The Bankart Performance Metrics Combined With a Cadaveric Shoulder Create a Precise and Accurate Assessment Tool for Measuring Surgeon Skill

Richard L. Angelo, M.D., Richard K. N. Ryu, M.D., Robert A. Pedowitz, M.D., Ph.D., and Anthony G. Gallagher, Ph.D., D.Sc.

Purpose: To determine if previously validated performance metrics for an arthroscopic Bankart repair (ABR) coupled with a cadaveric shoulder are a valid assessment tool with the ability to discriminate between the performances of experienced and novice surgeons and to establish a proficiency benchmark for an ABR using a cadaveric shoulder. **Methods:** Ten master/associate master faculty from an Arthroscopy Association of North America Resident Course (experienced group) were compared with 12 postgraduate year 4 and postgraduate year 5 orthopaedic residents (novice group). Each group was instructed to perform a diagnostic arthroscopy and a 3 suture anchor Bankart repair on a cadaveric shoulder. The procedure was videotaped in its entirety and independently scored in blinded fashion by a pair of trained reviewers. Scoring was based on defined and previously validated metrics for an ABR and included steps, errors, "sentinel" (more serious) errors, and time. **Results:** The inter-rater reliability was 0.92. Novice surgeons made 50% more errors (5.86 *v* 2.95, $P = .013$), showed more performance variability (SD, 1.86 *v* 0.55), and took longer to perform the procedure (45.5 minutes *v* 25.9 minutes, $P < .001$). The greatest difference in errors related to suture delivery and management (exclusive of knot tying) (1.95 *v* 0.45, $P = .024$). **Conclusions:** The assessment tool composed of validated arthroscopic Bankart metrics coupled with a cadaveric shoulder accurately distinguishes the performance of experienced from novice orthopaedic surgeons. A benchmark based on the mean performance of the experienced group includes completion of a 3-anchor Bankart repair, and enacting no more than 3 total errors and 1 sentinel error. **Clinical Relevance:** Validated procedural metrics combined with the use of a cadaveric shoulder can be used to assess the performance of an ABR. The methodology used may serve as a template for outcomes-based procedural skills training in general.

The traditional manner in which surgical trainees have acquired their operative skills is under considerable pressure. Concerns about patient safety,[1,2] pressures on operating room efficiency,[3] and the reduced availability of work hours[4,5] have resulted in fewer opportunities for in vivo operative experience. As a consequence, trainees are graduating from residency programs with considerably less operative experience and almost certainly less technical skill than residents graduating in the past who were exposed to greater surgical volumes. For example, Bell et al.[6] found that of the 121 surgical procedures that general surgery residency program directors believed residents should be competent in by the time of graduation, only 18 of them had been performed with sufficient frequency by residents for them to acquire competence during their training. They also found that the mode frequency with which the 121 procedures were performed was 0. The implications of these findings for surgical training are considerable and concerning. At a more practical level, it means that surgical skills training must be optimized and preparation for a surgical practice maximized.

Traditionally, surgical residents have been trained using the "apprenticeship" model, dependent in part on exposure to surgical cases, variable graduated participation in surgery, and time spent on specific clinical

From ProOrtho Clinic (R.L.A.), Kirkland, Washington; The Ryu Hurvitz Orthopedic Clinic (R.K.N.R.), Santa Barbara, California; Professor Emeritus, University of California (R.A.P.), Los Angeles, California, U.S.A.; and ASSERT, University College Cork (A.G.G.), Cork, Ireland.

The authors report the following potential conflict of interest or source of funding: R.L.A. receives support from DePuy Mitek. R.K.N.R. receives support from MedBridge, Mitek, and Rotation Medical.

Received June 25, 2014; accepted May 13, 2015.

Address correspondence to Richard L. Angelo, M.D., ProOrtho Clinic, 12911 120th Ave NE, Ste H-210, Kirkland, WA 98072, U.S.A. E-mail: rlamdortho@comcast.net

0749-8063/14538/$36.00

http://dx.doi.org/10.1016/j.arthro.2015.05.006

Table 1. Glossary

	Definition
Construct validity	A type of evidence that supports that specific test items identify the quality, ability, or trait they were designed to measure
Content validity	An estimate (opinion) by experts of the validity of a testing instrument based on a detailed examination of the contents of the test items
Damage to non-target tissue	Iatrogenic damage to tissues not intended to be addressed in the specific step (e.g., articular cartilage damage)
Definition	A definite, distinct, and clear objective characterization providing an accurate and reliable identification of whether an event was or was not observed to have occurred
Delphi Panel (modified)	A structured communication technique originally developed as a systematic, interactive forecasting method that relies on the opinion of a panel of experts; in modified form, experts answer queries or vote in 2 or more rounds (cycles) on the appropriateness of the metric-based operational definitions of detailed aspects of procedure performance with the goal of achieving consensus—voting is not anonymous
Error	A deviation from optimal performance
Face validity	An estimate (opinion) by experts who review the content of an assessment or tool to see if it seems appropriate and relevant to the concept it purports to measure
Inter-rater reliability	The extent of agreement between 2 raters on the occurrence of a series of observed events; it ranges between 0, no agreement, and 1.0, complete agreement
Metric	A standard of measurement of quantitative assessments used for objective evaluations to make comparisons or to track performance
Operational definition	Terms used to define a variable or event in terms of a process (or set of validation tests) needed to determine its existence, quantity, and duration
Performance metric	The features determining the accomplishment of a given task measured against preset known standards of accuracy and completeness
Procedure phase	A group or series of integrally related events or actions that, when combined with other phases, make up or constitute a complete operative procedure
Proficiency/proficient	A specific level of performance defined by a quantitative score (benchmark) or scores on a standardized test or other form of assessment
Proficiency-based progression	A training program that dictates that skill performance be demonstrated, to a predetermined benchmark level, by the trainee before advancement to more complex techniques
Sentinel error	An event or occurrence involving a serious deviation from optimal performance during a procedure that either (1) jeopardizes the success/desired result of the procedure or (2) creates iatrogenic insult to the patient's tissues
Step	A component task, the series aggregate of which constitutes the completion of a specific procedure
Task analysis	An assessment of how a procedure is accomplished, including a detailed (functional) description of the manual activities or tasks along with their duration, frequency, and complexity and any other unique and distinguishing factors
Task deconstruction	To break down a procedure into constituent tasks, steps, or components

rotations. At the outset, we sought to determine if a "proficiency-based progression" (PBP) method was potentially a more effective manner in which to train surgical skills than the apprenticeship model (Table 1 includes a glossary of terms used throughout the article). A PBP training program dictates that skill performance, to a predetermined benchmark level, be demonstrated by the trainee before advancing to more complex techniques. This method relies on a comprehensive and quantitative characterization of the skills to be learned. These performance characteristics, or "metrics," and their "operational definitions" (rather than descriptions) (Table 1) offer very specific goals and guidelines as part of the training curriculum. Previously, we reported on the development of "performance metrics" ("steps" and "errors")[7] (Table 1) for a standard reference approach to performing an arthroscopic Bankart repair (ABR).[8-12] Those metrics were derived from a careful "task analysis" and "deconstruction" (Table 1) using videos of complete Bankart procedures performed with patients in either the lateral decubitus or beach-chair orientation. The metrics were constructed in such a manner that they could be scored in an identical manner with the patient in either orientation. "Face validity" and "content validity" of the metrics were verified using a modified Delphi Panel methodology (Table 1). The Delphi Panel was composed of 27 experienced shoulder arthroscopists who have all served as master or associate master faculty for Arthroscopy Association of North America (AANA) shoulder courses at the Orthopedic Learning Center (Rosemont, IL). The Delphi Panel obtained excellent consensus on the metric-based characterization of the Bankart procedure.

In a subsequent report, we verified "construct validity" (the ability to discriminate between the performance of experienced and novice groups of surgeons) (Table 1) for the use of the ABR metrics with a shoulder

model simulator as a training tool.[13] In the present study we evaluate the construct validity of these exact same metrics on a much higher-fidelity platform, the human cadaveric shoulder. At present, an ABR in a cadaveric shoulder provides the closest approximation to a similar surgical repair in a live patient. Full-physics, high-fidelity virtual reality simulators with haptic feedback are likely to play a greater role in the future but are expensive to develop and not currently available. Even with such simulators, validated metrics will be needed to substantiate their effectiveness.

The purpose of this study was to determine if previously validated performance metrics for an ABR coupled with a cadaveric shoulder are a valid assessment tool with the ability to discriminate between the performances of experienced and novice surgeons. We also sought to establish a "proficiency" (Table 1) benchmark for this procedure using the cadaveric shoulder. The null hypothesis was that when using a cadaveric shoulder, the Bankart metrics would fail to discriminate between experienced and novice surgeon performance.

Methods

No institutional review board (IRB) approval was obtained for this study investigating the validity of the Bankart metrics coupled with the cadaveric shoulder. IRB approval was sought for the final Copernicus Study proper, which will compare 3 different training protocols evaluating surgical simulation and proficiency-based training methods. The Western IRB (No. 1-776362-1) opined that, as an educational curriculum study, this study was exempt from the need for full IRB approval [based on the criteria of 45 CFR 46.101(b)(1)]. The final study comparing the 3 training protocols was registered with the National Institutes of Health (ClinicalTrials.gov No. NCT01921621).

Study Groups

Two groups were compared in their performance of an arthroscopic Bankart procedure on a cadaveric shoulder. The experienced group consisted of all faculty members who served as master or associate master instructors for a standard 3-day AANA Resident Course conducted at the Orthopedic Learning Center. "Experienced" meant that they performed the procedure consistently in practice and taught the principles at the Orthopedic Learning Center during shoulder courses. An "expert" would not be possible to define without the surgeon meeting objective performance criteria (metrics) that achieved a specific benchmark (that some group or body determined meant "expert" performance). The novice group was limited to postgraduate year (PGY) 4 and PGY 5 orthopaedic residents who had registered for a Resident Course and who volunteered to participate in the investigation.

ABR Metrics

Metrics have been previously defined for a standard reference ABR.[7] Forty-five essential steps in 13 "procedural phases" (Table 1) (in Roman numerals) were defined with beginning points and endpoints (Table 2). Twenty-nine potential unique errors were specified (Table 3), 8 of which were designated as "sentinel" (Table 1). The more serious (sentinel) errors were defined as those expected to either (1) substantially compromise the outcome of the shoulder stabilization (e.g., "capsular penetration of the suture passing instrument is superior to the anchor hole" resulting in failure to achieve retention of the capsule and inferior glenohumeral ligaments) or (2) potentially lead to iatrogenic damage to the shoulder (e.g., "laceration of the intact labrum"). Some of the same errors could be enacted more than once during separate but similar phases of the procedure (e.g., suture delivery and management for each of 3 anchors). Thus a total of 77 potential errors, 20 of which were sentinel errors, were specified for the complete procedure. In addition, events that led to less consequential "damage to nontarget tissues" (DNTT) (Table 1) were simply recorded as standard errors (e.g., scuffing of the articular cartilage). A perfect score would indicate that all 45 steps were completed satisfactorily without committing any errors.

There are many ways to perform a Bankart procedure, but the task deconstruction was designed for a "reference" procedure (a routine Bankart repair), breaking it down into essential components. Each of the metrics was specifically crafted to accommodate different methods that can be used to accomplish the steps—for example, suture passage could be performed with a number of different instruments and techniques, but to accomplish re-tensioning of the capsulolabral tissue, the capsule must be purchased inferior to the anchor site (one of the metrics). The modified Delphi Panel procedure used to obtain face and content validity asks the following question of each panel member: "Is this metric (step or error) acceptable as written?," that is, "It is not incorrect" (although a particular panel member might perform the step in a different manner). The 45 steps and 77 errors were drafted, revised, and stress tested by the core group of primary investigators (R.L.A., R.K.N.R., R.A.P., A.G.G.) and then submitted to the Delphi Panel for comment, modification, and revision. The panel then obtained consensus for all of the 122 metrics, which were found to be acceptable in their final form.

Cadaveric Shoulder Study Specimens

Fresh-frozen cadaveric specimens with a complete shoulder girdle from the scapula and associated soft tissues to the mid humerus were used. After appropriate thawing, the scapula was mounted with a clamp

Table 2. Thirteen Phases of Bankart Procedure (in Roman Numerals) and Brief Summary of 45 Steps of Procedure

I. Portals
1. Posterior portal established
2. View posterior humeral head and extent of the Hill-Sachs when present
3. Introduce mid-anterior spinal needle immediately superior to the subscapularis and direct it toward the anteroinferior glenoid and labrum
4. Establish a cannula that abuts the superior border of the subscapularis near the lateral subscapularis insertion
5. Demonstrate instrument access to the anteroinferior glenoid/labrum
6. Introduce anterosuperior spinal needle at the superolateral aspect of the rotator interval and direct it toward the anterior glenoid
7. Establish an anterosuperior cannula, arthroscopic sheath, or switching stick
II. Arthroscopic instability assessment
View from posterior portal
8. View or probe the superior labral attachment onto the glenoid
9. View or probe articular surface of the rotator cuff
10. Probe anteroinferior glenoid/Bankart pathology including rim fracture, articular defect
View from anterosuperior portal
11. View or probe the midsubstance of the anterior-inferior glenohumeral ligaments
12. View or probe the insertion of the anterior glenohumeral ligaments onto the anterior humeral neck
III. Capsulolabral mobilization/glenoid preparation
13. Elevate the capsulolabral tissue from the glenoid neck and articular margin
14. View the subscapularis muscle superficial to the mobilized capsule
15. With an instrument, grasp and perform an inferior to superior shift of the capsulolabral tissue (to show tension is restored)
16. Obtain a view of the anterior glenoid neck
17. Mechanically abrade the glenoid neck
IV. Inferior anchor preparation/insertion
18. Seat the guide for the most inferior anchor hole at the inferior region of the anteroinferior quadrant
19. Drill anchor hole oblique to the glenoid articular face
20. Insert anchor
21. Test the anchor security by pulling on the suture tails
V. Suture delivery/management
22. Pass a cannulated suture hook or suture retriever through the capsular tissue—inferior to the anchor
23. Pass anchor suture limb through the capsular tissue and deliver out the anterior cannula
VI. Knot tying
24. Deliver an arthroscopic sliding knot
25. Back up with 3 or 4 half-hitches
26. Cut suture tails
VII. Second anchor preparation/insertion
27. Seat the drill guide for the second anchor superior to the first anchor and inferior to the equator
28. Drill anchor hole oblique to the glenoid articular face
29. Insert suture anchor
30. Test anchor security by pulling on the suture tails
VIII. Suture delivery/management
31. Pass a cannulated suture hook or suture retriever through the capsular tissue inferior to the suture anchor
32. Pass anchor suture limb through the capsular tissue and deliver out the anterior cannula
IX. Knot tying
33. Deliver an arthroscopic sliding knot
34. Back up with 3 or 4 half-hitches
35. Cut suture tails
X. Third anchor preparation/insertion
36. Seat the drill guide for the third anchor at or superior to the equator
37. Drill anchor hole oblique to the glenoid articular face
38. Insert suture anchor
39. Test anchor security by pulling on suture tails
XI. Suture delivery/management
40. Pass a cannulated suture hook or suture retriever through the capsular tissue at or inferior to the suture anchor
41. Pass anchor suture limb through the capsular tissue and deliver out the anterior cannula
XII. Knot tying
42. Deliver an arthroscopic sliding knot
43. Back up with 3 or 4 half-hitches
44. Cut suture tails
XIII. Procedure review
45. View and/or probe final completed repair

Table 3. Summary of 29 Different Bankart Procedure Metric Errors

Failure to maintain intra-articular position of the posterior cannula
Failure to maintain intra-articular position of the mid-anterior cannula
Failure to maintain intra-articular position of the anterosuperior cannula
Damage to the superior border of the subscapularis during placement of the mid-anterior portal
Damage to the anterior border of the supraspinatus during placement of the anterosuperior portal
Loss of intra-articular position of arthroscope/sheath or operating cannula (loss of each portal is scored only once for each Roman numeral, i.e., up to a total of 3 for arthroscope + 2 portals)
Lacerate intact capsulolabral tissue (sentinel error)
Failure to maintain control of a working instrument (sentinel error)
Guide is not located in the inferior region of the anteroinferior quadrant of the glenoid for the most inferior anchor
Entry of the completed tunnel lies outside safe zone of 0 to 3 mm from the bony glenoid rim (sentinel error)
Shallow undermining and deformation of articular cartilage (sentinel error)
Failure to maintain secure seating of the drill guide during anchor insertion
Breakage of the implant
Implant remains visibly proud (sentinel error)
Failure to insert the anchor with the inserter laser line (when present) to or beyond the laser line on the drill guide
Anchor fails to remain securely fixed within bone at the appropriate depth
Capsular penetration is at or superior to anchor hole (sentinel error)
Capsular penetration is not at or peripheral to the capsulolabral junction
Instrument breakage
Tearing of capsulolabral tissue
Uncorrected entanglement of shuttling device or suture
Offloading of suture anchor
Breakage of suturing device
Failure to create and maintain indentation of the capsule or labral tissue on knot completion (sentinel error)
Visible void is present between throws of the completed primary knot (sentinel error)
Completed knot abuts articular cartilage
Visible void is present between throws of the complete half-hitches
Suture breakage
Guide is inferior to the equator of the glenoid (for the third and final anchor)

NOTE. Metric errors can be associated with multiple phases and steps of the procedure (77 total errors).

in the subject surgeon's orientation of preference (lateral decubitus *v* beach chair). The cadaveric specimens were considered acceptable if (1) arthroscopic visibility of the target tissues was obtainable; (2) the specimen (flexibility) permitted adequate access to the target tissues; and (3) the integrity of the capsulolabral tissues was sufficient to permit mobilization, suture delivery, and knot tying. One of 3 designated AANA shoulder arthroscopy master instructors (the surgeon members of the group who created the arthroscopic Bankart metrics; R.L.A., R.K.N.R., R.A.P.) determined the acceptability of the cadaveric specimens.

Arthroscopic Bankart Repair

During a single weekend AANA resident arthroscopy course, the surgeons from both groups were instructed to establish portals (posterior, anterosuperior, and mid-anterior), complete a thorough diagnostic arthroscopy, and perform a 3-anchor ABR on the cadaveric shoulder. Furthermore, they were instructed to demonstrate/complete all of the steps for the Bankart repair that they would normally perform in clinical practice on a real patient. Equipment representatives from multiple different vendors served as surgical assistants and were randomly assigned to participating surgeons. The assistants were instructed to act only at the specific direction of the operating surgeon. Prompting and coaching (of technique) were prohibited (the procedures were proctored by staff from the Orthopedic Learning Center). A standard equipment tower with a 30° arthroscope was provided along with all instruments necessary to complete an ABR (Table 4).

The subject surgeon identified bony landmarks with a marking pen, established their desired portals, and performed a diagnostic examination. The arthroscope was then withdrawn from the shoulder joint. One of the 3 Master surgeons who evaluated the cadaver acceptability according to the criteria noted earlier, then

Table 4. Arthroscopic Instruments Used to Perform Bankart Procedure on Cadaveric Shoulder

5.5- and 8.5-mm obturator cannulas
Switching sticks
Hook probe
Regular and looped graspers
Liberator/elevator
Shaver
Drill guide/drill
Push-in anchor loaded with single suture
Mallet
Cannulated suture hook
Penetrator
Monofilament suture
Knot pusher
Arthroscopic scissors

Table 5. Copernicus Cadaver - Novice

Video #	33A	33B	33 ave	73A	73B	73 ave	123A	123B	123 ave	53A	53B	53 ave	83A	83B	83 ave	113A	113B	113 ave	13A	13B	13 ave	93A	93B	93 ave	43A	43B	43 ave	63A	63B	63 ave	23A	23B	23 ave	103 A	103 B	103 ave
I - Portals																																				
Steps uncompl.	1	1	1	1	1	1	3	3	3	0	0	0	1	1	1	0	3	1.5	1	1	1	3	1	2	1	0	0.5	0	0	0	0	0	0	0	0	0
Errors made	2	2	2	0	0	0	1	1	1	0	3	1.5	0	0	0	2	1	1.5	1	0	0.5	1	0	0.5	0	1	0.5	0	0	0	0	2	1	0	0	0
II - Instabl Asses.																																				
Steps uncompl.	3	5	4	3	5	4	3	3	3	0	1	0.5	2	3	2.5	1	2	1.5	1	1	1	3	3	3	1	0	0.5	1	0	0.5	2	0	1	1	1	1
Errors made	0	0	0	0	0	0	1	0	0.5	1	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
III - Caps/Gen Prep																																				
Steps uncompl.	1	2	1.5	3	3	3	4	4	4	1	1	1	5	5	5	4	4	4	1	2	1.5	3	2	2.5	3	1	2	5	3	4	3	3	3	1	3	2
Errors made	0	1	0.5	0	0	0	0	0	0	1	1	1	0	1	0.5	0	0	0	0	3	1.5	1	1	1	1	0	0.5	2	1	1.5	0	0	0	0	0	0
IV - 1st Inf Anch Prep																																				
Steps uncompl.	0	0	0	1	1	1	1	0	0.5	0	0	0	1	0	0.5	1	1	1	1	0	0.5	0	0	0	0	0	0	0	1	0.5	1	0	0.5	0	0	0
Errors made	0	0	0	0	0	0	1	2	1.5	0	0	0	1	0	0.5	1	2	1.5	0	0	0	3	2	2.5	1	1	1	0	0	0	0	1	0.5	0	0	0
V - 1st Sut Del / Mgmt																																				
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	1	1	1	0	0	0	2	1	1.5	1	1	1	2	1	1.5	1	1	1	0	0	0	2	3	2.5	0	0	0	4	1	2.5	0	0	0	2	2	2
VI - 1st Knot Tying																																				
Steps uncompl.	0	0	0	0	0	0	Procedure terminated by surgeon			0	0	0	1	0	0.5	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	1	0	0.5				0	1	0.5	2	1	1.5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5
VII - 2nd Anch Prep																																				
Steps uncompl.	1	1	1	1	0	0.5				1	1	1	2	1	1.5	0	0	0	1	0	0.5	0	1	0.5	0	1	0.5	1	1	1	1	1	1	2	1	1.5
Errors made	0	0	0	0	0	0				1	2	1.5	2	2	2	0	0	0	0	2	1	1	2	1.5	0	0	0	0	0	0	0	0	0	1	2	1.5
VIII - 2nd Sut Del/ Mgmt																																				
Steps uncompl.	0	0	0	0	0	0				0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	0	0	0				0	0	0	1	3	2	0	1	0.5	0	0	0	3	2	2.5	0	0	0	2	0	1	0	1	0.5	0	0	0
IX - 2nd Knot Tying																																				
Steps uncompl.	0	0	0	0	0	0				0	0	0	3	2	2.5	3	3	3	1	1	1	1	0	0.5	0	0	0	0	1	0.5	0	0	0	0	0	0
Errors made	0	0	0	0	0	0				0	0	0	0	5	2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
X - 3rd Anch Prep																																				
Steps uncompl.	0	0	0	1	0	0.5				1	1	1	1	1	1	4	4	4	1	0	0.5	4	4	4	0	1	0.5	1	1	1	1	1	1	1	0	0.5
Errors made	0	0	0	0	0	0				0	0	0	1	2	1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XI -3rd Sut Del/ Mgmt																																				
Steps uncompl.	1	0	0.5	0	0	0				0	0	0	0	0	0	2	2	2	0	0	0	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	0	0	0				1	0	0.5	1	0	0.5	0	0	0	0	0	0	0	0	0	0	1	0.5	2	1	1.5	0	1	0.5	0	0	0
XII - 3rd Knot Tying																																				
Steps uncompl.	0	0	0	0	0	0				0	0	0	0	0	0	3	3	3	0	0	0	3	3	3	0	0	0	1	1	1	0	0	0	0	0	0
Errors made	0	1	0.5	0	0	0				0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1.5	1	1	1	1	1	1	0	1	0.5
XIII - Eval Repair																																				
Steps uncompl.	0	0	0	0	0	0				0	0	0	0	0	0	2	2	2	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	0	0	0				0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Portal / Dx Time (B4 Bank)			6			13			26			14			17			36			5			19			9			34			23			7
Bankart Repair Time			50			59			15			45			75			35			31			68			47			39			32			33
Total Time (Dx+Rx)			56			72			41			59			92			71			36			87			56			73			55			40
									DNC																												

(continued)

Table 5. Continued

Video #	33A	33B	33 ave	73A	73B	73 ave	123A	123B	123 ave	53A	53B	53 ave	83A	83B	83 ave	113A	113B	113 ave	13A	13B	13 ave	93A	93B	93 ave	43A	43B	43 ave	63A	63B	63 ave	23A	23B	23 ave	103 A	103 B	103 ave
Rating Pairs																																				
Steps completed (45)	38	37	37.5	35	35	35	*12*	*14*	*13*	41	42	41.5	30	32	31	26	22	24	38	39	38.5	19	25	22	40	42	41	36	37	36.5	37	40	38.5	40	40	40
Errors made (77)	3	5	4	1	0	0.5	*5*	*4*	*4.5*	4	8	6	10	16	13	5	5	5	1	5	3	11	10	10.5	5	3	4	11	4	7.5	2	7	4.5	4	5	4.5
Sentinel errors	1	3	2	0	0	0	*1*	*1*	*1*	1	2	1.5	3	4	3.5	1	1	1	0	1	0.5	2	1	1.5	2	0	1	5	3	4	0	1	0.5	1	1	1
Steps completed																																				
Agreement	39			41						44			40			41			40			37			39			40			42			40		
Disagreement	6			4						1			5			4			5			8			6			5			3			5		
Step IRR	0.87			0.91						0.98			0.89			0.91			0.89			0.82			0.87			0.89			0.93			0.89		
Errors made																																				
Agreement	75			76						68			59			69			71			68			73			70			72			73		
Disagreement	2			1						9			18			8			6			9			4			7			5			4		
Error IRR	0.97			0.99						0.88			0.77			0.9			0.92			0.88			0.95			0.91			0.93			0.95		
Total score (S+E)																																				
Agreement	114			117						112			99			110			113			105			112			110			114			113		
Disagreement	8			5						10			23			12			9			17			10			12			8			9		
Total IRR	0.93			0.96						0.92			0.81			0.9			0.93			0.86			0.92			0.90			0.93			0.93		

NOTE. Data in italics indicate incomplete procedure.

anch, anchor; caps, capsule; DNC, did not complete; Dx, diagnostic; E, errors; eval, evaluate; inf, inferior; instabl, instability; IRR, inter-rater reliability; mgmt, management; Rx, treatment; S, steps; sut del, suture delivery.

reintroduced the arthroscope and created a standardized Bankart lesion from the 2- to 6-o'clock position and 6 to 9 mm deep (medial). While the lesion was clearly delineated, the capsulolabral tissues were not mobilized to any additional extent. Before the study, we attempted to use different techniques to create an anteroinferior capsular detachment from the glenoid (Bankart lesion) in a cadaveric shoulder specimen (i.e., a long-handled scalpel blade, a hook-tip cautery, and manual dissection with an elevator). It was determined that the most consistent lesion could be created using a liberator/elevator along with a mallet to provide a gentle, controlled impact force to the elevator to effectively "sculpt" the Bankart pathology. This method optimized preservation of the integrity of the capsulolabral tissues for subsequent repair.

Once the Bankart lesion was completed, the arthroscope was withdrawn and reintroduced by the subject surgeon, who operated for the duration of the procedure. A continuous video recording was made, beginning with the first arthroscopic view of the joint from the posterior portal and ending with the final examination of the completed procedure by the surgeon. In calculating the total operating time of the Bankart repair procedure for the subject surgeon, the segment of time required by the master faculty surgeon to create the Bankart pathology was subtracted from the total absolute running time. No time limit was imposed on the performance of the procedure in the cadaveric specimen.

Video Reviewer Training

Once the construction of the metrics for an ABR was completed and face and content validity verified,[7] a final version of a score sheet was formatted. Ten AANA master/associate master faculty surgeons (none belonging to the experienced group from this study) formed the panel of reviewers designated to score the videos. This group included the 3 members (R.L.A., R.K.N.R., R.A.P.) of the group who, in conjunction with a consultant experimental psychologist (A.G.G.), created the arthroscopic Bankart metric "definitions" (Table 1). The 10 reviewers were randomly assigned to form 5 fixed pairs, which remained constant throughout the scoring of all videos. Reviewer training was initiated with an 8-hour in-person meeting, during which time each metric was studied in detail. Multiple video examples of live patient cases were shown to illustrate each particular metric. Videos of the patients in both the lateral decubitus and beach-chair orientations were represented. Discussion helped to clarify how each step and error were to be scored, including the nuances and conventions to be used. Several weeks later, full-length practice videos 1 and 2 (one each in the lateral decubitus and beach-chair orientation) were sent to and independently

Table 6. Copernicus Cadaver - Experienced

Video #	12A	12B	12 ave	22A	22B	22 ave	112A	112B	112 ave	122A	122B	122 ave	32A	32B	32 ave	42A	42B	42 ave	52A	52B	52 ave	62A	62B	62 ave	82A	82b	82 ave	92A	92B	92 ave
I - Portals																														
Steps uncompl.	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	0.5	2	2	2	0	0	0	2	0	1
Errors made	1	1	1	1	2	1.5	1	1	1	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
II - Instabl Asses.																														
Steps uncompl.	1	3	2	0	2	1	1	1	1	0	1	0.5	1	2	1.5	3	3	3	1	1	1	2	0	1	1	1	1	2	0	1
Errors made	0	0	0	0	0	0	1	0	0.5	0	0	0	0	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0
III - Caps/Gen Prep																														
Steps uncompl.	2	3	2.5	4	3	3.5	2	1	1.5	0	1	0.5	1	1	1	2	2	2	0	1	0.5	1	1	1	3	0	1.5	4	3	3.5
Errors made	0	0	0	0	1	0.5	2	3	2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IV - 1st Inf Anch Prep																														
Steps uncompl.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0.5
Errors made	0	0	0	0	0	0	0	0	0	2	1	1.5	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0.5	0	1	0.5
V - 1st Sut Del/Mgmt																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	0	0	0	1	0	0.5	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5
VI - 1st Knot Tying																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
Errors made	0	1	0.5	0	0	0	0	1	0.5	0	0	0	1	2	1.5	0	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0
VII - 2nd Anch Prep																														
Steps uncompl.	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1	1	0	0.5	1	1	1	0	1	0.5	1	1	1
Errors made	0	0	0	2	1	1.5	0	0	0	0	0	0	0	0	0	0	1	0.5	0	2	1	1	1	1	0	0	0	0	0	0
VIII - 2nd Sut Del/Mgmt																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	1	0.5	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.5
IX - 2nd Knot Tying																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1
Errors made	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
X - 3rd Anch Prep																														
Steps uncompl.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0.5	1	1	1	1	1	1
Errors made	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	1	0.5	0	0	0	0	0	0	1	1	1	0	0	0
XI -3rd Sut Del/Mgmt																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
XII - 3rd Knot Tying																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	0	0	0	0	0	0	1	1	1	2	2	2
Errors made	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
XIII - Eval Repair																														
Steps uncompl.	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Errors made	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Portal/Dx Time (B4 Bank)			8			9			5			6			7			2			18			1			6			3
Bankart Repair Time			36			30			26			28			63			17			24			28			27			23
Total Time (Dx+Rx)			44			39			31			34			70			19			42			29			33			26
Rating Pairs																														

(continued)

Table 6. Continued

Video #	12A	12B	12 ave	22A	22B	22 ave	112A	112B	112 ave	122A	122B	122 ave	32A	32B	32 ave	42A	42B	42 ave	52A	52B	52 ave	62A	62B	62 ave	82A	82b	82 ave	92A	92B	92 ave
Steps completed (45)	37	35	36	39	38	38.5	40	41	40.5	41	40	40.5	36	37	36.5	36	36	36	43	42	42.5	38	41	39.5	38	38	38.5	30	36	33
Errors made (77)	1	4	2.5	4	5	4.5	5	5	5	2	1	1.5	3	9	6	3	6	4.5	0	3	1.5	1	1	1	1	2	1.5	1	2	1.5
Sentinel errors	0	1	0.5	1	1	1	3	2	2.5	1	0	0.5	0	2	1	1	3	2	0	0	0	1	1	1	0	0	0	1	1	1
Steps completed																														
Agreement		40			42			44			42			41			45			42			43			38			39	
Disagreement		5			3			1			3			4			0			3			2			7			6	
Step IRR		0.89			0.93			0.98			0.93			0.91			1			0.93			0.95			0.84			0.87	
Errors made																														
Agreement		73			74			71			76			69			74			74			77			76			74	
Disagreement		4			3			6			1			8			3			3			0			1			3	
Error IRR		0.95			0.96			0.92			0.99			0.9			0.96			0.96			1			0.99			0.96	
Total score (S+E)																														
Agreement		113			116			115			118			110			119			116			120			114			113	
Disagreement		9			6			7			4			12			3			6			2			8			9	
Total IRR		0.93			0.95			0.94			0.97			0.9			0.97			0.95			0.98			0.93			0.93	

anch, anchor; caps, capsule; Dx, diagnostic; E, errors; eval, evaluate; inf, inferior; instabl, instability; IRR, inter-rater reliability; mgmt, management; Rx, treatment; S, steps; sut del, suture delivery.

scored by each of the 10 reviewers, and the scores were tabulated. During 2 subsequent 2-hour group phone conferences, the differences and discrepancies among all reviewers were compared and discussed, seeking conformity in scoring. In addition, each designated pair of reviewers conducted 1 to 3 additional phone conferences to analyze the specific instances in which the 2 of them scored particular events differently. Subsequently, all reviewers scored practice videos 3 and 4, and the results were again tabulated (each patient orientation was again represented). The scores for each of the 5 designated pairs of reviewers were compared for the second set of practice videos. In only 1 of 10 comparisons (2 videos × 5 reviewer pairs) did the inter-rater reliability (IRR) (Table 1) calculation (as described later) fall below an acceptable level of 0.8,[14] with an IRR value of 0.76.

Video Scoring

The AANA research coordinator randomly assigned each of the 22 full-length study videos of the experienced and novice surgeons performing an ABR on a cadaveric shoulder to a single pair of reviewers. Other than the research coordinator and the study consultant, the primary investigators and all video reviewers remained blinded to the source of the video being reviewed. Each of the 22 videos was independently reviewed and scored by the 2 members of an assigned pair of reviewers. Each step and error metric was scored as either yes or no, designating whether the specific event was or was not observed to have occurred by the reviewer. In addition to scoring steps and errors, each event characterized as DNTT was scored. There was no limit to the number of individual instances DNTT could be scored, with each occurrence simply tallied as a single error event. The score sheet also contained a box for specific reviewer comments for each metric. The 2 individual scores from a pair of reviewers were averaged to obtain the overall score for each step, error, or DNTT event. Reviewer scores are tabulated in Table 5 (novice) and Table 6 (experienced). The number of uncompleted steps are listed for each phase rather than completed steps because they provided a more meaningful assessment of performance. Each phase was tabulated separately to determine which phases best discriminated between novice and experienced surgeon performance. The total time in minutes taken by the participant to perform the diagnostic and procedural components was documented for each video, beginning with the first view of the arthroscope from the posterior portal and ending with the withdrawal of the arthroscope after examination of the completed review. The time required by the designated faculty member to create the Bankart lesion was subtracted from the total running time of the video. The total number of steps, errors and sentinel errors are listed for the entire

procedure. In addition, the score agreement or disagreement between the specific pair of reviewers was tabulated seperately for the steps, errors, and total of steps and errors, and were used to calculate IRR correlations (as described in the "Statistical Methods" section).

Performance Benchmark

Prior research has used the mean performance (metric based) of a group of experienced or expert operators to objectively define proficiency.[15-19] To assess and ensure performance homogeneity among the group of experienced surgeons for establishment of an accurate benchmark, their performances were converted to *Z* scores. The standard score (more commonly referred to as a "*Z* score") is a very useful statistic because it (1) creates the ability to calculate the probability of a score occurring within a normal distribution and (2) enables one to compare more precisely the scores from 2 individuals on a standardized scale. It is then possible to objectively and transparently determine whether and precisely how much a given subject's score is above or below the mean of their peers. In the pre-study design phase of the project, a stipulation was made by the primary investigators to remove the data from analysis for an experienced surgeon performing more than 2 SDs from the mean of the experienced group for any of the 4 assessments: steps, errors, sentinel errors, and time. Any such performance by a participant in the experienced group would be deemed an "outlier," and the scores would be removed from the analysis so as not to skew the establishment of the reference benchmark. This stipulation applied to both the previously reported shoulder model construct validity study[13] and the cadaveric construct validity study presented in this report.

Incomplete Repairs

It was prospectively determined that if a surgeon did not substantially complete the 3-anchor Bankart repair, his or her partial scores would be removed from the analysis. This policy was established because if only a portion of a procedure was performed, it would not have been possible to accurately estimate or extrapolate how many errors, sentinel errors, or instances of DNTT the surgeon may have enacted had he or she completed the entire procedure. Furthermore, no estimate of total time for the procedure would be sensible.

Statistical Methods

For each of the 13 separate phases of the procedure, the numbers of "uncompleted steps" and "errors made" were tabulated and the scores for the 2 reviewers averaged (Tables 5 and 6). These data were used to determine which of the procedural phases showed the greatest differences in performance when comparing the experienced and novice surgeons (1-factor analysis of variance) (IBM SPSS statistical software program; IBM, Armonk, NY). Furthermore, for the entire procedure, the total numbers of steps completed, errors made, and sentinel errors enacted were also averaged for the pair of reviewers.

The 2 raw score sheets were compared for each of the individual steps (45 steps in total) and the number of "agreements" tabulated (either both reviewers documented that a step was performed or both scored the step as not being completed). In addition, the number of "disagreements" in scoring steps was tabulated (one of the reviewers indicated that the step had been completed and other scored that the step had not been completed). The IRR for the steps was calculated according to the following formula: Number of agreements/Number of agreements + Number of disagreements.

In a similar manner, there was either agreement or disagreement in the 2 scores for each of the potential errors (77 errors in total). The IRR for error scoring was calculated in the same manner as that for the steps. Finally, the IRR for scoring the entire procedure was calculated using both the step and error agreements/disagreements for the entire procedure (122 in total). An acceptable IRR is equal to or greater than 0.80.[14]

Results

Participants

Two groups were compared in their performance of an ABR on a cadaveric shoulder. The entire group of master and associate master instructors, serving as faculty for an AANA Resident Course, chose to participate and comprised the experienced group (n = 10). The faculty, all fellowship trained in arthroscopy or sports medicine, averaged over 16 years in clinical practice, with each having routine experience in performing arthroscopic shoulder techniques. All faculty members have been recognized nationally by the AANA for their talent and ability to teach and communicate shoulder arthroscopy skills to trainees. The novice group (n = 12) comprised 11 PGY 5 and 1 PGY 4 orthopaedic resident volunteers (from a total of 44 orthopaedic residents registered for the weekend course) who elected to participate in the study and perform an ABR on a cadaveric shoulder. These volunteers had previously registered for an AANA Resident Course with no prior knowledge of the Bankart repair assessment protocol. Other than their year in training, no information regarding their arthroscopic experience or surgical skill was obtained.

Cadaveric Specimens

One specimen for each of the study groups was rejected and replaced. One specimen for the novice group was large with very noncompliant tissues, substantially restricting the anterior working space and making the use of instruments difficult. One specimen from the experienced group was arthritic, which limited the ability to distract the humeral head from the glenoid; thus visualization and the use of instruments from the posterior portal were unacceptably restricted. Each of these 2 specimens was replaced by an acceptable fresh cadaveric shoulder.

IRR Assessments

The IRR calculations across each of the assessments were strong. Twenty-one videos could be scored completely. One novice completed only a single-anchor repair during the entire duration of the procedure, which provided incomplete data. The mean IRR for the paired scoring of the 21 videos for procedural steps was 0.91 (range, 0.82 to 1.00), and the mean IRR for errors including DNTT events was 0.93 (range, 0.77 to 1.00 [the value 0.77 for the error IRR calculation for 1 video was the single instance that fell below 0.80 among the 63 IRR calculations]). The mean IRR for the total of steps and errors was 0.93 (range, 0.81 to 0.98).

Outlier Performance

One novice subject completed only 13 of 45 steps and a 1-anchor repair before failing to progress and electively terminating the procedure. During that time, 4.5 errors and 1 sentinel error were created (the average of the pair of designated reviewers). Inclusion of the data for the relatively small number of errors enacted during the partial repair would bias and understate the average number of errors for the novice group. As a result, all scores for this outlier were removed. The number of steps could theoretically have been used because this number accurately reflected the number of steps that were actually completed, but we elected to use none of the data from this subject's limited repair.

Before the data for complete repairs were analyzed, score profiles were examined for significantly atypical performance in the experienced group. One subject in the experienced group took dramatically longer than the subject's colleagues to perform the procedure, primarily because of substantial difficulties with suture delivery and management. This subject required 63 minutes to complete the Bankart repair in comparison with the colleagues' mean of 29.5 minutes (SD, 12.6 minutes). For the experienced surgeon outlier, Z equaled 2.61 with an associated probability value of .005, which indicated that the difference in performance from this surgeon's peers was highly significant. Consistent with the prospectively established policy of removing an experienced subject's scores if his or her performance was more than 2 SDs from the group mean, this subject's data were removed from further statistical comparisons between the groups. Thus, the data for one subject from each of the two groups was removed from the comparative statistical analysis.

Experienced Group and Novice Group Comparisons

Comparisons were made separately for steps and errors for each of the 13 phases of the Bankart procedure, as well as the summary data for steps, errors, and sentinel errors (Table 5 shows data for the novice group and Table 6 shows data for the experienced group). The phases of anchor preparation/insertion, suture delivery and management, and knot tying were repeated for each of the 3 anchors. The 3 sets of data for the 3 similar phases (1 for each of 3 anchors) were combined. The novice surgeons made significantly more objectively scored overall procedure errors (Fig 1) than the experienced surgeons (5.68 errors for novice surgeons v 2.95 errors for experienced surgeons, $P = .013$). Not only did the novice surgeons make more errors but they also showed greater performance variability, as shown by the considerably larger standard deviation score (3.51 v 1.85). The greatest difference in the mean number of errors made occurred during the suture delivery and management phases of the procedure, which was statistically significant (1.95 errors for novice surgeons v 0.45 errors for experienced surgeons, $P = .024$) (Fig 1). The novice surgeons also made more sentinel errors than the experienced surgeons (1.5 v 0.95), but this difference was not statistically significant.

The most common errors and sentinel errors are shown in Table 7, with those errors common to all of the 3 anchors being summed. With respect to regular errors, failure to maintain intra-articular position of the cannulas was frequently observed for the novice group. Both groups experienced occasional instances of anchor pullout, the experienced group somewhat more often than the novice group. By far the most common sentinel error enacted by the novice group was improper introduction of the suture-delivery device into the capsule at or above the anchor hole, resulting in failure to achieve retentioning of the capsuloligamentous tissues superiorly. Damage (laceration) of the intact labrum during attempts to mobilize the capsulolabral tissues was also notably more common among the novice group compared with the experienced group. Overall, the novice surgeons also completed fewer steps than the experienced surgeons (35.04 v 38.15), but this difference was not statistically significant ($P = .187$).

Figure 2 shows the mean amount of time both groups of subjects took to perform the procedure. The novice

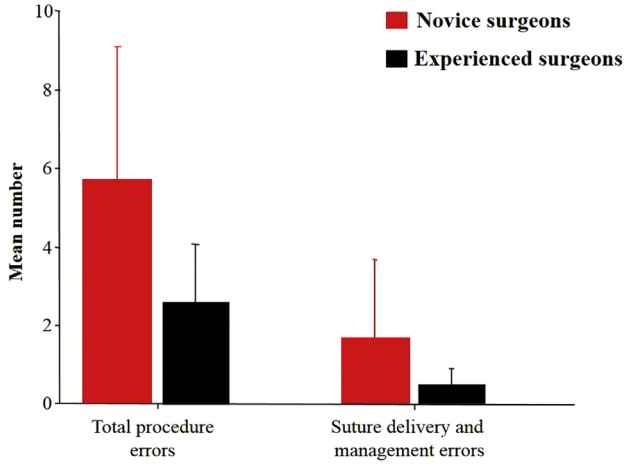

Fig 1. Mean total errors enacted by novice and experienced surgeon groups ($P = .013$) and mean suture delivery and management errors enacted by novice and experienced surgeon groups ($P = .024$).

surgeons took significantly more time to perform the repair than the experienced group (45.5 minutes [SD, 14.95 minutes] for novice surgeons *v* 25.9 minutes [SD, 5.33 minutes] for experienced surgeons; $P < .001$).

Discussion

Novice Versus Experienced Surgeon Performance

This study shows robust construct validity for the use of the arthroscopic Bankart procedure metrics with a cadaveric shoulder. The Bankart metrics are both precise (high IRR) and accurate (able to distinguish between novice and experienced surgeon performance).[20] Overall, experienced surgeons performed better than novice orthopaedic surgeons when evaluated using an objectively assessed and blinded review of video-recorded operative performance. Although the objectively assessed performance of the experienced surgeons was better than that of the novice surgeons across all of the measures, the metrics that best distinguished the 2 groups were procedure errors, particularly suture management errors. Operative time was also significantly different. We are unaware of previous similar studies using a detailed metric-based assessment of a complete surgical procedure with which to compare and contrast our results.

Tool Development

At the outset of the series of investigations termed the AANA Copernicus Initiative (a paradigm shift from the apprenticeship model to one of PBP training), we sought to study the effectiveness of PBP training for surgical skills. This investigation required the development and validation of 3 separate, specific tools to conduct the analysis. The first component to be created was a "metric tool" (steps, errors, and sentinel errors) for a specific procedure (an ABR was selected). This metric tool was shown to have face and content validity.[7] Second, a "training tool" (a shoulder model simulator coupled with the ABR metrics) was shown to have construct validity for an ABR with the ability to distinguish between experienced and novice surgeon performance.[13] Lastly, the current study shows construct validity for the cadaveric shoulder (coupled with the ABR metrics) as a valid "assessment tool" for comparing the performance of different surgeons.

Table 7. Common Errors and Sentinel Errors for Novice and Experienced Subjects

	Frequency of Errors by Group	
	Novice	Experienced
Error		
Failure to maintain intra-articular position of the posterior cannula	5	1
Failure to maintain intra-articular position of the mid-anterior cannula	4	0
Failure to maintain intra-articular position of the anterosuperior cannula	1	2
Elevate capsulolabral tissue from the glenoid neck	5	0
Anchor fails to remain securely fixed at appropriate depth*	4	6
Offloading of suture anchor	1	0
Sentinel error		
Lacerate intact capsulolabral tissue	4	1
Capsular penetration is at or superior to anchor hole*	13	3
Entry of completed tunnel lies outside safe zone of 0-3 mm from bony glenoid rim	1	1
Failure to create and maintain indentation of capsule or labral tissue	2	1

*Errors that were common to the 3 anchors were summed.

Inter-rater Reliability

The very high IRR for the scores from the reviewer pairs for the entire group of metrics (0.92) is reflective of the clarity and precision of the arthroscopic Bankart metrics drafted, as well as the thorough training of the 10 reviewers. The ability to score the steps and errors consistently is essential to obtain a reliable measure of the surgeon's performance and skill level for a particular procedure.

Shoulder Simulator Model Versus Cadaveric Shoulder

For the prior study undertaken to assess construct validity for the shoulder model simulator and Bankart metrics, surgeons in the experienced group made 63% fewer errors, committed 79% fewer sentinel errors, and performed the procedure in 42% less time than those in

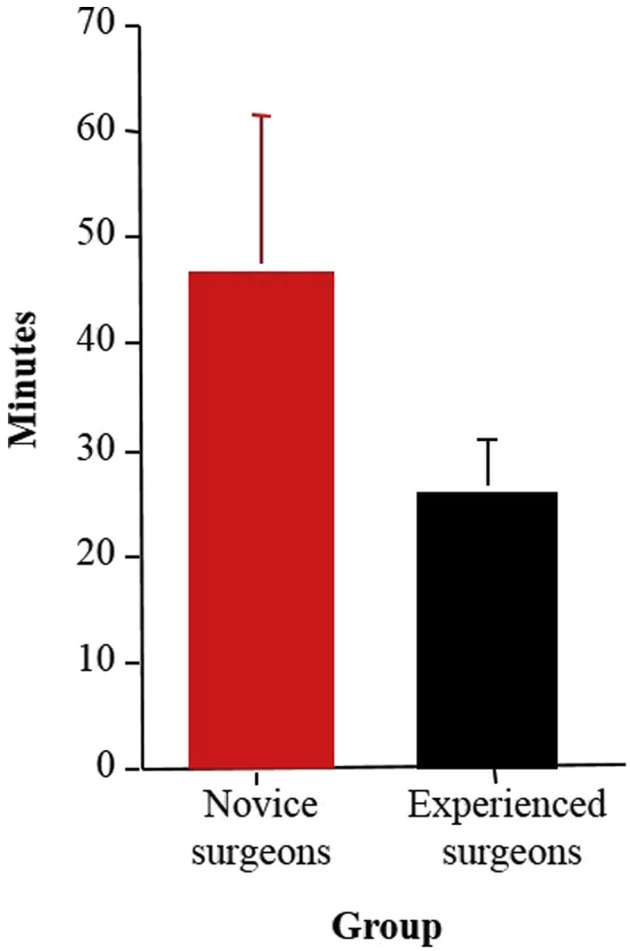

Fig 2. Mean time (in minutes) taken by novice and experienced surgeon groups to complete 3 suture anchor Bankart repair ($P < .001$).

the novice group (all differences being significant). The greatest difference in errors between the groups involved anchor preparation and insertion, suture delivery and management, and knot tying. In the current study using a cadaveric shoulder, experienced surgeons made 50% fewer errors and performed the procedure in 44% less time (both differences being significant). The number of sentinel errors was significantly less for the experienced group in the model validation study, and sentinel errors were also committed less frequently by the experienced surgeons in this cadaveric evaluation, although the difference was not statistically significant. With respect to specific phases of the procedure, the greatest discriminator in both investigations was for the phases of suture delivery and management. This finding is not surprising because the steps involved in those phases are among the most challenging for the Bankart repair. The number of steps performed did not differentiate between the novice and experienced groups in either the simulator model or cadaver studies. This result is not unexpected because the intent and effort to perform each of the steps predominantly reflect a familiarity and knowledge of the steps necessary to perform the procedure. Overall, in both studies, the experienced group showed less performance variability than the novice surgeons as demonstrated by their smaller standard deviation scores.

The benchmark was established based on the mean performance of the group of experienced surgeons. For this cadaveric study, it included completion of a 3-anchor Bankart repair with no more than 3 total errors and no more than 1 sentinel error. For the similar previous study on the simulator model using the identical metrics, the one difference in the benchmark was that no more than 4 (instead of 3) total errors were permitted.

Novice and Experienced Outliers

The partial data from 1 novice surgeon were removed from the analysis because the surgeon was not able to complete the suturing and knot tying for the first anchor before electively terminating the procedure because of the inability to make progress. A relatively large number of errors were enacted (mean, 4.5) (Table 5) for the portion of the procedure performed, but it was not possible to accurately estimate or extrapolate to the total number of errors that might have been created had the entire procedure been completed. The total number of errors enacted would likely have been substantial had the procedure been completed. Thus the average total numbers of errors and sentinel errors are likely to have been substantially understated for the novice group. A relatively small number of steps were accomplished (mean, 13) (Table 5), and had the data for this novice been included in that analysis, they would have affected the average total number of steps that the novices completed. Given that the diagnostic portion of the procedure took over 25 minutes for this subject, the overall time for completion of the 3-anchor repair would also likely have been much longer. Because the data for the majority of the analysis were incomplete, it was not possible to include any of this surgeon's performance data in the analysis.

One of the issues that emerged during this study, and indeed in one of our previous studies,[13] was atypical performance of 1 experienced subject. Atypical performance is an important issue as it relates to establishing benchmarks, which may have considerable implications for trainee progression. Since PBP training was first introduced and validated in 2002,[15] the average, or mean, performance of experienced operators has been used as the performance benchmark that trainees must meet and demonstrate before being allowed to progress in their training.[15-19] If an experienced individual's performance score was dramatically worse than his or her peers' scores and if this score were to be included in the establishment of the

benchmark, the reference level would clearly be lowered. The lowering of the performance threshold (benchmark) could have important patient safety implications. For example, in a study on bariatric surgery, it was found that surgeons performing at the lower end of the performance range had significantly poorer outcomes than surgeons performing at the upper end.[21] This was one of the first studies to quantitatively link objectively assessed surgeon skill performance with patient outcomes.

The criteria for removing outlier data from the group being used to create a performance benchmark must be established before the data are collected and should be objective, transparent, and fair. At the outset, before conducting this study, we identified and discussed the possibility of encountering atypical experienced surgeon performance. The core group of 4 primary investigators agreed that the resolution to this potential issue would be to remove all of a subject's scores from subsequent analysis if it could be unambiguously established that the subject's performance was statistically atypical (>2.0 SDs from the group mean). The experienced individual participating in this study performed considerably worse than that for operative time (2.61 SDs from the mean).

Performance Errors

The enactment of errors is emerging as one of the most important indicators of skill for operative performance.[13] While an individual may be able to perform all of the correct steps in an acceptable order with the appropriate instruments and score very well on those parameters, he or she may still perform the steps poorly. Procedural errors are operative behaviors that deviate from optimal performance. These metrics are a reliable measure of performance quality and are likely to be the most sensitive assessment tool in the evaluation of operative performance and safety.[22] Although simulation-based education and the resulting transfer of training will influence other performance parameters as well, such as procedural time, the greatest impact of such a training strategy appears to be on limiting performance errors.[22] It is therefore a necessity, at the outset, that the performance characterization of "deviations from optimal performance" (errors) must be particularly robust and well validated. It also implies that error performance assessed using a less rigid Likert scale (Table 1) (global rating scale) may result in a less focused approach to minimizing errors because the deviations from optimal performance have been less clearly defined.[23,24] A Likert-type scale is a method of ascribing a quantitative value to qualitative data to make the data amenable to statistical analysis. Likert scales (often with a range from 1 to 5 or from 1 to 7) are typically constructed with responses (opinion) around a neutral option (e.g., "suture delivery was 1, awkward, . . . 3, effective, . . . 5, highly efficient") and were originally designed to assess a range of respondent attitudes.[25] Given the inherent subjectivity in this method of attempting to rate objective performance, it may be difficult to obtain acceptable levels of IRR (≥80%) in the scoring of events.[24] In contrast, the approach to the assessment of performance used in this study uses precise definitions of performance and simply requires the video reviewer to determine whether the specific event did or did not occur. This binary approach to the measurement of performance has been shown to facilitate the reliable scoring of metric-based performance units across a variety of functions from skills training[18,26-28] at different experience levels.[29,30] It has also been shown to considerably enhance assessment reliability levels in comparison with Likert-scale scoring.[23]

The effectiveness of a deliberate practice, proficiency-based training curriculum using simulation relies on a clear and specific identification not only of the proper steps that the trainee should perform but also of what the trainee is doing wrong and how to prevent or correct his or her error. Other advantages of creating comprehensive procedure characterizations and explicit operational step and error definitions exist. Detailed metrics provide very clear guidance for the construction of simulation training platforms, specifying exactly what the simulator should be capable of emulating and, more importantly, measuring.[31] Comprehensive procedure characterization is challenging and time-consuming the first time it is undertaken and requires robust validation of all of the performance metrics. With experience, however, this methodology is considerably easier to apply to subsequent characterizations of different procedures by the same group.

Limitations

A limitation of this study relates to the use of cadaveric specimens for the arthroscopic Bankart lesion creation and repair. The specimens lacked some uniformity in the integrity of the capsule and labrum, soft-tissue compliance, shoulder mobility/distractibility, and bulk of the extra-articular tissues. In addition, although specific parameters were used for the creation of the Bankart lesions (i.e., 2- to 6-o'clock position on the glenoid rim and 6 to 9 mm deep/medial), the lesions could not be made absolutely uniform. Further variability existed in the presence of coexisting pathology (arthritis, synovial proliferation, rotator cuff partial tears, and so on). The "acceptability criteria" for the specimens (listed earlier) were used to minimize the impact of this potential problem.

An additional limitation of this study is that there was no confirmation that those serving as master/associate master surgeons and being representative of the

"experienced" group possessed a specified level of expert skill in performing an arthroscopic Bankart procedure. Nevertheless, the individual surgeons so identified have been recognized by the AANA as valuable educators either from lecture presentations with videos exhibiting skilled shoulder arthroscopy techniques or from repeated experience teaching in an arthroscopic laboratory setting with the ability to demonstrate and teach each of the key components of a Bankart procedure. Thus, "experienced" rather than "expert" is a reasonable description of the group. Similarly, other than identifying the year in training, no additional information was obtained to determine the extent of the residents' experience (novice group) with arthroscopic shoulder surgery (i.e., the number of arthroscopy/sports medicine rotations previously completed, the number of shoulder arthroscopic surgical procedures in which they served as assistant surgeons, and so on). Even with these data, accurate knowledge of the level of skill possessed by an individual resident would not be possible. Thus, although the arthroscopic skill sets of the subjects are representative of their respective groups and experience, those skills are highly likely to be somewhat heterogeneous.

We acknowledge that only a single operative procedure was analyzed for each of the subject surgeons. It is possible that data averaged over several procedures would be somewhat different from those obtained in this study. Cost and time considerations made the performance of a single ABR on a cadaveric shoulder most feasible. Finally, it should be noted that the participants in each group had no prior specific knowledge of the metrics to be scored in the review of their procedure, and we suspect that the experienced surgeons, in particular, might have performed and scored differently (better) for certain non-crucial parts of the procedure (e.g., the diagnostic steps at the beginning) had they been familiar with the metrics to be evaluated.

Conclusions

The assessment tool composed of validated arthroscopic Bankart metrics coupled with a cadaveric shoulder accurately distinguishes the performance of experienced from novice orthopaedic surgeons. A benchmark based on the mean performance of the experienced group includes completion of a 3-anchor Bankart repair, while enacting no more than 3 total errors and 1 sentinel error.

References

1. Committee on Quality of Health Care in America. *To err is human: Building a safer health system*. Washington, DC: National Academies Press, 2000;196-197.
2. Smith R. All changed, changed utterly. *BMJ* 1998;316:1917-1918.
3. Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. *Am J Surg* 1999;177:28-32.
4. Barden CB, Specht MC, McCarter MD, et al. Effects of limited work hours on surgical training. *J Am Coll Surg* 2002;195:531-538.
5. Temple J. Time for training: A review of the impact of the European Working Time Directive on the quality of training. Available at http://hee.nhs.uk/healtheducationengland/files/2012/08/Time-for-training-report.pdf. Accessed March 2, 2014.
6. Bell RH Jr, Biester TW, Tabuenca A, et al. Operative experience of residents in US general surgery programs: A gap between expectation and experience. *Ann Surg* 2009;249:719-724.
7. Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG. Metric development for an arthroscopic Bankart procedure: Assessment of face and content validity. *Arthroscopy* 2015;31:1430-1440.
8. Morgan CD, Bodenstab AB. Arthroscopic Bankart suture repair: Technique and early results. *Arthroscopy* 2010;26:819-820.
9. Streubel PN, Krych AJ, Simone JP, et al. Anterior glenohumeral instability: A pathology based surgical treatment strategy. *J Am Acad Orthop Surg* 2014;22:283-294.
10. Waterman BR, Burns TC, McCriskin B, et al. Outcomes after Bankart repair in a military population: Predictors for surgical revision and long-term disability. *Arthroscopy* 2014;30:172-177.
11. Shibata H, Gotoh M, Mitsui Y, Kai Y, Nakamura H, et al. Risk factors for shoulder re-dislocation after arthroscopic Bankart repair. *J Orthop Surg Res* 2014;9:53.
12. Ryu RK. Arthroscopic approach to traumatic anterior shoulder instability. *Arthroscopy* 2003;19:94-101.
13. Angelo RL, Pedowitz RA, Ryu RKN, Gallagher AG. The Bankart performance metrics combined with a shoulder model simulator create a precise and accurate training tool for measuring surgeon skill. *Arthroscopy* 2015;31:1639-1654.
14. American Educational Research Association. Standards for educational and psychological testing. Available at www.apa.org/science/programs/testing/standards.aspx. Accessed January 12, 2014.
15. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Ann Surg* 2002;236:458-464.
16. Gallagher AG, Ritter EM, Champion H, et al. Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg* 2005;241:364-372.
17. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007;193:797-804.
18. Van Sickle K, Ritter EM, Baghai M, et al. Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. *J Am Coll Surg* 2008;207:560-568.

19. Gallagher AG, O'Sullivan GC. *Fundamentals of surgical simulation; principles & practices*. London: Springer Verlag, 2011.
20. Rossi MJ, Lubowitz JH, Provencher MT, Poehling GG. Precision versus accuracy: A case for common sense. *Arthroscopy* 2012;28:1043-1044.
21. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013;369:1434-1442.
22. Gallagher AG, Seymour NE, Jordan-Black JA, et al. Prospective, randomized assessment of transfer of training (ToT) and transfer effectiveness ratio (TER) of virtual reality simulation training for laparoscopic skill acquisition. *Ann Surg* 2013;257:1025-1031.
23. Gallagher AG, O'Sullivan GC, Leonard G, Bunting BP, McGlade KJ. Objective structured assessment of technical skills and checklist scales reliability compared for high stakes assessments. *ANZ J Surg* 2014;84:568-573.
24. Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: Rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003;17:1525-1529.
25. Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932;140:44-53.
26. McClusky DA, Gallagher A, Ritter EM, et al. Virtual reality training improves junior residents' operating room performance: Results of a prospective, randomized, double-blinded study of the complete laparoscopic cholecystectomy. *J Am Coll Surg* 2004;199:73 (abstract, suppl).
27. Van Sickle K, Smith B, McClusky DA, et al. Evaluation of a tensiometer to provide objective feedback in knot-tying performance. *Am Surg* 2005;71:1018-1023.
28. Van Sickle K, Gallagher AG, Smith CD. The effect of escalating feedback on the acquisition of psychomotor skills for laparoscopy. *Surg Endosc* 2007;21:220-224.
29. Neary PC, Boyle E, Delaney CP, et al. Construct validation of a novel hybrid virtual-reality simulator for training and assessing laparoscopic colectomy; results from the first course for experienced senior laparoscopic surgeons. *Surg Endosc* 2008;22:2301-2309.
30. Nicholson WJ, Cates CU, Patel AD, et al. Face and content validation of virtual reality simulation for carotid angiography: Results from the first 100 physicians attending the Emory NeuroAnatomy Carotid Training (ENACT) program. *Simul Healthc* 2006;1:147-150.
31. Gallagher AG. Metric-based simulation training to proficiency in medical education:- What it is and how to do it. *Ulster Med J* 2012;81:107-113.