0.22	.097	−0.21–0.53
Systems-based practice	0.75	<.001	0.59–0.86
Overall	0.82	<.001	0.70–0.90

[a] This table details the ICC coefficients for each of the individual competencies assessed by the TPDSCI, as well as the overall ICC score of the TPDSCI. All domains excepting professionalism were significant at a *P* value of <.001. Overall tool ICC was significant at a *P* value of <.001.

correlated significant gaps. Of 220 possible correlations, individual domains were associated with a significant gap (either self-overappraisal or self-underappraisal) 136 times (62%). As data regarding self-appraisal in a given area must be interpreted in light of the participant's objective abilities in that area to be useful for feedback,[16] we examined these 136 instances further by correlating instances of self-overappraisal or self-underappraisal with average faculty scores in those areas. These results are shown in TABLE 2.

Discussion

There is strong evidence for the internal consistency and overall interrater reliability of the TPDSCI when used in a multirater context. Cronbach α for the instrument ranges between 0.69 and 0.72, indicating that the individual domains contribute to the description of a single construct, namely overall resuscitation competence. This further suggests that an overall score obtained by averaging individual domain scores is meaningful. Although higher alphas are typically sought, our tool contains a deliberate degree of heterogeneity that would render higher alpha levels suspect. The overall ICC of 0.82 further suggests that global TPDSCI scores are reproducible between multiple rater groups. As this study deployed the TPDSCI in the actual situation for which it was designed, these statistics provide strong evidence for the internal consistency and interrater reliability of the TPDSCI when used by multiple raters to assess team performance during simulated pediatric crises.

The ICC scores for most individual competency scores were similarly high, and strong evidence exists for the stability and reproducibility of these scores between rater groups. Professionalism is a significant exception, as the ICC for this competency was quite low. Further examination of the raw data for professionalism revealed that only 1 rater gave a professionalism score of fair, with all other scores ranging from good to excellent, suggesting a significant positive skew in this competency. This is not entirely surprising, as there is reason to believe that this type of assessment methodology may not be ideal for the assessment of professionalism.[27–30]

Construct validity was difficult to assess given the team-oriented nature of the tool. Although it would have been ideal to compare scores to some external anchor of expected performance, such as participant experience,[26] this was not possible because of the heterogeneous experience levels in the participant teams. In addition to this, we did not possess information regarding any other experience measures, such as previous code participation, that could function as a proxy for expected performance. Still, the contrast between the relatively large range of average session scores and relatively small range of individual rater scores within each session suggests that the tool is able to discriminate effectively between sessions. It is not unreasonable to assume that these differences at least partially correspond to real performance variance.

Examination of the gap analysis data revealed a high number of competencies in which team self-perception differed significantly from course faculty. This is not surprising given the reported unreliability of self-perception and is consistent with previously reported material regarding gap analysis in a medical context.[16,31,32] Although some gaps may be attributable to subjective factors such as differing perceptions or perspectives regarding the session, the TPDSCI assesses most competencies using objective, behavioral criteria, and thus most gaps likely represent truly inaccurate participant self-evaluation. Most (86%, 38/44) sessions had at least some degree of self-overappraisal, indicating that participants felt that they performed significantly better than they actually did, a concerning finding given the high stakes nature of pediatric resuscitations. A more worrisome observation is that 31% (43/136) of gaps indicating self-overappraisal occurred in competency domains rated as needing improvement, a combination indicating an area of weakness not recognized by the team. Such areas need focused intervention. This finding illustrates the value of gap analysis, as its use enables faculty raters to uncover ''blind spots'' and address them in a way that will enhance participant self-reflection.

Limitations

One potential limitation is the small number of raters (3) per session. However many simulation-based CRM programs will likely not have more than 3 faculty members present for any given session. Although sample size is a possible limitation, our subject number is equivalent to or greater than those of comparable pediatric CRM assessment tool validation studies.[7–10,13] Finally, we could not assess construct validity statistically and our observations are suggestive at best. Further study will be needed to examine this aspect of the tool.

Conclusion

Multirater feedback with gap analysis is a robust means for the assessment of team competence and self-appraisal in the context of simulation-based pediatric CRM training. The TPDSCI has demonstrated good internal consistency and overall interrater reliability. Good interrater reliability has also been demonstrated for all component competencies except professionalism. Data are encouraging with respect to construct validity, but significant further work will be needed to demonstrate this statistically. The use of gap analysis can further delineate the competency-specific self-perception of the resuscitation team, allowing for enhanced self-reflection and enabling faculty to intervene in areas in which inadequate self-appraisal coincides with a need for improvement.

References

1 Weinstock PH, Kappus LJ, Kleinman ME, Grenier B, Hickey P, Burns JP. Toward a new paradigm in hospital-based pediatric education: the development of an onsite simulator program. *Pediatr Crit Care Med*. 2005;6(6):635–641.
2 Weinberg ER, Auerbach MA, Shah NB. The use of simulation for pediatric training and assessment. *Curr Opin Pediatr*. 2009;21(3):282–287.
3 Nadel FM, Lavelle JM, Fein JA, Giardino AP, Decker JM, Durbin DR. Assessing pediatric senior residents' training in resuscitation: fund of knowledge, technical skills, and perception of confidence. *Pediatr Emerg Care*. 2000;16(2):73–76.
4 Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. *Acad Med*. 2003;78(8):783–788.
5 Welke TM, LeBlanc VR, Savoldelli GL, et al. Personalized oral debriefing versus standardized multimedia instruction after patient crisis simulation. *Anesth Analg*. 2009;109(1):183–189.
6 Miller KK, Riley W, Davis S, Hansen HE. In situ simulation: a method of experiential learning to promote safety and team behavior. *J Perinat Neonatal Nurs*. 2008;22(2):105–113.
7 Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Crit Care Med*. 2006;34(8):2167–2174.
8 Lockyer J, Singhal N, Fidler H, Weiner G, Aziz K, Curran V. The development and testing of a performance checklist to assess neonatal resuscitation megacode skill. *Pediatrics*. 2006;118(6):e1739–e1744.
9 Shayne P, Gallahue F, Rinnert S, Anderson CL, Hern G, Katz E. Reliability of a core competency checklist assessment in the emergency department: the Standardized Direct Observation Assessment Tool. *Acad Emerg Med*. 2006;13(7):727–732.
10 Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih MC, Kleinman ME. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics*. 2008;121(3):e597–e603.
11 Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). *Simul Healthc*. 2009;4(1):6–16.
12 Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. *Can J Anaesth*. 2006;53(1):33–39.
13 Potter TB, Palmer RG. 360-degree assessment in a multidisciplinary team setting. *Rheumatology (Oxford)*. 2003;42(11):1404–1407.
14 Wood J, Collins J, Burnside ES, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Acad Radiol*. 2004;11(8):931–939.
15 Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. *Acad Med*. 2004;79(5):458–463.
16 Calhoun AW, Rider EA, Meyer EC, Lamiani G, Truog RD. Assessment of communication skills and self-appraisal in the simulated environment: feasibility of multirater feedback with gap analysis. *Simul Healthc*. 2009;4(1):22–29.
17 Higgins RS, Bridges J, Burke JM, O'Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. *Ann Thorac Surg*. 2004;77(1):12–17.
18 Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof*. 2003;23(1):4–12.
19 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med*. 1997;72(10)(suppl 1):S82–S84.
20 Foster C, Law, M. How many perspectives provide a compass?: differentiating 360-degree and multi-source feedback. *Int J Select Assess*. 2006;14(3):288–291.
21 Cooper S, Wakelam A. Leadership of resuscitation teams: "lighthouse leadership". *Resuscitation*. 1999;42(1):27–45.
22 Marsch SC, Muller C, Marquardt K, Conrad G, Tschan F, Hunziker PR. Human factors affect the quality of cardiopulmonary resuscitation in simulated cardiac arrests. *Resuscitation*. 2004;60(1):51–56.
23 Accreditation Council for Graduate Medical Education. ACGME Outcome Project. http://www.acgme.org/outcome. Accessed March 15, 2007.
24 Accreditation Council for Graduate Medical Education. *Introduction to Competency-Based Residency Education*. 2006.
25 Full Circle Feedback Pty. Ltd. Full Circle feedback learning guide. http://www.acgme.org/outcome/e-learn/e_powerpoint.asp. Accessed February 7, 2011.
26 Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837.
27 Arnold L. Assessing professional behavior: yesterday, today, and tomorrow. *Acad Med*. 2002;77(6):502–515.
28 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269(13):1655–1660.
29 van Mook WN, van Luijk SJ, O'Sullivan H, et al. The concepts of professionalism and professional behaviour: conflicts in both definition and learning outcomes. *Eur J Intern Med*. 2009;20(4):e85–e89.
30 van Mook WN, van Luijk SJ, O'Sullivan H, Wass V, Schuwirth LW, van der Vleuten CP. General considerations regarding assessment of professional behaviour. *Eur J Intern Med*. 2009;20(4):e90–e95.
31 Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006;296(9):1094–1102.
32 Lockyer JM, Violato C, Fidler HM. What multisource feedback factors influence physician self-assessments?: a five-year longitudinal study. *Acad Med*. 2007;82(10)(suppl):S77–S80.