Training less-experienced faculty improves reliability of skills assessment in cardiac surgery

Xiaoying Lou, BS,[a] Richard Lee, MD,[b] Richard H. Feins, MD,[c] Daniel Enter, MD,[a] George L. Hicks, Jr, MD,[d] Edward D. Verrier, MD,[e] and James I. Fann, MD[f]

Objective: Previous work has demonstrated high inter-rater reliability in the objective assessment of simulated anastomoses among experienced educators. We evaluated the inter-rater reliability of less-experienced educators and the impact of focused training with a video-embedded coronary anastomosis assessment tool.

Methods: Nine less-experienced cardiothoracic surgery faculty members from different institutions evaluated 2 videos of simulated coronary anastomoses (1 by a medical student and 1 by a resident) at the Thoracic Surgery Directors Association Boot Camp. They then underwent a 30-minute training session using an assessment tool with embedded videos to anchor rating scores for 10 components of coronary artery anastomosis. Afterward, they evaluated 2 videos of a different student and resident performing the task. Components were scored on a 1 to 5 Likert scale, yielding an average composite score. Inter-rater reliabilities of component and composite scores were assessed using intraclass correlation coefficients (ICCs) and overall pass/fail ratings with kappa.

Results: All components of the assessment tool exhibited improvement in reliability, with 4 (bite, needle holder use, needle angles, and hand mechanics) improving the most from poor (ICC range, 0.09-0.48) to strong (ICC range, 0.80-0.90) agreement. After training, inter-rater reliabilities for composite scores improved from moderate (ICC, 0.76) to strong (ICC, 0.90) agreement, and for overall pass/fail ratings, from poor (kappa = 0.20) to moderate (kappa = 0.78) agreement.

Conclusions: Focused, video-based anchor training facilitates greater inter-rater reliability in the objective assessment of simulated coronary anastomoses. Among raters with less teaching experience, such training may be needed before objective evaluation of technical skills. (J Thorac Cardiovasc Surg 2014;148:2491-6)

See related commentary on pages 2497-8.

Supplemental material is available online.

From Northwestern University,[a] Chicago, Ill; St Louis University,[b] St Louis, Mo; University of North Carolina at Chapel Hill,[c] Chapel Hill, NC; University of Rochester Medical Center,[d] Rochester, NY; University of Washington Medical Center,[e] Seattle, Wash; and Stanford University,[f] Stanford, Calif.

This study was supported by the Thoracic Surgery Directors Association and the Joint Council for Thoracic Surgery Education, the Thoracic Surgery Foundation for Research and Education Simulation Grant (R.H.F. and J.I.F.), and the Western Thoracic Surgical Association Doty Award (J.I.F.). Corporate sponsors of this session include the Chamberlain Group, Sorin Group, St Jude Medical Inc, CryoLife Inc, Scanlan International, LifeLike BioTissue Inc, Teleflex Medical, and Ethicon Inc.

Disclosures: James I. Fann reports equity ownership in Apica. Richard H. Feins reports equity ownership in KindHeart Inc. The other authors have nothing to disclose with regard to commercial support.

Read at the 40th Annual Meeting of The Western Thoracic Surgical Association, Dana Point, California, June 25-28, 2014.

Received for publication June 23, 2014; revisions received Aug 25, 2014; accepted for publication Sept 4, 2014; available ahead of print Oct 11, 2014.

Address for reprints: James I. Fann, MD, Department of Cardiothoracic Surgery, Stanford University, 300 Pasteur Dr, Stanford, CA 94305 (E-mail: jfann@stanford.edu).

0022-5223/$0.00

Published by Elsevier Inc. on behalf of The American Association for Thoracic Surgery

http://dx.doi.org/10.1016/j.jtcvs.2014.09.017

Technical skill is a key component of surgical competence and a core component of cardiothoracic (CT) surgery training. For the last 2 decades, the use of surgical simulators has evolved as a way for trainees to learn and practice technical skills in a safe, cost-effective, and low-stress environment.[1] Simulation also affords opportunities for direct observation for formative and summative assessment. For such assessments to accurately reflect a trainee's level of technical skill, however, they must be standardized. As the role of simulation expands with the potential for incorporation in high-stakes settings, such as those used for promotion and certification, it is paramount that assessment tools demonstrate high inter-rater reliability and ease of execution.[2]

In CT surgery, the Joint Council on Thoracic Surgery Education (JCTSE) and the Thoracic Surgery Directors Association (TSDA) have developed instruments to evaluate trainee competence in common operative procedures.[3-6] For the JCTSE coronary artery anastomosis assessment tool, high inter-rater reliability among experienced educators and senior faculty members, even without rater training, has been demonstrated.[7] Because junior faculty members with less experience as educators are often charged with evaluating trainee competence, it is requisite that they achieve similar levels of inter-rater reliability.

Currently, inter-rater reliability among less-experienced educators has not been established. Moreover, although rater

Abbreviations and Acronyms

CT	= cardiothoracic
ICC	= intraclass correlation coefficient
JCTSE	= Joint Council on Thoracic Surgery Education
P/F	= pass/fail
TSDA	= Thoracic Surgery Directors Association

training has been recognized to improve inter-rater reliability, its effects have not been assessed in CT surgery. To address these needs, a skills assessment session was held at the JCTSE Educate the Educators program at the TSDA Boot Camp in 2013. The session included rater training for the JCTSE coronary artery anastomosis assessment tool. Rater training aims to improve rater performance by developing the necessary knowledge, skills, and attitudes to accurately evaluate skills and competencies.[8,9] The type of training used in this session can be characterized as performance dimension training with elements of frame of reference training. Performance dimension training teaches raters to recognize appropriate behaviors associated with each dimension targeted for evaluation using written or visual depictions. Examples representing expert consensus are provided to raters so that they associate similar behavioral cues with the dimension being evaluated. Frame of reference training involves recognition and expert-facilitated discussion of discrepancies between raters to provide feedback that improves rater performance.[9,10]

Although no standardized rater training techniques currently exist, it is generally agreed that jointly examining the sources of inter-rater variability and establishing a consensus to address any uncertainties enhances rater reliability.[11] In this study, we thus evaluated inter-rater reliability of less-experienced educators and the impact of focused training with a video-embedded coronary anastomosis assessment tool on improvement in inter-rater reliability.

MATERIALS AND METHODS

Nine CT surgery faculty members from different academic institutions participated as raters in the JCTSE Educate the Educators session on assessment at the TSDA Boot Camp at University of North Carolina, Chapel Hill. During coronary anastomoses training sessions, 4 individuals (2 medical students and 2 CT surgical residents) were recruited to perform a coronary artery anastomosis using a simulator; the individuals had a level of experience with coronary anastomoses consistent with their level of training. Approval for the study was obtained from the institutional review board at the University of North Carolina, Chapel Hill.

Model for Coronary Artery Anastomoses and Video Recordings

Coronary vessel anastomoses were performed using a synthetic graft task station and video recorded.[4] The medical students and residents anastomosed a 3-mm synthetic vein graft onto a 3-mm synthetic target vessel mounted in a portable chest model (HeartCase; Chamberlain Group, Great Barrington, Mass) using 6-0 polypropylene sutures and surgical instruments (Figure 1). The video recordings were edited to approximately 5 to 6 minutes, which included representative clips for subsequent evaluation of the assessment components. All video recordings were de-identified and limited to views of the simulation model and the participant's hands.

Joint Council on Thoracic Surgery Education Assessment Tool for Coronary Artery Anastomosis

The JCTSE assessment tool consists of 13 assessment components: arteriotomy, graft orientation, bite, spacing, needle holder use, use of forceps, needle angles, needle transfer, suture management, knot tying, hand mechanics, use of both hands, and economy of time and motion. Because of the limitations of the simulation model and the varying degree of aid of an assistant surgeon, 3 assessment components (arteriotomy, graft orientation, and economy of time and motion) could not be evaluated. The other components are scored on a Likert scale from 1 (poor) to 5 (excellent), with anchoring of 1, 3, and 5 ratings with behavioral descriptors (Appendix Table E1).

Training Protocol and Data Collection

After a brief introduction to the use of simulation of coronary artery anastomosis, raters were provided paper copies of the assessment tool and allowed 5 minutes to review the tool and behavioral anchors. No further explanation of the tool or its anchors was provided. Raters then consecutively viewed and evaluated 2 video recordings of 1 medical student and 1 resident performance of a coronary anastomosis on the simulator. For each anastomosis, a rating from 1 to 5 was assigned for 10 assessment components, yielding an average composite score. Each performance also received an overall pass/fail (P/F) rating. All evaluations were completed on paper, independently, and without knowledge of the subject's level of training. Assessment took place concurrently with video viewing. Afterward, raters used audience response clickers to input their ratings, which were captured by live polling software (TurningPoint 5.2.1; Turning Technologies, Youngstown, Ohio). This setup provided raters with immediate visual feedback that compared their ratings with those by the rest of the group.

Training consisted of 30 minutes of expert-facilitated discussion of the behavioral descriptors used to anchor the assessment tool. Raters were asked to review a series of 10- to 15-second video clips embedded into the assessment tool depicting the levels of skill corresponding to 1, 3, and 5 ratings for each of the 10 assessment components (Figure 2). The embedded video clips had been collected before the rating session and had been deemed to be representative samples of these anchors by the group of experienced raters involved in the development of the assessment tool.[7] All questions posed by raters were also answered, and areas of discrepancy were discussed. Immediately after the training session, all raters evaluated the remaining 2 videos of a different medical student and resident performing the task using the same procedure as outlined previously.

Statistical Analysis

Data are expressed as mean $\pm$ standard deviation. Inter-rater reliability of composite scores as continuous variables and assessment component scores as ordinal variables were assessed using intraclass correlation coefficients (ICCs), and overall P/F ratings as dichotomous variables using Fleiss' kappa of concordance (κ). Internal consistency reliability among assessment components was assessed with Cronbach's α. Reliability is an index ranging from 0 to 1. Although no consensus on index levels currently exists, it is generally accepted that tools with reliabilities in the 0.0 to 0.5 range are imprecise and those in the 0.5 to 0.8 range are moderately reliable. Tools with reliability indices greater than 0.8 exhibit