



Clinimetric and Psychometric Strategies for Development of a Health Measurement Scale

Robert G. Marx,^{1,3} Claire Bombardier,² Sheila Hogg-Johnson,³ and James G. Wright^{1,*}

¹DEPARTMENT OF SURGERY AND PUBLIC HEALTH SCIENCES, CLINICAL EPIDEMIOLOGY, AND HEALTH RESEARCH PROGRAM, UNIVERSITY OF TORONTO, THE HOSPITAL FOR SICK CHILDREN, TORONTO, ONTARIO, CANADA; ²DEPARTMENT OF MEDICINE, HEALTH ADMINISTRATION, UNIVERSITY OF TORONTO, INSTITUTE FOR WORK AND HEALTH, TORONTO, ONTARIO, CANADA; AND ³INSTITUTE FOR WORK AND HEALTH, TORONTO, ONTARIO, CANADA

ABSTRACT. Clinimetrics and psychometrics, two accepted methods for developing multiitem health measurement scales, have fundamentally different aims and methods that have seldom been compared and never prospectively. The purpose of this study was to determine whether these two methodologies provided comparable scales in the development of an upper extremity disability measure. Psychometric analysis involved field testing a 70-item questionnaire on 407 patients. Equidiscriminatory item total correlation (EITC) was used to select the top 30 items. Clinimetric testing used the mean importance and severity ratings of the 70 items by 76 patients to select the top 30 items. Clinimetric and psychometric analyses were performed independently. Cronbach's alpha was 0.97 for the top 30 items selected by EITC and 0.96 for the items selected based on patient's ratings. The two scales (after clinician modification to improve face validity) shared 16 items in common ($P = 0.10$). The intraclass correlation coefficient of the patient scores on the two 30-item scales was 0.93 before clinician input and 0.97 after. The mean (and standard deviation) difference between scales was 9.1 (8.8) before and 1.7 (5.2) after clinician input. A scale developed with a clinimetric strategy can measure a complex (so-called heterogeneous) clinical phenomenon (thought to be composed of several patient attributes) but still fulfill psychometric criteria for "homogeneity." Thus, these strategies for the development of health measurement scales, which have been considered potentially opposite or conflicting, may be complementary. J CLIN EPIDEMIOL 52;2:105–111, 1999. © 1999 Elsevier Science Inc.

KEY WORDS. Health status, psychometrics, clinimetrics, measurement

INTRODUCTION

The current focus on health care has resulted in a virtual explosion of health measurement scales used to measure phenomena such as quality of life and physical and psychological disability [1]. Clinimetric and psychometric strategies are the predominant techniques used for the development of multiitem health measurement scales [2]. Clinimetric strategies used in clinical medicine rely primarily on the judgements of patients and clinicians and aim to measure clinical phenomena that are generally believed to comprise several unrelated patient characteristics or attributes (scales measuring several attributes have been called "heterogeneous" scales) [3]. Psychometric strategies used in psychology and education rely more on mathematical techniques

and generally (although not exclusively) aim to develop one scale (or multiple scales) that measure single patient characteristics or attributes (scales measuring a single attribute have been called "homogeneous" scales) [4]. Thus, the two methods differ in aims and strategy.

These two methods have seldom been compared and never prospectively [2]. If scales developed with distinct strategies but intended to measure the same phenomenon yielded dissimilar results, this would have two important health care implications. First, dissimilar results would raise questions about the validity of one or both methods. Second, and more importantly, if one or both methods were of questionable validity, this would raise questions about the validity of any study that had assessed the efficacy of therapeutic interventions using a scale developed with these methods. Because important treatment and health care policy decisions are based on the patient outcomes measured by these scales, it is crucial to understand the impact of the technique chosen for scale development on the final scale. We compared clinimetric and psychometric strategies in the development of a health measurement scale.

*Address for correspondence: Dr. J.G. Wright, Associate Professor of Surgery and Public Health Sciences, University of Toronto, The Hospital for Sick Children, 555 University Avenue, S-107, Toronto, ON, M5G 1X8, Canada.

Accepted for publication on 13 October 1998.

METHODS

The two methods were compared in the development of a multiitem health measurement scale called the Disabilities of Arm, Shoulder and Hand (DASH) [5]. The DASH was intended to be a discriminative and evaluative [6] region-specific scale. Scale development was performed in two stages of item generation and item reduction [7,8].

Item generation defines the content of the index and ensures that all important variables are considered for inclusion in the scale [2]. Item generation is similar for the psychometric and clinimetric strategies [3,4]. The first step of item generation was to define the content of the scale. The DASH was intended to evaluate symptoms (such as pain, numbness, and tingling) as well as physical, social, and psychological disability for patients with upper-extremity disorders. The decision to focus on symptoms and disabilities when developing the DASH was based on a review of the disability literature [9,10], expert input, a review of the concepts covered by existing outcomes scales [11–13], and the fact that main determinants of quality of life in patients with musculoskeletal complaints are symptoms (such as pain, numbness, and tingling) and loss of physical function [1]. Therefore, a conceptual framework (intended to guide subsequent scale development) consisting of the following four “domains” was created to represent adequately both symptoms and disability [10]: physical disability, social disability, psychological disability, and symptoms. The four domains were divided into 17 subdomains to delineate accurately each phenomenon. In the item generation phase, items were identified from multiple sources including previous research, patients, and clinicians [3,8]. A total of 821 items was generated from 13 existing upper-extremity-related health status measures, expert clinician input, and patient focus groups [5].

Item reduction, the second stage of scale development, eliminates redundant or inappropriate items and decreases the number of items to a total that is feasible to administer to patients while ensuring that the scale measures the construct or clinical phenomenon of interest [7]. Item reduction for the DASH was performed in several steps. First, in the preliminary item reduction, items that were either redundant or not clinically related to the upper limb were eliminated to arrive at a total of 70 items [5]. Of the 70 items, 42 were physical disability, eight were social disability, four were physiologic disability, and 16 were symptoms. This process deliberately maintained items from all four domains (and subdomains) that were believed by clinicians to be largely unrelated. Thus, after preliminary item reduction, the 70 items were more directed toward preserving scale heterogeneity than achieving scale homogeneity [5].

The ultimate goal of item reduction for the DASH was to arrive at a self-administered 30-item scale. Because this is the stage at which the two strategies differ, clinimetric and psychometric strategies were used to reduce the number of

items from 70 to two separate 30-item scales. For the comparison of the two techniques, we selected standard and commonly applied clinimetric and psychometric techniques. The two methods were performed prospectively and independently. Each of the 70 items was assigned a number, and the subsequent analyses were initially performed blind to the identity of the items to ensure that clinicians' opinions would not unduly influence item selection. Informed consent was obtained from all patients, and ethical approval for the study was obtained from the University of Toronto as well as from each institution that participated.

Psychometric Strategy

Twenty centers from Canada, the United States, and Australia treating patients with upper-limb disorders were involved in the data collection. Demographic (age and gender) and clinical information (diagnosis, hand dominance, side affected, occupational status, and reason for unemployment) was recorded, and clinicians rated the severity of the patients' upper-limb conditions on a five-point rating scale of “very mild, mild, moderate, severe, very severe” (assigned numerical ratings of 1 to 5, respectively). The sample size was dictated by the factor analysis, which required a minimum of 350 subjects to satisfy the criteria of a minimum of five respondents per item for the 70 items [4,14].

Equidiscriminatory item total correlation (EITC) was the psychometric strategy chosen for item reduction. Equidiscriminatory item total correlation, a slight modification of the usual item total correlations methods, selected items that were correlated with each other but discriminated between individuals throughout a range of scores [4]. Patients completed all 70 items, each of which had five response options. (Patients rated difficulty with activities as “no difficulty, mild, moderate, severe difficulty, or unable.” Patients rated symptoms as “none, mild, moderate, severe, and extreme.”) Total scores were determined for each patient by adding together the responses for the 70 items. The sample was split based on the total scores for the 70 items at three percentile levels (i.e., at 25, 50, and 75). For example, for the distribution of people below the 25th percentile score, a score of zero was given to everyone below that score and a score of one to everyone above the 25th percentile. The correlation between the individual items and the total score was determined for each item using the new dichotomized score (of zero or one). The items were ranked in order of their item–total correlation for this dichotomized score, creating a first list of items. This procedure was repeated for the 50th and the 75th percentiles, creating second and third lists, so that three lists of item–total correlations were obtained (one of each percentile division). The top 10 items from each list were then chosen for a total of 30 items [4].

Cronbach's alpha is a measure of homogeneity or internal consistency where “increasing reliabilities beyond 0.80 in basic

research is often wasteful of time and money” [4]. Our *a priori* target was that the psychometric strategy would result in a scale or scales with a Cronbach’s alpha of at least 0.80.

We used factor analyses to determine whether the scale developed using the psychometric strategy was measuring a single or multiple factor that would require multiple scales. Factor analysis with varimax rotation [4,14] was performed on the 70 items as well as on the 30 selected by EITC. The number of factors was determined by the scree test [15]. The scree test plots successive eigen values against the eigen value number. The plot is used to locate a transition point in the curve [4]. Items were attributed to factors if they had a factor loading of at least 0.25 [16]. Items with loadings that differed by more than 0.05 on two factors were attributed to the factor with greater loading [17]. The psychometric analyses, as previously stated, were initially performed blind to the identity of the items.

Clinimetric Strategy

Three of the same 20 centers used for the psychometric testing were used to accrue a separate group of patients for the clinimetric testing. We collected the same demographic and clinical information from patients. A sample of more than 50 patients was sought for the clinimetric testing in order to obtain a 95% confidence interval (CI) of $\pm 15\%$ for the frequency of the complaint in the population [7]. For example, the percentage of patients unable to prepare a meal would be 30% with 95% CI $\pm 15\%$.

The clinimetric strategy relied on the ratings of patients to determine which items to include in the final scale. This clinimetric strategy has been used previously in the development of numerous scales [7,13,18,19]. Patients rated the importance and the severity of each of the 70 items on a five-point scale from “not at all important” to “extremely important” and from “not at all severe” to “extremely severe” [7]. The importance and severity scores were added together for each individual patient, and the mean importance-severity score was determined for each item. (Alternative methods, such as multiplication could have been used, but different methods provide similar results [20].) The 30 items with the highest scores were selected for the clinimetric scale [13]. The clinimetric analysis was initially performed blind to the identity of the items.

Comparison of Scales Development with Psychometric and Clinimetric Strategies

Patient scores were calculated (using the responses from the 407 patients who completed all 70 items) for the 30-items clinimetric and the 30-item psychometric scales. The scores on the two 30-item scales were compared with the random-effects intraclass correlation coefficient. The intraclass correlation coefficient (ICC) is an index of concordance for dimensional measurements ranging between zero and one,

where ≥ 0.75 represents “excellent” reproductibility [21]. We also used the methods described by Bland and Altman [22,23] to determine the mean, standard deviation, and limits of agreement (mean ± 2 standard deviation [SD]) for differences between scores. The limits of agreement show how much one method of measurement differs from another. If the differences between two methods of measurement are not “clinically important, then we could use the two measurement methods interchangeably” [23].

A comparison of the scales before any modifications by the clinicians was relevant because this represents a pure comparison of the two methods for item reduction. In practice, however, no scale developed either by psychometric or clinimetric methods would be acceptable without including clinical judgment to ensure face validity [6–8]. Even psychometric methods use clinical judgment to modify scales to improve face validity. “Face validity can be considered as one aspect of content validity, which concerns an inspection of the final product to make sure nothing went wrong in transforming plans into a completed instrument” [4 p 11]. The more practically relevant question was whether the two scales, based on varying methodologies, were similar after modification by clinicians. Thus, two groups of clinicians (with experience in upper-limb disorders and clinical epidemiology) were unblinded to the identity of the items and independently reviewed the 30-items questionnaires that were developed in the first stage of the research. The two groups of clinicians (who were provided with only one of the clinimetric or psychometric analyses) slightly modified the 30 items (by exchanging items from the rejected item pool) to improve face validity.

RESULTS

The mean age of the 407 patients used in the psychometric analysis was 45 years; 166 (41%) were men; and the mean clinician severity rating was 3.0. The mean age of the 76 patients used in the clinimetric analysis was 46 years; 30 (39%) were men, and the mean clinician severity rating was 3.0. As shown in Tables 1 and 2, the two groups of patients used in the psychometric and clinimetric analyses were comparable for age, gender, hand dominance, side affected, employment status, worker’s compensation status, diagnoses, and clinician severity rating ($P \geq 0.05$).

The 30 items selected by EITC in the psychometric analysis (see Appendix) had a Cronbach’s alpha of 0.97. The scree test suggested a single factor that accounted for 57% of the variance in the analysis of the 30 items. Factor rotation did not clarify the factor structure or suggest multiple factors. The clinimetric strategy, performed in parallel, selected 30 items (see Appendix) with the highest combined mean importance-severity score. The Cronbach’s alpha was 0.96. The two methods selected 15 items in common ($P = 0.21$; based on the hypergeometric distribution, “P” is the

TABLE 1. The demographics of the patients used for psychometric and clinimetric item reduction

	Psychometric (n = 407)	Clinimetric (n = 76)
Gender		
Male	166 (41%)	30 (39%)
Female	206 (51%)	46 (61%)
Missing	35 (8%)	0 (0%)
Dominant hand		
Right	365 (90%)	70 (92%)
Left	37 (9%)	6 (8%)
Missing	5 (1%)	0 (0%)
Employed		
Yes	198 (49%)	41 (54%)
No	161 (40%)	35 (46%)
Missing	48 (11%)	0 (0%)
Why not working?		
Upper limb problem	59 (15%)	22 (29%)
Other health problem	6 (2%)	4 (5%)
Unemployed	10 (2%)	0 (0%)
Retired	49 (12%)	9 (12%)
Other	21 (5%)	0 (0%)
Missing	16 (3%)	0 (0%)
Average age (y)	45.0 (SD = 16.7) (n = 365)	45.6 (SD = 15.9) (n = 76)
Worker's compensation	46 (11%)	9 (12%)
Side affected		
Right	175 (43%)	31 (41%)
Left	125 (31%)	24 (32%)
Both	103 (25%)	21 (28%)
Missing	4 (1%)	0 (0%)
Average clinician severity rating (rated 1–5)	3.02 (SD = 0.86) (n = 336)	3.01 (SD = 1.09) (n = 76)

For categories expressed as mean value, the standard deviations (SD) and the number of observations are listed. No statistically significant differences between the two groups for $P \leq 0.05$ for any characteristic.

probability that the two methods would share this number of items owing to chance) [24]. The clinimetric methods, however, selected a greater number of symptoms and psychological function items, whereas the psychometric strat-

TABLE 2. The diagnoses of the patients used for psychometric and clinimetric item reduction

Diagnosis ^a	Psychometric (n = 407)	Clinimetric (n = 76)
Colle's fracture	10 (2%)	5 (6%)
Humerus fracture	7 (1%)	1 (1%)
Thumb carpometacarpal arthritis	20 (4%)	2 (3%)
Carpal tunnel syndrome	42 (9%)	4 (5%)
Rotator cuff tendinopathy	87 (18%)	12 (16%)
Lateral epicondylitis	34 (5%)	1 (1%)
de Quervain's tenosynovitis	6 (1%)	0 (0%)
Osteoarthritis	34 (7%)	6 (8%)
Rheumatoid arthritis	30 (6%)	10 (13%)
Nonspecific	29 (6%)	1 (1%)
Other	151 (32%)	28 (36%)
Unknown	51 (10%)	8 (10%)

^aSome patients had more than one diagnosis.

egy selected a greater number of physical disability items. A random-effects ICC for the clinimetric and psychometric patient scores was 0.93 (Fig. 1), indicating "excellent" reproductibility [21]. The mean scores for the entire group of patients was 48.2 and 39.1 for clinimetric and psychometric scales, respectively. The mean difference in scores for the two scales was 9.1, with a SD of 8.8, and therefore, the limits of agreement were 9.1 ± 17.6 .

After clinician modification, the two methods agreed on 16 items ($P = 0.10$; based on the hypergeometric distribution) [24] (see Appendix 1). As shown in Appendix 1, clinical modification of the clinimetric scale substantially changed the item composition of the 30-item scale by interchanging 10 items. Clinical modification of the psychometric scale replaced only three items. Despite the modifications by clinicians of the item content of the two 30-item scales, the modifications resulted in only one additional item in common. The ICC increased to 0.97 after clinician input (see Fig. 2), and the scatter plot show the points clustering more closely around the line of perfect agreement. The mean scores for the entire group of patients was 40.9 and 39.2 for the clinimetric and psychometric scales, re-

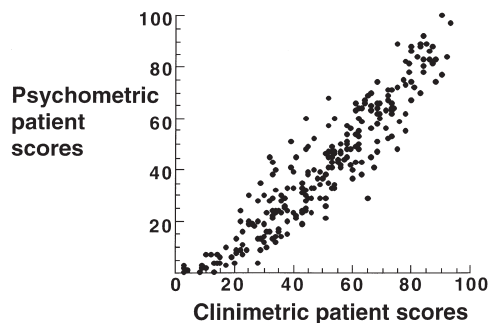


FIGURE 1. Psychometric and clinimetric patient scores (before clinician input).

spectively. The mean difference in scores for the two scales after clinician modification was 1.7, with a SD of 5.2, and therefore, the limits of agreement were 1.7 ± 10.4 .

DISCUSSION

Previous research has shown that individual patients present to clinicians with different spectrums of complaints and attach different importance of relief of their complaints. For example, some patients before total hip arthroplasty value improvement in recreational activities above all other complaints [18,19]. However, when examining the responses for groups of patients, as in this study, the responses to the items, as shown by high Cronbach alphas (and confirmed with the factor analyses) were highly correlated. In the case of upper-extremity disability and symptoms, it was possible to create scales that satisfied the psychometric criteria for “homogeneity” but measured phenomena that were considered by clinicians to be “heterogeneous.” Thus, when measures are developed with a clear conceptual framework and item generation is performed systematically and completely, as in this example, the two strategies lead to scales that provided similar overall scores. This result is reassuring given the number of interventions that are evaluated comparing groups using primary outcomes, such as health status or quality of life, developed with these two techniques [1].

From a qualitative point of view, the scales represented the domains differently. In this study, the clinimetric method ranked 12 symptom items and three psychological disability items in the top 30. The psychometric method selected almost exclusively physical disability items, with only two symptom items and no psychological disability items. Thus, this research confirms the results of Juniper *et al.* that the two methods select different types of items [25]. Juniper *et al.* [25] retrospectively applied factor analysis to a pool of items developed for the Asthma Quality of Life Questionnaire and found that factor analysis chose different types of items than those chosen by patients.

This study expands the research performed by Juniper *et al.* by comparing the two methods in a parallel, prospective,

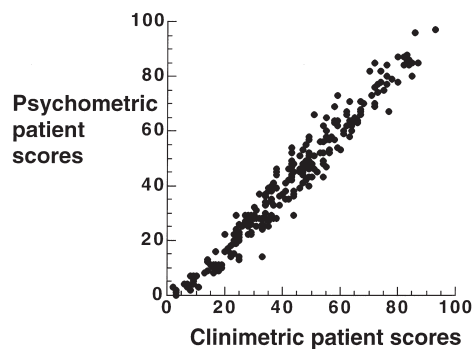


FIGURE 2. Psychometric and clinimetric patient scores (after clinician input).

and blinded fashion. Moreover, rather than just comparing the types of items chosen by the two methods, we compared scores patients received using the 30-items scale developed with the two methods. Despite an “excellent” ICC, the SD of the differences in scores patients received on the two scales (the limits of agreement) before clinician input was 8.8. The limits of agreement, described by Bland and Altman [23], is a statistical method for assessing agreement between two methods of clinical measurement to determine whether one measure can be substituted for another to evaluate individual patients. If the limits of agreement (mean \pm 2 SD) were small, we could “confidently” use the two methods of measurement interchangeably to evaluate individual patients [23]. The limits of agreement for the psychometric and clinimetric scales before clinician input were relatively large, suggesting that the two scales provided different results for individual patients. Health measurement scales, however, are not usually used to make decisions about individual patients, but are usually used to evaluate the responses of groups of patients to therapeutic intervention. The group scores patients received on the two scales after clinicians’ input were almost identical. Because the clinimetric technique identifies items that are relevant and important to patients (i.e., that represent the patients’ perspective), perhaps a combination of both methods may provide clinicians with the most appropriate scales for the measurement of health status in groups of patients.

The main potential limitation of this study is that the results may not be necessarily generalizable to all health measurements or all methods of clinimetric or psychometric scale development. Although we used standard and commonly applied clinimetric and psychometric techniques, this study will need replication in the development of other types of health measurements. Although the limits of agreement for the two scales after clinician input were more acceptable, further research should be done to determine whether changes in scores after a therapeutic intervention would be similar using the two scales.

In summary, a scale developed with a clinimetric strategy can measure a complex (so-called heterogeneous) clinical

phenomenon (thought to be composed of several patient attributes) but still fulfill psychometric criteria for "homogeneity." Thus, these strategies for the development of health measurement scales, which have been considered potentially opposite or conflicting, may be complementary.

This research was supported by the Institute for Work and Health and the American Academy of Orthopaedic Surgeons. Dr. Wright is supported by a Scientist Award of the Medical Research Council of Canada. Dr. Marx was supported by an Arthritis Society Fellowship, the Surgeon Scientist Program at the University of Toronto, and the Institute for Work and Health.

References

1. Wright JG. Quality of life in orthopaedics. In: Spilker B, Eds. **Quality of Life and Pharmacoeconomics in Clinical Trials**. Philadelphia: Lippincott-Raven Publishers; 1996.
2. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. **J Clin Epidemiol** 1992; 42: 1201-1218.
3. Feinstein AR. **Clinimetrics**. Westford, MA: Murray Printing Company; 1987.
4. Nunally JC, Bernstein IH. **Psychometric Theory**. New York: McGraw-Hill; 1994.
5. Hudak PL, Amadio PC, Bombardier CB, Beaton D, Cole D, Davis A, et al. Development of an upper extremity outcome measure: The DASH (Disabilities of the Arm, Shoulder and Hand). **Am J Ind Med** 1996; 29: 602-608.
6. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: What are the necessary measurement properties? **J Clin Epidemiol** 1992; 45: 1341-1345.
7. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. **Can Med Assoc J** 1986; 134: 889-895.
8. Stenier DL, Norman GR. **Health Measurement Scales. A Practical Guide to their Development and Use**. Oxford: Oxford University Press; 1995.
9. Jette AM. Physical disablement concepts for physical therapy research and practice. **Phys Ther** 1994; 74: 380-386.
10. Verbrugge LM, Jette AM. The disablement process. **Soc Sci Med** 1994; 38: 1-14.
11. Bombardier C, Tugwell P. A methodological framework to develop and select indices for clinical trials: Statistical and judgmental approaches. **J Rheumatol** 1982; 9: 753-757.
12. Tugwell P, Bombardier C. A methodological framework for developing and selecting endpoints in clinical trials. **J Rheumatol** 1982; 9: 758-762.
13. Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith L, McIlroy W, et al. Measuring quality of life in the frail elderly. **J Clin Epidemiol** 1993; 46: 1433-1444.
14. Gorsuch RL. **Factor Analysis**. Hillsdale, NJ: Erlbaum Associates; 1983.
15. Cattell RB. The scree test for the number of factors. **Multivariate Behav Res** 1966; 1: 245-276.
16. Stevens J. **Applied Multivariate Statistics for the Social Sciences**. London: Hillsdale, NJ; 1986.
17. Norman GR, Streiner DL. **PDQ Statistics**. Hamilton: B.C. Decker Inc.; 1986.
18. Wright JG, Rudicel S, Feinstein AR. Ask patients what they want. Evaluation of individual complaints before total hip replacement. **J Bone Joint Surg Br** 1994; 76-B: 229-234.
19. Wright JG, Young NL. The patient-specific index: Asking patients what they want. **J Bone Joint Surg Am** 1997; 79-A: 974-983.
20. Marx RG, Bombardier C, Hogg-Johnson S, Wright JG. How should importance and severity ratings be combined for item reduction in the development of health status instruments? **J Clin Epidemiol** In Press
21. Rosner B. **Fundamentals of Biostatistics**. Toronto: Duxbury Press; 1995.
22. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. **Comput Biol Med** 1990; 20: 337-340.
23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. **Lancet** 1986; 1: 307-310.
24. Kalbfleisch JG. **Probability and Statistical Inference**. New York: Springer-Verlag; 1985.
25. Juniper FE, Guyatt GH, Streiner DL, King KR. Clinical impact versus factor analysis for quality of life questionnaire construction. **J Clin Epidemiol** 1997; 50: 233-238.

APPENDIX 1. Items selected by the clinimetric and psychometric techniques

Clinimetric items	Psychometric items
3. Opening a tight or new jar	3. Opening a tight or new jar
5. Writing	7. Preparing a meal
7. Preparing a meal	9. Typing
13. Doing heavy household chores (e.g., washing walls or floors)	10. Washing and drying dishes
18. Carrying a shopping bag or briefcase	11. Pushing open a heavy door
19. Carrying a heavy object (>10 lb)	13. Doing heavy household chores (e.g., washing walls or floors)
27. Washing your back	14. Sweeping
38. Recreational activities in which you take some force through your arm, shoulder or hand (e.g., golf, hammering, tennis)	15. Gardening or doing yard work
39. Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton)	16. Shoveling
42. Taking care of your family	17. Making a bed

(Continued)

APPENDIX 1. Continued

Clinimetric items	Psychometric items
43. Getting comfortable for sleep	18. Carrying a shopping bag or brief case
*46. Spending your usual amount of time on work or other regular daily activities	19. Carrying a heavy object (>10 lb)
*47. Accomplishing as much as you would like at work or other regular daily activities	21. Changing a lightbulb overhead
48. Being limited in the kind of work you do, or other activities	23. Washing or blow drying hair
49. Having difficulty at work or with other activities (i.e., requiring extra effort)	27. Washing your back
50. Arm, shoulder or hand pain experienced at the present time	28. Putting on makeup or shaving
51. Your worst pain in the past 3 months	37. Recreational activities that require little effort (e.g., card playing, knitting, etc.)
52. Your usual pain, on average, in the past three months	38. Recreational activities in which you take some force through your arm, shoulder, or hand (e.g., golf, hammering, tennis)
*53. Being kept from your usual activities (work, school or housework) because of your arm, shoulder, or hand problem	39. Recreational activities in which you move your arm freely (e.g., playing frisbee, badminton)
*54. Having pain in your arm, shoulder, or hand when performing a specific activity	40. Using a screwdriver or similar tool
*55. The amount of time you have pain in your arm, shoulder, or hand when performing a specific activity	41. Managing transportation needs (getting from one place to another)
*59. Having severe difficulty sleeping because of arm, shoulder, or hand pain	42. Taking care of your family
*60. Having frequent difficulty sleeping because of arm, shoulder, or hand pain	45b. Social activities with family, friends, neighbors, or groups (severity)
61. Having a feeling of severe weakness in your arm, shoulder, or hand	45a. Social activities with family, friends, neighbors, or groups (frequency)
*62. Having a frequent feeling of weakness in your arm, shoulder, or hand	*46. Spending your usual amount of time on work or other regular daily activities
63. Having a feeling of severe stiffness in your arm, shoulder, or hand	47. Accomplishing as much as you would like at work or other regular daily activities
*64. Having a frequent feeling of stiffness in your arm, shoulder, or hand	48. Being limited in the kind of work you do, or other activities
66. Feeling less confident using your affected arm, shoulder, or hand	*49. Having difficulty at work or with other activities (i.e., requiring extra effort)
67. Feeling disabled even though you may look fine to others	50. Arm, shoulder, or hand pain experienced at the present time
*68. Feeling less capable, less confident, or less useful	55. The amount of time you have pain in your arm, shoulder, or hand when performing a specific activity
<hr/>	
Items added by the clinicians	Items added by the clinicians
6. Turning a key	44. Sexual activities
11. Pushing open a heavy door	56. Having severe tingling (pins and needles) in your arm, shoulder, and hand
12. Placing an object on a shelf above your head	68. Feeling less capable, less confident, or less useful
15. Gardening or doing yard work	
22. Wiping yourself on the toilet	
23. Washing or blow drying hair	
34. Using a knife	
36. Pouring from a jug or tea pot	
41. Managing transportation needs (getting from one place to another)	
44. Sexual activities	

*Items that were replaced by other items during the "clinician input" phase. The items added by the clinicians are listed in this appendix under "items added by the clinicians."