Editorial

Is Occam's Razor Meaningful for Selecting Significant Outcome Items and to Narrow Down Question Numbers in a Psychometric Scale?

Assessment of interventional results based on patient-reported outcomes brings greater understanding of patients' value judgments of therapeutic effectiveness, and in turn requires development of accurate psychometric instruments[1]. Though patient-reported outcome measures are very important for clinical practice, we cannot measure the function or disability of patients directly. It is absolutely important, therefore, to obtain the information on functional status, health-related quality of life (HRQOL), and other related data such as patients' values and perceptions, through valid and reliable psychological assessments[2].

How can we measure a patient's health condition? "Measuring health" or "measuring disease" are necessary steps in outcome research. A patient-centered questionnaire is a widely used method to collect necessary information from subjects with a targeted condition. It is a core procedure to measure HRQOL with such an assessment. And it is essential to assess the difference in the patient's condition before and after medical intervention, to determine its effectiveness. This is the key reason we must understand the psychometric principles.

Parkes and colleagues, in this issue of *The Journal,* discuss the sensitivity to change of pain measures in knee osteoarthritis (OA)[3]. They conducted a comparative study to investigate the increased sensitivity to change of combining outcomes compared to single measures of pain[3]. They have previously published an article focused on the same topic[4].

How can we manage the number and content of outcome items to sharpen our measuring aim? When applying a psychometric scale to a certain condition, the process of selecting outcome items for research is a very important and interesting topic. A comprehensive approach means many items could cover a wide range of conceptual constructs, but the weakness is in the feasibility, or the statistical handling needed to apply those items to real subjects.

This topic is related to the so-called Occam's razor. Occam's (or Ockham's) razor, also called the law of economy or the law of parsimony, is a principle stated by the Franciscan philosopher William of Occam (1285–1347?): *pluralitas non est ponenda sine necessitate*, "plurality should not be posited without necessity." The principle gives precedence to simplicity: of 2 competing theories, the simpler explanation of an entity is to be preferred. The principle is also expressed as "Entities are not to be multiplied beyond necessity"[5].

To select the most appropriate content for new assessment items, an initial set of questions and items has been changed several times through repeated clinical application. Even as a simplified case, a short version of a certain psychometric scale is often necessary in various aspects of clinical practice[6]. Goetz, *et al* described the methodology currently used to shorten measurement scales through a literature review and compared it with a previous review for proposing updated and structured guidelines for a short version of measurement scales[7]. Factor analysis or item response theory is often used to reduce the number of putative underlying factors and to maintain a similar conceptual architecture framework of a targeted condition[8,9,10].

On the other hand, construction of a psychometric instrument is basically a polysemous assessment. In a tradeoff situation between changing the number of items and sharpening analytic capability, we will be seeking a simpler formula or assessment scale. For example, Beck and Gable described how the *a priori* approach of specifying an instrument's content domain is addressed along with the *a posteriori* procedure of having a panel of judges assess the validation of the items[11]. The problem occurs when narrowing down question numbers.

Three items stand out as significant in the article by Parkes and colleagues[3]:

• The study attempts to evaluate meaningful ways of combining single outcomes to improve responsiveness and gain more power to detect treatment effects without collecting more data.

• Combining outcomes can improve efficiency in future

See Sensitivity to change of pain outcomes, *page 1308*

The Journal of Rheumatology 2018; 45:9; doi:10.3899/jrheum.180264

clinical trials, because it helps improve detection of smaller treatment effects with fewer participants.

• Combining outcomes appears to produce composites with greater sensitivity to change than constituent parts.

In the article, pain and rescue medication outcomes were standardized and combined into 3 composite outcomes through principal components analysis to produce 1 score (composite outcome), and their responsiveness was compared to Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain, the standard. While improvements in sensitivity were modest, the required sample size for trials using composites was 20–40% less than trials using WOMAC pain alone. Parkes and colleagues conclude that combining information from related but distinct outcomes considered relevant to particular treatments improved responsiveness, could reduce sample size requirements in OA trials, and might offer a better way to detect treatment efficacy in OA trials[3]. Fries, *et al* reported the use of computerized adaptive testing to select the best items to sharpen the estimate of a person's functional ability[12,13].

In Parkes, *et al*, the authors used the term *sensitivity* to mean responsiveness (sensitivity to change) of a scale[3]. Their target was pain assessment. Pain usually has a direct effect on daily living activities and has a special position in orthopedic problems such as knee OA. For example, pain is more variable because of disease condition than mood or perception of other related scores. Pain measurement appears particularly suited to the item reduction approach, given its complexity. Regarding WOMAC score, Stratford and Kennedy pointed out that activity overlap on the pain and function subscales plays a causal role in limiting the WOMAC physical function subscale's ability to detect change[14].

Combining information from several different domains may improve a composite's ability to detect a change when one truly occurs, and therefore responsiveness may also be improved. In Parkes, *et al*, repeated measurements are carried out using the SAS PROC MIXED procedure[3]. It is a method of getting the results without reducing the amount of information. However, I am afraid that the relationship between statistical power and responsiveness (sensitivity to change) of outcome measures has a tradeoff response to other aspects of psychometric measurement as well: confounding, minimum clinically important difference (MCID), and response shift[15,16].

Among the scale items to measure psychometric properties, it is inevitable to get some confounding factors mixed in. If there is a certain strong item having close connection with others, the change of such an item directly influences the relationship among items. It is important to consider this issue. The remaining problem on MCID is also important. My last concern is response shift of the participants during followup, such as a change in an individual's values, internal standards, and conceptualization of QOL on QOL assessments.

I do not know whether a single pain score (WOMAC) is the optimal, standard measure. WOMAC pain score as well as stiffness score are just categorical ones. As previously documented, categorical scores are less sensitive than continuous ones; especially, its distribution is relatively narrow. Therefore, the predictive power is usually lower for categorical scores than for continuous ones. When both scores are collected on the same individuals, it could be possible to compare. Ultimately, continuous and categorical scores serve different purposes.

But a carefully tested measure that covers many aspects of validities could be the most appropriate one.

As with any use of mathematical models, it is important to assess the fit of the data to the model. In item response theory, item characteristic curve is a step to identify the meaning of each item. The results provided in Parkes and colleagues[3] on WOMAC pain scale show that it is a clear confounding factor in the scale. Apart from conventional factor analysis or principal component analysis, covariance structure analysis or indices of model fit can make the domain structure clear. Akaike information criterion (AIC) for model fitting is a way to find the appropriate combination of explanatory variables to explain the objective variable (i.e., the most suitable combination of items through a mathematical method)[17]. It could be a powerful procedure to investigate the status of confounding factors using statistical analysis. Several indices of model fit including the AIC are also available to identify the domain architecture of the scale, and the stepwise method of multiple variate analysis can identify the contribution of each item.

Iwaya, *et al* reported on the relationship between subjective assessment and objective evaluations of locomotive function in the elderly[18]. A self-reported scale provides precise information on disabilities affecting activities of daily life and proportionally reflects physician-judged dysfunction grade. A carefully organized psychometric questionnaire could have powerful analytic capability equal to a physician's assessment. We have to continue our efforts to identify important items contributing to the main construct, to sharpen analytic capability.

MASAMI AKAI, MD, PhD,
Vice-dean and Professor,
Graduate School,
International University of
Health and Welfare,
Tokyo, Japan.

Address correspondence to Dr. M. Akai, Graduate School, International University of Health and Welfare, 4-1-26 Akasaka, Minato-ku, Tokyo 107-8402, Japan. E-mail: akai-masami@iuhw.ac.jp

REFERENCES

1. McDowell I. Measuring health: a guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.

2. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622-9.
3. Parkes MJ, Callaghan MJ, Tive L, Lunt M, Felson DT. Responsiveness of single versus composite measures of pain in knee osteoarthritis. J Rheumatol 2018;45:1308-15.
4. Parkes MJ, Callaghan MJ, O'Neill TW, Forsythe LM, Lunt M, Felson DT. Sensitivity to change of patient-preference measures for pain in patients with knee osteoarthritis: data from two trials. Arthritis Care Res 2016;68:1224-31.
5. Duignan B. Occam's razor. Encyclopedia Britannica. [Internet. Accessed March 21, 2018.] Available from: www.britannica.com/topic/Occams-razor
6. Stanton JM, Sinar EF, Balzer WK, Smith PC. Issues and strategies for reducing the length of self-report scales. Pers Psychol 2002;55:167-94.
7. Goetz C, Coste J, Lemetayer F, Rat AC, Montel S, Recchia S, et al. Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. J Clin Epidemiol 2013;66:710-8.
8. Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. J Clin Epidemiol 1997;50:247-52.
9. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007;16 Suppl 1:5-18.
10. Beck CT, Gable RK. Item response theory in affective instrument development: an illustration. J Nurs Meas 2001;9:5–22.
11. Beck CT, Gable RK. Ensuring content validity: an illustration of the process. J Nurs Meas 2001;9:201–15.
12. Fries JF, Witter J, Rose M, Cella D, Khanna D, Morgan-DeWitt E. Item response theory, computerized adaptive testing, and PROMIS: assessment of physical function. J Rheumatol 2014;41:153-8.
13. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther 2011;13:R147.
14. Stratford PW, Kennedy DM. Does parallel item content on WOMAC's pain and function subscales limit its ability to detect change in functional status? BMC Musculoskelet Disord 2004;5:17.
15. Lee MK, Yost KJ, McDonald JS, Dougherty RW, Vine RL, Kallmes DF. Item response theory analysis to evaluate reliability and minimal clinically important change of the Roland-Morris Disability Questionnaire in patients with severe disability due to back pain from vertebral compression fractures. Spine J 2017;17:821-9.
16. McPhail S, Haines T. Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. Health Qual Life Outcomes 2010;8:65.
17. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory. Petrov BN, Caski F, editors. Budapest: Akadémiai Kiadó; 1973:267-81.
18. Iwaya T, Doi T, Seichi A, Hoshino Y, Ogata T, Akai M. Relationship between physician-judged functioning level and self-reported disabilities in elderly people with locomotive disorders. Qual Life Res 2017;26:35-43.

J Rheumatol 2018;45:1208–10; doi:10.3899/jrheum.180264