Metric Development for an Arthroscopic Bankart Procedure: Assessment of Face and Content Validity

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 establish the metrics (operational definitions) necessary to characterize a reference arthroscopic Bankart procedure, and to seek consensus from experienced shoulder arthroscopists on the appropriateness of the steps, as well as errors identified. **Methods:** Three experienced arthroscopic shoulder surgeons and an experimental psychologist (comprising the Metrics Group) deconstructed an arthroscopic Bankart procedure. Fourteen full-length videos were analyzed to identify the essential steps and potential errors. Sentinel (i.e., more serious) errors were defined as either (1) potentially jeopardizing the procedure outcome or (2) creating iatrogenic damage to the shoulder. The metrics were stress tested for clarity and the ability to be scored in binary fashion during a video review as either occurring or not occurring. The metrics were subjected to analysis by a panel of 27 experienced arthroscopic shoulder surgeons to obtain face and content validity using a modified Delphi Panel methodology (consensus opinion of experienced surgeons rendered by cyclical deliberations). **Results:** Forty-five steps and 13 phases characterizing an arthroscopic Bankart procedure were identified. Seventy-seven procedural errors were specified, with 20 designated as sentinel errors. The modified Delphi Panel deliberation created the following changes: 2 metrics were deleted, 1 was added, and 5 were modified. Consensus on the resulting Bankart metrics was obtained and face and content validity verified. **Conclusions:** This study confirms that a core group of experienced arthroscopic surgeons is able to perform task deconstruction of an arthroscopic Bankart repair and create unambiguous step and error definitions (metrics) that accurately characterize the essential components of the procedure. Analysis and revision by a larger panel of experienced arthroscopists were able to validate the Bankart metrics. **Clinical Relevance:** The ability to perform task deconstruction and validate the resulting metrics will play a key role in improving surgical skills training and assessing trainee progression toward proficiency.

The intent of any surgical training program, both for residents and for established surgeons acquiring a new procedural skill, is to enable the trainee to acquire the requisite skill sets necessary to perform the designated surgical procedure well and safely. To accomplish this mission, a clearly defined endpoint or set of skill proficiencies must be identified. Furthermore, it must be verified that mastery of those skill sets can accurately be measured during the trainee's progress. It must also be confirmed that the acquisition of those skills is predictive of the ability to perform an effective surgical procedure. Many experienced surgeons who are "proficient" in the performance of a specific procedure and are able to perform it well are also able to identify and agree on the essential "steps" to be completed, as well as the "errors" to be avoided, for that procedure. The reader is referred to Table 1 for a glossary of terms used throughout the article. One challenge, however, in identifying those key features is that surgeons rarely think about the procedures they perform with this level of detail. Surgeons who are proficient in the performance of a specific operation will exhibit many if not all of the important "performance characteristics" (Table 1) that contribute to actually performing the procedure well. They may, however, have automated to many of these steps and how they are performed and, as a consequence, may be less cognizant of the details and more granular elements of the techniques they use.[1-3]

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 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 and Mitek.

Received December 26, 2013; accepted April 15, 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

0749-8063/13907/$36.00
http://dx.doi.org/10.1016/j.arthro.2015.04.093

Table 1. Glossary

	Definition
Behavioral scientist	A professional who engages in any discipline concerned specifically with the subject of human actions and behavior
Concurrent validity	A type of evidence in which there is a positive relation between the test scores from one instrument and the scores from another instrument purporting to measure the same construct
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 (by expert/experienced opinion) of the validity of a testing instrument based on a detailed examination of the contents of the test items
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 an experienced panel; in the modified form, the members of the panel answer queries/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
Description	A qualitative characterization of certain or salient aspects or features of an event
Error	A deviation from optimal performance
Face validity	An estimate by an experienced panel that reviews the content of an assessment or tool to see if it seems appropriate and relevant to the concept it purports to measure
Iterative process	A process for calculating or progressing toward a desired result by means of repeated cycles of operations (deliberations); an iterative process should be convergent, that is, it should come closer to the desired result as the number of iterations increases
Likert-like scale	A method of ascribing a quantitative value to qualitative data to make them amenable to statistical analysis
Metric	A standard of measurement of quantitative assessments used for objective evaluations to make comparisons or to track performance
Metric stress testing	A method for determining how specific metric definitions fare during their application and use in scoring in vivo or video-recorded performances
Metric unit	A method of measurement in which the basic parts or components are discrete performance elements
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 characteristics	The features determining the accomplishment of a given task measured against preset known standards of accuracy and completeness
Predictive validity	A type of evidence that determines the extent to which the scores on a test are predictive of actual performance
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
Reference procedure	A straightforward operative procedure; an agreed on/accepted approach to the performance of an uncomplicated surgical procedure
Reliability of identification (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
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 a 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

The units of performance that constitute skill can be elucidated with a "task analysis" (Table 1), or breakdown and detailed description of the steps or actions necessary to perform the procedure. In attempting to characterize specific skills, psychologists have subjected them to a detailed task analysis and then "operationally defined" (Table 1), rather than simply described, the resulting steps. A definition specifies the order, duration, and result of the specific action and provides precise parameters such that it can be unambiguously determined whether that specific event did or did not occur. A "description" (Table 1), on the other hand, only offers a general characterization of an event or behavior in qualitative terms. Definitions are the preferred foundations of measurement science. The definitions, or "metrics" (Table 1), for a specific procedure provide a quantitative standard of measurement that can be used to objectively assess performance. These metrics must then be validated with respect to whether their characterization fits with what is known about the skill being analyzed. The task analysis—derived characterizations, or "metric units" (Table 1), of skilled performance do not have to capture every aspect of performance but should at least allow for ordinal differentiation between different levels of performance as described by Dreyfus and Dreyfus.[4] The metrics created from this analysis can serve as a tool to evaluate the effectiveness of different training protocols for a particular surgical procedure.

"Face validity" (Table 1) is verified by the opinion of an experienced panel that reviews the content of an assessment or tool to determine if it is appropriate and relevant to the concept it purports to measure. "Content validity" (Table 1) of a testing instrument is similarly obtained and based on the opinion of an experienced panel that performs a detailed examination of the contents of the test items. Thus the face validity and content validity of tools assessing procedural skill are not verified by statistical analysis but, rather, by the summary opinion of an experienced panel of surgeons. An additional question that relates to establishing the validity of the metric definitions for a particular procedure asks, "Do more skilled individuals perform better on the defined metrics than less skilled or experienced individuals, and do the specific metrics identify the quality, ability, or trait they were designed to measure ('construct validity' [Table 1])"? In contrast to face and content validity, the establishment of construct validity requires sufficient data and statistical analysis to prove that it exists.

Task analysis for a particular operation should be performed initially for a "reference procedure" (Table 1)[5-7]—one that is straightforward with a generally accepted or agreed on method that is uncomplicated under ideal circumstances. An optimal approach to learning should ensure that trainees are capable of performing a routine procedure before they have to deal with the technique variations necessary to address more complex pathology.

We sought to study the effectiveness of "proficiency-based progression" (Table 1) training plus simulation for the acquisition of surgical skills. Proficiency-based progression dictates that the learner must demonstrate the ability to meet specific performance benchmarks before he or she is permitted to progress in training. This investigation required the development and validation of specific tools to conduct the analysis. The first component needed was a "metric tool" (Table 1) that could objectively and accurately characterize an arthroscopic Bankart repair. The development of this tool is the focus of this study. Future investigations will report on the establishment of additional tools.

The purpose of this study was to establish the metrics (operational definitions) necessary to characterize a reference arthroscopic Bankart procedure and to seek consensus from experienced shoulder arthroscopists on the appropriateness of the steps, as well as errors identified. The null hypothesis was that face and content validity for the step and error metrics derived from task deconstruction of an arthroscopic Bankart procedure would not be demonstrated.

Methods

Arthroscopic Bankart Metric Development

Three experienced arthroscopic shoulder surgeons (R.L.A., R.K.N.R., R.A.P.), each with over 25 years of clinical practice, and an experimental psychologist (A.G.G.) formed the Metrics Group that characterized an arthroscopic Bankart repair. A detailed task analysis and deconstruction process (described in detail elsewhere)[6] was used to identify the units of performance that are integral to the skilled performance of the instability repair. The goal was to characterize a "reference" arthroscopic Bankart repair and not one attempting to manage unusual or complex instability pathology. Procedure performance characterization (task deconstruction) was guided by (1) decades of practice and teaching experience by the Metrics Group, (2) published studies on arthroscopic Bankart repair,[8-10] and (3) manufacturer guidelines on device use. Two 2.5-day face-to-face meetings and eight 1.5- to 2-hour online conferences were conducted, along with countless E-mail exchanges, to craft the procedural metrics. For the online sessions, the use of Skype videoconferencing (Microsoft, Redmond, WA; available at www.skype.com) enabled the investigators (R.L.A., R.K.N.R., R.A.P., A.G.G.; who reside in different geographic locations) to simultaneously review arthroscopic videos in real time with acceptable resolution. One investigator initiated a standard Skype video connection for a group call using a laptop computer. A second computer (desktop) with a high-resolution screen was used to play the arthroscopic video being studied. An independent USB camera (Ipevo, Sunnyvale, CA) was connected to the USB port of the laptop to which the Skype video input was directed, instead of the resident camera on the laptop screen ("settings" tab in Skype). Thus all of the members on the group Skype call viewed the arthroscopic image rather than the call initiator's image.

Fourteen video recordings of a complete in vivo Bankart procedure, performed by surgeons with varying levels of experience (Table 2), were reviewed by the Metrics Group in detail to assist in the creation and stress testing of the metrics. The videos represented surgeons with practice experience ranging from 3 to 33 years. Both the lateral decubitus (n = 10) and

Table 2. Source Videos for Bankart Metric Creation and Stress Testing

Video No.	Surgeon Time in Practice, yr	Patient Orientation
1	25	LD
2	17	LD
3	25	LD
4	26	LD
5	25	LD
6	17	BC
7	18	BC
8	26	LD
9	28	LD
10	21	LD
11	24	LD
12	25	LD
13	3	BC
14	4	BC

BC, beach chair; LD, lateral decubitus.

beach-chair (n = 4) orientations of the patients were represented. All metrics were constructed to be applicable to and able to be scored for surgical procedures performed with patients in both the lateral decubitus and beach-chair orientations. During the series of video reviews, each metric unit was identified and the definition refined so that it could be unambiguously scored as either occurring or not occurring, with a high degree of reliability, by an independent group of raters. Each step was further defined by identifying beginning points and endpoints during the procedure for that metric. The aim was that these detailed metric units would accurately capture the essence of procedure performance, as well as serve as a sound and comprehensive training guide for persons learning the procedure. The metrics included the specific operative steps, the general order in which they should be accomplished, and the instruments and the manner in which they should be used. "Procedural phases" (Table 1) were specified for groups of related steps. In addition to specifying each procedural step, metrics were also created to identify potential "errors" (Table 1), or actions that deviate from optimal performance and should not be done.[11] The intent, again, was to create unambiguous operational definitions (rather than descriptions) for each metric error. A special designation was made for more serious, or "sentinel," errors defined by events that, by themselves, could either (1) jeopardize the outcome of the procedure or (2) lead to significant iatrogenic damage to the shoulder joint. An additional error characterization was termed "damage to non-target tissue." This occurrence defined an event that was injurious to tissues not intentionally being addressed during the defined task, such as "scuffing of articular cartilage by an instrument" or "lacerating the intact labrum."

By agreed on convention, an event (step or error) must have been observed on the video to be scored. Thus inference that an event was "likely to have occurred" was eliminated. For example, if comparable views of the anterior humeral head showed relatively healthy or pristine articular cartilage early in the procedure, with scuffing and abrasion later during the repair, but the injurious event was not observed on the video, it was not scored as an error (or damage to non-target tissue).

Metric Stress Testing and Reliability of Identification

After the 4 members of the Metrics Group were satisfied that the entirety of the procedure had been well characterized, they "stress tested" (Table 1) the metrics by subjecting them to a robust assessment of how reliably they could be independently scored in blinded fashion. Eight video recordings of complete arthroscopic Bankart procedures that were performed by surgeons possessing a wide range of technical skill were independently reviewed and scored. Both the lateral decubitus and beach-chair orientations were represented by the videos studied. Each metric was scored in binary fashion as either yes or no (occurring or not occurring). After each video review, differences in the scoring of each metric by the reviewers (R.L.A., R.K.N.R., R.A.P.) were compared and discussed. Where necessary, operational definitions were clarified, modified, or dropped and new ones added to optimize the functionality of the characterizations as a whole. This process of independent viewing, scoring, and revising the step and error metrics was continued until the Metrics Group was satisfied that the metrics accurately and unambiguously characterized the specifics of an arthroscopic Bankart procedure and could be "reliably identified" (Table 1) by independent reviewers. The extent of agreement between 2 raters for the entire group of step and error Bankart metrics could potentially range between 0, no agreement, and 1.0, complete agreement.

Face and Content Validation of Bankart Metrics by Modified Delphi Panel

The Delphi Panel method[12] (Table 1) is a process that provides an interactive communication structure between researchers (i.e., the Metric Group authors) and an experienced panel (as described later) in a field or discipline to provide systematic feedback on a given topic (i.e., the accuracy of the metrics developed for a reference approach to a Bankart procedure). The Delphi method uses an "iterative process" (Table 1) for progressing toward a desired result by means of repeated cycles of deliberations. The iterative process should be convergent, that is, it should come closer to the desired result as the number of iterations or cycles of review increases. For the Bankart characterization, the desired result (consensus on the appropriateness of a particular metric) was obtained by means of repeated cycles of questioning, deliberation, metric modification, and voting on the appropriateness of each refined metric definition. The methodology assumes that good-quality knowledge evolves from the process. The Delphi method was modified to the extent that the voting cycles, with each new iteration, were not anonymous.

The determination of face and content validity for the Bankart characterization was made by subjecting each metric to an appraisal by a group of surgeons who were very experienced in the performance of an arthroscopic Bankart repair. Twenty-seven board-certified orthopaedic surgeons (the 3 Metrics Group surgeons and 24 additional Arthroscopy Association of North America shoulder faculty instructors) with an average of over 23 years in clinical practice involving shoulder arthroscopy attended a Delphi Panel. Four of the panelists are full-time academicians, 9 are in private group practice and have direct involvement in teaching fellows, and 14 are in private practice with a clinical affiliation with a university orthopaedic department. Each member of the Delphi Panel is a master or associate master faculty member of the Arthroscopy Association of North America and has taught the technique for arthroscopic Bankart repair during shoulder courses conducted at the Orthopedic Learning Center (Rosemont, IL). An experimental psychologist facilitated the meeting.

Delphi Panel Procedure

An overview of the project and meeting objectives was presented. Background information regarding proficiency-based progression training, prior literature showing the validity of this training approach for procedural specialties, and the specific objectives of the current Delphi Panel[12] were reviewed. It was explained that the Bankart metrics had been developed by the Metrics Group for a reference approach to arthroscopic anterior shoulder stabilization for unidirectional anterior glenohumeral instability.[6,7] It was acknowledged that the designated reference procedure might not reflect the exact techniques used by individual panelists but that the operative steps that were presented accurately embodied the essential and key components of the procedure. An affirmative vote by a panel member indicated that the metric definition presented was accurate and acceptable as written but not necessarily that it was the manner in which that particular panelist might have chosen to complete the step. "Consensus" meant that there was unanimity in voting among the panelists and that a particular metric definition was "not wrong or inappropriate." Each of the procedural steps and potential errors were evaluated individually. After each metric definition was presented, panel members voted on whether or not the metric was acceptable as written. If the panel could not achieve consensus because of lack of clarity or differences in opinion, the metric definition was revised accordingly and a new vote conducted on the acceptability of the modified metric. This process was repeated until the metric was accepted. If consensus could not be achieved through a series of modifications, the metric was deleted. When it was deemed necessary, a new metric was defined and added.

Results

Bankart Procedure Metrics

The step metrics resulting from task deconstruction were grouped into 13 separate phases of the procedure (in Roman numerals). Each phase (e.g., "arthroscopic instability assessment" or "inferior anchor preparation/insertion") contains a series of related, unambiguously defined, observable procedure events (steps) with specific beginning and ending points. All potential errors identified had been noted to occur during the stress testing of the metrics. Some of the identical errors and sentinel errors could occur during different phases of the procedure that recurred during the 3-anchor repair (e.g., "uncorrected entanglement of shuttling device or suture").

Modified Delphi Panel

All phases of the procedure were accepted as identified. Only a minority of procedure phases and their associated metrics were accepted without discussion. At the conclusion of the deliberations, consensus among the Delphi Panel was reached for 45 steps, 77 errors (29 unique), and 20 sentinel errors (8 unique) (Tables 3 and 4). During the panel deliberations, 2 metrics were deleted, 1 was added, and 5 were modified before consensus was achieved (Table 5).

Summary of Points Raised and Voting Outcomes of Bankart Delphi Panel

The minutes of the metric validation meeting (Copernicus Study/Delphi Panel), held November 18, 2011, are presented herein (recorded by Robert Pedowitz, M.D., Ph.D.). The meeting chair was Rick Angelo, M.D. The Project Leadership Team comprised Rick Angelo, Rick Ryu, Rob Pedowitz, and Tony Gallagher. The attendees comprised R. Angelo, R. Ryu, R. Pedowitz, J. Tokish, R. Bell, R. Hunter, K. Nord, V. Goradia, A. Barber, S. Snyder, B. Beach, J. Abrams, B. Shaffer, J. Tauro, L. Higgins, S. Weber, S. Koo, D. Richards, J. Esch, J. Dodds, J. Randle, J. Richmond, A. Curtis, J. Burns, N. Sgaglione, J. Kelly, and S. Powell (27 voting attendees), as well as T. Gallagher (meeting facilitator)

The meeting overview is as follows:

1. Dr. Angelo presented a brief overview of the project and meeting objectives.
2. Dr. Gallagher presented the background of proficiency-based training, as well as some prior literature demonstrating the validity of this training

Table 3. Thirteen Phases of Arthroscopic 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 anchor security by pulling on 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 of the glenoid
- 28. Drill anchor hole oblique to the glenoid articular face
- 29. Insert suture anchor
- 30. Test anchor security by pulling on 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 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 4. Summary of 29 Different Bankart Procedure Metric Errors

1. Failure to maintain intra-articular position of the posterior cannula
2. Failure to maintain intra-articular position of the mid-anterior cannula
3. Failure to maintain intra-articular position of the anterosuperior cannula
4. Damage to the superior border of the subscapularis during creation of the midanterior portal
5. Damage to the anterior border of the supraspinatus during creation of the anterosuperior portal
6. 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)
7. Laceration of intact capsulolabral tissue (sentinel error)
8. Failure to maintain control of a working instrument (sentinel error)
9. Guide is not located in the inferior region of the anteroinferior quadrant of the glenoid for the most inferior anchor
10. Entry of the completed tunnel lies outside safe zone of 0 to 3 mm from the bony glenoid rim (sentinel error)
11. Shallow undermining and deformation of articular cartilage (sentinel error)
12. Failure to maintain secure seating of the drill guide during anchor insertion
13. Breakage of the implant
14. Implant remains visibly proud (sentinel error)
15. Failure to insert the anchor with the inserter laser line (when present) to or beyond the laser line on the drill guide
16. Anchor fails to remain securely fixed within bone at the appropriate depth
17. Capsular penetration is at or superior to anchor hole (sentinel error)
18. Capsular penetration is not at or peripheral to the capsulolabral junction
19. Instrument breakage
20. Tearing of capsulolabral tissue
21. Uncorrected entanglement of shuttling device or suture
22. Off-loading of suture anchor
23. Breakage of suturing device
24. Failure to create and maintain indentation of the capsule or labral tissue on knot completion (sentinel error)
25. Visible void is present between throws of the completed primary knot (sentinel error)
26. Completed knot abuts articular cartilage
27. Visible void is present between throws of the completed half-hitches
28. Suture breakage
29. Guide is inferior to the equator of the glenoid for the third anchor position

NOTE. Metric errors can be associated with multiple phases and steps of the procedure.

approach for procedural specialties, and he explained the specific objectives of the current Delphi meeting.

3. Dr. Angelo presented each procedural step and explained the associated metrics that have been developed by the Metrics Group for a reference approach to anterior shoulder stabilization for glenohumeral instability. The comments and recommendations for each of the steps, with associated vote, are presented in Table 6.

Discussion

The principal findings of this study are that (1) an arthroscopic Bankart procedure can be deconstructed

Table 5. Delphi Panel Metric Changes

Modification	Issue	Deliberation	Resolution
Deleted (2)	Should failure of anchor purchase remain an error?	Cadaveric bone variability too great to score accurately	Delete error
	Alternating posts for knot tying	Arthroscope views not consistent enough to score reliably	Delete error
Added (1)	Completed knot position	May cause iatrogenic damage if it abuts articular cartilage	Add error to each of the 3 knots tied
Modified (5)	Diagnostic steps—probe or view (how long?)	Does “looking” equal “ascertaining”?	View must be held long enough to determine pathology
	Adequacy of capsular mobilization	Should demonstrate effort at capsular mobility	Must take instrument (grasper) to demonstrate
	Whether to ascribe “critical” to “laceration of labrum”	Labral variability too great in cadavers; can still see violation of “hoop” integrity	Retain as a sentinel error
	Consider deleting the term “sliding” from knot description	Sliding knot would be acceptable for a reference procedure	Retain the term “sliding”
	Should diagnostic steps be included in the procedure metrics?	May skew results if there is excessive influence of diagnostic steps in procedure	Include only steps directly related to instability assessment

Table 6. Comments, Recommendations, and Associated Vote

Comments and Recommendations Regarding Procedural Steps and Metrics	Vote on Steps and Metrics
I. Portals (steps 1-7)	Unanimous affirmative
Agreement that this is an outside-in reference approach for portal placement, though some surgeons use an inside-out approach	
Importance of pre-surgery setup, though assessment of this phase would be very difficult using arthroscopic videotapes	
II. Diagnostic arthroscopy (steps 8-12)	Unanimous affirmative
Clarified metric “view or probe,” not “view and probe”	
Discussion about whether we should include metrics for view or probe of posterior labrum, superior labrum, biceps, and rotator cuff	
Recommendation: Limit diagnostic elements for the current procedural assessment, and consider creation of a diagnostic arthroscopy reference procedure (vote to drop diagnostic elements failed—because one component of the metrics is teaching essential components of the procedure)	
III. Capsulolabral mobilization (steps 13-15)	Unanimous affirmative
Should a laceration of the labrum be defined as a “critical error”?	
Should the pre-existing tissue quality of the labrum be assessed so that laceration of poor tissue does not qualify as a critical error?	
Consider adjustment of metric definition to describe grasping of the anatomic structure of the anterior inferior glenohumeral ligament	
IV. Glenoid neck preparation (steps 16-17)	Unanimous affirmative
V. Insertion of first anchor (steps 18-21)	Unanimous affirmative
Failure of anchor from bone should not be considered a critical error because loss of fixation could be related to bone quality	
VI. Suture management—first anchor (steps 22-23)	Unanimous affirmative
Should we include retrieval of broken suturing device (generally thought this would be quite rare, so not a useful metric)?	
The definition for adequacy of capsulolabral tissue capture seems adequate	
VII. Knot tying—first anchor (steps 24-26)	Unanimous affirmative
Consider deleting the term “sliding” from the knot description	
Add error of a knot completed and left on the articular surface	
Need to drop “alternating posts” metric	
VIII. Insertion of second anchor (steps 27-30)	Unanimous affirmative
Failure of anchor from bone should not be considered a critical error because loss of fixation could be related to bone quality	
IX. Suture management—second anchor (steps 31-32)	Unanimous affirmative
X. Knot tying—second anchor (steps 33-35)	Unanimous affirmative
Consider deleting the term “sliding” from the knot description	
Add error of a knot completed and left on the articular surface	
Probably need to drop “alternating posts” metric	
XI. Insertion of third anchor (steps 36-39)	Unanimous affirmative
Failure of anchor from bone should not be considered a critical error because loss of fixation could be related to bone quality	
XII. Suture management—third anchor (steps 40-41)	Unanimous affirmative
XIII. Knot tying—third anchor (steps 42-44)	Unanimous affirmative
Consider deleting the term “sliding” from the knot description	
Add error of a knot completed and left on the articular surface	
Need to drop “alternating posts” metric	
XIV. Final assessment (step 45)	Unanimous affirmative

into the essential steps necessary for the effective completion of the repair, (2) the potential errors related to the procedure are able to be identified and characterized, and (3) face and content validity for the resulting step and error metrics can be obtained through use of the modified Delphi Panel technique. Traditionally, surgeons have been trained using the “apprenticeship” model, which is related to “process” or time, based (a certain variety of rotations, exposure to numbers of specific cases, etc.). A paradigm shift toward proficiency-based progression training, which is outcomes based, is occurring and mandates that the trainee be able to demonstrate the ability to meet specific skills benchmarks to progress in training. These benchmarks must have specific, clear, objective, and fair standards of performance. Validated metrics will be essential in defining these standards. In addition, as the move toward including surgical skills credentialing and procedural competency occurs for licensing, the same validated standards will be needed. The methodology used in this study provides a framework for the development of those metrics and standards.

An arthroscopic Bankart repair (index procedure) was selected as the reference surgical procedure to study for several reasons. For the patient with unidirectional anterior instability due primarily to a Bankart lesion without significant bone loss, a suture anchor repair with 3 implants is a commonly accepted method used to obtain a successful patient outcome.[13-16] In addition, the essential components of the procedure are well outlined regardless of whether the patient is placed in the lateral decubitus or beach-chair orientation.[17-21]

The task analysis stage of metric development is crucial because metrics are the fundamental building blocks of a good training program. Metrics, thus, not only define how the training should be characterized and the procedure performed by the trainee but also must afford the opportunity for meaningful assessment of the trainee's performance and progress. The entire process of metric development should be as transparent, objective, and unambiguous as possible. Metric definitions should be characterized in such a way that they are sufficiently complete and detailed for an individual—not associated with the initial development process—to use them to score performance reliably. Metric definitions should include behavioral markers that indicate the beginning points and endpoints of the performance characteristics (steps) to be assessed. These parameters will become particularly important in the future as the procedural metrics are used with higher-fidelity simulators. The details of the metric definitions will be necessary for the simulator to be appropriately programmed to provide the trainee with performance assessments and accurate feedback.

Other approaches to the measurement of surgical performance use qualitative descriptions of performance and require the user to rate items on a graduated Likert-type scale (Table 1), which ascribes a quantitative value to qualitative data to make them 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 attitudes. Because of the component of subjectivity, this method of attempting to rate objective performance can render it difficult to obtain acceptable levels of inter-rater reliability (>80%) in the scoring of events. In contrast, the approach to the assessment of performance in our study uses precise definitions of performance and simply requires the reviewer to report whether the specific event occurred or not. This binary approach to the measurement of individual events has been shown to facilitate the reliable scoring of metric-based performance units across a variety of functions during skills training[22-25] of individuals with different experience levels.[26,27] This approach has also been shown to be more reliable than Likert-scale scoring.[28]

Behaviors that deviate from optimal performance (errors) can be characterized, including those of a "more serious nature." The issue of whether those more serious errors should be termed "critical errors" or some alternative label was raised at the outset during the metric definition process. It was agreed on by the Metrics Group that use of the term "critical error" could imply that the event was life-threatening or might have serious medicolegal implications. It was elected, instead, to use the term "sentinel" (Table 1) to connote an error that should be carefully "watched for and to avoid." Sentinel errors involve a serious deviation from optimal performance during a procedure because they can either jeopardize the success/desired result of the procedure or create iatrogenic insult to the patient's tissues. A single specific sentinel error may not always lead to a poor outcome but should stringently be avoided.[6] The underlying philosophy of this approach to errors is that suboptimal outcomes do not happen by accident but usually result from the coalescence of deviations from optimal procedure performance.

The face and content validity of the metric-based procedure characterization by subject specialists can be verified using the modified Delphi Panel methodology reported in this study. The metrics developed were informed by research studies, professional guidelines, clinical experience, and manufacturers' guidelines.[6,7] Although the surgeons in the Metrics Group are very experienced in the performance of the Bankart procedure, the Delphi process provided an excellent method to ensure that the procedure characterization is appropriate, represents best practice, and is acceptable to a larger group of experienced master and associate master Bankart faculty. As anticipated, many surgeons pointed out that they might perform a specific step in a different manner but that the approach outlined by the Metrics Group was "not incorrect" or inadvisable. The members of the panel made very helpful suggestions for improving the definitions.

Assuming that the Bankart metric identifications and definitions represent a real-world surgical procedure, these performance characteristics should be able to distinguish between experienced (skilled) surgeons and novices, that is, provide construct validity. Future studies regarding construct validity will seek to provide information about which metrics best distinguish between experienced and novice surgeon performance. This information will facilitate the establishment of a benchmark to define the "proficiency level" that trainees should acquire before progressing to in vivo practice.[6,29]

Limitations

A limitation of this study resides in the fact that every potential error, regardless of how rare the occurrence, might not have been included. The Delphi Panel, however, confirmed that the errors listed were those most likely to occur and that should be avoided in the safe performance of an arthroscopic Bankart repair. Although common errors may be relatively easy to agree on, it is somewhat more challenging to decide which errors should be designated as "sentinel," without a specific weighting methodology. Although the issue of using this designation for events that cause iatrogenic damage is more straightforward, the concept of also using the term for events that might "potentially lead to a suboptimal outcome" is more subject to the opinion of the Metrics Group and the Delphi Panel.

Furthermore, data are not available to confirm that the specific steps identified by the Metrics Group and the Delphi Panel directly correlate with a successful surgical outcome for patients with unidirectional shoulder instability. Therefore the metrics created remain predominantly based on the opinion of experienced surgeons and instructors. An outcomes study will be needed to fully establish the predictive validity for the Bankart metrics as authored. In addition, the surgeons comprising the Metrics Group and Delphi Panel were all North American surgeons. International arthroscopists may have created somewhat different metrics for an arthroscopic Bankart repair.

Conclusions

This study rejects the null hypothesis and confirms that a core group of experienced arthroscopic surgeons is able to perform task deconstruction of an arthroscopic Bankart repair and create unambiguous step and error definitions (metrics) that accurately characterize the essential components of the procedure. Analysis and revision by a larger panel of experienced arthroscopists were able to validate the Bankart metrics.

References

1. Adams JA. A closed-loop theory of motor learning. *J Mot Behav* 1971;3:111-149.
2. Adams JA. Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychol Bull* 1987;101:41-74.
3. Adams JA. The changing face of motor learning. *Hum Mov Sci* 1990;9:209-220.
4. Dreyfus HL, Dreyfus SE. *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: The Free Press, 1986;50.
5. Andersen DK. How can educators use simulation applications to teach and assess surgical judgment? *Acad Med* 2012;87:934-941.
6. Gallagher AG, O'Sullivan GC. *Fundamentals of surgical simulation; principles & practices*. London: Springer Verlag, 2011.
7. 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.
8. Burks RT, Presson AP, Weng H-Y. An analysis of technical aspects of the arthroscopic Bankart procedure as performed in the United States. *Arthroscopy* 2014;30:1246-1253.
9. Morgan CD, Bodenstab AB. Arthroscopic Bankart suture repair: Technique and early results. *Arthroscopy* 2010;26:819-820.
10. 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.
11. Seymour NE, Gallagher AG, Roman SA, et al. Analysis of errors in laparoscopic surgical procedures. *Surg Endosc* 2004;18:592-595.
12. Cuschieri A, Francis N, Crosby J, Hanna GB. What do master surgeons think of surgical competence and revalidation? *Am J Surg* 2001;182:110-116.
13. Ee GW, Mohamed S, Tan AH. Long term results of arthroscopic Bankart repair for traumatic anterior shoulder instability. *J Orthop Surg Res* 2011;6:28.
14. van der Linde JA, van Kampen DA, Willems JW. How should we define failure after surgical shoulder stabilization? *Arthroscopy* 2013;29:1602-1603.
15. 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.
16. Shibata H, Gotoh M, Mitsui Y, et al. Risk factors for shoulder re-dislocation after arthroscopic Bankart repair. *J Orthop Surg Res* 2014;9:53.
17. Angelo R. Arthroscopic Bankart repair for unidirectional shoulder instability. *Instr Course Lect* 2009;58:305-313.
18. Angelo RL. Controversies in arthroscopic shoulder surgery: Arthroscopic versus open Bankart repair, thermal treatment of capsular tissue, acromioplasties—Are they necessary? *Arthroscopy* 2003;19:224-228.
19. Ryu RK. Arthroscopic approach to traumatic anterior shoulder instability. *Arthroscopy* 2003;19:94-101.
20. Kim S-H, Ha K-I, Cho Y-B, et al. Arthroscopic anterior stabilization of the shoulder: Two to six-year follow-up. *J Bone Joint Surg Am* 2003;85:1511-1518.
21. Cole BJ, L'Insalata J, Irrgang J, Warner JJ. Comparison of arthroscopic and open anterior shoulder stabilization. A two to six-year follow-up study. *J Bone Joint Surg Am* 2000;82:1108-1114.
22. 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).
23. Van Sickle K, Smith B, McClusky DA III, et al. Evaluation of a tensiometer to provide objective feedback in knot-tying performance. *Am Surg* 2005;71:1018-1023.
24. 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.

25. 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.
26. 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.
27. 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.
28. 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* 2012;84:568-573.
29. 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.