METHODS

Dynamic Health Assessments:
The Search for More Practical and More Precise Outcomes Measures

John E Ware, Jr., Jakob Bjorner and Mark Kosinski
Quality Metric Inc., Lincoln, RI, and Health Assessment Lab, The Health Institute, New England Medical Center, Boston, MA

Background

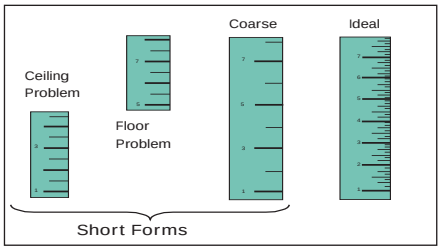
The widespread adoption of short forms underscores the importance of practical considerations in determining whether patient-based questionnaires are used to measure health outcomes. For example, the SF-12 - a subset of 12 questions from the SF-36 Health Survey - is widely used because it fits on a single questionnaire page that most adults can fill out within two minutes. An even shorter form, the SF-8, is currently being tested. Unfortunately, however, the very features underlying the popularity of such short forms, render them less precise[1]. The loss of precision with most short forms is greatest when a score is estimated for an individual patient. A short form can distinguish only the very largest changes in a patient's health status from fluctuations due to measurement error. Accordingly, the use of short forms in clinical practice has been criticized[2]. Although these critics have failed to note that scores based on a short form are likely to be precise enough for some patients at some scale levels, they make an important point. On average, short-form questionnaires (e.g., Dartmouth COOP Charts, Duke Health Profile, Functional Status Questionnaire, Nottingham Health Profile[3], and the SF-36[4]) represent a compromise in precision and other desirable attributes in favor of practicality. Today's short forms offer a good compromise for purposes of monitoring the health of large samples from general and specific populations, but not for clinical practice.

The problem with today's printed short-form questionnaires is that they rely on a fixed set of questions that can't possibly be the best set for all respondents and purposes. The advantage of such standardization is that results can be compared. The disadvantage is that, regardless of their answers, all patients are asked the same questions and at least some are likely to appear redundant, illogical or unnecessary.

There are many tradeoffs involved in constructing a short-form questionnaire. What concepts should be left out? How many questions should be included to measure the highest and lowest levels of health? How much precision is necessary at each level? The tradeoffs involved in covering a wide range of levels and maintaining adequate precision at each level are illustrated in Figure 1, which characterizes measures as rulers.

The marks on the rulers are defined by questionnaire items. As illustrated by the first "ruler," most widely-used short forms include questions that define only the lower levels where

Figure 1.: Short-Forms Versus Ideal Health Status Measures



the sickest patients score. Accordingly, they yield a concentration of scores at the higher levels, particularly in general populations (ceiling problem). Other short forms focus on higher levels and have a concentration of scores at the floor, particularly among those who are most ill (floor problem). A third short-form strategy ("Coarse" in Figure 1) is to spread questions over a wider range, resulting in larger gaps and less precision at any one level. The "ideal" measure in Figure 1 (a very long form) has enough questions to cover the full range with a high degree of precision at all observed levels.

The only way to achieve a high degree of precision with a short form would be to focus all of the questions on a particular level of health. If that were the respondent's level, there would be no compromise. However, each person would require their own short form. Is it possible to match questions to the respondent's level of health? This strategy has been used to achieve short and precise educational and psychological tests for decades. They are called "computerized adaptive tests"[5]. The result is a simple form of artificial intelligence that selects questions tailored to the test-taker, shortens or lengthens the test to achieve the desired precision, scores everyone on a standard metric so that results can be compared, and displays results instantly. Examples of such tests are national licensing exams for nurses and pharmacists. Many paper-pencil aptitude tests and admissions tests for graduate students will be replaced by computerized adaptive tests in the U.S. in 1999. However, these tests require computers and "modern" psychometric methods that have only rarely been applied to health questionnaires.

Preliminary Studies of Health Questionnaires

To determine whether computerized testing methods are applicable to health assessments, the Health Assessment Lab at New England Medical Center's Health Institute began applying modern, as opposed to "classical," psychometric methods to widely-used health questionnaires more than five years ago. The Lab's initial studies focused on whether the assumptions underlying these methods can be satisfied for measures of health status. (For more information about modern psychometric methods, see the box on page 13)

Initial results for English-language physical functioning measures were promising (e.g., Haley, McHorney and Ware, 1994)[6]. The work has been expanded to multiple languages[7], other item analysis methods[8], and to items "pooled" from different questionnaires[8]. At least preliminary calibrations for items from nearly two dozen widely used measures have been estimated. Thus, although there are many differences between "tests" and health status measures, modern psychometric methods are likely to be very useful in measuring health. It should be noted that many others have independently reached the same conclusion (e.g., Fisher, Eubanks and Marier, 1997)[9]. Further, there are good software programs available for use in such analyses. The features of some of our favorites are discussed elsewhere[10] and on the Internet (www.qmetric.com).

Unfortunately, however, the software programs most useful in constructing and calibrating items measuring health, don't have provisions for administering and scoring them dynamically. Public domain and commercially-available software programs that can administer and score questions in a dynamic fashion don't appear to have the right features for assessing health. For example, unlike tests, health measures don't have "right" and "wrong" answers. Further, health scores don't need to be "corrected" for guessing. After several years of evaluating the strengths and weaknesses of available computerized testing software, one of us (JW) founded QualityMetric, Inc. in Lincoln, RI, to develop new software with the specific features necessary for the next generation of health assessments, which will be administered dynamically. QualityMetric is working with other technology companies to take full advantage of the latest advances in communications technology. They will use the new software and item pools to administer dynamic health assessments using computers, interactive television, telephones and other devices.

The Logic of Dynamic Health Assessments

Regardless of which technology is used to administer a dynamic health assessment, the logic is the same. As illustrated in Figure 3.

(continued on p 12)

MAPI RESEARCH INSTITUTE

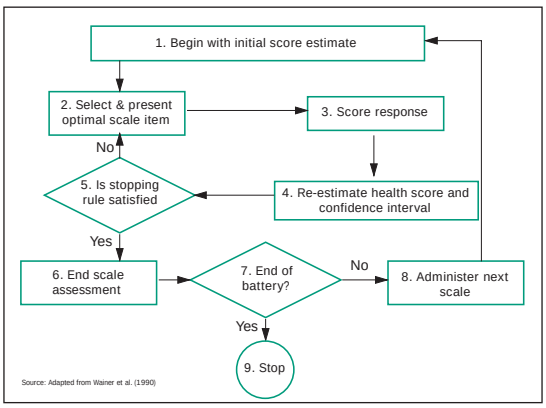
METHODS

Dynamic Health Assessments

(continued from p 11)

The process begins at Step 1 with an initial estimation of the respondent's score. (One of the mental health demo versions begins with the population average). That estimate is used to select the most informative item, which is administered at Step 2. The answer is used at Step 3 to re-estimate the score. At Step 4, a respondent-specific confidence interval (CI) is computed. At Step 5, the computer determines whether the score has been estimated within a preset standard of precision based on the CI. This unique feature makes it possible to match the level of score precision to the specific purpose of measurement for each patient, as illustrated in the first evaluation of the current software, discussed below. If the estimate is not precise enough the cycle is repeated. Once the precision standard is met, the computer either begins assessing the next concept or ends the battery, as shown in Figure 3.

Figure 3.: Logic of Dynamic Health Assessment



Results of the First Test: Dynamic Mental Health Assessments

The first evaluations of the new software were performed to test the accuracy of dynamic scores and whether dynamic assessments reduced respondent burden. For these purposes, we compared scores based on dynamic assessments with scores based on all of the information available from 31 of the items from the Mental Health Inventory (MHI), which were in the first item pool. Data were analyzed for 2,753 patients who participated in the Medical Outcomes Study (MOS). The MHI and the MOS sample are described in detail elsewhere[11].

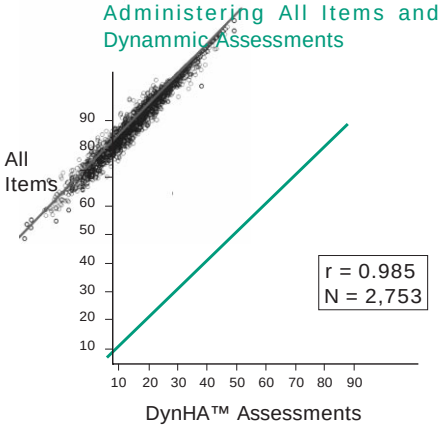
To simulate the most difficult application, a clinical setting involving the interpretation of scores for individual patients, we required a high level of precision. The 95% confidence interval around each patient's score was set at +/- 5.4 points or less for the lowest scoring patients (the bottom third of the score distribution). These patients scored near or below an established cut-point used in screening patients for psychiatric disorders[12]. A high level of precision was set for these patients so that the decisions as to whether to manage them differently and whether each patient's health was improved could be made reliably. For patients above that cut-point, the precision standard was relaxed to +/- 7.9 or less, which is the 95% confidence used to define changes in mental health in the MOS[13]. For patients at or above the 90th percentile, the precision standard was relaxed further because differences in their scores were assumed to lack clinical relevance. For each patient at all levels, the dynamic software was programmed to rule out a positive screen with 98% accuracy before relaxing any precision standard.

How accurate were scores based on these dynamic assessments of mental health? A scatterplot showing the degree of correspondence between MHI-31 scores and scores from dynamic assessments is presented in Figure 4 for 2,753 chronically ill patients who participated in the MOS.

The product-moment correlation between scores was very high ($r = 0.985$, $p < 0.0001$) and the means for the two estimates were nearly identical (48.78 and 48.85 for MHI-31 and dynamic scores, respectively). Other criteria were also evaluated. For example, scores based on dynamic assessments using only three questions for each patient correlated 0.932 with scores based on all 31 MHI questions. Thus, scores from dynamic mental health assessments were virtually interchangeable with MHI-31 scores throughout the scale range in this initial test.

The practical implications of dynamic assessments are clearly apparent from analyses of administrative data logs. Respondent burden was dramatically reduced. For example, for the lowest-scoring patients (bottom third of mental health scores) for whom the highest standard of precision was set for dynamic assessments, 92% of patients required only four or five questionnaire items to satisfy the precision standard. With the next round of additions of items and calibrations to the item pool, the efficiency of the system is expected to increase. In the meantime, our goal is to increase understanding of how dynamic assessments work. Those interested are encouraged to try the current mental health demo on the Internet (www.qmetric.com) and the program in the Understanding Health Outcomes educational series which focuses on the methods and applications of dynamic health assessments (www.healthstatprod.com). More detailed results from these initial tests will be the subject of a forthcoming report.

Figure 4.: Correspondence Between Mental Health Scores Computed by Administering All Items and Dynammic Assessments



The First Application: Dynamic Assessments of Headache Sufferers

The first application of the new dynamic health assessment software will measure the impact of headache on the lives of headache sufferers. Because an estimated 26 million Americans suffer from migraine headaches and many of them are undiagnosed, a priority is being placed on monitoring how this disease affects individual patients' in terms of their work productivity, social function, and family relationships. This project, which is sponsored by Glaxo Wellcome, will help individual patients determine their overall health status. In addition to generic measures, headache-specific questions will focus on the severity of a patient's headache and disability.

Even with proper diagnosis, many people with migraine find it a challenge to communicate with their physicians the impact of these headaches on their lives. It is hoped that headache sufferers will be able to easily complete a dynamic assessment either by phone or Internet. After completing the assessment, patients and their physicians will receive a report that can be used to develop an effective treatment plan and then monitor how much the patient is helped. This system is unique in that it will be the first to measure each patient's disability on a standardized metric, with a high degree of precision, on the basis of a very brief survey. The first version is expected to be available to headache sufferers worldwide during the second quarter of 1999. Dynamic health assessments will be administered at the point of care or in a patient's home. Multiple options for assessments will be offered, including a simple web browser and telephone, to ensure widespread access during the initial evaluations of this breakthrough technology. ●

(continued on p 13)

METHODS

Dynamic Health Assessments

(continued from p 12)

Modern Psychometric Methods

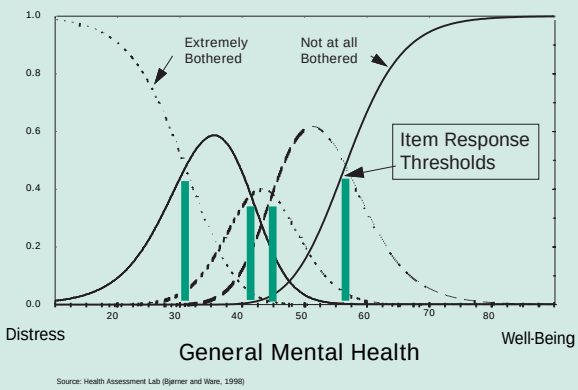
The psychometric methods that make it possible to calibrate questionnaire items on a standard metric ("ruler") also yield the algorithms necessary to run the "engine" that powers dynamic assessments. These statistical models tell us how likely a person at each level of health is to choose each response to each survey question. Figure 2 illustrates how these probabilities differ for responses to a widely used question about emotional distress.

The horizontal axis in this figure is a bipolar general mental health concept ranging from emotional distress to well being. A comprehensive pool of items from widely used questionnaires has defined the concept. We have calibrated this axis so that mental health has a mean score of 50 and a standard deviation of 10 in the general US population. The five curves in the figure show the probability of selecting each response choice at each level of mental health. For example, for those who are most distressed the probability of choosing "extremely bothered" approaches 100% whereas those at the highest levels of well being are most likely to choose "Not at all bothered." The curves illustrating these probabilities are referred to as trace lines or item characteristic curves. The estimation methods assume that these item characteristics hold true regardless of the health status of the population.

The four bold vertical lines in Figure 2 mark the scores on the "ruler" at which the probability curves for adjacent response categories intersect for this question. For example, at a mental health score of about 30, the probability of choosing the first two response categories is equal. These four values are important because they define the item difficulties or thresholds associated with this question. They are the "marks" on the ruler that makes it possible to measure mental health. As a person's mental health increases beyond one of these thresholds, he or she is more likely to choose the response category above the threshold rather than the category below the threshold. As explained in greater detail elsewhere[10] and on the Internet (www.qmetric.com), we reverse this logic to estimate the probability of each mental health score from a particular pattern of item responses. The resulting likelihood function makes it possible to estimate each person's score along with a person-specific confidence interval. In principle, we can get an unbiased estimate of mental health, i.e., an estimate without systematic error, from any subset of items that fits the model. The number of items administered can be increased to achieve the desired level of precision. The likelihood function can also be used for purposes of monitoring the quality of data for each respondent.

Most statistical models for estimating such item parameters can be traced to one of two measurement traditions. The first originates from the work of George Rasch[14]. The partial credit model we used in the studies summarized here belongs to the Rasch family of models. We are also testing models based on a second tradition - Item Response Theory (IRT) - originating from the work of Thurstone, Lord, and Birnbaum (e.g., Lord and Novick, 1968[15]; Lord, 1980[16]; Van der Linden et al., 1997[17]; Wainer and Mislevy, 1990[18]). These models place greater emphasis on fitting the data at hand. For questionnaires that use a categorical rating scale like the one shown in Figure 2, the two approaches have an important difference. Whereas the Rasch family of models requires that items have equal discrimination (the slope of the trace lines are equal across items), IRT models include a parameter for item discrimination that allows some items to have a steeper slope than others do. Both approaches to modeling assume unidimensionality, i.e., that the items included on a particular scale measure only one concept.

Figure 2.: Item Response Model: Emotional Distress Question



Please, address all communications to: John E. Ware Jr., QualityMetric Inc., 640 George Washington Hwy, Lincoln, RI 02865. Phone: +1 401-334-8800, x242. Fax: +1 401-334-8801. E-Mail: jware@qmetric.com.

1. Ware JE, Kosinski M, and Keller S. A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, 1996;34(3):220-233.
2. McHorney CA and Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research*, 1995;4:293-307.
3. 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;35:522-537.
4. Ware JE, Jr. The SF-36 and SF-12 Health Surveys: How to Use Them. Study Guide. Woodbridge, NJ: *HealthStat Productions*, Inc, 1997.
5. Wainer H, Dorans NJ, Flaugher R, et al. Computerized Adaptive Testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.
6. Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10): II Comparison of relative precision using likert and rasch scoring methods. Boston, MA: The Health Institute, 1994.
7. Raczek A, Haley SM, Aaronson NK, Apolone G, Bech P, et al. Tests of Scaling Assumptions and Improved Scoring Algorithms for the SF-36 Physical Functioning Scale in Seven Countries: Results from the International Quality of Life Assessment Project. J Clin Epidemiol 1998 (in print).
8. Bjorner JB, Essink-Bot ML, Kosinski M, Ware JE. Different questionnaires on common health "rulers". Quality of Life Research, 1998;7:572.
9. Fisher WP, Jr., Eubanks RL, Marier RL. Equating the MOS SF36 and the LSU HSI Physical Functioning Scales. J Outcome Measur 1997;1:329-362.
10. Bjorner JB, Ware JE. Using modern psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor* 1998;3(2):11-16.
11. Stewart AL, Ware JE. Measuring Functioning and Well-Being. The Medical Outcomes Study Approach. Durham, No Duke University Press, 1992.
12. Ware JE, Kosinski M, Keller SK. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute, 1994.
13. Ware JE, Bayliss MS, Roger WH, et al. Differences in four-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service: results from the medical outcomes study. *Journal of American Medical Association* 1996;276(13):103947.
14. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980.
15. Lord FM, Novick MR. Statistical Theories of Mental Test Scores Applications of Item Response Theory to Practical Testing Problems. Reading, MA: Addison-Wesley, 1968.
16. Lord FM. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum Associates, 1980.
17. Van der Linden WJ, Hambleton RK. Handbook of modern Item Response Theory. Berlin: Springer, 1997.
18. Wainer H, Mislevy RJ. Item Response Theory, Item Calibration, and Proficiency Estimation. In: Wainer H, Dorans NJ, Flaugher R, et al, Computerized Adaptive Testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.