Inter-rater Reliability for Metrics Scored in a Binary Fashion—Performance Assessment for an Arthroscopic Bankart Repair

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

Purpose: To determine the inter-rater reliability (IRR) of a procedure-specific checklist scored in a binary fashion for the evaluation of surgical skill and whether it meets a minimum level of agreement (≥0.8 between 2 raters) required for high-stakes assessment. **Methods:** In a prospective randomized and blinded fashion, and after detailed assessment training, 10 Arthroscopy Association of North America Master/Associate Master faculty arthroscopic surgeons (in 5 pairs) with an average of 21 years of surgical experience assessed the video-recorded 3-anchor arthroscopic Bankart repair performance of 44 postgraduate year 4 or 5 residents from 21 Accreditation Council for Graduate Medical Education orthopaedic residency training programs from across the United States. **Results:** No paired scores of resident surgeon performance evaluated by the 5 teams of faculty assessors dropped below the 0.8 IRR level (mean = 0.93; range 0.84-0.99; standard deviation = 0.035). A comparison between the 5 assessor groups with 1 factor analysis of variance showed that there was no significant difference between the groups ($P = .205$). Pearson's product-moment correlation coefficient revealed a strong and statistically significant negative correlation, that is, −0.856 ($P < .000$), indicating that as intra-operative error rate scores increased, the IRR decreased. **Conclusions:** Arthroscopy Association of North America shoulder faculty raters from across the United States showed high levels of IRR in the assessment of an arthroscopic 3-anchor Bankart repair procedure. All paired assessments were above the 0.8 level and the mean IRR of all resident assessments was 0.93, indicating that they could be used for high-stakes decisions. **Clinical Relevance:** With the move toward outcomes-based performance evaluation for graduate medical education, high-stakes assessments of surgical skill will require robust, reliable measurement tools that are able to withstand challenge. Surgical checklists employing metrics scored in a binary fashion meet the need and can show a high (>80%) IRR.

See commentary on page 2199

Program directors and leaders in surgical training face a significant challenge in addressing the substantial knowledge and experience gaps of trainee surgeons exiting residency programs.[1] Surgical trainees on completion of their general surgical training program have themselves expressed concern about their readiness to enter independent clinical practice and approximately 80% of them opted for additional training. In their 2009 paper, Bell et al.[1] identified that surgical trainees lacked exposure to, never mind training to competency in many surgical procedures, fundamental to general surgery training. The Institute of Medicine in 2013 proposed a radical change to medical education and training that necessitated a move from a

From the ASSERT Centre, University College Cork (A.G.G.), Cork, Ireland; The Ryu Hurvitz Orthopedic Clinic (R.K.N.R.), Santa Barbara; Department of Orthopedics, University of California (R.A.P.), Los Angeles, California, U.S.A.; ASSERT Centre and School of Medicine, University College Cork (P.H.), Cork, Ireland; and ProOrtho Clinic (R.L.A.), Kirkland, Washington, U.S.A.

The authors report the following potential conflicts of interest or sources of funding: A.G.G. was paid directly as a consultant by AANA; and was reimbursed for travel and accommodations directly related to the AANA Copernicus Initiative. R.K.N.R. receives travel expenses from AANA; receives consultancy fees from Medbridge and Rotation Medical; receives payment for lectures including service on speakers bureaus from Mitek; and receives royalties from AANA Education Foundation. R.A.P. receives consultancy fees from Virtamed; and is employed with Virtamed. R.L.A. receives travel and housing expenses for meetings/OLC study components from AANA; and is a consultant for DePuy/Mitek (education/product review). Full ICMJE author disclosure forms are available for this article online, as supplementary material.

Received August 13, 2017; accepted February 1, 2018.

Address correspondence to Richard L. Angelo, M.D., Ph.D., ProOrtho Clinic, Kirkland, WA 98072, U.S.A. E-mail: rlamdortho@comcast.net

0749-8063/17966/$36.00

https://doi.org/10.1016/j.arthro.2018.02.007

process-driven approach to education and training, which relies on time in training or numbers of procedures performed, to an outcome-based approach, where competency and patient outcomes must be shown rather than assumed.[2]

The implications of this change will be a dramatic increase in the number and types of high-stakes assessments for graduate medical education.[3-6] This change dictates therefore that if the continuation or cessation of a residents' training is to be decided based on their objectively assessed performance, those assessments need to be objective, transparent, and fair. Consequently, assessments must demonstratively achieve the internationally agreed on validation standards required of a high-stakes assessment.[7]

There are different types of validities (e.g., face, content, concurrent, construct, and predictive validity).[8,9] However, an assessment that shows validity but is unreliable is by default not valid.[7] Inter-rater reliability (IRR) is a fundamental benchmark of the reliability or lack thereof for an assessment strategy. It compares the scores between 2 raters on their assessment of the performance of an individual. IRR estimates should reach at least 0.8 or 80% agreement to be considered of value for assessing performance, particularly if that assessment is for high-stakes judgments such as training progression or the assurance of proficiency.

The most robust and reliable assessment methods of surgical procedural performance are procedure checklists. The step metrics, which make up a procedure checklist, are derived from a systematic and detailed characterization of what experienced and "good" practitioners do during optimal procedural performance.[10-12] Suboptimal and deviations from optimal procedural performance (metric errors) are also identified. Performance metrics are defined rather than described.[9,13] Furthermore, performance benchmarks are quantitatively defined based on the performance of experienced and "good" practitioners. Previous studies have shown that checklists are scored more reliably than Likert-type scales[14] and that they show good levels of IRR.[15-20] However, to date, these studies have all been single site evaluations. The purpose of this investigation was to determine the IRR of a procedure-specific checklist scored in a binary fashion for the evaluation of surgical skill and whether it meets a minimum level of agreement (0.8) required for high-stakes assessment. We hypothesized that a checklist for an arthroscopic Bankart repair scored in binary fashion would be able to show an IRR of 0.8 or better.

Methods

Participants/Subjects

Ten Arthroscopy Association of North America (AANA) Master/Associate Master faculty arthroscopic surgeons were designated to participate in one of 5 fixed pairs of raters for the purpose of determining IRR in scoring. Video performance of 44 subjects who were postgraduate year 4 or 5 residents from 21 Accreditation Council for Graduate Medical Education accredited orthopaedic residency training programs from across the United States was assessed for an arthroscopic Bankart repair. The residents were being evaluated as part of the AANA Copernicus Investigation.[21,22]

The mean age of the Master/Associate Master surgeon raters was 55 years (standard deviation [SD] = 7 years) with a mean of 21 years in practice (SD = 7 years). All raters were fellowship trained in Arthroscopy and Sports Medicine. In this multicenter investigation, faculty surgeons practiced in (1) Seattle, (2) Los Angeles (n = 3), (3) Jackson, MS (n = 2), (4) New York, (5) Lansing, MI, (6) Santa Barbara, CA, and (7) Richmond, VA. All the surgeons were considered shoulder specialists.

All subjects to be assessed were assigned a unique identifying number that gave no indication of their postgraduate year, residency program, or study group.

Procedure

The format of the study is shown in Figure 1. The metrics, including 45 steps (grouped into 13 phases of the procedure) and 77 potential errors, were developed during the procedure characterization/task deconstruction phase.[10] The metrics were shown to have face, content, and construct validity,[18,23,24] and used to build a proficiency-based progression (PBP) education and training curriculum (Phase 1). During Phase 2 of this study, the attending surgeons were trained to score the metrics reliably (outlined in more detail below). In Phase 3, the attending surgeons objectively assessed the video-recorded performance of the 3 groups of trainees who progressed along 3 different training pathways: Group A (control) received the standard AANA resident training curriculum with didactic lectures, knot tying practice, and skills training on a cadaver shoulder; Group B (simulator) was offered similar training with the additional opportunity to use a medium fidelity shoulder model simulator to practice each step of the Bankart procedure; and Group C (PBP + simulation) who had similar training to Group B, but were also required to show a passing score on the cognitive understanding of the previously created Bankart procedural metrics, and show the ability to meet previously established proficiency benchmarks for knot tying and an arthroscopic Bankart repair on the model simulator before progressing to train in the cadaver lab. The details of this investigation are reported in elsewhere.[19]

Video Reviewer Training

After the construction of the metrics, which included verification of face, content, and construct validity[10,18,23]

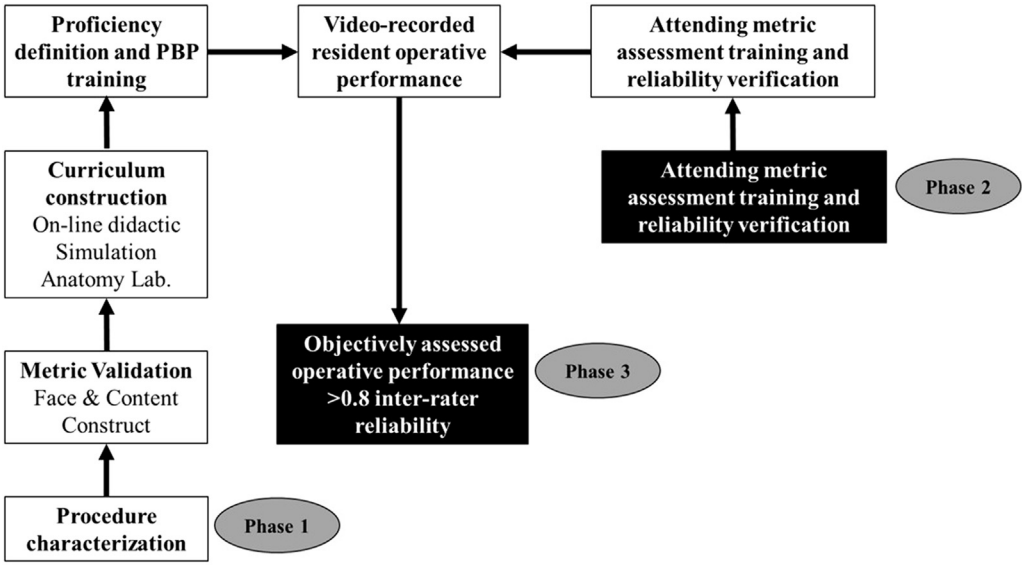

Fig 1. The 3 different phases of the study leading to the reliable assessment of intraoperative assessment. (PBP, proficiency-based progression.)

for an arthroscopic Bankart repair, a final version of a performance score sheet was formatted.

Ten AANA Master/Associate Master faculty surgeons formed the panel of assessors designated to review and score the videos using the validated Bankart metrics. This group included the 3 arthroscopic surgeons who, in conjunction with a consultant experimental psychologist, developed the arthroscopic Bankart metric definitions (Table 1 = steps, Table 2 = potential errors). The 10 reviewers were randomly assigned to form 5 fixed pairs, which remained constant throughout the scoring of all videos.

Training in Reliable Assessment

Training of the assessor group was initiated with an 8-hour in-person meeting during which each metric was studied in detail. Multiple video examples of live patient cases were shown to illustrate each metric. Videos of the patients in both the lateral decubitus and beach chair orientations were represented (as both patient orientations are in common use during shoulder arthroscopy). A discussion helped to clarify how each step and error was to be scored, including the nuances and conventions to be used.

Assessment Reliability Verification

After training, full-length practice videos 1 and 2 (one each in the lateral decubitus and beach chair orientation) were sent to and independently scored by each of the 10 raters, and the scores tabulated. In 2 subsequent 2-hour group phone conferences, the differences and discrepancies amongst all assessors were compared and discussed seeking conformity in scoring. In addition, each designated pair of raters conducted 1 to 3 additional phone conferences to analyze the specific instances in which the 2 of them scored events differently. Subsequently, all assessors scored practice videos 3 and 4 and the results were tabulated (both patient orientations again represented). The scores for each of the 5 designated pairs of assessors were compared for the second set of practice videos. In only 1 of 10 comparisons (2 videos * 5 assessor pairs) did the IRR calculation (agreements/agreements + disagreements) fall below an acceptable level[9,25] of 0.8 at 0.76.

Surgical Resident Video Scoring

The AANA research coordinator randomly assigned each of the 44 full-length study videos (with a unique identifying number) to a single pair of assessors. All video assessors remained blinded to the source of the video being reviewed. Each video was independently reviewed and scored by the 2 members of an assigned pair of assessors. All scores were tabulated for each of 122 metrics for the procedure (45 steps, 77 potential errors; Tables 1 and 2). Each step and error metric was scored as either a "yes" or "no," designating whether the specific event was or was not observed to occur by the reviewer. In addition to scoring steps and errors, each event characterized as "damage to nontarget tissue" was scored (e.g., gouging the articular cartilage, or tearing of the capsule). There was no limit to the number of individual instances damage to nontarget tissue could be scored, with each occurrence tallied as a single error event.

Score Tabulation

The 2 raw score sheets from the designated pair of assessors were compared for each of the individual steps (N = 45) and the number of "agreements" tabulated (either both assessors documented that a step was

Table 1. The 13 Stages of the Bankart Procedure (in Roman Numerals) and a Brief Summary of the 45 Steps of the Procedure (AANA, 2013)

Scope Bankart Repair Steps
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. Show instrument access to the anteroinferior glenoid/labrum
6. Introduce an 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 the posterior portal
8. View or probe the superior labral attachment onto the glenoid
9. View or probe the articular surface of the cuff
10. Probe anteroinferior glenoid/Bankart pathology including rim fracture, articular defect
View from the anterosuperior portal
11. View or probe the mid-substance 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 suscapularis muscle superficial to the mobilized capsule
15. With an instrument, grasp and perform an inferior to superior shift of the capsulolabral tissue (to restore tension)
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 suture anchor
V. Suture delivery/management
22. Pass a cannulated suture hook or suture retriever through the capsular tissue inferior to the anchor
23. Pass the 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 4½ hitches
26. Cut the 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 the anchor hole oblique to the glenoid articular face
29. Insert a 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 the 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 4½ hitches
35. Cut the 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 the suture tails
XI. Suture delivery/management
40. Pass a cannulated suture hook or suture retriever through the capsular tissue
41. Pass the 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 4½ hitches
44. Cut the suture tails
XIII. Procedure review
45. View and/or probe the final completed repair

Table 2. A Summary of the 29 Different Bankart Procedure Metric Errors; Metric Errors Can Be Associated With Multiple Phases and Steps of the Procedure (N = 77 Total Errors) (AANA, 2013)

Bankart Repair Metric Errors
1. Failure to maintain the intra-articular position of the posterior cannula
2. Failure to maintain the intra-articular position of the mid-anterior cannula
3. Failure to maintain the intra-articular position of the anterosuperior cannula
4. Damage to the superior boarder of the subscapularis
5. Damage to the anterior boarder of the supraspinatus
6. Loss of the intra-articular position of scope/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 scope + 2 portals)
7. Lacerate intact capsulolabral tissue (SENTINEL ERROR)
8. Failure to maintain the control of the working instrument (SENTINAL ERROR)
9. Guide is not located in the inferior region of the anteroinferior quadrant of the glenoid
10. Entry of the completed tunnel lies outside the safe zone of 0 to 3 mm from the bony glenoid rim (SENTINEL ERROR)
11. Shallow undermining and deformation of the 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 the bone at the appropriate depth
17. Capsular penetration is at or superior to the anchor hole (SENTINEL ERROR)
18. Capsular penetration is not at or peripheral to the capsulolabral junction
19. Instrument breakage
20. Tearing of the capsulolabral tissue
21. Uncorrected entanglement of the shuttling device or suture
22. Off-loading the suture anchor
23. Break the suturing device
24. Failure to create and maintain indentation of the capsule or the labral tissue (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 complete half hitches
28. Suture breakage
29. Guide is inferior to the equator of the glenoid (for the third and final anchor)

performed, or both scored the step as not being completed). In addition, the number of "disagreements" in scoring steps was tabulated (one of the assessors indicated that the step had and other scored that the step had not been completed). The IRR for the steps was calculated according to the following formula:

$$\frac{\text{Agreements}}{\text{Agreements} + \text{disagreements}}$$

In a similar manner, there was either agreement or disagreement in the 2 scores for each of the potential errors (N = 77). The IRR for error scoring was calculated in the same manner as that for the steps. The IRR for scoring the entire procedure was calculated using both the step and error agreements/disagreements for the complete procedure (N = 122). Acceptable IRR[9] is $\geq$0.80. Finally, we assessed the relation between IRR levels and objectively assessed intraoperative error rates with Pearson's product moment correlation coefficient.

Results

The resident performance scores for steps completed ranged from 31.5 to 44 (of 45 total steps). Errors committed ranged from 0 to 13.5. Three resident subjects were unable to complete the arthroscopic Bankart repair. Figure 2 shows the IRR scores for each individual resident assessed by the 5 teams of assessors. Four teams scored 9 videos each and 1 team scored 8 videos. The highest IRR score for the assessment of a participant was for a video recording scored by Team 2 (IRR = 0.99), and the lowest score was for a subject scored by Team 5 (IRR = 0.84). The mean IRR for the 44 assessments was 0.93. No subject scored by the 5 teams of assessors dropped below the 0.8 IRR level. The median and interquartile range for the assessments completed by each of the rater pairs are shown in Figure 3. Team 1 assessments showed the greatest variability between subjects' assessment reliability, and Team 2 showed the greatest assessment concordance. A comparison between the groups with one-factor analysis of variance (ANOVA) showed that there was no significant difference between the 5 rater pairs, $F = (4, 37), 1.56, P = .205$ (Fig 3).

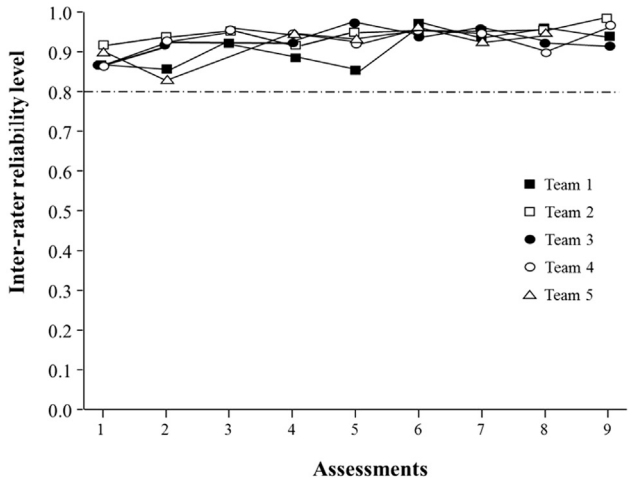

Fig 2. The inter-rater reliability for the assessment of each resident surgeon's performance (each scored by 1 of the 5 teams of assessors). The graph reveals that the inter-rater reliability for all 44 videos assessed was above the 0.8 level (and clustered around the 0.9 level).

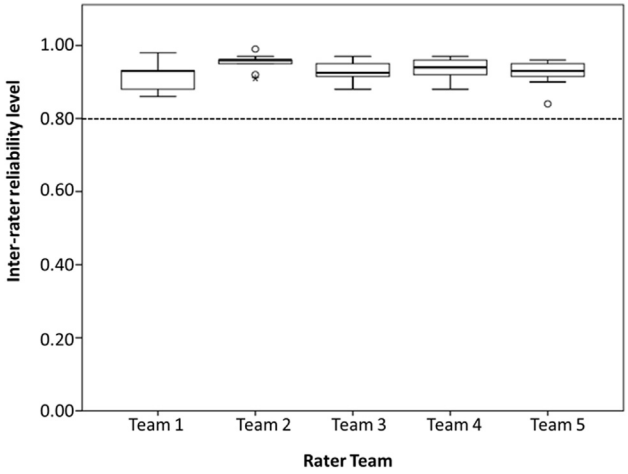

Fig 3. The median and interquartile range of the inter-rater reliability calculations for the surgeons assessed by the 5 teams of assessors.

Figure 4 shows the median and interquartile range for the assessments on the 3 different groups of trainees. The paired assessments evaluating the videos of Group A (control) showed the lowest IRR levels (even though it was still above the 0.8 IRR level). The assessments of Group C (PBP + simulation) showed the highest levels of IRR. When compared with one-factor analysis of variance this difference was found to be statistically significant ($F = (2, 39), 5.99, P = .005$). Contrasts between the groups with Scheffe F-tests showed that the difference between the IRR assessments for residents in Group A (control) and those in Group C (PBP + simulation) was statistically significant ($P = .005$) but the difference between those in Group A and Group B (simulation) was not statistically significant ($P = .14$).

The overall mean IRR for the paired assessments of the 44 resident videos was 0.93 (SD = 0.035). Calculation of Pearson's product moment correlation coefficient revealed a strong and statistically significant negative correlation, that is, −0.856 ($P < .000$), indicating that as intraoperative error rate scores increased, the IRR decreased (Fig 4).

Discussion

In this study, we have shown that a procedure checklist measurement instrument is a very reliable way to assess the surgical procedure performance of senior orthopaedic residents in a multicenter context. All of the assessments by all assessor pairs were >0.8 for IRR, which is the fundamental requirement of an evaluation that is to be used for high-stakes decisions such as training progression. These results replicate previous findings from single institution and smaller scale studies.[10-12] Our results also showed that there was no significant difference between the 5 pairs of raters even though they assessed video-recorded performances independently (in a blinded fashion) and the assessors were dispersed across the United States.

One finding of note was the significant difference in reliability assessments between the 3 groups of trainees. Although the paired assessments of the Control group were still above the 0.8 IRR level, they were significantly lower than those for Group C (PBP + simulation). The reason for this is probably what you would expect in that the errors are generally more difficult to score than the steps and would have a higher probability of producing differences between the raters. Group A residents enacted significantly more errors than Group C, which most likely explains the difference in IRR assessments between Groups A and C. The large and statistically significant negative correlation between error scores and IRR appears to corroborate this hypothesis. In addition, the poorer the error metrics drafted (i.e., explicitly operationally defined), the poorer the discrimination between the novice and experienced groups and the lower the IRR is likely to be. It is therefore of critical importance to carefully draft robust error metrics.

It should be stressed that the high IRR observed in this and other proficiency-based progression studies is almost certainly a function of the attention to detail in (1) procedure characterization of optimal and suboptimal performance, (2) the detailed operational definitions of these performance characteristics, (3) validation efforts made to ensure that the attributes identified truly represent important aspects of the surgical procedure being characterized, and (4) the thoroughness of the

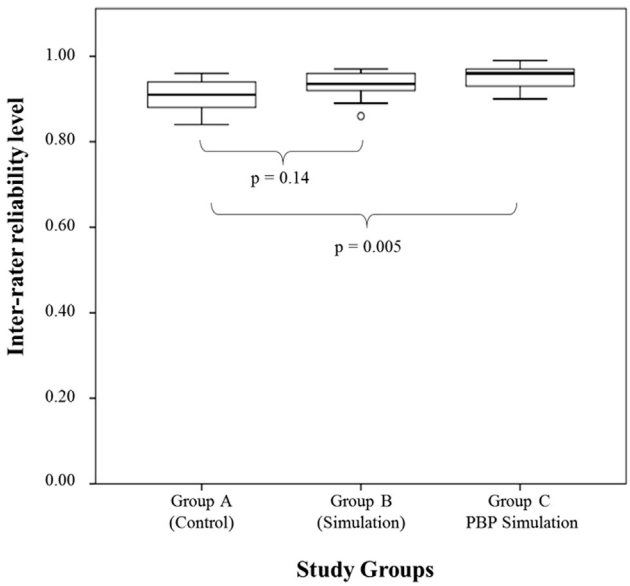

Fig 4. The median and interquartile range of the inter-rater reliability calculations for the different trainee groups surgeons assessed—Groups A (control), B (simulation), and C (proficiency based progression + simulation).

assessor training. It should also be acknowledged that these efforts take considerable time. It is however our view that these efforts constitute a worthy investment. If the metrics are to be used for high-stakes assessments, then the users are duty-bound to ensure that the assessments meet the highest and most rigorous validation standards.

An integral and fundamental part of this assessment is the training of the assessors. As described in the Methods section, considerable effort was invested in training the assessors to use the metrics reliably. It is usually assumed that because of their intellectual caliber and work ethic, those surgeons will be able to pick up and use an assessment methodology very quickly. This is an unwarranted assumption and assessors should be required to show that they could score the metrics reliably before they use them in a high-stakes context. After all, we would expect no less if the metrics were being used to assess our own performance.

Finally, although the total number of videos (44) is relatively small, the large number of metrics scored for each video (122) makes the assessment robust. The number of metric errors exceeded the number of steps and, if disproportionately large, could affect the IRR. Errors that clearly do not occur would typically be scored uniformly as a "No" by both raters (an "agreement"). If a large number of potential errors that were never observed to occur were included in the metrics, the IRR calculation (agreements/agreements + disagreements) would be spuriously inflated. For that reason, in the current investigation, only those errors that were observed to occur in the videos of novice performance used to draft the metrics were included in the final set of Bankart metrics.

Limitations

The reference procedure selected for task deconstruction and metric development was straightforward with a commonly accepted technique for the Bankart repair. It is possible that it would have been more challenging to draft clear, unambiguous metrics for a procedure with multiple, accepted methods of performing the same task. To draft metrics, which are uniform and at the same time are able to recognize different techniques, could lead to a lack of clarity and create disparate scoring amongst rater pairs. To create metrics, which are more generalizable (to score different techniques), could risk the inability to discriminate between levels of performance with a resulting loss of construct validity for the assessment tool. Therefore, a uniform, reference procedure is best suited to task deconstruction and metric characterization.

Conclusions

AANA shoulder faculty raters from across the United States showed high levels of IRR in the assessment of an arthroscopic 3-anchor Bankart repair procedure. All paired assessments were above the 0.8 level and the mean IRR of all resident assessments was 0.93, which means they could be used for high-stakes decisions.

References

1. 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.
2. Asch DA, Weinstein DF. Innovation in medical education. *N Engl J Med* 2014;371:794-795.
3. Gallagher AG, Neary P, Gillen P, et al. Novel method for assessment and selection of trainees for higher surgical training in general surgery. *ANZ J Surg* 2008;78:282-290.
4. Gallagher AG, Leonard G, Traynor OJ. Role and feasibility of psychomotor and dexterity testing in selection for surgical training. *ANZ J Surg* 2009;79:108-113.
5. Carroll SM, Kennedy AM, Traynor O, Gallagher AG. Objective assessment of surgical performance and its impact on a national selection programme of candidates for higher surgical training in plastic surgery. *J Plast Reconstr Aesthet Surg* 2009;62:1543-1549.
6. Gallagher AG, O'Sullivan GC, Neary PC, et al. An objective evaluation of a multi-component, competitive, selection process for admitting surgeons into higher surgical training in a national setting. *World J Surg* 2014;38:296-304.
7. APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). *Standards for educational and psychological tests*, 1999.
8. 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.
9. Gallagher AG, O'Sullivan GC. *Fundamentals of surgical simulation; principles and practices*. London: Springer Verlag, 2011.
10. Angelo RL, Ryu RK, Pedowitz RA, Gallagher AG. Metric development for an arthroscopic Bankart procedure: Assessment of face and content validity. *Arthroscopy* 2015;31:1430-1440.
11. Cates CU, Gallagher AG. The future of simulation technologies for complex cardiovascular procedures. *Eur Heart J* 2012;33:2127-2134.
12. O'Sullivan O, Aboulafia A, Iohom G, O'Donnell BD, Shorten GD. Proactive error analysis of ultrasound-guided axillary brachial plexus block performance. *Reg Anesth Pain Med* 2011;36:502-527.
13. 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.
14. 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.
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;236:458-463. discussion 63-64.
16. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Andersen DK, Satava RM. Analysis of errors in laparoscopic surgical procedures. *Surg Endosc* 2004;18:592-595.
17. 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.
18. Angelo RL, Ryu RK, Pedowitz RA, Gallagher AG. The Bankart performance metrics combined with a cadaveric shoulder create a precise and accurate assessment tool for measuring surgeon skill. *Arthroscopy* 2015;31:1655-1670.
19. Angelo RL, Ryu RK, Pedowitz RA, et al. A proficiency-based progression training curriculum coupled with a model simulator results in the acquisition of a superior arthroscopic Bankart skill set. *Arthroscopy* 2015;31:1854-1871.
20. Cates CU, Lönn L, Gallagher AG. Prospective, randomised, and blinded comparison of proficiency-based progression full-physics virtual reality simulator training versus invasive vascular experience for learning carotid artery angiography by very experienced operators. *BMJ* 2016;2:1-5.
21. Angelo RL. Magellan and Copernicus: Arthroscopy Association of North America seeking excellence in education. *Arthroscopy* 2015;31:1428-1429.
22. Lubowitz JH, Provencher MT, Brand JC, Rossi MJ. The apprenticeship model for surgical training is inferior. *Arthroscopy* 2015;31:1847-1848.
23. Angelo RL, Pedowitz RA, Ryu RK, 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.
24. Pedowitz R, Nicandri GT, Angelo R, et al. Objective assessment of knot-tying proficiency with the Fundamentals of Arthroscopic Surgery Training (FAST) program workstation and knot tester. *Arthroscopy* 2015;31:1872-1879.
25. Kazdin AE. *Behavior modification in applied settings*. Pacific Grove, CA: Brooks: Cole Publishing Co., 1994.