An examination of bias in shoulder scoring instruments among healthy collegiate and recreational athletes

Mark R. Brinker, MD,[a,b,c] Jeffrey S. Cuomo, MD,[a] George J. Popham, MD,[a] Daniel P. O'Connor, MS,[c] and Robert L. Barrack, MD,[a] *New Orleans, La, and Houston, Tex*

This investigation examined whether gender, activity level, hand dominance, or age displayed bias for 5 common shoulder scores among 120 healthy collegiate or recreational athletes. Data were collected for 5 instruments: Constant-Murley, UCLA rating, pre-1994 American Shoulder and Elbow Surgeons, Shoulder Pain and Disability Index, and Oxford. Normalized total and subscale scores and effect sizes were analyzed to determine how each demographic variable affected the scores. The normalized scores for the 5 instruments were not equivalent. The normalized total scores were significantly lower for the Constant-Murley score ($P < .0001$) compared with those for the other instruments. Men had significantly higher Constant-Murley scores, primarily because of the strength subscale. The Constant-Murley score also displayed gender and age biases. The other instruments had mostly negligible variance attributable to gender, activity level, hand dominance, or age. In the absence of a universal validated method, shoulder scoring instruments should be carefully chosen to match the population and the purpose of the study. (J Shoulder Elbow Surg 2002;11:463-9.)

INTRODUCTION

Numerous instruments have been designed to assess the preoperative and postoperative status of patients undergoing reconstructive procedures of the shoulder.* Because these instruments are convenient, they are widely used for clinical investigations of shoulder conditions and reconstructive procedures.[5,36] At present, however, there is no single standard for a shoulder rating system.[8,24,32,36]

From the Department of Orthopaedic Surgery, Tulane University School of Medicine, New Orleans, La[a], and the Fondren Orthopedic Group LLP[b] and the Joe W. King Orthopedic Institute,[c] Texas Orthopedic Hospital, Houston.[a]

Reprint requests: Mark R. Brinker, MD, Fondren Orthopedic Group LLP, Texas Orthopedic Hospital, 7401 S Main, Houston, TX 77030.

1058-2746/2002/$35.00 + 0 **32/1/126209**
doi:10.1067/mse.2002.126209

Recent concerns related to the rising cost of health care, evaluation of cost-effectiveness, and determination of treatment effectiveness have resulted in interest in outcome studies.[19,21,23] As a result, the need for validated instruments as a means of measuring outcomes has been emphasized,[21,23,25] and the need for normative data has been recognized.[6,7,34]

Administration of questionnaires to healthy individuals to establish normative data may reveal that scores vary significantly because of factors other than shoulder pain, disability, or impairment.[36] These nuisance factors introduce systematic variance that is not attributable to the domain of interest. This effect is known as instrument bias. If bias is present, normative scores need to be stratified, and the scores of patients who are injured or recovering need to be interpreted relative to their subgroup matched to the respective biasing factor.

The purpose of this investigation was to examine the effect of gender, activity level, age, and hand dominance on scores for 5 shoulder scoring instruments[4,9,11,14,40] among healthy collegiate and recreational athletes. We examined a group of subjects who had no prior history of shoulder or neck injuries, problems, or surgeries, in order to eliminate the effects of pathologic conditions or treatments. The demographic variables were selected based on their potential effect on physical function in patient and healthy populations, as established by previous research[3,16,29,33,36-39] and by consensus of a research committee of orthopaedic surgery faculty at our institution.

Theoretically, healthy subjects should receive virtually the same high score on all instruments if those instruments are valid for evaluating normal function. Recovery from an injury or surgery could then be inferred from achievement of a high score. Conversely, should an instrument or subscale differ among healthy persons according to some demographic factor (eg, gender, age), the instrument or subscale displays bias. Should such bias exist, analysis of the subscale and instrument should reveal the bias in interpretation of scores. To date, the presence of bias due to demographic characteristics for the

*References 4, 9, 11, 12, 14, 24, 26, 27, 30, 31, 35, 39.

total and subscale scores of these shoulder scoring instruments has not been widely examined.

The hypotheses were that the different shoulder scoring instruments would represent the same underlying dimensions of shoulder function among healthy, physically active subjects and that the shoulder scores would vary by gender, hand dominance, differences in activity level, and age. To test these hypotheses, normalized (ie, adjusted for scale differences) total and subscale scores of 4 groups of healthy subjects, who differed with respect to activity level and age, were compared. Scores of men and women were contrasted within and across the groups, and the effect of hand dominance was tested within and across the groups.

MATERIALS AND METHODS

This study was approved by our institutional review board. A total of 142 volunteers were recruited without prior knowledge of the intent, purpose, or methods of the study. Four groups of volunteers were recruited according to age and activity differences. Group I consisted of National Collegiate Athletic Association (NCAA) throwing athletes (baseball, volleyball, and tennis), group II consisted of NCAA nonthrowing athletes (basketball and soccer), group III consisted of recreational athletes aged under 40 years, and group IV consisted of recreational athletes aged 40 years or older. This was a convenience sample drawn from the relevant intercollegiate teams and from the recreational and intramural programs of our institution. Sampling continued until equal group sizes (n = 30) and equal gender distribution within groups were obtained.

Twenty-two volunteers were excluded from the study because they (1) had previously sought treatment for a shoulder complaint, (2) had a known abnormal condition of the shoulder or upper extremity, (3) had a history of shoulder or neck surgery, or (4) had a prior injury or abnormal condition of the upper extremity or cervical spine. One hundred twenty volunteers met the eligibility criteria. There were 60 men and 60 women with a mean age of 28.8 years (range, 17-81 years).

To obtain demographic information, a detailed medical history was obtained for all 120 subjects. No subject in this investigation had a major medical condition, as defined by Brinker et al,[6,7] and all subjects were above poverty-level income.[13]

Physical examinations were performed at 4 examination-specific stations by 2 senior-level orthopaedic residents and 2 senior medical students. Each of the 4 examiners performed only one of the following tasks to avoid introducing intertester error. Range-of-motion measurements were made with a goniometer and were recorded to the nearest 5°. Muscular strength was measured with a tensiometer, as described by Constant and Murley.[9] Strength was also assessed by a standard manual motor testing scale (0, absent; 5, normal). Shoulder stability was evaluated as described by Barrett et al.[4]

Data were recorded on standardized flow sheets so that shoulder scores could be calculated for the following 5 instruments: the shoulder score of Constant and Murley[9] (Constant-Murley score); the University of California at Los Angeles rating for the shoulder (UCLA rating)[1,14]; the "pre-1994" American Shoulder and Elbow Surgeons' self-evaluation form for the shoulder as used by Barrett et al,[4] Gartsman,[18] Romeo et al,[32] and Williams et al[39] (pre-1994 ASES form); the Shoulder Pain and Disability Index (SPADI)[40]; and the Oxford shoulder questionnaire of Dawson et al[11] (Oxford questionnaire). These shoulder scoring instruments were selected for their prevalence in the published literature. A recent MEDLINE search (August 1, 2001) of "shoulder" and "score" as keywords revealed that 233 (40.6%) of 573 articles located used one or more of the instruments included in the present study.

The Constant-Murley score was developed to evaluate functional recovery of the shoulder after injury or surgery by means of a standardized method.[9,10] Conboy et al[8] reported that the Constant-Murley score was highly unreliable among patients with shoulder instability. The total Constant-Murley score ranges from 0 to 100 points, with higher scores indicative of better function. The UCLA rating was first described in a study of patients undergoing total shoulder arthroplasty.[1] The current version was presented in a report on rotator cuff repair[14] and has since been used in patients with shoulder instability.[17,24]

The pre-1994 ASES form was developed primarily as a standard method of evaluation for patients with shoulder pathology rather than to produce a score.[31,32,39] Because numerical ratings are used for each pre-1994 ASES form subscale,[4,32] a score can be computed by summing within the subscales and summing subscale scores for a total score.[39]

The SPADI was developed to measure current subjective perception of shoulder function in persons with painful shoulder syndromes.[31,40] The original SPADI used visual analog scales to rate 13 items related to shoulder-specific pain (5 items) and disability (8 items).[31] The current form of the SPADI was used in the present study.[40] This form required subjects to rate the 13 items on a scale ranging from 0 to 11, with 11 indicating the greatest severity or disability.[40] The SPADI total and subscale scores are reported as the proportion of items to which the subject responded affirmatively.

The 12-item Oxford questionnaire was developed to evaluate the outcome of shoulder surgery, excluding patients with instability.[11] The 12 questions are used to rate pain or function. They are each scored 1 to 5 points, with higher scores indicating greater pain or disability. The sum of the item scores is the total score, which ranges from 12 to 60 points.[11]

Total and subscale scores were calculated for each of the 5 scoring instruments (Table I). Scores were calculated for the left and right shoulders for all subjects. Total and subscale scores were normalized by dividing the observed score by the maximum possible score to obtain a proportion, similar to the method of scoring for the SPADI. To compare among the 4 groups, between genders, and by dominant extremity, a summary score was calculated as the total of the individual scores of the 5 shoulder instruments, as well as the total of the individual scores of similar subscales from the 5 instruments.

Means and SDs were computed for the total and subscale scores of each shoulder scoring instrument. To quan-

Table I Total and subscale scores of the 5 instruments

Scoring systems (total points possible)	Subscale scores (total points possible)
Constant-Murley (100)	Subjective (35)
	Pain (15)
	Function (20)
	Objective (65)
	Range of motion (40)
	Strength (25)
UCLA (30)	Subjective (20)
	Pain (10)
	Function (10)
	Objective (10)
	Range of motion (5)
	Strength (5)
ASES (100)	Subjective (65)
	Pain (5)
	Function (60)
	Objective (35)
	Instability (15)
	Strength (20)
SPADI (100)	Subjective (100)
	Pain (50)
	Disability (50)
	Objective (0)
	NA
Oxford (60)	Subjective (60)
	Pain (20)
	Function (40)
	Objective (0)
	NA

NA, Not applicable.

tify the magnitude of the effect of each variable on the various subscale and total instrument scores, effect sizes (proportion of total score variance explained) were computed. The use of effect size rather than simple significance testing allows an analysis of how much influence an independent variable has on the dependent variable, allowing a judgment about whether the effect is clinically relevant.[15,20,22] Effect size in the present study was computed as η^2, which is analogous to r^2 or multiple R^2 in that it indicates the proportion of variance in shoulder instrument score accounted for by the variables. It is calculated as $\eta^2 = 1 - (SS_{within}/SS_{total})$, where SS_{within} is the sum of squares within subjects and SS_{total} is the total sum of squares. The advantage of η^2 over r^2 or multiple R^2 is that it can be used to describe both linear and nonlinear relationships. Repeated-measures analysis of variance and Newman-Keuls tests were used to obtain the necessary sums of squares and to estimate the statistical significance of the effect sizes.

RESULTS

Table II shows subject demographics. Group IV (mean age, 52.4 years) was significantly older than each of the other groups ($P < .005$). Because of research design, men and women did not differ significantly by age either within or across groups. Table III shows the means and ranges of normalized total and subscale scores by group for each of the 5 shoulder scoring instruments. Table IV shows the effect sizes of each of the variables (group, gender, and hand dominance) on the instruments' total and subscale scores.

Across groups, the Constant-Murley score yielded normalized total scores that were significantly lower than all 4 other instruments ($P < .005$). The SPADI total scores were significantly higher than both the Oxford questionnaire total scores ($P = .003$) and the UCLA rating total scores ($P = .001$). Newman-Keuls analyses revealed that the lower Constant-Murley total scores across groups could be attributed principally to lower scores on its strength subscale, with little variance present in the other Constant-Murley subscales.

The largest effect in this sample was that of gender on the Constant-Murley score. Gender accounted for 57% (9.1 out of a possible 100 points, $P = .0001$) of the variance observed in the total Constant-Murley score. Most of this was explained by the strong gender effect in the strength subscale of the Constant-Murley score, accounting for 66% of its observed variance ($P = .0001$). On average, men scored 8.3 points higher (out of 25 points) than did women on the strength subscale of the Constant-Murley score. Figure 1 shows that the gender effect on the strength subscale of the Constant-Murley score was consistent across all 4 groups. Gender also significantly affected the range of motion subscale of the Constant-Murley score ($P = .0005$) but accounted for only 4 points (out of 40 points) of the observed variance in that subscale score. Thus, the gender effect on the Constant-Murley score was primarily a result of bias in the strength subscale.

Gender also had a significant effect on the range of motion subscale score of the UCLA rating ($P = .004$) and the pain subscale of the Oxford questionnaire ($P = .04$) but accounted for only 7% and 3%, respectively, of the observed variance in these subscales. Expressed in scale units, men scored 0.15 higher (out of 5 points) on the range of motion subscale of the UCLA rating and men rated less than 0.5 lower (out of 50 points) on the pain subscale of the Oxford questionnaire.

Group differences had several small but significant effects on the shoulder scoring instruments. The largest effects for group were observed in the range of motion, strength, and pain subscales of the Constant-Murley score. The magnitudes of these effects were much smaller than those observed for gender. For example, the largest group difference on the range of motion subscale, which was between groups II and III, demonstrated a difference of only 1.2 points (out of 40 points). The largest group difference on the pain subscale, between groups II and IV, was 0.9 points (out of 15 points). The largest group difference for

Table II Demographics of the activity groups (N = 120)

Group	Gender	n	Mean age (range)	Sport
Group I: NCAA throwing athletes	Men	15	18.9 (18-20)	Baseball
	Women	15	19.5 (18-21)	Tennis or volleyball
Group II: NCAA nonthrowing athletes	Men	15	19.8 (17-22)	Basketball
	Women	15	18.7 (18-21)	Soccer
Group III: Recreational athletes <40 y	Men	15	26.7 (24-31)	NA
	Women	15	21.8 (18-26)	NA
Group IV: Recreational athletes ≥40 y	Men	15	51.3 (40-81)	NA
	Women	15	53.4 (40-80)	NA

NA, not applicable.

Table III Means (ranges) of normalized total scores for 5 instruments

Shoulder score	Group I	Group II	Group III	Group IV	All groups
Constant-Murley	92.0% (69.3%-100%)	94.1% (79.7%-100%)	94.6% (86.5%-100%)	90.4% (71.9%-100%)	92.7% (69.3%-100%)
UCLA	97.6% (75.0%-100%)	99.0% (89.3%-100)	98.1% (92.9%-100%)	97.1% (85.7%-100%)	98.0% (75.0%-100%)
ASES	98.7% (95.0%-100%)	98.2% (94%-100%)	98.9% (95.5%-100%)	98.1% (88.5%-100%)	98.5% (88.5%-100%)
SPADI	99.2% (91.6%-100%)	100.0% (100%-100%)	994.% (92.7%-100%)	98.2% (77.6%-100%)	99.2% (77.6%-100%)
Oxford	97.5% (86.5%-100%)	99.0% (97.9%-100%)	98.8% (91.7%-100%)	97.4% (77.1%-100%)	98.2% (86.5%-100%)
Total for all instruments	97.0% (87.5%-100%)	98.1% (94.1%-99.6%)	97.9% (94.1%-100%)	96.2% (80.2%-100%)	95.4% (69.3%-100%)

Table IV Effect sizes for each variable on the instruments' total and subscale scores (N = 120 subjects)

Shoulder instrument	Subscale	Group	Gender	Dominant extremity
Constant-Murley	Total score	0.07*	0.57*	0.03
	Pain	0.07*	0	0.01
	Function	0	0	0
	Range of motion	0.12*	0.10*	0
	Strength	0.09*	0.66*	0.03
UCLA	Total score	0.03	0	0.01
	Pain	0.07*	0	0
	Function	0.02	0	0
	Range of motion	0.05	0.07*	0
	Strength	0.03	0.01	0
ASES	Total score	0.03	0.01	0
	Pain	0.07*	0	0
	Function	0.10*	0	0
	Instability	0.09*	0.01	0
	Strength	0	0	0
SPADI	Total score	0.06	0	0.01
	Pain	0.05	0	0
	Function	0.08*	0	0
Oxford	Total score	0.05	0	0.01
	Pain	0.04	0.03*	0
	Function	0.03	0.01	0

Numbers represent effect size (η^2).
Zero indicates that less than 1% of the variance in the respective score is accounted for by the variable.
*$P < .05$ (statistically significant).

strength, between groups II and III, was only 3.8 points (out of 25 points). Figure 1 shows that, although the groups differed slightly, gender rather than group was the dominant factor contributing to bias on the strength subscale of the Constant-Murley score. Similarly, group contributed significant but small effects for the pain subscale of the UCLA rating; the function subscale of the SPADI; and the pain, function, and instability subscales of the pre-1994 ASES form (Table IV).

Group IV (recreational athletes aged 40 years or older), in whom the mean age was more than 28 years older than in the other 3 groups, had a significantly lower mean Constant-Murley score ($P < .001$),

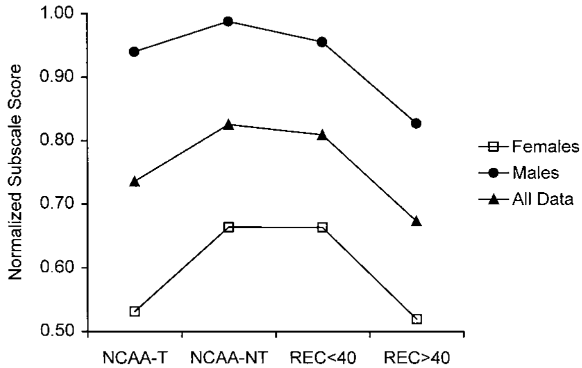

Figure 1 Normalized mean Constant-Murley strength subscale scores by gender and group. *NCAA-T,* Collegiate throwing athletes; *NCAA-NT,* collegiate nonthrowing athletes; *REC<40,* recreational athletes aged under 40 years; *REC>40,* recreational athletes aged 40 years or older.

again primarily as a result of a significantly lower mean strength subscale Constant-Murley score ($P < .001$). Group IV did not differ significantly from the other groups on any other total or subscale shoulder instrument score.

Finally, the effects of dominant extremity were trivial, accounting for less than 3% of the observed variance for any subscale or total score. They were not significant for the 5 shoulder scoring instruments used.

DISCUSSION

Our comparison of normalized scores (proportion of total possible points) demonstrated that only the Constant-Murley score systematically differed from the other instruments in scoring shoulder function among healthy, physically active persons. The SPADI also differed significantly from two other subscales. These shoulder scoring instruments are intended to measure shoulder functional outcomes but do not measure the same parameters in the same way. In other words, a normalized score of 98 out of 100 points on one instrument does not translate into a score of 98 out of 100 points on another. Therefore, the results of surgical outcomes for various instruments do not readily lend themselves to intersystem comparison.

Hand dominance had no significant effect on any total or subscale score for any of the 5 instruments. Significantly higher normalized total scores were observed for men. Similarly, a significant gender effect was observed for the normalized total scores, range-of-motion scores, and strength scores for the Constant-Murley instrument. No effect of gender was observed for the other 4 instruments.

The effect of gender appears to be unique to the Constant-Murley score, and this must be appreciated when the results of outcome studies that report Constant-Murley scores are reviewed. The most gender bias was present on the strength subscale of the Constant-Murley score, similar to previous findings.[2,8] The mean difference in Constant-Murley strength subscale score in our subjects ranged from 7.3 points for the recreational athletes aged under 40 years (group III) to 9.8 for the NCAA throwing athletes (group I). The mean gender difference in Constant-Murley strength score across all groups between men and women was 8.3 points. These differences stem from the method by which strength is rated to arrive at the Constant-Murley score. An absolute measure of strength is used (pounds of force) rather than a relative measure (eg, gender-specific norms). It is well known that women, on average, cannot produce the same muscular force as men because of differences in absolute lean body mass.[28] Thus, any measure of absolute strength is biased in favor of men.

A significant effect of group on normalized total scores was observed for the Constant-Murley score, with group IV scoring below the others. Because group IV was older than the other 3 groups, the differences in scores may have resulted from an age effect. This age-group effect can also be attributed to the strength subscale of this system. In their original report, Constant and Murley[9] specifically stated that the strength subscale of the Constant-Murley score "diminishes with advancing age" as a result of the method of measurement (absolute strength), suggesting normal Constant-Murley scores will differ by age. Romeo et al[33] found a significant effect of age on Constant-Murley score only for female subjects undergoing rotator cuff repair. When normalized total scores were analyzed by group for the other 4 shoulder scoring instruments, no significant effect was observed. A significant age effect has been reported for many functional outcome instruments, particularly those requiring a rating of physical abilities.[6,9,29,33,37]

With the exception of the Constant-Murley score, the scoring instruments did not demonstrate systematic bias based on activity level, age, gender, or hand dominance. Romeo et al[33] also reported that, with the exception of the Constant-Murley score, total scores for various instruments do not display such bias for patients undergoing rotator cuff repair.

Different instruments weigh various subjective and objective subscales differently to obtain a total instrument score.[33] For outcomes research, the patient's subjective perception about his or her own condition should take priority.[23,32,39] Several shoulder scoring instruments use only the subjective report of the patient and thus do not allow clinical measurements or clinician input to influence the score.[11,27,31,40]

It should be noted that none of the 5 instruments was designed for use in subjects with healthy shoulders. It is possible that these shoulder scoring instruments may be valid for differentiating patients with different shoulder conditions or may be able to discriminate recovery over time but may be unreliable, and therefore not valid, when used among healthy persons. In addition, not all instruments comprise similar subscales (eg, pain, instability, range of motion), and they are thereby not necessarily valid across all shoulder conditions. Finally, an implicit assumption of these shoulder scoring instruments is homogeneous reliability and validity across all levels of shoulder disability, including healthy, normal function. If this assumption is violated, simple comparisons of scores between patients or over time in the same patient become problematic. These questions are beyond the scope of the current study but are worthy of further research.

It was not the intent of this investigation to make specific recommendations for the use of specific shoulder scoring instruments. With the growing interest in outcome studies in orthopaedics, follow-up assessments must not only measure traditional parameters, such as range of motion and instability, but must also measure the patient's perceptions of his or her health status.[19,23] The optimal methods of measurement for patients undergoing reconstructive shoulder surgery has not yet been determined.[8,24,32,36]

We examined 240 shoulders in 120 healthy subjects who had never sought treatment for any shoulder-related or cervical spine–related complaint. We found that both age and gender should be taken into account when reporting or interpreting Constant-Murley scores. In addition, normalized total and subscale scores for the 5 shoulder scoring instruments were not equivalent. Thus, the investigator or clinician should carefully choose an instrument that matches the population and the purpose of the study. This information should be useful for investigators studying patients undergoing reconstructive procedures of the shoulder.

We wish to thank Michele L. Clowers for secretarial preparation of this manuscript, Michelle DeSilvio, MS, for assistance with the statistical analysis, and Drs Spencer Guinn and Stephen Timon for their assistance in performing this investigation.

REFERENCES

1. Amstutz HC, Sew Hoy AL, Clarke IC. UCLA anatomic total shoulder arthroplasty. Clin Orthop 1981;155:7-20.
2. Bankes MJ, Crossman JE, Emery RJ. A standard method of shoulder strength measurement for the Constant score with a spring balance. J Shoulder Elbow Surg 1998;7:116-21.
3. Barnes CJ, Van Steyn SJ, Fischer RA. The effects of age, sex, and shoulder dominance on range of motion of the shoulder. J Shoulder Elbow Surg 2001;10:242-6.
4. Barrett WP, Franklin JL, Jackins SE, Wyss CR, Matsen FA III. Total shoulder arthroplasty. J Bone Joint Surg Am 1987;69:865-72.
5. Beaton DE, Richards RR. Measuring function of the shoulder. A cross-sectional comparison of five questionnaires. J Bone Joint Surg Am 1996;78:882-90.
6. Brinker MR, Lund PJ, Barrack RL. Demographic biases of scoring instruments for the results of total knee arthroplasty. J Bone Joint Surg Am 1997;79:858-65.
7. Brinker MR, Lund PJ, Cox DD, Barrack RL. Demographic biases found in scoring instruments of total hip arthroplasty. J Arthroplasty 1996;11:820-30.
8. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. J Bone Joint Surg Br 1996; 78:229-32.
9. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop 1987;214:160-4.
10. Constant CR. An evaluation of the Constant-Murley shoulder assessment [letter]. J Bone Joint Surg Br 1997;79:695-6.
11. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br 1996; 78:593-600.
12. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. J Bone Joint Surg Br 1999;81:420-6.
13. Department of Health and Human Services. Federal poverty income guidelines for fiscal year 1993. In: Federal Register. Washington, DC: Government Printing Office; 1992. p. 5455-7.
14. Ellman H, Hanker G, Bayer M. Repair of the rotator cuff. End-result study of factors influencing reconstruction. J Bone Joint Surg Am 1986;68:1136-44.
15. Freedman KB, Bernstein J. Sample size and statistical power in clinical orthopaedic research. J Bone Joint Surg Am 1999;81: 1454-60.
16. Gartsman GM, Brinker MR, Khan M, Karahan M. Self-assessment of general health status in patients with five common shoulder conditions. J Shoulder Elbow Surg 1998;7:228-37.
17. Gartsman GM, Roddey TS, Hammerman SM. Arthroscopic treatment of multidirectional glenohumeral instability: 2- to 5-year follow-up. Arthroscopy 2001;17:236-43.
18. Gartsman GM. Arthroscopic acromioplasty for lesions of the rotator cuff. J Bone Joint Surg Am 1990;72:169-80.
19. Greenfield S. The state of outcome research: are we on target? [editorial]. N Engl J Med 1989;320:1142-3.
20. Hays WL. Statistics. Fort Worth: Holt, Rinehart and Winston; 1988.
21. Johnson L. Outcomes analysis in spinal research. How clinical research differs from outcomes analysis. Orthop Clin North Am 1994;25:205-13.
22. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178-S189.
23. Keller RB, Rudicel SA, Liang MH. Outcomes research in orthopaedics. J Bone Joint Surg Am 1993;75:1562-74.
24. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med 1998;26:764-72.
25. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.
26. L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MG. A self-administered questionnaire for assessment of symptoms and function of the shoulder. J Bone Joint Surg Am 1997;79:738-48.
27. Matsen FAI, Ziegler DW, DeBartolo SE. Patient self-assessment of health status and function in glenohumeral degenerative joint disease. J Shoulder Elbow Surg 1995;4:345-51.
28. McArdle WD, Katch FI, Katch VL. Measuring and improving muscular strength. In: Exercise physiology. Energy, nutrition, and human performance. Philadelphia: Lea & Febiger; 1991. p. 452-9.
29. O'Connor DP, Brinker MR, Woods GW. Preoperative health

status of patients with four knee conditions treated with arthroscopy. Clin Orthop 2002;395:164-73.
30. Richards RR, An KN, Bigliani LU, et al. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg 1994; 3:347-52.
31. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. Arthritis Care Res 1991;4:143-9.
32. Romeo AA, Bach BRJ, O'Halloran KL. Scoring systems for shoulder conditions. Am J Sports Med 1996;24:472-6.
33. Romeo AA, Hang DW, Bach BRJ, Shott S. Repair of full thickness rotator cuff tears. Gender, age, and other factors affecting outcome. Clin Orthop 1999;367:243-55.
34. Rose MS, Koshman ML, Spreng S, Sheldon R. Statistical issues encountered in the comparison of health-related quality of life in diseased patients to published general population norms. Problems and solutions. J Clin Epidemiol 1999;52:405-12.
35. Rowe CR, Patel D, Southmayd WW. The Bankart procedure: a long-term end-result study. J Bone Joint Surg Am 1978;60: 1-16.
36. Soldatis JJ, Moseley JB, Etminan M. Shoulder symptoms in healthy athletes: a comparison of outcome scoring systems. J Shoulder Elbow Surg 1997;6:265-71.
37. Ware JE, Kosinski M, Keller SD. SF-12: how to score the SF-12 physical and mental health summary scales. Boston: The Health Institute, New England Medical Center; 1995.
38. Ware JE, Kosinski M, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston: The Health Institute; 1994.
39. Williams GN, Gangel TJ, Arciero RA, Uhorchak JM, Taylor DC. Comparison of the single assessment numeric evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. Am J Sports Med 1999;27:214-21.
40. Williams JW, Holleman DR, Simel DL. Measuring shoulder function with the Shoulder Pain and Disability Index. J Rheumatol 1995;22:727-32.