J Shoulder Elbow Surg (2019) 28, 1273–1280

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

A concise shoulder outcome measure: application of computerized adaptive testing to the American Shoulder and Elbow Surgeons Shoulder Assessment

Otho R. Plummer, PhD[a,*], Joseph A. Abboud, MD[b], John-Erik Bell, MD[c], Anand M. Murthi, MD[d], Anthony A. Romeo, MD[e], Priyanka Singh, PhD[a], Benjamin M. Zmistowski, MD[b]

[a]*Universal Research Solutions, LLC, Columbia, MO, USA*
[b]*Department of Shoulder and Elbow Surgery, Rothman Institute and Jefferson University, Philadelphia, PA, USA*
[c]*Department of Shoulder and Elbow Surgery, Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA*
[d]*Department of Shoulder and Elbow Surgery, MedStar Union Memorial, Baltimore, MD, USA*
[e]*Department of Orthopaedic Surgery, Rush University Medical Center, Chicago, IL, USA*

Background: Patient-reported outcome measures enable quantitative and patient-centric assessment of orthopedic interventions; however, increased use of these forms has an associated burden for patients and practices. We examined the utility of a computerized adaptive testing (CAT) method to reduce the number of questions on the American Shoulder and Elbow Surgeons (ASES) instrument.

Methods: A previously developed ASES CAT system was applied to the responses of 2763 patients who underwent shoulder evaluation and treatment and had answered all questions on the full ASES instrument. Analyses to assess the accuracy of the CAT score in replicating the full-form score included the mean and standard deviation of both groups of scores, frequency distributions of the 2 sets of scores and score differences, Pearson and intraclass correlation coefficients, and Bland-Altman assessment of patterns in score differences.

Results: By tailoring questions according to prior responses, CAT reduced the question burden by 40%. The mean difference between CAT and full ASES scores was −0.14, and the scores were within 5 points in 95% of cases (a 12-point difference is considered the threshold for clinical significance) and were clustered around zero. The correlation coefficients were 0.99, and the frequency distributions of the CAT and full ASES scores were nearly identical. The differences between scores were independent of the overall score, and no significant bias for CAT scores was found in either a positive or negative direction.

Conclusion: The ASES CAT system lessens respondent burden with a negligible effect on score integrity.

No institutional review board approval was required.

*Reprint requests: Otho R. Plummer, PhD, Universal Research Solutions, 414 Broadway, Ste 102, Columbia, MO 65201, USA.

E-mail address: otho.plummer@oberd.com (O.R. Plummer).

https://doi.org/10.1016/j.jse.2018.11.068

Level of evidence: Basic Science Study; Development of Validation of Outcome Instruments

Keywords: ASES; shoulder; patient-reported outcomes; CAT; MCID; response burden

Patient-reported outcome measures (PROMs) are an integral tool for assessing the impact of orthopedic treatments. Outcome measures used include general health and social scores, disease-specific scores, and region (eg, hip, knee, and shoulder)–specific measures. These tools are an important (and often the only) way to obtain quantitative, meaningful assessments of the effect of treatment on the quality of life and level of function experienced by patients. In addition, these measures can be used to assess the cost-effectiveness of interventions and may become a prerequisite for reimbursement for care.[6,15] The American Shoulder and Elbow Surgeons (ASES) developed the ASES Shoulder Score (Table I) to encourage such patient-centered assessment of outcomes for shoulder ailments[3,24]; the instrument has been validated for a range of shoulder and elbow problems, including shoulder instability, rotator cuff disease, osteoarthritis, and shoulder arthroplasty.[33]

Although outcome measures are ubiquitous in assessing treatment success in research publications,[10,14,30] they can also be clinically valuable for gauging individual patient outcomes[2,7] and potentially beneficial for calculating outcome-based reimbursement.[15] General adoption of these tools in everyday clinical practice has been limited, however, and the burden on physicians and patients cannot be ignored. For patients, completing these often numerous forms is time-consuming and must be repeated frequently. For elderly patients, they can be physically and mentally exhausting. For physicians, capturing outcome measures can impact the clinical workflow, be time-consuming and financially prohibitive, and present the challenge of secure data storage.

Use of computer-administrated rather than paper-and-pencil forms has reduced these burdens significantly,[21,23] particularly when patients can choose between modalities such as the Internet, voice response systems, mobile operating systems, and text messaging, which enable remote responses from home. Time saving and convenience for the patient must be continually pursued to achieve meaningful response rates for longitudinal studies, and advances in computer technology must continue to be leveraged to improve the efficiency and reliability of outcome assessments.

In view of this aim, computerized adaptive testing (CAT) is beginning to be applied in medicine. By tailoring questions to the specific traits of the respondent, CAT technology allows accurate assessments to be obtained with fewer questions. The defining characteristic of CAT is that, at any point, the next question to be asked is chosen based on the information already obtained. The adaptive algorithm is designed to elicit the most significant and relevant information and to forgo questions that need not be asked at all. This technology has been developed, refined, and successfully applied in academic testing and other fields.[8,11] The idea that patient-derived assessments are a valuable, and perhaps the most pertinent, measure of health outcomes has widespread support in the clinical and medical research community. Ultimately, creating the most concise versions possible of each outcome instrument in use lowers the perform burden on patients so that multiple forms and comprehensive assessments are feasible.

Within the OBERD software system (Universal Research Solutions, Columbia, MO, USA)*—a clinical practice tool for outcome measure collection—a CAT version of the ASES score has been recently developed. This study investigates the ASES CAT system, with a focus on the reduced question burden and the ability of CAT to provide accurate assessments of patient's perceived shoulder and elbow function.

The OBERD CAT system, which was developed before this study, uses so-called machine learning programs that analyze how response patterns affect the overall score for completed forms to construct rules for optimal ways to ask the questions. This process is referred to as ''training'' the CAT, and the CAT is said to ''learn'' how to question effectively. The set of completed forms used to develop a CAT is called the ''training set.'' Accuracy improves and the question burden is eased as the CAT system learns to be more efficient. The CAT is deemed to be valid if it produces accurate results when applied to cases (a ''test set'') it has not seen before. Thus, development of a CAT model proceeds in 2 steps. First, a large, random sample of completed forms is used to build a CAT model; in the present instance, the training set contained 9229 cases. Then, the success of the model is judged by the accuracy with which it reproduces full-form scores and the extent to which it reduces the question burden in a large number of independent cases. This report assesses the accuracy and efficiency of the ASES CAT in an independent test group.

Methods

A total of 2763 patients who presented to clinics for shoulder evaluation at 3 separate sites were included in this analysis (Table II). We included different collection sites, diagnoses, and

* OBERD is a software system that administers PROMs for research and clinical use. It is a product of Universal Research Solutions.

ages and both preoperative and postoperative assessments in our test set to ensure that the generalizability of the ASES CAT was being evaluated. Each patient completed a full ASES survey using OBERD as routinely used in the clinic; the algorithms constructed by the OBERD CAT system were then retrospectively applied to the actual patient responses. For each patient, when the CAT version was implemented, responses were supplied from the stored instrument rather than the live patient (which does not affect CAT functionality), and a score was calculated according to the CAT-specified algorithm.

The medical significance of the accuracy achieved by the CAT model was viewed in the context of the minimal clinically important difference (MCID) for the ASES form, which is the minimum change in score that must occur for a patient to notice a difference in functional outcome. Previous reports have identified a 12-point change in score as the threshold of noticeability for the ASES score.[29,31]

Several analyses were performed to confirm the validity of the ASES CAT score for assessing shoulder outcomes, inspired by the recommendations of Bland and Altman[4] for quantifying the agreement between 2 methods of measurement.[1] First, the mean and standard deviation of test scores were compared for the CAT vs. full ASES format. Second, the Pearson correlation coefficient (R) was calculated to determine the strength of a linear relationship between scores. Third, the intraclass correlation coefficient (ICC) was calculated to evaluate the extent to which the difference in scores was explained by inherent variability in the ASES instrument, rather than as a consequence of the full vs. CAT methods. Fourth, frequency distributions of the CAT and full ASES scores were plotted and compared to ensure similar distributions. Fifth, the distribution of the score differences (full ASES instrument minus CAT) for each case was plotted; this enabled predictions about the uncertainty of the CAT score at the individual patient level, which is the crux of the evaluation. Sixth, a Bland-Altman plot was generated to assess patterns in the differences between the CAT and full ASES scores.[13,22] For each case, the difference between the full and CAT scores was plotted against the mean of the 2 scores. Various aspects of the data could be quickly appreciated, such as whether the differences concentrated near zero or whether differences grew with the magnitude of the score. Analyses were performed with the R software suite (version 3.4.3; R Foundation for Statistical Computing, Vienna, Austria), the Python programming language (version 3.4.5; Python Software Foundation, Beaverton, OR, USA), or spreadsheets.

Results

The ASES full form has 11 questions, with 1 question about pain and 10 questions about shoulder function. The CAT model determined that the pain question needed to be asked first, no doubt because it accounts for 50% of the score. In this data set, regarding the 10 remaining questions, CAT required 55% of patients (1520 of 2763) to answer 5 questions, 23% (635 of 2763) to answer 6 questions, 11% (304 of 2763) to answer 7 questions, and 11% (304 of

Table I Scored questions and response options for American Shoulder and Elbow Surgeons Shoulder Score

	Answer
Patient self-evaluation	
Pain description	
How bad is your pain today?	VAS ranging from 0 (no pain at all) to 10 (pain as bad as it can be)
Activities–of–daily living questionnaire	
Choose the answer that indicates your ability to do the following activities:	
1. Put on a coat	Unable to do Very difficult to do Somewhat difficult Not difficult
2. Sleep on your painful or affected side	Unable to do Very difficult to do Somewhat difficult Not difficult
3. Wash back/do up bra in back	Unable to do Very difficult to do Somewhat difficult Not difficult
4. Manage toileting	Unable to do Very difficult to do Somewhat difficult Not difficult
5. Comb hair	Unable to do Very difficult to do Somewhat difficult Not difficult
6. Reach a high shelf	Unable to do Very difficult to do Somewhat difficult Not difficult
7. Lift 10 lb above the shoulder	Unable to do Very difficult to do Somewhat difficult Not difficult
8. Throw a ball overhand	Unable to do Very difficult to do Somewhat difficult Not difficult
9. Do usual work	Unable to do Very difficult to do Somewhat difficult Not difficult
10. Do usual sport	Unable to do Very difficult to do Somewhat difficult Not difficult

VAS, visual analog scale.

Table II Demographic information for patients whose full-form ASES responses were applied for validation of ASES CAT

Provider*	Diagnosis areas	Gender	Age range, yr	No. of patients
Site 1†	Rotator cuff sprain, osteoarthritis of shoulder, adhesive capsulitis of shoulder	Female: 44.17% Male: 55.82%	18-60 61-82	245 320
Site 2†	Osteoarthritis, complete rupture of rotator cuff, sprain or strain	Female: 33.73% Male: 35.79% Unknown: 30.49%	18-60 61-92	474 710
Site 3†	Osteoarthritis (overweight), osteoporosis	Female: 35.68% Male: 58.79% Unknown: 5.53%	18-60 61-82	730 284
Overall		Female: 34% Male: 49% Unknown: 17%	18-92	2763

ASES, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *CAT*, computerized adaptive testing.
* Identifies source of patients.
† Identifies clinic of particular surgeon or author.

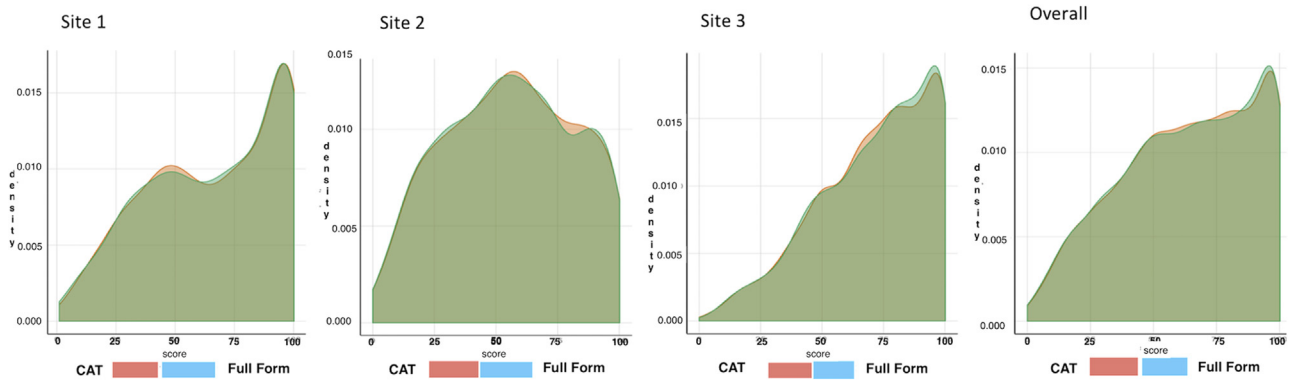

Figure 1 Distribution of full American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form scores (*blue*) overlaid with distribution of computerized adaptive testing scores (*orange*) by site and overall. *Green* shows where the two scores are the same. Although score distributions varied considerably for each site (as expected, given the range of respondent demographic characteristics), the computerized adaptive testing score distributions were nearly identical matches to the full-form distributions for each site and overall.

2763) to answer 8 questions. Counting the pain question, this produced an average of 6.6 questions per case.

The mean CAT score was 0.14 points higher than the mean full-form score, with a similar spread in values (63.53 ± 26.51 vs. 63.67 ± 26.42); both the Pearson correlation coefficient and ICC were higher than 0.99 (Table II). There was nearly complete identity between the distributions of the CAT scores and the full ASES scores (Fig. 1). In addition, the distribution of the differences in each pair of CAT and full-test ASES scores was clustered around zero. The maximum difference between CAT and full scores was 12.6 points, and only 2 cases (0.07%, 2 of 2763) had differences greater than the ASES MCID of 12 points. The CAT result was within 5 points of the full test result in 95% of cases (2625 of 2763) (Fig. 2).

Examination of the Bland-Altman plot (Fig. 3) showed that the difference between CAT and full ASES scores was independent of the overall score (eg, the differences did not show a percentage error). Furthermore, there was no bias of the CAT in either a positive direction (greater difference at higher scores) or negative direction (less difference at lower scores). Summary statistics are provided in Table III.

Discussion

The ASES shoulder score is one of the most common PROMs used in shoulder research[12,19,26] and has been validated for such use.[17,20,23] Many institutions have adopted routine, periodic collection of the ASES score to provide a consistent assessment of clinical success, a basis for scientific investigations of treatment methods, and documentation of the functional status of the shoulder. This commitment requires significant institutional resources and patients' time. Therefore, a new CAT method has been developed to lessen the burden. Here, we show that application of CAT to the ASES questionnaire has a negligible impact on the outcome score while significantly minimizing the number of required questions.

Two metrics are essential for intelligent use of the ASES score as a measuring instrument and for interpretation of

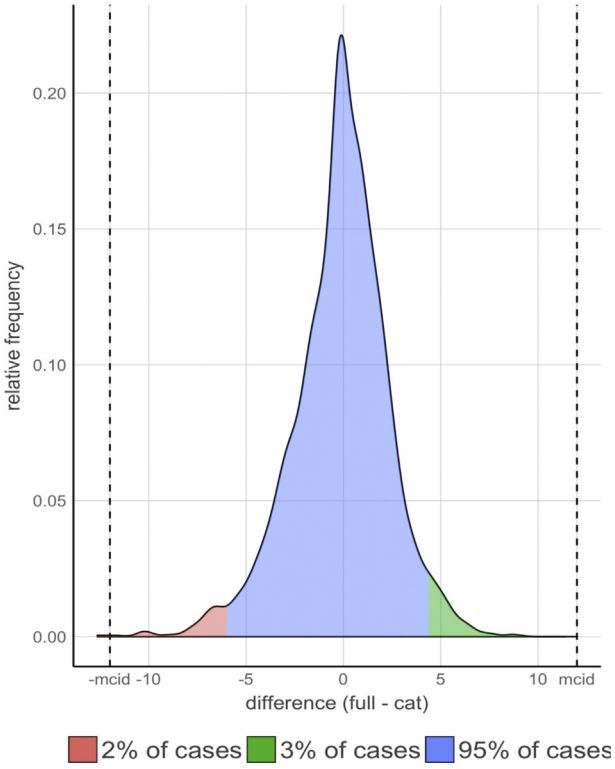

Figure 2 Distribution of differences between scores. Ninety-five percent of the computerized adaptive testing scores were within 5 points of the full American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form scores (which is less than half of the minimal clinically important difference [*mcid*] of 12 points). Most of the differences were clustered around zero, with very little bias toward higher or lower computerized adaptive testing scores relative to full-form scores.

Figure 3 Bland-Altman plot of difference between computerized adaptive testing (*CAT*) and full American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form (ASES) scores vs. mean of the 2 scores for each case. For most cases, the magnitude of difference between CAT and full-form ASES scores was lower than 5 points, with very few cases showing a difference of 10 points or higher. The minimal clinically important difference (*mcid*) for the ASES instrument is 12 points, indicating that use of the ASES CAT in place of the full form would not affect clinical interpretation of outcomes. In addition, the difference between scores is independent of the overall score (no bias toward larger differences at higher scores).

concordance between the CAT and the full-form versions: repeatability and the MCID. ''Repeatability'' refers to measurement error and is usually addressed by a test-retest strategy; the instrument is administered twice to each individual in a representative group, and consistency in scores is expressed as the ICC.[15] The ICC for test-retest reliability of the full ASES form was reported to be 0.84,[19] which is viewed as strong agreement between the initial and retest scores and justification that the ASES instrument is reliable. In comparison, the ICC for full-form vs. ASES CAT shoulder scores was 0.99. As such, the negligible difference seen with application of the CAT methods for the ASES score is substantially less than would be seen by simply retesting with the full test.

The MCID refers to the capability of an instrument to detect changes in clinical function and pain. The MCID is typically established by an ''anchor'' method,[25,28,32] in which a group of patients are asked (at the same follow-up interval) whether their condition has noticeably improved since treatment. The average change in score (or, in some studies, the minimum, 25th percentile, or median score) for all patients who indicated improvement is taken to be the MCID. On the basis of a number of such studies,[29,31] we considered a change in ASES score of 12 points to indicate true improvement for most cohorts, and this value was used as the MCID. Because the MCID reflects the patient's experience, we believe that the most appropriate way to understand the inherent uncertainty in the ASES instrument is to consider 2 measurements to be indistinguishable within the margin of error of the instrument when the observed difference is less than the MCID. Our data showed that the CAT and full ASES scores were within the MCID for 99.75% of cases and were within one-half of the MCID for 95% of cases.

Our analysis showed that OBERD CAT accurately reproduces the entire frequency distribution of the ASES full-form scores, not merely summary statistics such as the mean and standard deviation. This close match in score distributions is particularly important given that the distributions were not normal. Furthermore, the inherent variability of the ASES shoulder survey, as reflected by the MCID of 12 points for the ASES score,[18,29,31] is substantially greater than the uncertainty introduced by the CAT prediction. The paired score differences were symmetrically and randomly distributed (ie, clustered around zero, with only insignificant bias toward the CAT or full-form score being consistently higher), and only 0.07% of the differences were greater than the MCID.

Table III Summary statistics for full-form vs. ASES CAT scores by site and overall

	No. of patients	Full ASES scores, mean (SD)	ASES CAT scores, mean (SD)	*R*	ICC
Site 1	565	66.02 (27.72)	66.23 (27.65)	>0.99	>0.99
Site 2	1184	55.43 (26.33)	55.75 (26.30)	>0.99	>0.99
Site 3	1014	71.59 (23.10)	71.49 (23.06)	>0.99	>0.99
Overall	2763	63.53 (26.51)	63.67 (26.42)	>0.99	>0.99

ASES, American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form; *CAT*, computerized adaptive testing; *SD*, standard deviation; *R*, Pearson correlation coefficient; *ICC*, intraclass correlation coefficient.

The findings of this study must be considered in light of certain difficulties and limitations. In a test-retest activity or in the comparison of instruments that address the same topic, it is usually important to expose each instrument to equivalent subjects, especially when patient-level comparisons are needed, as in this study. Typically, in such cases, both instruments will be administered to each subject. However, they cannot be administered at the same time, posing the following difficulty[8]: If both instruments are administered in the same setting, the patient may be remembering his or her answers rather than providing an independent information. If time is inserted between administrations, it must be long enough to mitigate the memory effect, but if such a period is too long, the patient's condition may have changed. Balancing these effects is generally left to the judgment of the investigator. In the present case, the particular relationship of the 2 instruments at hand permitted this problem to be avoided: A single administration of the standard form to a patient allowed calculation of the CAT score as well.

To wit, in our study, instead of asking the patient to answer the CAT, the response is delivered from his or her full form. The CAT behavior is not affected by the origin of the response. Moreover, a subject's environment does not differ between a CAT administration and administration of any other type of questionnaire in OBERD. Although a CAT is required to show 1 question at a time, because it does not know what the next question will be until it obtains the current answer, OBERD also presents 1 question at a time for all instruments as a feature of its interface design. The subject sees no difference whether the question is taken from a fixed list, selected at random, or provided by any other selection scheme, such as a CAT system.

Because the CAT and a list will generally present the questions in a different order, there remains the important matter of whether the order is significant. This is actually an issue for any PROM because, in modern testing theory (item response theory [IRT]), a measure is distinguished from other types of surveys in part by the requirement that each question can be regarded as an independent observation unaffected by other questions. When the ASES instrument was constructed, methods designed to ensure this were rarely used for PROMs, and the validity of the ASES instrument has been questioned on these grounds.[8] These doubts were removed when studies of the psychometric properties of the ASES instrument found excellent agreement with the goals of IRT.[17] In addition, 2 of the present authors (O.R.P. and P.S.) had examined ASES data sets in terms of Rasch theory,[8] the progenitor and strictest form of IRT,[5,9] even before the construction of the CAT (unpublished data, December 2016). Satisfactory results were found for the usual indicators such as "infit" and "outfit" statistics, as well as good person and item separation. Of special importance, the residual correlations were all in the range (<0.3) generally held[16,23,27] to indicate independence—the final requirement to establish that the same set of patient responses can be used for both instruments.

It should be noted that the ASES development process was exemplary in the following regard. The research committee examined existing questionnaires, considered a large number of questions, prepared and refined a first draft, circulated this draft at the annual meeting, evaluated and incorporated the comments and suggestions obtained, circulated a second draft for comment at the next annual meeting, and only then produced the final result. This scrutiny by so many reviewers suggests a low likelihood of any remaining ambiguities or infelicities. The result is shown in Table I, in which the simple structure and short clear questions can be seen directly, without need of measurement theory.

A limitation of the present work is that neither this study nor the training of the ASES CAT considered differences in level of agreement that might exist between cohorts differing in demographic details, diagnosis details, treatment details, or comorbidities. Results from prospective studies that examine such cohorts can be readily incorporated into the CAT system to improve its performance should weaknesses be revealed. It is recommended that such factors be evaluated for use as explanatory variables in future CAT versions even though they are not considered by the standard ASES instrument.

Conclusion

The analyses performed here provide strong evidence for reliability of the OBERD CAT. This study shows that the CAT system can be used interchangeably with the

full ASES instrument for both research and clinical purposes. Furthermore, there is a significant reduction in question burden with CAT, which should translate into saved patient time and practice resources. Our surgeons and those in clinical practice who obtain PROMs often note that form completion, whether paper or digital format, drops off with repeated assignments, precluding the accumulation of valuable longitudinal outcome data. The fatigue factor will decrease with the abbreviated and more personalized CAT format, thus allowing improved engagement with the ASES outcome score. The goal of 80% retention of respondents over the long term then becomes more achievable. The ASES CAT can also improve the rate of obtaining complete, validated preoperative scores to allow comprehensive prospective scoring for future clinical trials.

Disclaimer

No outside sources contributed funding or grants to support the study or were involved in data collection, data analysis, or the preparation or editing of this manuscript.

Otho R. Plummer is Chief Scientific Officer of Universal Research Solutions–OBERD.

Joseph A. Abboud receives research support from DePuy Synthes, Zimmer, Tornier, Arthrex, OREF, Integra, and OrthoSpace; receives royalties from Wolters Kluwer Health–Lippincott Williams & Wilkins, Integra, DJO, Cayenne, and Globus; is a paid speaker for Tornier; owns stock in Aevumed, Parvizi Surgical Innovation, and Marlin Medical Alliance; is on the scientific advisory board of Mininvasive and the board of directors of Mid Atlantic Shoulder and Elbow Society; is a cofounder of Shoulder JAM; and is a shareholder in OBERD.

John-Erik Bell is a cofounder of Shoulder JAM.

Anand M. Murthi receives research support from DePuy Synthes, Arthrex, OREF, OrthoSpace, and Smith & Nephew; receives royalties from Wolters Kluwer Health–Lippincott Williams & Wilkins, Ignite Orthopaedics, and Globus; is a paid consultant for Ignite Orthopaedics; is on the advisory board of Catalyst Scientific and the board of directors of Mid Atlantic Shoulder and Elbow Society; is a cofounder of Shoulder JAM; is a founder and executive board member of the Association of Clinical Elbow and Shoulder Surgeons; and is on the editorial boards of *Current Orthopaedic Practice* and the *Journal of Orthopaedics*.

Anthony A. Romeo receives support from Arthrex, DJO Surgical, *Orthopedics Today*, Ossur, SAGE, Saunders/Mosby-Elsevier, SLACK, Smith & Nephew, Wolters Kluwer Health, and Lippincott Williams & Wilkins.

Priyanka Singh is an employee of the Data Science Unit of Universal Research Solutions–OBERD.

The other author, his immediate family, and any research foundations with which he is affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.

References

1. Altman DG, Bland JM. Measurement in method: the analysis of method comparison studies. Statistician 1983;32:307-17.
2. Andrawis J, Akhavan S, Chan V, Lehil M, Pong D, Bozic KJ. Higher preoperative patient activation associated with better patient-reported outcomes after total joint arthroplasty. Clin Orthop Relat Res 2015; 473:2688-97. http://doi.org/10.1007/s11999-015-4247-4
3. Angst F, Schwyzer HK, Aeschlimann A, Simmen BR, Goldhahn J. Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and its short version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society standardized shoulder assessment form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ), and Western Ontario Shoulder Instability Index (WOSI). Arthritis Care Res (Hoboken) 2011;63:S174-88. http://doi.org/10.1002/acr.20630
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
5. Boone WJ. Rasch analysis for instrument development: why, when, and how? CBE Life Sci Educ 2016;15:rm4. http://doi.org/10.1187/cbe.16-04-0148
6. Brogan AP, DeMuro C, Barrett AM, D'Alessio D, Bal V, Hogue SL. Payer perspectives on patient-reported outcomes in health care decision making: oncology examples. J Manag Care Spec Pharm 2017;23: 125-34. http://doi.org/10.18553/jcmp.2017.23.2.125
7. Carragee EJ, Cheng I. Minimum acceptable outcomes after lumbar spinal fusion. Spine J 2010;10:313-20. http://doi.org/10.1016/j.spinee.2010.02.001
8. Chien TW, Wu HM, Wang WC, Castillo RV, Chou W. Reduction in patient burdens with graphical computerized adaptive testing on the ADL scale: tool development and simulation. Health Qual Life Outcomes 2009;7:39. http://doi.org/10.1186/1477-7525-7-39
9. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ 2003;37:739-45. http://doi.org/10.1046/j.1365-2923.2003.01587
10. Franklin JM, Gebski V, Poston GJ, Sharma RA. Clinical trials of interventional oncology-moving from efficacy to outcomes. Nat Rev Clin Oncol 2015;12:93-104. http://doi.org/10.1038/nrclinonc.2014.199
11. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol 2009;36:2061-6. http://doi.org/10.3899/jrheum.090358
12. Gagnier JJ, Robbins C, Bedi A, Carpenter JE, Miller BS. Establishing minimally important differences for the American Shoulder and Elbow Surgeons score and the Western Ontario Rotator Cuff Index in patients with full-thickness rotator cuff tears. J Shoulder Elbow Surg 2018;27:e160-6. http://doi.org/10.1016/j.jse.2017.10.042
13. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015;25:141-51. http://doi.org/10.11613/BM.2015.015
14. Hossack T, Woo H. Validation of a patient reported outcome questionnaire for assessing success of endoscopic prostatectomy. Prostate Int 2014;2:182-7. http://doi.org/10.12954/PI.14066

15. Jenkinson C, Morley D. Patient reported outcomes. Eur J Cardiovasc Nurs 2016;5:112-3. http://doi.org/10.1177/1474515115623407
16. Kim HY. Statistical notes for clinical researchers: evaluation of measurement error 1: using intraclass correlation coefficients. Restor Dent Endod 2013;38:98-102. http://doi.org/10.5395/rde.2013.38.2.98
17. Kocher MS, Horan MP, Briggs KK, Richardson TR, O'Holleran J, Hawkins RJ. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. J Bone Joint Surg Am 2005;87:2006-11. http://doi.org/10.2106/JBJS.C.01624
18. Maltenfort M, Díaz-Ledezma C. Statistics in brief: minimum clinically important difference—availability of reliable estimates. Clin Orthop Relat Res 2017;475:933-46. http://doi.org/10.1007/s11999-016-5204-6
19. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. J Shoulder Elbow Surg 2002;11:587-94. http://doi.org/10.1067/mse.2002.127096
20. Moser AD, Knaut LA, Zotz TG, Scharan KO. Validity and reliability of the Portuguese version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form. Rev Bras Reumatol 2012;52:348-56. http://doi.org/10.1590/S0482-50042012000300005
21. Movsas B, Hunt D, Watkins-Bruner D, Lee WR, Tharpe H, Goldstein D, et al. Can electronic web-based technology improve quality of life data collection? Analysis of Radiation Therapy Oncology Group 0828. Pract Radiat Oncol 2014;4:187-91. http://doi.org/10.1016/j.prro.2013.07.014
22. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. Br J Anaesth 2007;99:309-11. http://doi.org/10.1093/bja/aem214
23. Piitulainen K, Paloneva J, Ylinen J, Kautiainen H, Häkkinen A. Reliability and validity of the Finnish version of the American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section. BMC Musculoskelet Disord 2014;15:272. http://doi.org/10.1186/1471-2474-15-272
24. Richards RR, An K-N, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG, et al. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg 1994;3:347-52.
25. Tashjian RZ, Hung M, Keener JD, Bowen RC, McAllister J, Chen W, et al. Determining the minimal clinically important difference for the American Shoulder and Elbow Surgeons score, Simple Shoulder Test, and visual analog scale (VAS) measuring pain after shoulder arthroplasty. J Shoulder Elbow Surg 2017;26:144-8. http://doi.org/10.1016/j.jse.2016.06.007
26. Unger RZ, Burnham JM, Gammon L, Malempati CS, Jacobs CA, Makhni EC. The responsiveness of patient-reported outcome tools in shoulder surgery is dependent on the underlying pathological condition. Am J Sports Med, 2018. 363546517749213, http://doi.org/10.1177/0363546517749213
27. Vrotsou K, Cuéllar R, Silió F, Rodriguez MÁ, Garay D, Busto G, et al. Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. Health Qual Life Outcomes 2016;14:147. http://doi.org/10.1186/s12955-016-0552-1
28. Werner BC, Chang B, Nguyen JT, Dines DM, Gulotta LV. What change in American Shoulder and Elbow Surgeons score represents a clinically important change after shoulder arthroplasty? Clin Orthop Relat Res 2016;474:2672-81. http://doi.org/10.1007/s11999-016-4968-z
29. Werner BC, Wong AC, Mahony GT, Craig EV, Dines DM, Warren RF, et al. Causes of poor postoperative improvement after reverse total shoulder arthroplasty. Shoulder Elbow Surg 2016;25:217-22. http://doi.org/10.1016/j.jse.2016.01.002
30. Wolfe F, Michaud K. Proposed metrics for the determination of rheumatoid arthritis outcome and treatment success and failure. J Rheumatol 2009;36:27-33. http://doi.org/10.3899/jrheum.080591
31. Wong SE, Zhang AL, Berliner JL, Ma CB, Feeley BT. Preoperative patient-reported scores can predict postoperative outcomes after shoulder arthroplasty. J Shoulder Elbow Surg 2016;25:913-9. http://doi.org/10.1016/j.jse.2016.01.029
32. Wright A, Hannon J, Hegedus EJ, Kavchak AE. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). J Man Manip Ther 2012;20:160-6. http://doi.org/10.1179/2042618612Y.0000000001
33. Wylie JD, Beckmann JT, Granger E, Tashjian RZ. Functional outcomes assessment in shoulder surgery. World J Orthop 2014;5:623-33. http://doi.org/10.5312/wjo.v5.i5.623