



How Should Importance and Severity Ratings Be Combined for Item Reduction in the Development of Health Status Instruments?

Robert G. Marx,^{1,2} Claire Bombardier,^{2,3} Sheilah Hogg-Johnson,²
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; ²INSTITUTE FOR WORK AND HEALTH, TORONTO, ONTARIO, CANADA; AND ³DEPARTMENT OF MEDICINE, HEALTH ADMINISTRATION, UNIVERSITY OF TORONTO, TORONTO, ONTARIO, CANADA

ABSTRACT. Patients' ratings of the severity and importance of items are often used to select items for health status instruments. The purpose of this study was to compare six different methods of combining severity-importance ratings. Two different patient groups separately rated the importance and severity of their complaints; (i) 76 patients with upper-extremity disorders rated 70 upper-extremity-related questions; and (ii) 86 patients with hip arthrosis rated 22 questions relating to their hip problem. The rank ordering of the items using the six different methods in the two populations were very similar ($\tau_{bi} = 0.91$ and 0.87 , respectively). Furthermore, the six methods when used to choose 30 upper-extremity items shared 25 items in common and shared 9 (of 10) hip items in the second group. In conclusion, the results of item reduction were not affected by the method of creating importance-severity ratings. *J CLIN EPIDEMIOL* 52;3:193–197, 1999. © 1999 Elsevier Science Inc.

KEY WORDS. Importance, severity, health, measurement, scale, instrument

INTRODUCTION

Health status is an important patient outcome used to evaluate many medical and surgical interventions. Many health-status instruments have been developed using clinimetric techniques that rely primarily on the opinions of patients and clinicians [1]. Scale development is usually performed in two stages of item generation and item reduction [2,3].

Item generation defines the content of the index and ensures that all important variables are considered for inclusion in the scale [4]. Items are usually generated from other scales, previous research, and the opinions of patients and clinicians. The input of patients is particularly important to ensure content validity and because they may identify new items that have not been used previously [5]. Item reduction eliminates 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 attribute, construct, or clinical phenomenon of interest [2,3]. Clinimetric and psychometric strategies are the predominant techniques used for health measurement scale development [4].

Clinimetric strategies, used in clinical medicine, are based on patient and clinician judgment and aim to measure clinical phenomena that are deemed the most relevant [1]. Psychometric strategies used in psychology and education rely more on mathematical techniques, such as Cronbach's alpha and factor analysis. These methods aim to develop one scale (or multiple scales) that measure single patient characteristics or attributes. The advantage of clinimetric scales over psychometric scales is that patient opinion is used in the development, thus ensuring that the instrument is relevant to the individuals for whom it will be used.

One of the most popular clinimetric techniques for item reduction, which has been used extensively in the development of clinical measurement scales, is based on patients' ratings of the severity and importance of the items identified in the item generation phase [2]. Patients rate the severity of the item (example: how severe is your arm pain?) as well as the importance (i.e., how important is your arm pain?). Patients' ratings for each item are combined to create severity-importance scores, and the items' rankings are used to decide which items to include in the scale.

Two main methods have been used to create severity-importance scores for items. The first method selects items based on "frequency-importance" scores. Frequency was defined as the proportion of patients for whom a given item is "troublesome" [6–10]. In these studies, patients rated items

*Address correspondence to: Dr. James G. Wright, The Hospital for Sick Children, 555 University Avenue, S-107, Toronto, Ontario, Canada M5G 1X8.

Accepted for publication on 6 November 1998.

in binary categories of “troublesome” or “not troublesome.” Patients separately rated the importance of each item on four-point (or five-point) category rating scales. The mean importance rating for the entire group (for an item) was multiplied by the frequency of that item for the entire group to create a frequency-importance score for each item [6–10].

In a second method, patients rated the severity of their complaints and separately rated the importance of their complaints using visual analogue or ordinal scales [5]. Then each subject’s ratings of importance and severity for a given item were multiplied together, and the resulting products were averaged over subjects [5]. Thus, the two methods differ in two ways: (i) one method uses a binary rating of frequency and the other uses a multicategory rating of severity, and (ii) the first method multiplies together the group means, whereas the second uses the average of individual severity-importance scores. Moreover, no rationale was provided in either method for multiplying the ratings rather than adding the ratings. The validity of the scales constructed by these methods may therefore be questioned. Furthermore, if these 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 decisions regarding therapy and health policy are based on the patient outcomes measured by these scales, it is crucial that the methods used to create such scales be well understood. The purpose of this study was to compare six different methods of creating severity-importance ratings.

METHODS

Six methods for combining severity and importance ratings were tested using two different patient populations. For both groups, consecutive patients were recruited from tertiary care centers. The severity of disease and the type of patient were representative of the patients for whom the instruments are intended. The first group of patients participated in research directed toward developing an evaluative [11] scale to measure disability and symptoms in patients with upper-limb disorders. The details of items generation and preliminary item reduction has been published previously [12]. An item pool of 821 items was generated from previous literature, clinicians, and patient focus groups [12]. After eliminating redundant items using clinicians’ ratings, the number of items was reduced to 70. There are no definitive guidelines with respect to how many items should be included in a given scale, and the decision may be somewhat subjective. The ultimate aim was to reduce the number of items to a 30-item scale. Each question was worded separately with five response options, thus the 30 items were answerable in 10 to 15 minutes, allowing simultaneous administration of other health status measures [12]. Seventy-six patients with upper-extremity disorders, recruited

from two orthopedic and one rheumatology clinic, separately ranked the severity and importance of the seventy items on five-category ordinal scales with response options of “not at all troublesome” to “extremely troublesome” and “not at all important” to “extremely important,” respectively.

The second group of patients participated in research directed toward the development of a scale to measure disability and symptoms in patients undergoing total hip arthroplasty for hip arthrosis. Twenty-two items had been generated from clinician interviews, previous literature, and interviews with 78 patients before total hip arthroplasty [13]. Eighty-six patients, before total hip arthroplasty, rated the severity and importance of 22 items. Although 22 items is a reasonable number for a measurement scale, we elected to reduce the number of items to 10 for the purpose of this study. Severity was rated for each item using one of three different seven-category ordinal response options depending on which was most appropriate for the complaint: “not bothersome” to “most bothersome imaginable”; “not severe” to “most severe imaginable”; or “not difficult” to “most difficult imaginable.” Importance was rated for each item using seven response options, from “not important” to “most important imaginable.”

Patients’ ratings were combined using addition, multiplication, and an adjusted mean technique (Table 1). The adjusted mean, not previously described, is the product of the severity and importance scores multiplied by two and divided by the sum of the severity and importance scores (Table 1). The advantage of this method is that it is less severely skewed by low values (for either importance or severity ratings) compared with the multiplication method. For each of the three methods, scores were created in two ways: (i) the mean severity score and the mean importance scores for the entire group (called the group-value method) were determined for each item, and then the severity-importance scores were determined (using the addition, multiplication, or the adjusted mean methods); and (ii) individual severity-importance scores were calculated for each patient (called the individual-value method), and then the mean severity-importance scores were determined (using addition, multiplication, or the adjusted mean).

TABLE 1. Methods used for combining importance and severity scores

Formula	Group values	Individual values
Addition	$I + S$	$\Sigma(i + s)/n$
Multiplication	$I \times S$	$\Sigma(i \times s)/n$
Adjusted mean	$2(I \times S)/(I + S)$	$\Sigma(2(i \times s)/(i + s))/n$

Abbreviations: I = mean importance value for the entire group; S = mean severity value for the entire group; i = individual importance value for a patient (where there are n patients in the group); s = individual severity value for a patient (where there are n patients in the group).

The lowest response category for ratings of both severity and importance are commonly scored as either zero or one. The choice made impacts on the results when using individual values for the multiplication or adjusted mean techniques. If the lowest response is accorded the value zero (for either severity or importance), then the combined score is nil, no matter what the corresponding severity or importance score. Alternatively, if the lowest response is scored one, this would not nullify the corresponding severity or importance score when the two are multiplied. The addition method, however, is not affected by the decision of whether to use zero or one because the intervals between the responses are maintained.

The decision of which score to assign to the lowest category is affected both by the wording of the response options and by patients' interpretations of the responses. For example, a response option of "minimally important" is not literally equivalent to the value zero. In this study, the response option of "not important" may be literally interpreted as "zero" or "not at all." However, we assigned the lowest responses a value of one because we were unsure whether the patients who selected the lowest response truly believed that their answer represented "no" (or zero) importance.

The items were rank ordered by severity-importance scores for each technique. A statistic recommended by Stavig [14], τ_{bi} , for total monotonic agreement of ranked data based on Kendall's τ was used to assess the level of

ranking agreement among the six methods. We also examined the top 30 severity-importance scores (of the 70 upper-limb questions) and the top 10 severity-importance scores (of the 22 hip questions) for each method to determine whether there was a practical difference in item selection among the methods.

RESULTS

Seventy-six patients with upper-extremity disorders were interviewed. Their average age was 45 years, and 39% were male. The diagnoses were varied with the most common being rotator cuff tendinopathy (16%), rheumatoid arthritis (13%), osteoarthritis (8%), and distal radius fracture (6%). The hip patients had an average age of 63 years, and 53% were male. Their diagnoses included osteoarthritis (70%), avascular necrosis (9%), childhood hip disease (9%), inflammatory arthritis (7%), and other (5%).

For each of the six methods, the rank order of the items was generally similar but not identical (see Table 2, for the hip patients' results; the upper-extremity results were similar and are not provided). When the difference between the severity-importance scores of items was small, the methods clustered the items similarly but ranked them in a slightly different order. When the difference between the severity-importance scores of items was large, the methods tended to agree completely on the rank ordering.

TABLE 2. Importance-severity rankings by the addition, multiplication and adjusted mean techniques for the hip patients

Item	A-m	M-m	AM-m	A-i	M-i	AM-i
Daytime pain	16	16	14	15	12	15
Night time pain	9	8	8	7	7	8
Limp	18	18	18	17	17	17
Hip stiffness	17	17	17	16	15	16
Taking medication for the hip	2	2	2	2	2	2
Having to use walking aids	12	12	12	14	13	10
Difference in leg lengths	10	10	10	12	16	13
Fear of falling	14.5	15	16	18	21	20
Loss of independence	20	20	20	20	19	18
Walking	22	22	21	22	20	21
Going up and down stairs	19	19	19	19	18	19
Putting on shoes or stockings	13	13	13	11	10	11
Sitting	1	1	1	1	1	1
Using public transportation	5	4	3	9	9	9
Driving	4	5	5	3	3	3
Job/Housework	11	11	11	10	11	12
Recreational activities/Hobbies	21	21	22	21	22	22
Sexual activity	3	3	4	5	5	4
Getting onto and off of toilet	6	6	6	4	4	5
Picking things off the floor	14.5	14	15	13	14	14
Standing for five minutes	7	7	7	6	6	6

Abbreviations: A-m = addition using the mean importance and severity scores; M-m = multiplication using the mean importance and severity scores; AM-m = adjusted mean using the mean importance and severity scores; A-i = addition using the individual patient scores; M-i = multiplication using the individual patient scores; AM-i = adjusted mean using the individual patient scores.

The statistic τ_{bi} , measuring total monotonic agreement of the rankings by the six methods, was 0.91 for the upper-extremity data and 0.87 for the hip data, both indicating substantial agreement between the six methods (where 1.0 indicates perfect agreement). The six methods had 25 items in common among their top ranked 30 items (of 70) for the upper-limb patients. Similarly, the six methods ranked 9 of the hip items among their top 10 (of 22).

DISCUSSION

Clinimetric techniques of item reduction rely heavily on the opinions of patients. This study showed, using six different methods to select items based on patient's severity-importance ratings in two independent patient groups, that the rank ordering of the severity-importance scores was similar for all methods. Furthermore, on inspection, the rank order of the items by the six methods was generally similar. More importantly, the selection of the highest ranked items led to scales that had greater than 80% of the items in common. Because the methods of combining importance and severity scores all led to similar results, the exact method is less important.

The preferred method of combining patients' ratings may be affected by three issues. First, missing data led to a difference in the rank ordering of the items when comparing the individual-value and group-value methods. For the group-value method, all data were used to create the group means, whereas for the individual-value method, both the severity and importance responses of a patient had to be available for that patient to contribute to the score. The latter strategy for dealing with missing values is probably more appropriate because the goal is to combine severity and importance for each patient, not to only use one of the two. Although the impact of the effect of missing data in this study was small because neither data set had many missing values, this issue might be important with more missing values.

A second issue is the different methods of rating severity. By asking patients if a given item is troublesome or not [6–10], patients are essentially describing severity in a binary fashion. Because the precision of the information gathered with a binary scale is limited compared to a multiple category ordinal rating system, both severity and importance should probably be measured with multicategory ordinal scales. At least five to seven response categories should be used if there is an underlying continuum [3].

The final issue is whether to use a multiplicative, additive, or adjusted mean method to combine severity and importance scores. If the multiplication or adjusted mean techniques are used for the individual-value method, the anchor value given to the lowest response category is important. If there is truly no importance accorded to the item, a score of zero more accurately represents that perspective than a score of one. In general, it is unlikely that many patients would rate an item as not at all important or

severe. However, some items may potentially be not important or not severe in a given patient, despite the fact that clinicians or other patients believe the item is important. If a patient answered the lowest response for either the severity or importance of a given item, and a score of zero is assigned, then the two individual scores yield the overall score of zero when multiplied together for that item in a patient. Thus, this method could potentially underestimate the final importance-severity score of a given item if the ratings were multiplied for each individual (i.e., the individual value method). This issue is avoided by the additive method because the value of the origin of the scale is not important and intervals are maintained.

The limitation of this study is that both groups of patients suffered from musculoskeletal disorders, and, therefore, the results may not be generalizable to patients with other health problems. However, it is unlikely that a different population would affect the results to a large enough degree to alter the final results of the study.

In conclusion, the choice of which method to use is not of great importance here because there were no major discrepancies between the rankings of the severity-importance scores by the various techniques. However, the addition of the severity and importance ratings avoids the dilemma of how to score the lowest response. Furthermore, to measure severity and importance as precisely as possible, we suggest using multicategory ordinal scales with at least five categories for both.

Dr. Marx is supported by the Arthritis Society, the Institute for Work and Health, and the Surgical Scientist Program of the Department of Surgery of the University of Toronto. Dr. Wright is supported by a MRC scholarship.

References

1. Feinstein AR. *Clinimetrics*. Westford, MA.: Murray Printing Company; 1987.
2. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. *Can Med Assoc J* 1986; 134: 889–895.
3. Streiner DL, Norman GR. *Health Measurement Scales. A Practical Guide to their Development and Use*. Oxford: Oxford University Press; 1989.
4. Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol* 1992; 45: 1201–1218.
5. Wright JG, Rudicel S, Feinstein AR. Ask patients what they want. Evaluation of individual complaints before total hip replacement. *J Bone Joint Surg* 1994; 76-B: 229–234.
6. Guyatt GH, Berman LB, Townsend M, Pugsley OS, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987; 42: 773–778.
7. 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.
8. Juniper EF, Guyatt GH. Development and testing of a new measure of health status for clinical trials in rhinoconjunctivitis. *Clin Exp Allergy* 1991; 21: 77–83.

9. Juniper EF, Guyatt GH, Epstein RS, Ferrie PJ, Jaeschke R, Hiller TK. Evaluation impairment of health related quality of life in asthma: Development of a questionnaire for use in clinical trials. **Thorax** 1992; 47: 76–83.
10. Levine MN, Guyatt GH, Gent M, De Pauw S, Goodyear MD, Hryniuk WM, *et al.* Quality of life in stage II breast cancer: An instrument for clinical trials. **J Clin Oncol** 1988; 6: 1798–1810.
11. Kirshner B, Guyatt G. A methodological framework for assessing health indices. **J Chron Dis** 1985; 38: 27–36.
12. 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.
13. Wright JG, Young NL. The patient specific index: Asking patients what they want. **J Bone Joint Surg** 1997; 79-A: 974–983.
14. Stavig GR. Monotonic measures of agreement for ranked data. **Br J Math Stat Psychol** 1984; 37: 283–287.