Conceptualization and Measurement of Health-Related Quality of Life: Comments on an Evolving Field

John E. Ware Jr, PhD

ABSTRACT. Ware JE Jr. Conceptualization and measurement of health-related quality of life: comments on an evolving field. Arch Phys Med Rehabil 2003;84 Suppl 2:S43-51.

This article summarizes personal views on the rapidly evolving field of functional health assessment and comments on their implications for advances in assessment methods used in rehabilitation medicine. Topics of strategic importance included (1) a new formulation of the structure of health status designed to distinguish role participation from the physical and mental components of health for purposes of international studies; (2) applications of item response theory that offer advantages in constructing better functional health measures and cross-calibrating their underlying metrics; (3) computerized dynamic assessment technology, well proven in education and psychology, which may lead to more practical assessments and more precise score estimates across a wide range of functional health levels; and (4) intellectual property issues involved in standardizing and promoting readily available assessment tools, ensuring their scientific validity, and achieving the best possible partnership between the scientific community and those developing commercial applications. Promising results from preliminary attempts to standardize and improve the metrics of functional health assessment constitute grounds for optimism regarding their potential usefulness in rehabilitation medicine. Someday, all tools used to measure each functional health concept, including the best single-item measure and the most precise computerized dynamic health assessment, will be scored on the same metric and their results will be directly comparable. To achieve this goal in rehabilitation medicine, we have much work to do.

Key Words: Computers; Health status; Outcome assessment (health care); Quality of life; Questionnaires; Rehabilitation; Treatment outcome.

THE FIELD OF FUNCTIONAL health assessment appears to be evolving at a more rapid rate, and I would like to comment on some of these developments and their implications. I address 4 topics of strategic importance, including (1) a new formulation of the structure of health status, (2) applications of item response theory (IRT), (3) computerized dynamic health assessment, and (4) standardization and intellectual property issues.

My objectives are to stimulate research and to interest others in what I believe to be promising opportunities for advances in patient-based methods of health status assessment in rehabilitation medicine. Preliminary results from our attempts to standardize the metrics of functional health assessment, including those used in rehabilitation medicine, and to develop more practical methods of data collection and processing make me optimistic regarding the potential contributions from efforts in the pursuit of the 4 topics noted above.

A NEW CONCEPTUAL FRAMEWORK

Investigators worldwide are considering new formulations of the structure of health, including new approaches to the conceptualization of physical functioning and social and role participation.[1] Because such reformulations have substantial implications for the blueprints we follow in formulating hypotheses about the major components of health, the domains that best represent each component, as well as our approaches to crafting specific operational definitions, it is very timely to at least briefly consider them. One in particular—the functional health domain of *role participation*—has implications for the scoring and interpretation of aggregate measures of health.

For more than 20 years, my colleagues and I have found it useful to make distinctions among the physical, mental, and social dimensions of *individual* health status and to distinguish measures of social and role participation from the many other indicators of the functional aspects of those domains.[2] As early as 1984,[3] we were using concentric circles and the metaphor of health as an onion in making such conceptual distinctions and in discussing the interrelationships among the layers of health; and I still find this metaphor useful today. At the core are biologic health and the hundreds of disease-specific measures of the physiology and functioning of various organ systems commonly used in diagnosis and treatment evaluation. The outermost layer—quality of life (QOL)—is still regarded as a much broader concept reflecting the dozen or more domains of life, including community, family, and work.[4] As discussed previously, a disruption in any of the multiple layers of the health onion could impact an inner or outer layer of health. For example, a disease or bodily injury could impair physical and mental functioning leading to problems at home or at work. Dissatisfaction with life could lead to organ-level dysfunction and so on. Neither the disease-specific core nor the outermost QOL layer will be discussed here so that I may focus on the domains of health-related quality of life (HRQOL) that are most affected by disease or injury and by treatment.[3,5]

As shown in figure 1, we now distinguish between 2 principal components of HRQOL—physical and mental—that can be thought of as 2 multilayered health onions. The distinction between them makes both conceptual sense and is strongly supported by empirical studies: for example, factor analyses of 12 Sickness Impact Profile (SIP) scales,[6] of the 28 scales from the Health Insurance Experiment[7] (HIE), of the 19 scales from the Medical Outcomes Study[8] (MOS), and of other widely used

From QualityMetric Inc, Lincoln, RI; and Health Assessment Lab, Boston, MA.

Supported in part by the National Institute on Disability and Rehabilitation Research, Rehabilitation Research and Training Center on Measuring Rehabilitation Outcomes (grant no. H133B990005).

Presented at the Quality of Life Measurement: Building an Agenda for the Future Conference, November 16, 2001, Parsippany, NJ.

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the author(s) is/are associated.

Reprint requests to John E. Ware Jr, PhD, QualityMetric Inc, 640 George Washington Hwy, Lincoln, RI 02865, e-mail: *jware@qualitymetric.com.*

0003-9993/03/8404-7694$30.00/0
doi:10.1053/apmr.2003.50246

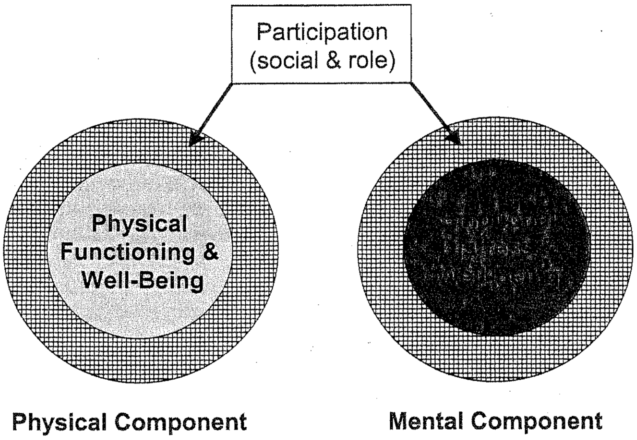

Fig 1. Current conceptualization of HRQOL.

generic health surveys, including analyses of different permutations of Dutch translations of 21 scales from the Dartmouth COOP chart system, EuroQOL, Nottingham Health Profile, and the MOS 36-Item Short-Form Health Survey (SF-36).[9] In addition to establishing that physical and mental components of health account for the great majority of the variance in comprehensive generic health surveys, such empirical studies suggest that a variety of different indicators (eg, functional impairments, subjective ratings, personal evaluations) can be meaningfully aggregated for purposes of estimating scores for each health component.

In contrast to the HIE, in the MOS we changed our approach to measuring the role dimension of participation by constructing distinct scales for each of the 2 principal components of health.[8] In the MOS, instructions to respondents to make attributions to "physical health" versus "emotional problems," as opposed to "health" in general, markedly changed the validity of role functioning scales in the full-length MOS battery of measures,[8] as well in the much shorter SF-36 that was developed from that battery.[10] As shown in figure 1, the result was physical- and mental-specific measures of limitations in role participation characterized as the outer layers of the physical and mental components of health, in our current conceptualization. We are currently testing measures of limitations in role participation that make no attributions to health or any other causes. These may be the ultimate outcomes measures for use in evaluating the broadest array of interventions.

In the meantime, there are good reasons to consider a new conceptualization of HRQOL, such as that shown in figure 2. In the new conceptualization, the distinction between the 2 principal components of health—physical and mental—is retained. However, a third principal component—participation (role, social)—is hypothesized. There are at least 4 good reasons for measuring and interpreting role participation separately as proposed in this new conceptualization: (1) the new *International Classification of Functioning, Disability and Health*[1] advocates for a separate measurement and interpretation of this domain of health; (2) each individual's performance of his/her social role has been favored as an ultimate "bottom line" in judging HRQOL for decades,[11] (3) recent empirical findings suggesting that the distinction between physical and mental causes of role limitations are not made or are made differently in Japan and other countries in Asia,[12] and (4) the fact that economists make no distinction between a physical and an emotional cause of a restriction in an individual's participation because each restriction has the same utility regardless of its cause.[13]

Accordingly, the proposed new conceptualization calls for the construction, scoring, and interpretation of role participation as a component distinct from the physical and mental components of health. It is time to formally test this hypothesized new higher-order structure, for example, by using structural equation modeling and correlation matrices from the United States, Japan, and other countries. In addition to addressing the conceptual and methodologic issues listed previously, an added advantage of the new conceptualization is that it would facilitate studies of the implications of differences in physical and mental capacities for an individual's participation in life activities. For example, imagine a causal model in which these capacities predict overall participation. Further, governmental agencies seem to be moving in this direction, for example, the National Institute on Disability and Rehabilitation Research (NIDRR), with its call for improved measures of participation for use in evaluating rehabilitation facilities. Currently, a 5-year NIDRR-sponsored effort is underway to develop and validate participation measures for use in evaluating rehabilitation facilities.

APPLICATIONS OF ITEM RESPONSE MODELS

Our first applications of IRT were made in the early 1990s, with the goal of examining the correlation between "modern" (IRT-based) and "classical" approaches to scale construction and scoring, such as the Likert[14] method of summated ratings. We adopted the latter in scoring the functional health and well-being measures used in the HIE[7] and in the MOS.[8] Our 1994 article[15] testing the unidimensionality and reproducibility of the 10-item physical functioning scale is to our knowledge the first published application of IRT to functional health assessment. From that study, we concluded that Rasch–IRT scoring is an alternative to the current Likert-based summative score approach. In light of subsequent developments, that conclusion was an understatement.

A thorough discussion of the Rasch and more general IRT models we have used since then is beyond the scope of this commentary. However, a brief summary of the steps followed

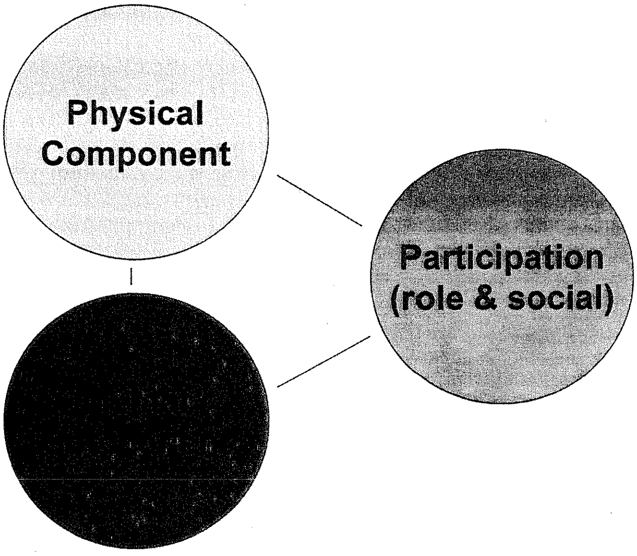

Fig 2. Proposed new conceptualization of HRQOL.

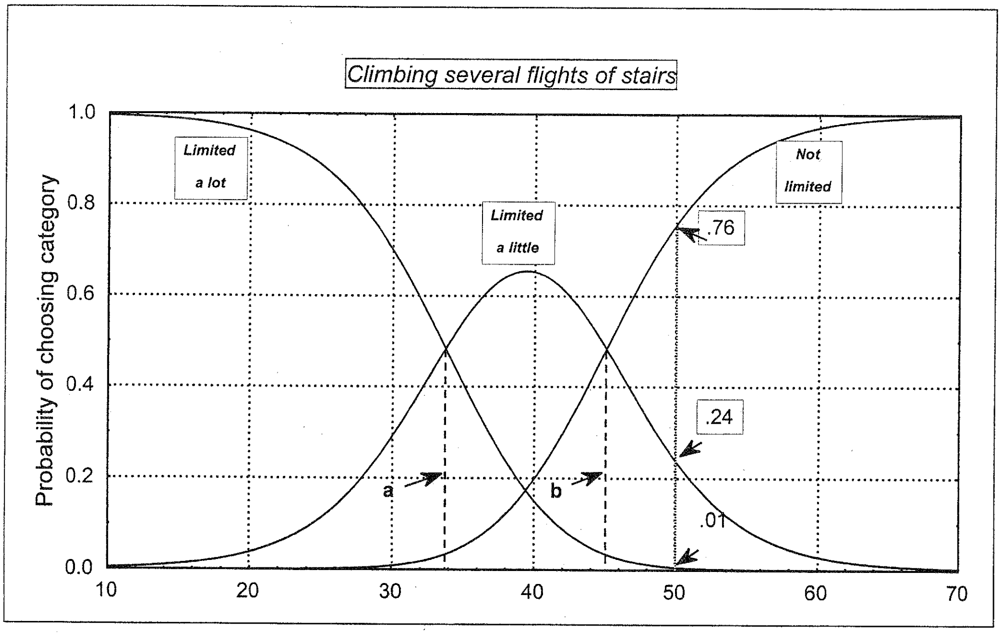

Fig 3. IRT model for physical functioning–10, item 4. Reprinted with permission.[40]

and an example of an IRT model for 1 physical functioning item may set a better stage for explaining what my colleagues and I believe are the potential advantages of such models as well as the challenges involved in their applications. Briefly, I begin with (1) basic descriptive analyses of the relationship between each item and a crude estimate of the underlying latent variable it is hypothesized to measure using TestGraf software,[16] (2) formal tests of the dimensionality of items in each "pool" using methods of factor analysis for categorical data (eg, using Mplus software[17]), and (3) a generalized partial credit IRT model[18] and the marginal maximum likelihood estimation procedures of the PARSCALE software[19] to calibrate the items. We recommend these methods to measure physical activity and role participation in rehabilitation medicine.

IRT models are statistical models applicable to virtually all categorical rating scale variables measuring the same domain (eg, the 10 physical functioning items). As shown in figure 3, an IRT model is a probability (vertical axis) of the selection of each item response category as a function of the score (horizontal axis) on the underlying latent variable. In this example, the question (item 4 in the physical functioning scale) is about limitations in "climbing several flights of stairs," and the categorical rating scale offers 3 response choices: "limited a little," "limited a lot," and "not limited." As documented elsewhere,[20] the probabilities shown in figure 2 were estimated from a partial credit IRT model for physical functioning scores with a mean of 50 and a standard deviation (SD) of 10 in the general US population by using norm-based scoring.[21] The curves (trace lines) in figure 3 define important characteristics of this item. For example, according to the model, a person with an average score (a score of 50) has a .76 probability of choosing "not limited," slightly less than .24 probability of choosing "limited a little," and less than .01 probability of choosing "limited a lot." Other noteworthy features of figure 2 include the vertical dashed lines (labeled a, b) at 2 of the intersections of the item trace lines, where respondents are equally likely to choose the 2 adjacent response categories. For example, at a score of about 45, choices of the second and third response categories are equally likely. These thresholds are important because they define the "difficulty" of each response category. This information is used in estimating each person's physical functioning score and in selecting items for dynamic administrations of physical functioning items based on computerized adaptive test logic.

As illustrated elsewhere,[20] item characteristic curves can be combined to determine the probability of all patterns of item responses and to estimate a likelihood function for each pattern. The resulting likelihood functions are the basis for unbiased estimates of scores for the underlying latent physical functioning variable, as well as estimates of the reliability and confidence intervals (CIs) associated with that score. These estimates represent a difference between classical psychometric methods, which yield 1 reliability coefficient that we apply to scores at all levels. In contrast, modern methods yield a score estimate and a reliability coefficient that is specific to a given score level. This important distinction will come up in the discussion of computerized adaptive testing applications to functional health assessment.

Clearly, we are witnessing increased interest in applying IRT methodology in the analysis of surveys of functional health status and well-being.[22-24] There are advantages to the IRT methodology that account for this interest, including the following: (1) the models yield more detailed evaluations of the measurement properties of each questionnaire item, (2) estimates of score precision (eg, reliability, standard errors) are specific to the score level, (3) the contribution of each item to the overall test precision and to a score in a specific score range can be estimated, (4) the models yield unbiased estimates of the underlying latent variable score from any subset of items in the "pool," (5) the models can be used to evaluate whether each person is responding consistently across items, (6) the models make it possible to cross-calibrate scores for different scales measuring the same concept, and (7) the models make dynamic health assessment—using the computer to select the most appropriate items and to determine the optimal test length—

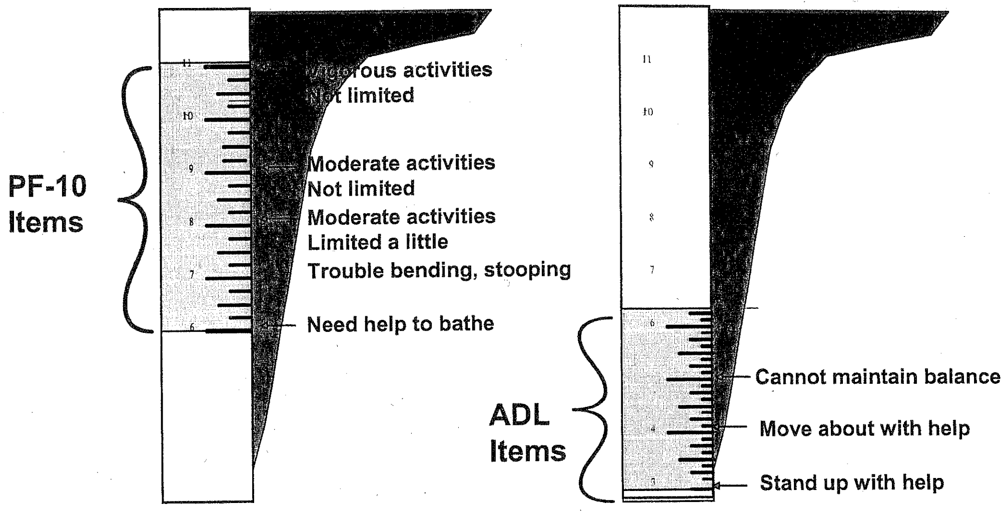
PF-10 Items
Vigorous activities Not limited
Moderate activities Not limited
Moderate activities Limited a little
Trouble bending, stooping
Need help to bathe
ADL Items
Cannot maintain balance
Move about with help
Stand up with help

Fig 4. Combining measures to lower the floor for the physical functioning–10 (PF - 10) items[8] and ADL items.[41]

possible with results that can be compared even when different items are administered.

Raising the Ceiling and Lowering the Floor

Large datasets are providing opportunities to explore approaches to extending the range of functional health assessment and have provided some previews of the practical implications of doing so. For example, the ongoing Medicare Health Outcomes Survey (HOS), which is the largest outcomes survey ever undertaken by the Centers for Medicare and Medicaid Services, formerly the Health Care Financing Administration, is a source of preliminary calibrations of physical functioning items and items measuring activities of daily living (ADLs; eg, ability to stand up with help) on a common metric. As shown in figure 4, analyses of the first 100,000 respondents to the HOS survey confirmed our expectations about the relative placements of the physical functioning (fig 4, left panel) and ADL items (right panel) on a common metric. These items "fit" a common physical functioning IRT generalized partial credit model. (Item calibrations and a summary of estimation methods used are available on request.)

The shaded part in each panel characterizes the highly skewed distribution of physical functioning scores in the general US population. The practical implication is that about 3% of Medicare HOS respondents score at the floor of the physical functioning scale, which is defined as the lowest physical functioning item threshold ("Need help to bathe"). As shown in the right panel of figure 4, IRT thresholds estimated for the ADL items fielded in the HOS were well below those estimated for the physical functioning items and, thus, extended substantially the range measured. The practical implication is that more than 90% of those scoring at the floor of the physical functioning scale can be meaningfully measured by using a combined physical functioning–ADL metric. That combined scale has already proven useful in predicting mortality in the HOS.[25] It should be noted that figure 4 also shows a substantial ceiling effect. In the United States and other developed countries, nearly 40% of adults score above the highest physical functioning threshold ("vigorous activities").[26] Work in progress suggests that the ceiling of the next generation item pool for the physical functioning domain will be more than 1 SD higher than the ceiling shown in figure 4. Questionnaires used in sports medicine and items from other sources have proven useful in measuring these higher levels of physical functioning.

A noteworthy theoretical advantage of IRT models is that they are "scale" free. In other words, the addition of items and the new "marks on the ruler" that they define do not affect the calibrations of the items that are already there. Specifically, for example, lowering the physical functioning floor in figure 4 by using the ADL items in the right panel will not change the relative placements of the physical functioning items. However, to apply this advantage in practice we must abandon one of our most widespread approaches to the scoring of functional health scales, namely, 0 to 100 transformations. In the past,[8] we have transformed the physical functioning scale and other summated rating scales so that their lowest and highest measured levels are scored as 0 and 100, respectively. Scores in between indicated the percentage of the range in between those extremes. Obviously, if we lower the floor of the physical functioning scale using the ADL items as shown in figure 4 and transform the new scale to 0 to 100, the resulting levels that the new and old metrics have in common will no longer be comparable. Our solution, which is norm-based scoring, has proven quite useful in addressing this issue in educational and psychologic testing for about 100 years. By using norm-based scoring, scores are expressed as deviations from a measure of central tendency (eg, the average) instead of the extremes of the range. Accordingly, as we raise the ceiling and lower the floor, the placements (and scoring) of the item thresholds in relation to the average do not change. This simple linear transformation, with a mean of 50 and SD of 10, and their practical implications are illustrated elsewhere.[27]

Other Advantages of IRT Models

Before discussing the ultimate practical implication of IRT models—that they are the psychometric basis for computerized adaptive testing applications to health assessment—I take this opportunity to explain briefly other noteworthy advantages, including missing data estimation and the cross-calibration of health metrics.

The same models that make it possible to estimate meaningfully and compare scores for respondents who answer different sets of questions also make unbiased estimates of health scores possible even when some responses are missing. There are 2 advantages of the IRT-based approach to the missing data

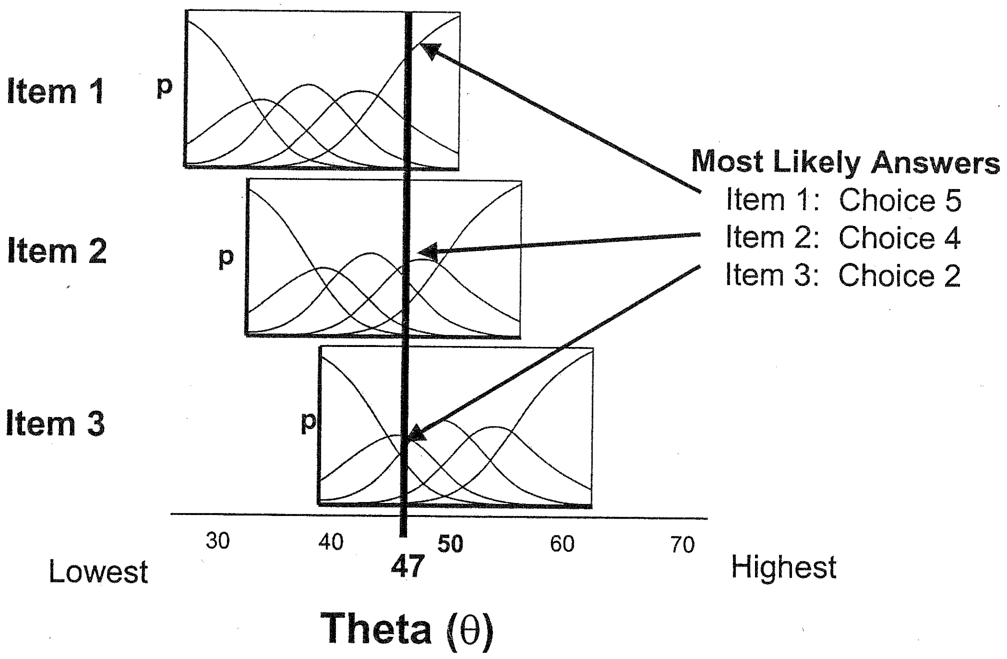

Fig 5. Responses to all calibrated items are predictable from θ.

problem. First, many scores that would have been missing using classical methods can be estimated. For example, the percentage of computable SF-36 scores that could be estimated for elderly MOS participants by using standard scoring and missing data estimation methods was increased from 82.95% to 94.74% using IRT-based methods in reanalyses of data from the MOS. In the Medicare HOS,[25] scores for nearly 40,000 respondents with 1 or more missing SF-36 responses at baseline were recovered and estimated without bias. For this reason, the National Committee for Quality Assurance uses the new scoring software for the SF-36 that incorporates missing data estimation algorithms data for its Health Plan Employer Data & Information Set Medicare outcomes survey[25] and makes this scoring service available to the approximately 200 participating health care plans on the Internet.

Second, IRT models will ultimately make it possible to cross-calibrate measures of each functional health concept, much like how the Celsius and Fahrenheit thermometers have been cross-calibrated. To show the implications of this potential advance, we used IRT models and nonlinear regression to cross-calibrate 4 widely used measures of headache impact and published a preliminary table for use in equating scores across those instruments, and in relation to a norm-based "criterion" score (θ) based on all items from all instruments.[20] More recently, Bjorner and Kosinski[28] replicated and extended this work by using an IRT model for the items in 7 measures of headache impact and virtually all scores possible for each of the 6 measures and for the criterion (θ) score. These models will ultimately make it possible to estimate scores for widely used measures without actually administering them. Once the "true" score, θ, has been estimated, the most likely response choice is "known" in a probabilistic sense, for all items in the calibrated "pool." For example, as shown in figure 5, it is most likely that those with θ equal to 47 will select choice 5 for item 1, choice 4 for item 2, and choice 2 for item 3. Accordingly, once θ has been estimated, scores for all measures sufficiently represented in the same item pool along with their associated CIs can be estimated. As documented elsewhere,[28] these IRT-based estimates of a headache impact scale are sometimes more accurate than those based on all of the original items and the developer's scoring algorithms. Given that θ estimates based on computerized dynamic health assessments are often reliable enough to interpret after asking only 5 or 10 items, as discussed later, it may someday be possible to estimate ADL, FIM™ instrument, SIP mobility, and the physical functioning scale scores for most respondents after only 1 to 2 minutes of data collection over a wide range of functional health levels.

Not All Items Fit IRT Models

Some generic instruments such as the SIP,[6] the Health Perceptions Questionnaire[29] (HPQ), and the Nottingham Health Profile[30] have been constructed by using verbatim comments from patients under care and from consumers in general. These questionnaires have the potential advantages of better representing the domains of health, as well as the vernacular actually used in describing them. A more recent example of a disease-specific instrument constructed from patients statements is the Headache Disability Inventory[31] (HDI). However, when such statements were combined with 5-choice categorical rating scales, such as the "definitely true" to "definitely false" categories used with the HPQ and those tested more recently[32] with the disease-specific HDI, they did not fit the pattern assumed by the IRT model. To fit the IRT model, the neutral and middle response categories for both the generic HPQ and disease-specific HDI items had to be collapsed into 3-choice categorical rating scales. We are still learning about the tradeoffs involved in different approaches to constructing questions and response categories.

COMPUTERIZED DYNAMIC HEALTH ASSESSMENT

Rehabilitation medicine and most other applications of patient-based assessments are asking for more *practical* solutions to their data collection and processing requirements so that the screening of patient needs and outcomes monitoring efforts can

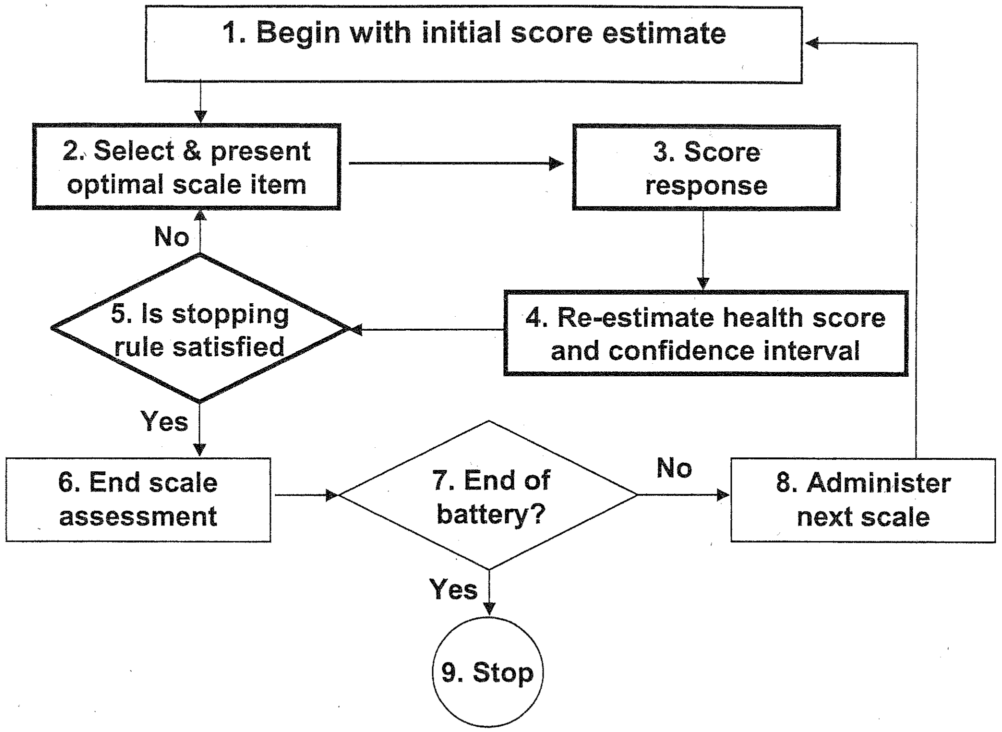
1. Begin with initial score estimate
2. Select & present optimal scale item
3. Score response
4. Re-estimate health score and confidence interval
5. Is stopping rule satisfied
No
Yes
6. End scale assessment
7. End of battery?
No
8. Administer next scale
Yes
9. Stop

Fig 6. Logic of computerized dynamic health assessment. Adapted from Wainer.[33]

take place on a much larger scale. Short-form surveys such as the SF-36 are steps in this direction. However, the very features of short-form surveys that make them more practical are typically achieved by restricting the range that they measure or by settling for an unacceptably large amount of noise at one or more scale levels. To the contrary, tools that are more precise are required for some applications, for example, tools used to inform medical decision making at the individual patient level. In educational testing[33] and in other applications,[34] computerized adaptive testing–based methods have proven useful in achieving assessments that are both more precise and more practical. Work in progress in rehabilitation medicine and related fields suggests that computerized adaptive testing can achieve the same advantages for measures of functional health. Finally, we need to standardize the concepts and cross-calibrate the metrics of functional health assessment to meet the needs of assessment across diverse populations and purposes so that results can be meaningfully compared and interpreted.

Logic of Computerized Adaptive Testing

In contrast to a traditional "static" survey administration, in which a score is estimated after all items have been answered, computerized adaptive testing uses a preliminary score estimate to select the next most informative question to administer. Figure 6 shows the sequence of steps inherent in computerized adaptive testing administrations. First, an initial score estimate (step 1) is the basis of selecting the most informative item for each respondent from a pool of items that have been calibrated by using item response modeling. The most informative item, according to that model, is administered (step 2) and the item is scored (step 3). At step 4, the respondent's level of health is reestimated along with a respondent-specific CI. At step 5, the computer determines whether the score has been estimated with sufficient precision by comparing the CI with a preset standard. If not, the cycle (steps 2–5) is repeated until the precision standard is met or a preset maximum number of questions have been answered. When the stopping rule is satisfied, the computer either begins assessing the next health concept or ends the battery.

For both simulated and actual computerized adaptive testing–based functional health assessments to date, we have programmed our software to administer the same initial global question to all respondents for purposes of "priming" the computerized adaptive testing "engine." For purposes of assessing headache impact, this initial question was selected to discriminate well over a wide range (ie, item thresholds extending across nearly 60% of the range covered by the entire item pool). DYNHA® Health Assessment[a] was programmed to match the precision standard to the specific purpose of measurement for each patient. For example, for moderate and severe scores, which are most likely to qualify a patient for disease management, items are administered until the highest level of precision is achieved. The DYNHA system is documented and dynamic versions of the Headache Impact Test™ and generic health measures are shown elsewhere (*http://www.qualitymetric.com, http://www.amIhealthy.com*).

Lessons From Computerized Adaptive Testing Applications to Date

Most of our experiences with computerized adaptive testing–based functional health assessments to date have come from studies of adult headache sufferers. We have developed dynamic impact assessment software for them because the causes of headaches are often undiagnosed and untreated. It was hoped that more accurate and user-friendly information about the disability and distress caused by migraines and other headaches would be useful to patients and health care providers. We began by calibrating items from widely used surveys of headache impact using IRT as described in detail elsewhere.[20,35] Second, simulated computerized adaptive testing administrations were evaluated to approximate the accuracy of dynamic estimates and to determine the extent of reductions in respondent burden likely to be gained using computerized adaptive testing logic. In the first simulation, 1016 headache sufferers were administered all 53 items from 4 headache impact instruments. A single headache impact score, which was estimated from 47 items that fit an IRT model, was compared with the estimate from computerized adaptive testing simulations. For purposes of the simulations, responses to up to 5 items selected on the basis of item information functions were fed to the simulation software for each respondent. When scores estimated from these 5 or fewer responses were compared with estimates based on all 47 items, a product-moment correlation of .938 was observed, indicating a high degree of agreement. The practical implications of computerized adaptive testing were apparent in the counts of respondents who achieved a reliable score, that is, with a measurement error of 5 points or less, using only 5 items in the simulation study. That standard of precision was met for 99% and 98% of those with migraine headache and severe headache in the initial simulations. Such simulations are possible whenever datasets include responses to all items in an item pool under the assumption that answers to a subset of those items selected using computerized adaptive testing would have been identical to the answers given when they were embedded in the source instruments.

As reported elsewhere,[36] the substantial reductions in respondent burden estimated from computerized adaptive testing simulations were replicated during the first 20,000 computerized adaptive testing administrations using DYNHA on the Internet in the fall and winter 2000. As expected, for 75% of respondents, mostly women migraine sufferers between the ages of 25 and 54, a precise estimate of headache impact was achieved within 5 or fewer questions. Those estimates were strongly related to headache severity and frequency, as hypothesized. On the strength of these findings, Bayliss et al[36] recommended tests of computerized adaptive testing–based dynamic health assessments among patients with other diseases and conditions.

Work in progress suggests that more efficient estimates of physical activity and role participation may also be possible using computerized adaptive testing logic among adults in rehabilitation settings. Preliminary results from the first computerized adaptive testing–based simulations by our NIDRR-sponsored research team are very promising. For example, substantial reductions in respondent burden at acceptable levels of score precision are likely from computerized adaptive testing–based administrations. The first tests focused on items measuring physical activity, including those from the FIM instrument, the Outcome Assessment Information Set, and new items from the Activity Measure for Post-Acute Care. Briefly, we administered 101 such items to 485 patients sampled from inpatient, skilled nursing, and outpatient facilities and calibrated them using a generalized partial credit IRT model. Simulations were performed by using the same DYNHA software described earlier. When scores were estimated dynamically from 6 or fewer items, a very high level of agreement was observed in comparison with criterion scores estimated from an IRT model for all 101 items (r=.95, N=485). CIs for scores estimated dynamically were 5 points or less with 6 or fewer items for 94% of patients, despite the substantial reduction in respondent burden (nearly 95%). The preliminary results, which are encouraging, are currently being replicated by using item pools representing role participation and other functional health domains. Telephone administrations with speech recognition by computer are also being evaluated in other populations. It remains to be determined whether these technologies will be useful among those who are visually impaired and how best to assess rehabilitation patients whose physical disabilities or socioeconomic backgrounds make it difficult for them to complete self assessments.

STANDARDIZATION AND INTELLECTUAL PROPERTY ISSUES

With the widespread standardization of metrics for quantifying the major domains of functional health and well-being, intellectual property issues must be addressed. First, for the standards to be well understood and accepted, they must be well documented and readily available. Second, we must use registered trademarks and other strategies to protect the use of labels assigned to measures, to distinguish between those that do and do not reproduce accepted standards. In a nutshell, standardization is essential to achieving results that can be meaningfully interpreted. The importance of these issues is reflected in the fact that one of the longest chapters in the textbook currently used in the introductory clinical research course at the National Institutes of Health is about intellectual property and technology transfer[37]; that text also includes a chapter about HRQOL.[38]

The new phase we are entering requires the transfer of measurement technology from the scientific community to the health care industry and requires agreement on the most appropriate business model for widespread adoption of standards and rapid advances in the technology of health status assessment. How do we promote public metrics that are also protected to ensure their scientific validity across a variety of applications? What partnerships between the scientific community and those who profit from commercial applications of the new tools will work best?

The relationship among the Medical Outcomes Trust, Health Assessment Lab, and QualityMetric Inc is an example of how one might attempt to resolve these crucial issues. These 3 organizations established common policies for granting permissions for the use of the SF-36 Health Survey and other widely used tools and merged their licensing programs for commercial applications. Their goals, which are explained on the Internet (eg, *http://www.sf-36.com*, *http://www.qualitymetric.com*), include (1) maintaining the scientific standards for surveys and scoring algorithms that make results directly comparable and interpretable, (2) making surveys available royalty free to individuals and organizations who collect their own data for academic research, and (3) a commercial licensing program that includes royalty payments by those who profit from the use of the intellectual property in support of the research community that is advancing the state of the art. This triumvirate of organizations will attempt to establish still another milestone when it begins posting IRT calibrations for the SF-36 and other widely used functional health measures on their websites in 2002. The initial response from both the scientific community and industry has been very favorable, as evidenced by more than 1000 applications for licenses, including government agencies and health care survey firms. It is hoped that this example will prove useful in addressing these important issues and that others will share their ideas for promoting health assessment standards and the public-private partnerships required to make them more available to all.

CONCLUSION

The parallels between the issues and controversies involved in scale development in the health outcomes field and the history of the evolution of thermometers are noteworthy and deserve brief mention here. Like health, temperature is a *hypothetical construct* that cannot be observed directly and, therefore, must be inferred from such observables as the expansion of a liquid in a glass tube or a metal coil connected to a dial. There are pros and cons in using alcohol as opposed to mercury as the liquid and in using open as opposed to sealed tubes. Temperature readings based on open tubes are influenced more by environmental circumstances (eg, differences in atmospheric pressure) much like the environment influences performance on some functional health measures. According to Klein's historical survey,[39] temperature scales were initially called thermoscopes because they lacked the precision necessary to do more than rank order the objects they measured, and over- and underestimations of hot and cold based on these scopes were initially accepted because of the corresponding subjective bias known to prevail in the individual experience of temperature.

The first thermometers used different scales, and hot and cold were not even scored in the same direction. Both Daniel Gabriel Fahrenheit and Anders Celsius attempted to quantify the freezing and boiling points of water, which were labeled, respectively, as 32° and 212° by Fahrenheit and as 100° and 0° by Celsius. Much like today's truncated 0 to 100 functional health scales, the Celsius thermometer focused on an arbitrary ceiling and floor and an arbitrary, if not counterintuitive, direction of scoring. It was not until after his death that the "upside down" scale of temperature constructed by Celsius was reversed so that water freezes at 0°; he had scored the freezing point at 100°. Although his metric was clearly not the best, Fahrenheit's thermometer became very popular because of a crucial interpretation guideline he provided, namely, the temperature of blood in a healthy person. The availability of norms and other interpretation guidelines are also likely to be crucial factors in the adoption and usefulness of functional health scales. With the introduction of thermodynamics, physicists now have "true" units of temperature. When will we have those units for functional health?

Perhaps, a good place to start would be with the cross-calibration of representative items sampled from widely used forms. Although IRT models require special skills and unfamiliar software, they have the potential to take rehabilitation medicine, and the health outcomes field in general, to a much higher plateau. The advantages of standardizing the metrics used to assess a core set of health concepts can be substantial as evidenced by the SF-8 and the SF-36. For physical functioning, and for every other important health and well-being concept, alternative forms of measures that vary in length according to the precision required for each application, should be calibrated and scored on the same metric. It is no longer necessary or even desirable to limit ourselves to short forms that are embedded in longer forms. Someday all forms, including the best single-item measure of a particular concept and the most precise computerized dynamic health assessment will be scored on the same metric and the results will be directly comparable. To achieve this goal in rehabilitation medicine, we have much work to do.

References

1. World Health Organization. ICF: international classification of functioning, disability and health. Geneva: WHO; 2001.
2. Ware JE Jr, Brook RH, Davies-Avery A, et al. Conceptualization and measurement of health for adults in the Health Insurance Study: volume I, model of health and methodology. Santa Monica (CA): RAND Corp; 1980. Publication No. R-1987/1-HEW.
3. Ware JE Jr. Conceptualizing disease impact and treatment outcomes. Cancer 1984;53:2316-23.
4. Campbell A, Converse PE, Rodgers WL. The quality of American life. New York: Sage Foundation; 1976.
5. Patrick DL, Erickson P. Health status and health policy: allocating resources to health care. New York: Oxford Univ Pr; 1993.
6. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.
7. Brook RH, Ware JE Jr, Davies-Avery A, et al. Overview of adult health status measures fielded in Rand's health insurance study. Med Care 1981;17(7 Suppl):iii-x, 1-131.
8. Stewart AL, Ware JE Jr, editors. Measuring functioning and well-being: the Medical Outcomes Study approach. Raleigh-Durham (NC): Duke Univ Pr; 1992.
9. Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997;35:522-37.
10. Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute; 1993.
11. Levine S, Croog SH. What constitutes quality of life? A conceptualization of the dimensions of life quality in healthy populations and patients with cardiovascular disease. In: Wenger NK, Mattson ME, Furberg CF, Elinson J, editors. Assessment of quality of life in clinical trials of cardiovascular therapies. New York: Le Jacq Publishing; 1984. p 46-66.
12. Fukuhara S, Ware J Jr, Kosinski M, Wada S, Gandek B. Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. J Clin Epidemiol 1998;51:1045-54.
13. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;27:271-92.
14. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;No 140.
15. Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671-84.
16. Ramsay JO. TestGraf—a program for the graphical analysis of multiple choice test questionnaire data. Montreal (QC): McGill Univ; 1995.
17. Muthen BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 1984;29:177-85.
18. Muraki E. A generalized partial credit model. In: van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. Berlin: Springer; 1997. p 153-64.
19. Muraki E, Bock RD. Parscale: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago: Scientific Software; 1996.
20. Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38(9 Suppl):II73-82.
21. Ware JE Jr, Kosinski M. SF-36 Physical and Mental Health summary scales: a manual for users of version I. 2nd ed. Lincoln (RI): Quality Metric Inc; 2001.
22. Fisher WP Jr, Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HIS Physical Functioning Scales. J Outcome Meas 1997;1:329-62.
23. Raczek A, Ware JE Jr, Bjorner JB, et al. Comparison of Rasch and summated rating scales constructed from SF-36 Physical Functioning items in seven countries: results from the IQOLA Project. J Clin Epidemiol 1998;51:1203-14.

24. Patrick DL, Chiang YP, editors. Health outcomes methodology: symposium proceedings. Med Care 2000;38(9 Suppl):II1-210.
25. NCQA. HEDIS 2002. Specifications for the Medicare Health Outcomes Survey. Washington (DC): National Committee for Quality Assurance; 2002.
26. Gandek B, Ware JE Jr, editors. Translating functional health and well-being: International Quality of Life Assessment (IQOLA) project studies of the SF-36 Health Survey. J Clin Epidemiol 1998;51:891-1214.
27. Ware JE Jr, Kosinski M, Dewey JE. How to score version 2 of the SF-36 Health Survey (standard & acute forms). Lincoln (RI): QualityMetric Inc; 2000.
28. Bjorner JB, Kosinski M. Cross-calibration of widely-used headache impact scales using item response theory. Qual Life Res. In press.
29. Ware JE Jr. Scales for measuring general health perceptions. Health Serv Res 1976;11:396-415.
30. Hunt S, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med [A] 1981;15(3 Pt 1):221-9.
31. Jacobson GP, Ranadan NM, Norris L, Newman CW. Headache disability inventory (HDI): short term test-retest reliability and spouse perceptions. Headache 1995;35:534-9.
32. Bjorner J, Kosinski M, Ware JE Jr. Calibration of a comprehensive headache impact item pool: the Headache Impact Test. Qual Life Res. In press.
33. Wainer H. Computerized adaptive testing: a primer. 2nd ed. Mahwah (NJ): Lawrence Erlbaum Associates; 2000.
34. Sands WA, Waters BK, McBride JR. Computerized adaptive testing: from inquiry to operation. Washington (DC): American Psychological Association; 1997.
35. Ware JE Jr, Kosinski M, Bjorner JB, et al. Application of computerized adaptive testing to the assessment of headache impact. Qual Life Res. In press.
36. Bayliss M, Dewey J, Cady R, et al. Using the Internet to administer a computerized adaptive measure of headache impact. Qual Life Res. In press.
37. Haight JC. Technology transfer. In: Gallin JI, editor. Principles and practice of clinical research. San Diego: Academic Pr; 2002. p 329-60.
38. Gerber LH. Measures of function and health-related quality of life. In: Gallin JI, editor. Principles and practice of clinical research. San Diego: Academic Pr; 2002. p 267-74.
39. Klein HA. The science of measurement. A historical survey. New York: Dover; 1974.
40. Bjorner J, Ware JE Jr. Using modern psychometric methods to measure health outcomes. Med Outcomes Trust Monitor 1998; 3(2):11-6.
41. Katz SW, Ford AB, Moskowitz RW, et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914-9.

Supplier

a. QualityMetric Inc, 640 George Washington Hwy, Lincoln, RI 02865.