MEDICAL CARE
Volume 38, Number 9, Supplement II, pp II-102–II-124


# Health-Related Quality-of-Life Assessments in Diverse Population Groups in the United States

ANITA L. STEWART, PHD,* AND ANNA NÁPOLES-SPRINGER, PHD†

**BACKGROUND. Effectiveness research needs to represent the increasing diversity of the United States. Health-related quality-of-life (HRQOL) measures are often included as secondary treatment outcomes. Because most HRQOL measures were developed in nonminority, well-educated samples, we must determine whether such measures are conceptually and psychometrically equivalent in diverse subgroups. Without equivalence, overall findings and observed group differences may contain measurement bias.**

**OBJECTIVES. The objectives of this work were to discuss the nature of diversity, importance of ensuring the adequacy of HRQOL measures in diverse groups, methods for assessing comparability of HRQOL measures across groups, and methodological and analytical challenges.**

**RESULTS. Integration of qualitative and quantitative methods is needed to achieve measurement adequacy in diverse groups. Little research explores conceptual equivalence across US subgroups; of the few studies of psychometric comparability, findings are inconsistent. Evidence is needed regarding whether current measures are comparable or need modifications to meet universality assumptions, and we need to determine the best methods for evaluating this. We recommend coordinated efforts to develop guidelines for assessing measurement adequacy across diverse subgroups, allocate resources for measurement studies in diverse populations, improve reporting of and access to measurement results by subgroups, and develop strategies for optimizing the universality of HRQOL measures and resolving inadequacies.**

**CONCLUSIONS. We advocate culturally sensitive research that involves cultural subgroups throughout the research process. Because examining the cultural equivalence of HRQOL measures within the United States is somewhat new, we have a unique opportunity to shape the direction of this work through development and dissemination of appropriate methods.**

**Key words: Race; ethnicity; socioeconomic status; age; gender; culture; measurement; psychometrics; questionnaires; health surveys. (Med Care 2000;38[suppl II]:II-102–II-124)**

The US population is becoming increasingly diverse in ethnic composition and age. Demographic projections indicate that between 1996 and 2050, the proportion of Latinos will grow from 10.5% to 24.5% of the total population, blacks from 12.7% to 15.4%, Asians and Pacific Islanders from 3.7% to 8.7%, and those ≥65 years of age from 13% to 20%.[1] Including American Indians

*From the University of California San Francisco, Institute for Health and Aging; the Center for Aging in Diverse Communities; and Medical Effectiveness Research Center for Diverse Populations, San Francisco, California.

†From the University of California San Francisco, Division of General Internal Medicine; the Center for Aging in Diverse Communities; and Medical Effectiveness Research Center for Diverse Populations, Department of Medicine, San Francisco, California.

Supported by the Resource Center on Minority Aging Research program sponsored by the National Institute on Aging (P30 AG15272), the National Institute of Nursing Research, and the Office of Research on Minority Health.

Address correspondence to: Anita L. Stewart, PhD, Professor in Residence, University of California San Francisco, Institute for Health and Aging, Box 0646, San Francisco CA 94143-0646. E-Mail: anitast@itsa.ucsf.edu

and Alaska natives, nonwhites will make up about half of the population in 2050. In 1997, 13% of the population lived below the poverty level,[2] and the gap between the poor and the wealthy continues to widen.[3–5] Despite these important population changes, ethnic minorities, women, those with lower socioeconomic status (SES), and older persons have been underrepresented in epidemiological and clinical research.[6–10] In response, the National Institutes of Health (NIH) now mandate inclusion of minorities and women in research, and public policy now focuses on aging and minority health. As investigators attempt to include underrepresented groups in treatment effectiveness research, a new issue has emerged. Measures of health-related quality of life (HRQOL), often included as secondary outcomes, may not be relevant, appropriate, reliable, and valid in these groups, because most were developed and tested primarily in nonminority, well-educated samples. It is questionable whether measures developed in one cultural group (mainstream) can be used to assess phenomena in another without understanding the implications of doing so.[11–14]

Although it is unrealistic and conceptually inappropriate to expect any HRQOL measure to be free of cultural influence,[15] the use of group- or culture-specific measures is impractical in large effectiveness studies. With sufficient effort, however, we can minimize bias, maintain sensitivity to diversity, and produce comparable measures.[15,16]

In this article, we address 4 questions related to the use of measures of HRQOL to assess treatment effectiveness in diverse samples: (1) How does one define diverse population groups? (2) Why is it important to ensure conceptual and psychometric adequacy in diverse subpopulations in the United States? (3) How does one assess the applicability and comparability of HRQOL measures across diverse subpopulations? (4) What are the methodological and analytical challenges in assessing and interpreting such measurement constructs in the context of evaluating treatment effectiveness?

## Defining Diverse Population Groups

Investigating group differences in perceptions of health and HRQOL requires considering the economic, social, and cultural contexts that shape those perceptions.[17] Values, traditions, and beliefs within communities interact with environmental conditions and availability of opportunities to influence the health and HRQOL of individuals. Culture prescribes views on what constitutes health and HRQOL and influences health at both the group and individual level.[18] At the individual level, HRQOL refers to "the physical, psychological, and social domains of health, seen as distinct areas that are influenced by a person's experiences, beliefs, expectations, and perceptions."[19]

Historically in the United States, cultural groups have been defined mainly in terms of race and ethnicity. Given the increasing diversity, the definition of a cultural or diverse group for the purposes of effectiveness research may need to be broadened to include the extent to which individual belief systems are shared by members of any ethnic, religious, or social group.

The most relevant diverse populations in studies of HRQOL outcomes of medical effectiveness are groups at risk of poor outcomes. These groups, sometimes referred to as vulnerable populations, include women and children, the elderly, ethnic people of color, persons with lower SES, immigrants, gay men, lesbians, and the homeless.[20] As defined by the Agency for Healthcare Research and Quality (AHRQ), vulnerable populations are groups of people "made vulnerable by their financial circumstances or place of residence; health, age, or functional or developmental status; or ability to communicate effectively . . . [and] personal characteristics, such as ethnicity and sex."[21] Other population groups falling within the scope of this definition that have been largely overlooked in HRQOL research are the cognitively impaired, the physically and mentally disabled,[22] those with limited English language skills, those with low literacy,[23] and rural populations.[24]

The problem with subgroup analyses is heterogeneity within classifications. Labels such as "African American," "gay," or "low SES" represent crude categorizations. Presenting conceptual and psychometric differences (or similarities) in measures by group can mask the complex and multidimensional nature of socioeconomic and cultural factors, including discrimination.[13,25,26] However, such classifications can serve as a place to begin in studies of potential differences in HRQOL outcomes in vulnerable groups. Indeed, substantial variations in health, medical procedures, and treatments have been found through these standard racial and ethnic classifications.[27–37]

## Importance of Ensuring Conceptual and Psychometric Adequacy in Diverse Population Groups in the United States

### Increased Medical Effectiveness Research in Diverse Populations Requires Comparable Measures

There are a number of reasons why we might see an increase in medical effectiveness studies that include diverse groups. As a result of underrepresenting many vulnerable groups in research, results of treatment effectiveness trials cannot be generalized to these groups. In addition, little is known about whether the effect of a given treatment on HRQOL (eg, in a randomized trial) varies across these groups.[38,39] Both of these gaps are likely to be addressed in new studies.

Another reason for increased research with diverse groups is the accumulating evidence that treatments and treatment recommendations vary by ethnic group,[27–37] SES,[40–42] gender,[43–49] age,[50] and gender and race.[51,52] Studies may thus begin to evaluate whether these observed treatment disparities lead to similar disparities in HRQOL outcomes.

Studies such as these that examine group differences in HRQOL in diverse subpopulations require evidence of the adequacy of conceptual and measurement properties across groups. The reason is that nearly all self-reported and directly assessed measures of HRQOL have been developed and tested on primarily nonminority, English-speaking populations. Although some data on the adequacy of HRQOL measures for use in nonwhite and older populations are emerging,[53–58] few large-scale studies of the adequacy of these measures have been conducted in groups differing in educational level, ethnicity, language, and level of acculturation, in part because available data sets rarely have sufficient numbers to perform adequate subgroup measurement analyses. The risks of interpreting results of medical effectiveness research without assurances of equivalence of concepts and measurement properties are great in light of potential applications of such findings. Such applications include managed care reimbursement policies, policy and funding decisions, development of practice guidelines, and establishing priorities for medical effectiveness research.

### Observed HRQOL Differences: True Differences or Bias?

Evidence suggests that there are ethnic differences in levels of HRQOL: blacks, Mexican-Americans, and whites differ in reporting levels of illnesses and disability,[59] symptoms,[60,61] disease labeling,[62] and self-rated health.[63,64] For some health measures, ethnic differences remain after adjustment for social class.[25]

These observed disparities in mean HRQOL scores could be true group differences or may instead reflect cultural bias in the instruments. Even if translation is not required, when using an instrument in a nonmainstream population, it is important to consider that "the particular world view, the relevance of any particular construct, and the way in which the construct is defined may vary by gender, ethnicity, age, occupation, education, or other factors."[65]

Bias can be introduced into measures through culturally mediated differences in perceptions of the meaning of items and health constructs.[14] Bias can also occur as a result of cultural or group differences in the cognitive processes involved in forming responses, ie, processes underlying the recognition, labeling, and reporting of psychological and physical states.[60,66] For example, observed lower levels of self-rated health in Latinos compared to other ethnic groups[62,64] may be a function of a response pattern specific to Latinos that emphasizes comparing one's health to that of others rather than of real ethnic differences in health.[62] Similarly, self-rated health differences between Latinos and non-Latino whites were not explained by more objective indicators of health status (eg, illness reports, prescription medication, hospitalizations),[64] suggesting that measurement bias could be a problem. Establishing that HRQOL concepts and measures are cross-culturally equivalent is a prerequisite for investigations of cultural or group differences in HRQOL.[67]

The key distinction is between universal (etic) concepts of health and concepts that are group- or culture-specific (emic).[12,16,62,68–70] To the extent that a health concept captures both elements, one can refer to "derived etics," where an etic concept can be defined and measured in ways that address some of the group-specific issues and concerns.[12,69] Because the goal of large-scale studies is to have HRQOL measures that can be applied in many diverse groups, we should aim for measures that are derived etics. These are measures of

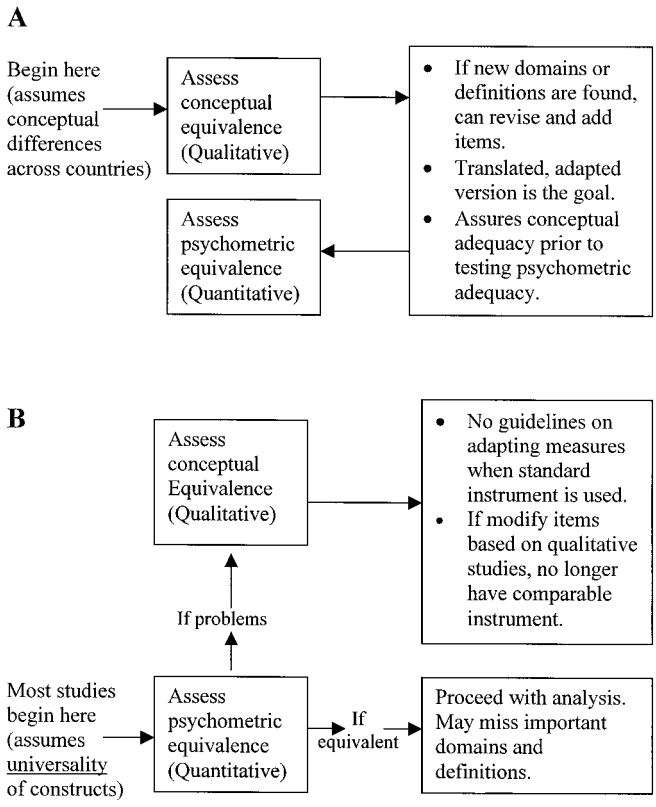


FIG. 1. A, Typical international approach; B, US subgroup approach when no translation is done.

HRQOL concepts that can be empirically demonstrated to be relevant across specified groups, are culturally appropriate for all groups, reflect the concerns of all groups, and minimize the potential for bias. We are not aware of any HRQOL measure in English that reflects the standard of a derived etic.

### Assessing the Applicability and Comparability of HRQOL Measures Across Diverse Population Groups

Most of the methodology for assessing the comparability of measures and instruments across cultural groups comes from international and US studies that translate HRQOL measures into ≥1 language. These studies usually assume that concepts and measures are not universal because of large cultural differences by language and nation. Their aim is to develop a modified instrument that is as comparable as possible to the original; adaptations are built into the process.[71] From preliminary qualitative work, items can be added or modified to reflect conceptual differences, substitute more appropriate idioms and colloquialisms, and change examples to accommodate the new cultural situation and population (the Figure, part A).

Our goal is to understand how well instruments developed in mainstream groups work in diverse US population segments. In the United States, it is common to select a standard profile instrument or set of measures and apply these to a study population, taking advantage of the tremendous resources and expertise invested in HRQOL measures to date. However, this approach assumes conceptual equivalence (universality) and ignores the perspectives of diverse subgroups, what Guyatt[72] refers to as the cultural hegemony of the US middle class and Rogler[14] refers to as middle-class ethnocentrism (the Figure, part B). The problem with this approach is that if measures have similar psychometric properties across groups, investigators may proceed with group comparisons

TABLE 1. Cross-Cultural Methods of Adaptation of Existing Instruments: An Integration and Modification of Traditional Approaches

| Cross-Cultural Adaptation Terminology | Methods |
|---|---|
| Conceptual equivalence of constructs and items[67,68,74,76,77,81] | Methods[14–16,67,74,76,77,88,90] |
| • Construct exists and is relevant and acceptable in all cultures; instrument measures same construct in each culture.<br>• Contains all relevant constructs for all groups (none missing).<br>• Value or emphasis placed on different domains is equivalent.<br>• Items represent well the definition of the construct. | • Review literature, especially ethnographic and anthropological, in target cultures for ways in which constructs are operationalized.<br>• Conduct interviews and focus groups of persons from target group to learn how they think about and define construct.<br>• Consult broad range of experts from target group to rate items and constructs in terms of relative importance, equivalence, relevance, appropriateness, and acceptability, and identify missing items. |
| Semantic equivalence[16,67,68,74,76,77,81] | Methods[67,74,76–78,91,92] |
| • Items mean same thing to people from different groups and in target and original language.<br>• Same expression exists in the target culture.<br>• Situations or examples given fit target culture and equivalent expressions found for idioms and colloquialisms.<br>• Level of language used is appropriate to target population.<br>• Technical features of language are equivalent, ie, complexity, syntax, grammar, and level of abstraction. | • Use structured qualitative methods with target population to identify meaning they ascribe to constructs (focus groups, expert panels).<br>• Resolve discrepancies using cognitive testing with probes to determine what subjects think items mean.<br>• Apply semantic differential techniques across groups to define semantic space in which word is located.<br>• Translation methods include forward and backward translation. Translate into the idiom of the culture. Have multidisciplinary, bilingual, bicultural laypersons rate equivalence of the original, translated, and backtranslated versions for meaning of each item. |
| Operational equivalence[67,71,76,77,79,81] | Methods[39,67,71,77,78,81] |
| • Ensure that standardized methods of survey administration are appropriate for target culture, ie, mode of administration, questionnaire format, reading level, instructions, item format, and respondent burden.<br>• Identification of respondent groups is standardized in each culture. | • Pretest and debrief; include probes about difficulty and appropriateness of survey.<br>• Use cognitive testing methods (eg, in-depth interviewing or think-aloud interviews) to identify whether cognitive processes involved in interpreting and answering questions differ across groups.<br>• Compare effects of different methods of administration on scores.<br>• Assess cultural norms regarding ways to address people and ways of framing questions.<br>• Have expert panel consider whether data-gathering approach is consistent with culture to which it is being applied. |
| Psychometric or measurement equivalence[16,67,68,77,78,80,81,83] | Psychometric or measurement equivalence[94–98,155] |
| Comparable psychometric properties, including item equivalence. These include variability (floor and ceiling effects), missing data, internal consistency and test-retest reliability, factor structure (including factor loadings), and construct validity (including comparability of effect sizes and responsiveness). | Variability, internal-consistency reliability, hypothesized structure, item convergent and discriminant validity are determined in Multitrait Scaling Analysis, via MAP software. Structural equation modeling can ascertain factor structure and construct validity across groups. Construct validity is evaluated by testing comparability of hypothesized patterns of association with similar variables. Responsiveness has alternative approaches. |

(*Continued*)

TABLE 1. (Continued)

| Cross-Cultural Adaptation Terminology | Methods |
|---|---|
| Item equivalence (item bias)[67,71,74,78–82,84]<br>● Items are not differentially more difficult (eg, biased) in target culture than in original, or across groups.<br>● Item weights reflect comparative importance of items in all groups.<br>● Meaning of and distance between response categories is similar across cultures. | Item equivalence[71,74,78,79,81,83,85–87,89,93,99]<br>● Differential item functioning analysis using item response theory methods or log linear models.<br>● Reexamine item weights in target culture via ratings by experts or laypeople, or use mathematical approach.<br>● Scale items relative to the central tendencies of the culture studied. Rank items in both cultures using an external scale or referent to compare intervals between ranks. Compare ranking of measures by subgroups to determine comparability across cultures.<br>● Thurstone's method of equal-appearing intervals or Stevens' magnitude estimation method. |
| Criterion equivalence[67,76,77]<br>● Interpretation of scores is same across groups and when compared with norms for each group. When norms are available, pertains to ensuring equivalent norms across cultures.<br>● Translated version demonstrates same relations to a previously established independent criterion as that obtained during validation of the original version.<br>● For classification measures: classification criteria measure the same phenomenon in both cultures. | Methods[76]<br>● Need to establish norms in each culture and establish comparable cutoff scores for when the trait or disorder is said to exist. Cutoff score for second culture can be adjusted to achieve this.<br>● For classification measures, determine whether sensitivity and specificity are comparable in each group. |

even though the measures may lack conceptual equivalence. That is, if a measure is defined too narrowly for some subgroups but is psychometrically equivalent, the conceptual differences would be missed.[14] This approach thus carries the risk that measures developed in selected groups may be biased when applied in diverse groups.[73]

## Approaches for Ensuring Cross-Cultural Equivalence of Translated Surveys

For translated instruments, there are several sources of methods for assessing cross-cultural equivalence.[71,74,75] The process includes assessing numerous dimensions of equivalence, although the terminology and definitions vary substantially. Because of the inconsistency, we present in Table 1 a framework for assessing equivalence that integrates and modifies the published dimensions and approaches of numerous investigators.[16,67,68,71,74,76–84] Table 1 includes 6 dimensions of equivalence: conceptual, semantic, operational, psychometric, item, and criterion. We also note briefly the basic methods for addressing the steps suggested by these investigators and others.[14,15,39,85–99]

## Approaches for Assessing Comparability Across Diverse US Groups

When standard HRQOL measures are used in diverse groups, the task is to assess both the conceptual and psychometric equivalence of the measures across subgroups when the groups are thought to be sufficiently different from the population on which the measures were developed.[71] This task can build on the cross-cultural methodology used in international studies in which quantitative and qualitative approaches are considered complementary.[13,71,100] To the extent that one believes that the concepts are equivalent (universality assumption) but the measures may be problematic, one can begin with psychometric testing and proceed to qualitative studies only if problems are identified.[71] Conversely, and more widely accepted by researchers experienced in multiethnic studies, one can begin with qualitative studies of

the concepts and measures.[12,88,90] By doing qualitative studies first, serious omissions of concepts (and items) or differences in their meaning can be detected.[14,67,68]

## Qualitative Approaches to Testing Measures in Diverse US Subgroups

Qualitative studies to assess the conceptual equivalence of existing HRQOL measures could explore how individuals from diverse backgrounds describe the concept, whether any elements are missing from standard definitions, and reasons why items may have been problematic during psychometric testing. Furthermore, some in-depth qualitative approaches elucidate how people construct their answers, eg, the cognitive processes of reporting.[66,101]

Three commonly used qualitative methods in measurement studies are cognitive testing, focus groups, and expert panels. Cognitive testing uses theories and methods of cognitive psychology to understand processes used by respondents to understand and answer questions, and to design questions to increase comprehension.[101–103] Three in-depth methods for investigating these processes are focus groups, behavioral coding, probe techniques, and think-aloud interviews.[101,102,104] These techniques aim to identify questions that pose problems for either interviewers or respondents and determine the nature and source of the problem to find solutions. Think-aloud interviews, for example, require that respondents verbalize their thought processes as they answer items. They are particularly useful in identifying cognitive processes.

Focus groups are in-depth interviews of small, homogeneous groups. They provide researchers with access to the language and concepts used by participants to think and talk about particular topics.[11,105,106] Hearing participants use their own vocabulary, language, and communication patterns facilitates development and evaluation of optimal item wording for different groups.[107] The saturation level (point at which no new relevant data emerge) is typically reached after ~6 to 8 (or more) groups, depending on the homogeneity of the groups and the research goals.[105]

Consultation with "experts" is often recommended as a way to learn efficiently about a concept. Presumably, the experts (ie, on the cultural issues and concepts of the group being studied) would have a range of experience, and may include individuals who represent the target group.

## Psychometric Testing of Measures

To test existing measures in diverse subgroups requires applying traditional psychometric approaches, but by subgroup. These approaches include conducting tests: (1) within a particular subgroup to determine whether measures are adequate and/or comparable to published studies and (2) across multiple groups to compare measurement properties simultaneously. The methods include examination of content validity (a form of conceptual equivalence), missing data, variability, reliability, measurement or factor structure, differential item functioning, item weights, use of response scales, and construct validity (including response bias and responsiveness to change). Criterion validity is seldom examined in the HRQOL field because of the lack of true criteria.[68]

Table 2 presents definitions of these psychometric approaches and findings from HRQOL studies that reported results by population subgroups. We focused on studies of adults that reported findings for specific diverse populations and that had sample sizes of ≥50 persons in each group. Results are presented separately for English and translated instruments. When comparisons across languages were made, they appear in the "translations" column.

Only 1 study was found that addressed the content validity of HRQOL measures in a diverse sample,[15] indicating that this is an area for expanded research efforts. Missing data analyses in diverse groups have been conducted with 3 commonly used HRQOL measures: the Center for Epidemiological Studies Depression Scale (CES-D), various Medical Outcomes Study (MOS) measures, and the Rosow-Breslau physical functioning scale. Problems were noted in older age groups, and results by gender and ethnicity were mixed.

In general, internal consistency results were good for the English and Spanish versions of various MOS measures, the CES-D, Health Assessment Questionnaire, General Well-Being Index, and the Functional Assessment of Cancer Treatment (FACT) in several ethnic subgroups. Results of studies examining the factor structure of Spanish and English versions of the CES-D were mixed; some replicated the original 4-factor structure, and others found differences by age, gender, and ethnic subgroups. Multitrait scaling results of Chinese and Japanese versions of the SF-36, although limited to 1 study in each language, were generally good. Analyses of the

TABLE 2. Psychometric Approaches to Assessing the Adequacy of HRQOL Measures for Use With Diverse Populations, Including Examples Representing Each Approach

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| *Content validity–conceptual equivalence:* Judgmental evidence regarding the extent to which relevant constructs of HRQOL are represented in the instrument, as well as the extent to which the theoretical construct definition is fully represented by items in the measure.[108] Requires determination of whether specified constructs are relevant to subgroups and whether the instrument contains all relevant constructs for particular subgroups, ie, that no relevant constructs or portions of constructs are missing.[68,109] | None. | For Spanish FACT-G, of 28 items, only 1 had a low "relevance" rating by bilingual/bicultural advisory committee; spirituality domain identified by patients during pilot study as missing (being addressed in next study) (n = 92 cancer patients, 23–80 y).[15] |
| *Missing Data:* Some subgroups may have more difficulties with standard survey methods in general and particularly with some types of questions in HRQOL surveys. Because missing data tend to indicate problems with particular items, it tends to be nonrandom.[109] Thus, to the extent that missing data are more common in particular subgroups, scores that are imputed will be more biased in those groups.[109] | For CES-D, no differences in percent missing >4 of 20 items by gender. White men (n = 317) and women (n = 741) and black men (n = 597) more likely to have missed 1–4 items (39% to 43%) than black women (n = 1,392, 33%). Primary care patients (n = 3,047, 60–102 y).[110]<br>For 14 MOS long-form measures in baseline MOS longitudinal panel, missing data increased with age across 4 age groups (18–44, 45–64, 65–74, 75+ y) (n = 2,546, 18–97 y).[55]<br>For 4 MOS SF-20 measures in random half of MOS screening sample, comparable missing data across 5 ethnic groups: Asian, white, black, Latino, "other" (n = 10,293, ≥18 y).[54]<br>For MOS SF-36 in MOS baseline sample, data completeness was lower for older than younger, <9th grade education than ≥9th grade, black than white and other, and poverty than nonpoverty subgroups (n = 3,445, ≥18 y).[112]<br>For Rosow-Breslau physical functioning scale, mean percent of "don't know" (DK) responses increased by age group with significant differences between 65–74, 75–84, and >84 y (group n not provided); women (n = 1,942) had significantly more DK responses than men (n = 1,155). (Rural sample, n = 3,097, ≥65 y).[113] | For CES-D, no differences in percent of missing data in Anglos (n = 254), blacks (n = 270), Chicanos completing English version (n = 144), and Chicanos completing Spanish version (n = 37) of the CES-D (n = 705, 20–59 y).[111] |
| *Variability:* Instrument yields comparable distributions. Scale scores should represent approximately the full range (or comparable ranges) with comparable (and minimal) floor and ceiling effects to enable worsening or improvement to be detected within each group. Extensive floor and ceiling effects can attenuate reliability and in turn validity, by reducing variation in the measure.[83,109,114] | Over 50% at ceiling on FSQ BADL, modified Katz ADL, OARS-IADL, and Role-Physical and Role-Emotional scales of SF-36 (n = 83, 64–92 y).[117]<br>For 4 MOS SF-20 measures in a random half of MOS screening sample, floor effects on all measures were 0% to 12% in each group: Asians, whites, blacks, Latinos, and "other"; ceiling effects more common (3% to 36%) and generally similar across groups (n = 10,283, ≥18 y).[54] | In Chinese version of MOS SF-36, floor effects for the 8 scales occurred in 0% to 32% of the elderly Chinese sample (n = 210, 55–96 y) and 0% to 27% of the general Chinese sample. Ceiling effects were observed in 0% to 46% of the elderly sample and 0% to 52% of the general sample.[116]<br>For Japanese version of MOS SF-36, floor effects for 8 scales ranged from 0% to 49% and ceiling effects from 16% to 96% of elderly Japanese sample (n = 80, 64–100 y).[118] |

(*Continued*)

TABLE 2. (Continued)

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| | For MOS SF-36, SIP, and QWB in health plan enrollees (n = 200, ≥65 y), ceiling effects common for Role-Physical and Physical Functioning but not for General Health and SIP. The QWB and, to a lesser degree, the SF-36 General Health scale had wider distributions of scores than SIP. Not stratified by age subgroups.[115] | |
| *Reliability:*<br>Extent to which a measure is free of random error. Internal-consistency reliability pertains to the interitem consistency of multi-item scales and is the most commonly reported coefficient. Reproducibility is as important given that treatment effectiveness is evaluated over time; thus, test-retest reliability should be known as well. Poor reliability in a particular subgroup is indicative of measurement problems and will attenuate estimates of treatment effectiveness for that subgroup. | For CES-D, internal consistency >0.70 for white men (n = 317), white women (n = 741), black men (n = 597), and black women (n = 1,392) (60–102 y).[110]<br>For CES-D, internal consistency 0.91 for women (n = 588) and 0.86 for men (n = 412) (n = 1,000, 18–92 y).[121]<br>For CES-D, internal consistency 0.87 in US-born non-Latino whites (n = 1,063) and US-born Mexican Americans (n = 500) (n = 1,563, 20–74 y).[123]<br>For FSQ, modified Katz ADL, OARS-IADL, performance-based measures of physical function, and SF-36 internal consistency >0.70 except for Katz ADL and OARS-IADL (n = 83, 64–92 y).[117]<br>Internal consistency >0.70 for all scales in GWB, Rosenberg Self-Esteem, and Pearlin Mastery scale in white 9th grade students (n = 907–950) and for 4 of 5 scales in Laotian Hmong 9th grade students (n = 61–95 y).[122]<br>For 4 MOS SF-20 measures, internal consistency >0.70 for Asians, whites, blacks, Latinos, "others" in random half of MOS screening sample of patients (n = 10,283, ≥18 y).[54]<br>For 14 MOS long-form scales in baseline MOS longitudinal panel, internal consistency >0.70 for all age groups (18–44, 45–64, 65–74, 75+ y) (n = 2,546, 18–97 y).[55]<br>For MOS SF-36, internal consistency <0.70 for Social Functioning in blacks (n = 357) and for no scales in whites (n = 803) in sample of adults (n = 1,160, ≥30 y).[119]<br>For MOS SF-36 in baseline sample, internal consistency >0.70 for all subgroups including poverty status (n = 253); blacks (n = 481); <9th grade education (n = 209); and age ≥65 y (n = 987).[112]<br>For MOS SF-36, only Pain Scale had internal consistency <0.70 in Pima Indians (n = 54, aged 24 to 78 y).[124] | For CES-D, internal consistency >0.70 for Anglos (n = 254), blacks (n = 270), Chicanos completing English version (n = 144), and Chicanos completing Spanish version (n = 37) (n = 705, 20–59 y).[111]<br>For Spanish CES-D in HHANES, internal consistency = 0.86 among US-born Mexican Americans (n = 1,918) and 0.85 among Mexico-born Mexican Americans (n = 1,167) (n = 3,085, 20–74 y).[123]<br>For Spanish CES-D in probability sample of adults, internal consistency = 0.87 in US-born Mexican Americans (n = 500) and 0.78 among Mexico-born Mexican Americans (n = 637) (n = 1,137, 20–74 y).[123]<br>In Spanish FACT-G, internal consistency 0.66–0.83 for 5 subscales (4/5 >0.70) and 0.89 summary index (n = 92 cancer patients, 23–80 y).[15]<br>For Spanish and English GWB (60% English; results not stratified by language), internal consistency >0.70 in 3 of 4 scales and in total score in combined sample of Spanish- and English-speaking Mexican American women (combined because no mean differences by language) (n = 122, 18–65 y).[145]<br>For Spanish version of HAQ and MOS pain scales, internal consistency and test-retest reliability >0.70; same results by national origin (n = 272, ≥52 y).[120]<br>For the MOS SF-36, internal consistency of Chinese version from 0.38–0.90 (3 <0.70) and 0.54–0.92 (1 <0.70) (n = 219 Chinese, 55–96 y).[116]<br>For Japanese version of MOS SF-36, 6 of 8 scales had internal consistency >0.70 in older Japanese sample (n = 80, 64–100 y).[118] |
| *Measurement or factor structure:*<br>Instrument yields comparable confirmatory factor structure in defined subgroups.<br>Method: Multitrait scaling analysis[94] in which hypothesized scales are tested regarding extent to which items meet Likert scaling criteria of convergent and discriminant validity. Software is available.[95] | For CES-D, equivalent hypothesized 3-factor solution and equivalent factor loadings observed across 3 generations of Mexican Americans; measurement error variances differed across all 3 groups (n = 362, 371, and 372 in older, middle, and younger generations).[128] | For Affect Balance Scale, similar factor structure found across Laotian (n = 193), Cantonese (n = 756), and Vietnamese (n = 399) translations using SEM. Factor loadings comparable for combined Southeast Asian and English groups (n = 319). (n = 1,667, ≥18 y).[130] |

(*Continued*)

TABLE 2. (Continued)

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| Method: Multiple group confirmatory factor analysis using SEM methods.[125,126] Multigroup SEM enables test of significance of equivalence between 2 or more groups. Factor structure in multigroup SEM studies can be examined at several levels, each of which provides additional evidence of equivalence: (1) the same factors are observed using the same items (factorial or configural invariance); (2) the item loadings on the factors are equivalent (item loading or factorial loading invariance); (3) the mean scores (item intercepts) on the items are equivalent, showing evidence of strong factorial invariance; and (4) the residual item variances are equivalent (evidence of strict factorial invariance). | For CES-D, original 4-factor structure[152] confirmed with original items loading on each respective factor and equivalent factor loadings in 2 older (55–80 y, n = 230; n = 278) and 1 younger (20–54 y, n = 217) mainly white samples.[129]<br>For CES-D, 3-factor structure found in women (n = 588) and 4-factor structure in men (n = 412) and different factor loadings; in men, higher loadings on somatic and positive affect factors; in women, factor loadings were higher than for men on all 4 factors in sample of adults (n = 1,000, 18 to 92 y).[121]<br>For CES-D, different factor structures found for black and white men and for black and white women in older adults in primary care practice (n = 3,057, 60–102 y).[110]<br>For 4 MOS SF-20 measures in random half of MOS screening sample of patients, comparable item-scale correlations found for Asians, whites, blacks, Latinos, and "others" (n = 10,283, ≥18 y).[54]<br>For MOS SF-36, a 9-factor model fit data best in at-risk black and white men and women compared with 8 hypothesized dimensions (original SF-36 not developed with factor analysis). Same number of factors and same item clusters were found for all 4 race-by-gender subgroups; some differences in magnitude of the factor loadings (n = 1,051; 50–74 y with comorbidities or ≥75 y).[127]<br>For MOS SF-36 in baseline sample, item-scale correlations for General Health were <0.40 in all age groups (<65 y 0.39 to 0.73; 65–74 y 0.38 to 0.69; ≥75 y, 0.34 to 0.77); in women (0.38 to 0.74); in blacks (0.35 to 0.68), and in "others" (0.33 to 0.66). Scaling success rates <0.90 on 3 scales for ≥75 y (n = 287), on 3 scales for blacks (n = 481), on 5 scales for "others" (n = 221), on 5 scales for <9th grade education (n = 209), on 2 scales for 9–11 y of education (n = 313), and on 4 scales for poverty status (n = 253) (n = 3,455, 18 to 98 y).[112]<br>For MQOLS-CA, identified 2 factors (psychological well-being and physical competence) in women (n = 254) and 2 different factors (vitality and personal resources) in men (n = 222), with 4 unique items for women and 3 for men. No cancer-specific items loaded on any factors for either gender in sample of oncology outpatients (mean age, 58 y for women; 60 y for men).[132]<br>Using SEM, confirmed 3-factor structure of self-reported physical health (chronic illness, functional status, and self-rated health) originally hypothesized by Whitelaw and Liang[136] in a one-fourth random sample of the NHIS 1984 Supplement on Aging (n = 2,942, 60–99 y).[137] | For Spanish CES-D in HHANES, found similar 3-factor solution (rather than original 4 factors in English) in 3 ethnic subgroups: Cuban (n = 808), Puerto Rican (n = 1,266), Mexican American (n = 3,117) with combined affective and somatic factor; items loading on interpersonal factor differed between Cuban Americans and other groups. Factor structure differed by gender with Latino men resembling original 4-factor structure and 3-factor structure in women. For men, factor structure differed by origin and language, but not for women (≥20 y).[131]<br>4-Factor structure of CES-D[152] confirmed with original items loading on each respective factor in Anglos (n = 254), blacks (n = 270), Chicanos completing English version (n = 144), and Chicanos completing Spanish version (n = 37) (aged 20–59 y).[111]<br>For CES-D, using SEM, confirmed hypothesized 2-factor model (depression and well-being) in elderly Mexican Americans ≥65 y in 2 samples, the Hispanic EPESE (n = 2,536) and a previous study (n = 330), across gender and English vs Spanish language subgroups.[134]<br>For GWB, found 4-factor structure of psychological distress, well-being, general health, and vitality in combined sample of Spanish- and English-speaking Mexican American women (combined because found no mean differences by language on items) compared with originally hypothesized 6-factor structure[135] (n = 122, 18–65 y).[145]<br>For Chinese translation of MOS SF-36, item-scale correlations >0.40 for all 8 scales. Items discriminated between own hypothesized scale and other scales in all but Social Functioning (n = 52, aged 64–87 y).[133]<br>For Japanese version of MOS SF-36, 78% of item-scale correlations ≥0.40 (n = 80, 64–100 y).[118] |

(*Continued*)

TABLE 2. (Continued)

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| | Measurement model of 4 functioning concepts (eg, ADL, lower body functioning) fit data equally well in men (n = 2,447) and women (n = 4,023) and by ethnicity (Mexican American, n = 193; other Latino, n = 338; black, n = 864; and white, n = 5,268) (>70 y).[58] For measurement model of health that includes measures of lower body disability, upper body disability, basic ADLs, household ADLs, advanced ADLs, and perceived health, found differences in factor loadings between whites (n = 4,494) and blacks (n = 530) on 7 of 19 items and gender differences on 6 of 19 items. Latent variable of basic ADLs had the most differences in factor loadings (n = 5,024, ≥70 y).[57] | |
| *Differential item functioning (DIF):* Formerly known as item bias analysis, DIF refers to an item that has a different meaning in subgroups of the population,[138] ie, that the probability of a positive response to an item for any given value of a measured underlying attribute is different for different subgroups.[139] The absence of DIF is a necessary condition for establishing the lack of bias of a scale.[138] Methods for testing DIF across groups include analysis of variance and tests of the measurement model using SEM, as well as methods using item response theory, although the latter is preferred.[139] Item response theory, or latent trait analysis, evaluates DIF through examination of internal item characteristics; item characteristic curves provide the probability of certain answers at different levels of the underlying trait.[81,86,153,154] | For CES-D, "had crying spells" item appears to mean different things to men and women, given that women are more likely to cry. For a man to endorse this item appears to reflect a more severe state than for a woman to endorse it.[138] For CES-D, 2 items were gender biased, and 3 others had other psychometric problems. Sample included 708 cancer patients and 504 caregivers of chronically ill elderly.[165] Some of the somatic symptoms and enjoyment in the SHORT-CARE depression scale were found to be less severe indicators of depression for symptoms such as headache, crying, somatic symptoms for Latino compared with non-Latino white elderly subjects.[114,140] | |
| *Item weights:* In the special case of utility measures and/or preference-weighted scales, item weights should be comparable across subgroups.[78,93] | For QWB, Latinos (n = 52) rated items 0.05 points on a 0–1 scale below non-Latinos, controlling for demographic characteristics ($P < 0.01$) during development of item weights in 1974, although explained variance was small. No differences for blacks (n = 29).[141] | |
| *Use of response scale:* Two relevant issues: (1) extent to which different groups use the response scale quantifiers similarly, and (2) extent to which distances between response scale quantifiers are similar across groups. There is interindividual variation in the use of these vague quantifiers, and it is sometimes systematic.[138] Most items do not have equal intervals between response choices, but the point is to ensure that the intervals do not vary greatly across groups.[83] | | Mexican Americans (n = 7,462) and Puerto Ricans (n = 2,834) in HHANES less likely than physician to rate health as excellent or very good and more likely to rate it as good, suggesting that "good" response has different meaning for subgroups; discrepancy especially pronounced for less acculturated and Spanish version (20–74 y).[138,142] |

(*Continued*)

TABLE 2. (Continued)

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| | | In Latinos interviewed in Spanish, self-rated health item identified Latinos as having highest need across 5 groups (blacks, Asians, Latinos interviewed in English, Latinos interviewed in Spanish, and whites), whereas 5 other measures of need identified Latinos as having lowest level of need. Results suggest this item (in Spanish) measures something different (n = 1,045 Latinos interviewed in Spanish; total, n = 7,264) (18–64 y).[63] |
| *Construct validity:* The primary type of validity study that is done with respect to HRQOL measures is construct validity, because criteria against which to validate these measures seldom exist. Construct validity pertains to the extent to which measures are related to other measures to which they should be related (convergent validity) and not related to other measures that are different (discriminant validity). To the extent that hypotheses are confirmed in different groups, there is evidence that the instrument measures the same theoretical construct in each of the groups. | For AQLQ, interscale correlations for 5 of 10 pairs of scales ≥0.80 for blacks (n = 46; mean age, 33.1 y) and 3 of 5 pairs of scales for whites (n = 66; mean age, 33.6 y) indicating discriminant validity problems with emotional function, activity limitation, and symptoms. Correlation coefficients between scale scores and measures of disease severity, MOS SF-36, Cantril's Ladder, and HUI all in expected directions and not significantly different for blacks and whites.[38]<br>Correlations between self-administered (FSQ), interviewer-administered (Katz ADL and OARS-IADL), and performance-based measures of physical function were inconsistent and weak in community-dwelling elderly (n = 83, 64–92 y).[117]<br>For 3 MOS long-form measures in baseline longitudinal panel, known-groups validity generally comparable across age groups (18–44, 45–64, 65–74, 75+ y) (n = 2,546, 18–97 y).[55]<br>For English language MOS SF-36, all 8 scales successfully discriminated between Cuban patients with (n = 85) and without (n = 105) benign prostatic hyperplasia.[143]<br>For MOS SF-36 in diabetic Pima Indians, subjects with more comorbid chronic conditions had significantly lower scores on 6 of 9 dimensions; those taking insulin (n = 22) had lower scores on all scales than those who were not (n = 27) (n = 49, 24 to 78 y).[124] | Both English and Spanish versions of AUA symptom index, UCLA Prostate Cancer Index, a pain inventory, and SF-36 administered to bilingual men in Florida. Urological disease-specific items had κ between the 2 language versions 0.69–0.96. κ lower for 2 SF-36 social functioning items (κ = 0.26 and 0.51) (n = 100; mean age, 61 y).[146]<br>Spanish FACT scores appeared to perform at least as well as the original English language version in terms of relationships between the scales and other health measures. (n = 92 cancer patients, 23–80 y).[15]<br>Interscale correlations ≤0.65 between 4 scale scores identified through exploratory factor analysis of GWB in combined sample of Spanish- and English-speaking Mexican American women (combined because no mean differences were found by language on items) (n = 122, 18–65 y).[145]<br>For Langner Scale of Psychiatric Symptoms, residual ethnic differences in reported symptoms between Puerto Ricans (n = 47, 38 of whom took Spanish version) and whites (n = 106), controlling for social class, were no longer significant once 3 types of response biases were controlled for: social desirability, acquiescence to attitude statements, and acquiescence to health-related statements) (18 to 84 y).[148]<br>All scales of Spanish MDRS discriminated between normal elderly Mexican American sample (n = 102; mean age, 71 y) and neurologically and neuropsychiatrically impaired sample (n = 20; mean age, 73 y). Significant mean differences found on all scales between normal elderly Mexican American sample and original normative group of 85 nonimpaired English-speaking elders, indicating need for Spanish norms.[144]<br>For Spanish MOS SF-36, all 8 scales discriminated between Cubans with (n = 84) and without (n = 103) benign prostatic hyperplasia.[143] |

(*Continued*)

TABLE 2. (Continued)

| Type of Psychometric Approach | Example in English Language Version | Example in Alternative Language Version |
|---|---|---|
| | | For Japanese version of MOS SF-36, poor discriminant validity for 8 items (22%) because more highly correlated with other scale than own hypothesized scale. Clinical measures of balance and gait were more highly correlated with physical functioning than others, and lowest correlations were with mental health (n = 80, 64–100 y).[118] |
| *Response bias:* Response biases are systematic errors that are not detected through reliability coefficients (ie, systematic errors are reliable). Instrument should yield minimal or comparable response biases (systematic error) across groups. Three main forms of response bias are relevant with respect to particular subgroups, acquiescent response bias, socially desirable responding, and preference for extreme response categories. | On 14 MOS long-form measures, older patients in MOS baseline longitudinal panel were more likely to exhibit socially desirable responding than younger patients. No differences in acquiescent response set by age group (n = 2,546, 18–97 y).[55] | Spanish-heritage subjects had greater tendency to "yes saying" than nonwhites and whites on attitudes toward health care items in Spanish version of 1976 CHAS-NORC survey (n = 7,787, all ages).[147] |
| *Responsiveness to change:* A form of longitudinal validity—the extent to which a measure is sensitive to known changes over time,[149] although alternative definitions have been proposed.[155] Lack of change in an intervention may or may not be indicative of responsiveness, as the lack of change could be due to an ineffective intervention.[84] Evidence of equivalence in responsiveness across groups needs to be obtained from other studies before interpreting differences in treatment effectiveness by subgroup in a particular study. | Assessed baseline to last follow-up visit (7-year) changes in MOS SF-36 in blacks with hypertensive nephrosclerosis (n = 84); patients randomized to usual mean arterial blood pressure (MAP) goal or a low-MAP goal and 1 of 3 hypertensive drugs. Mean scores significantly increased on Physical Functioning (9.2), Role-Physical (19.0), Social Functioning (9.0), and Vitality (5.6) in usual-MAP goal group; none of the scores changed for low-MAP group or any of the drug regimens.[150]<br>No significant differences between 3 antihypertensive drug treatment groups among black men (n = 154) and black women (n = 152) in changes in total scores on all HRQOL measures from baseline to 8 weeks of active therapy. Within treatment groups, men improved on 6 of 10 and women on 4 of 10 measures (n = 306; 18–70 y).[151] | |

ADL indicates activities of daily living; AQLQ, Asthma Quality of Life Questionnaire; AUA, American Urological Association Symptom Index; FSQ BADL, Functional Status Questionnaire Basic Activities of Daily Living; CHAS-NORC, Center for Health Administration Studies and the National Opinion Research Center; DIF, differential item functioning; EPESE, Established Populations for Epidemiologic Studies of the Elderly; FSQ, Functional Status Questionnaire; GWB, General Well-Being Index; HAQ, Stanford Health Assessment Questionnaire; HHANES, Hispanic Health and Nutrition Examination Survey; HUI, Health Utilities Index; MDRS, Mattis Dementia Rating Scale; MQOLS-CA, Multidimensional Quality of Life Scale—Cancer Version; NHIS, National Health Interview Survey; OARS-IADL, Older Americans Resources and Services Instrumental Activities of Daily Living; SEM, structural equation modeling; SIP, Sickness Impact Profile; SF-20, Medical Outcomes Study Short-Form General Health Survey; and SF-36, Medical Outcomes Study Short-Form 36.

English SF-36 found ethnic, gender, SES, and age differences in item-scale correlations. Analyses of more complex models of health that included several instruments and constructs also produced mixed results, with 1 study finding comparable measurement structures across ethnic groups[58] and another finding noncomparability.[57]

Few HRQOL studies have applied differential item functioning techniques to subgroups. One study found possible gender differences in the interpretation of a CES-D item on crying,[138] and another found ethnic differences in items assessing somatic symptoms.[140]

English versions of various MOS measures have demonstrated good construct validity across age groups[156] and selected ethnic subgroups, as have both the Spanish and English versions of the SF-36 in Cubans.[143] A few studies of Spanish versions of the FACT, General Well-Being Sched-

TABLE 3. Proposed Recommendations for Evaluating the Conceptual and Psychometric Adequacy of HRQOL Measures in Diverse Populations

| Recommendation | Rationale |
|---|---|
| Improve representation of diverse groups | |
| • Encourage community partnerships with diverse communities during all phases of research project before conceptual or measurement studies. | • Need to be more sensitive to and respectful of diverse communities; need to maximize research benefits accrued by community to be studied. |
| • Disseminate information on effective recruitment and data collection methods in vulnerable groups. | • Increases representation of vulnerable groups in HRQOL research. |
| Expand evidence base | |
| • Include description of methods used for recruitment, data collection, and instrument translation and adaptation in published studies. | • Enables others to evaluate comparability and relative usefulness of methods across studies. |
| • Include description of basic psychometric properties and qualitative measurement study results by relevant population subgroups in published studies, eg, age, gender, race, ethnicity, and socioeconomic status. | • Accumulates systematic evidence about conceptual and psychometric adequacy of HRQOL measures in US subpopulations. |
| • Encourage pooling of operationally equivalent HRQOL data sets for measurement studies in subgroups. | • Single data sets rarely have sufficient subgroup sample sizes to provide an adequate test of the measures. |
| • Dedicate resources to measurement studies in diverse groups through special program announcements and funding of specialized measurement studies. | • Necessary for implementation of further HRQOL measurement studies in subgroups. |
| Coordinate measurement studies | |
| • Establish an infrastructure to coordinate HRQOL measurement studies and centralize findings. | • Will assist with systematic development of consensus as to criteria and guidelines for the testing, reporting, analysis, modification, and development of new measures of HRQOL for use in diverse populations. |
| Optimize universality of existing HRQOL measures | |
| • Integrate qualitative and quantitative methods to ensure conceptual validity of HRQOL measures for the intended groups before psychometric testing. | • Establishes whether measures are measuring what we intend to measure without the omission of important concepts. Ensures that HRQOL concepts are relevant before items and scales are tested for psychometric adequacy. |
| • Omit or replace items that are emic in nature with more etic or pancultural items (unless retained as items for further analysis acknowledging their lack of comparability).[62] | • Allows for valid group comparisons on HRQOL measures. |
| Advance field of measurement in diverse populations | |
| • Raise awareness of importance of conducting measurement studies when diverse groups are included in treatment effectiveness studies. | • To address methodologically necessary steps often overlooked by investigators that strengthen confidence in validity of study findings. |
| • Continue development of strategies for constructing and scoring measures to be as equivalent as possible. | • Optimizes validity of HRQOL research conducted in diverse populations. |
| • Continue development of analytic strategies to interpret differences in treatment effectiveness when HRQOL measurement properties differ across subgroups. | • Optimizes validity of HRQOL research conducted in diverse populations. |
| • Further develop consensus on methods for dealing with nonequivalence of measurement properties of standard HRQOL measures. | • Allows for standardization of methods and increased comparability across studies. |
| • Study the relative usefulness of a combination of universal and culture-specific modules for measuring changes in HRQOL. | • Culture-specific measures could facilitate understanding of whether these are more likely to detect change than universal measures in particular subgroups. |

ule, Mattis Dementia Rating Scale, American Urological Association Symptom Index, and UCLA prostate cancer index have also shown good preliminary construct validity. Using a Japanese version of the SF-36, researchers found problems with discriminant validity in almost one fourth of the items,[118] indicating that more measurement studies are needed in this group. The few studies of response bias point to a greater likelihood of socially desirable responding in older age groups and mixed results in Spanish-speaking respondents. In 1 study of blacks, the SF-36 demonstrated responsiveness to change.[150]

Another measurement issue pertains to the validity of norm-based comparisons. The validity of norms is determined largely by the degree to which the population on which the norms were developed approximates the sample being compared. Thus, use of norms based on populations that do not represent well the diversity of the US population may result in invalid subgroup comparisons of samples that include minority or other diverse groups.

Clearly, measurement studies need to be conducted more routinely in research on diverse groups.[138] Given the relatively few studies examining the comparability of standard HRQOL measures across and within diverse groups and the conflicting results, more work is clearly needed.

### Research Methods Are a Crucial Context for Interpreting Measurement Characteristics

Psychometric properties are a function of the setting as well as the recruitment and data collection methods used. Thus, interpretation of measurement studies requires knowing these parameters.[88,90,109,157] Such methods affect the nature of the final "diverse" group and the extent to which it represents the target diverse population. Studies that use culturally sensitive recruitment and interviewing strategies may yield more representative samples of the target group and more accurate responses to survey items than those incorporating traditional methods.[158,159] To best achieve a representative sample, a community partnership approach is useful in which community members become actively involved in the process of developing the survey and recruitment methods.[160]

## Methodological and Analytical Challenges in Assessing and Interpreting Measures in the Context of Evaluating Treatment Effectiveness

Given the plethora of standard generic and disease-specific HRQOL measures and the variety of diverse subgroups of interest, the task at hand is daunting. We need direction as to how to proceed with research that includes diverse groups and simultaneously begin to build an evidence base. We summarize here some challenges facing investigators and the field of measurement of HRQOL in this effort.

### Raise Awareness of Need for More Studies and Better Reporting of Measurement Results

We need to increase awareness of the importance of exploring measurement comparability in treatment effectiveness studies that include diverse populations. This requires disseminating information on the importance of and techniques for assessing comparability, not only when translating a measure but also when an English instrument will be used. Reporting measurement findings by subgroup is also necessary; these findings may be unpublished either because noncomparable measurement findings are considered as a measurement failure or because comparability is assumed and tests are never conducted. In addition, Rogler[161] suggests reporting in detail the nature of cross-cultural modifications made, to enhance awareness of the types of conceptual inadequacies that occur and the methods for adaptation. If developers of measures and researchers conducting measurement studies in diverse groups consistently reported conceptual and measurement comparability findings, the evidence base would grow.

As noted above, most studies of English measures focused on psychometric properties; only 1 qualitative study was found that explored whether the measures were missing relevant constructs and items or whether the definitions and items had similar meaning to the subgroups.[15] Thus, qualitative studies should be particularly emphasized as an essential and important complementary approach.[67]

### Need for More Resources to Conduct Studies

Comprehensive measurement testing in diverse groups requires large studies of representative samples[71] as well as sophisticated quantitative and qualitative techniques. Studies need to be conducted by multidisciplinary research teams skilled in these methods. Such studies take time, given the sequential nature of the testing. Thus, targeted funding initiatives are needed for measurement projects. This could take the form of the international projects designed to develop comparable measures across countries.[162,163] A model for accomplishing these goals in the United States is the NIH-funded Resource Centers for Minority Aging Research, which have cores dedicated to the development of appropriate measures for aging research with minority populations. Other models have been developed by AHRQ[164] and NIH.[165] Funding efforts could consist of interagency collaborations.

In addition, agencies could consider funding measurement studies as a preliminary phase of treatment effectiveness studies. Even more useful would be to allocate funds for exploring HRQOL measurement issues in diverse populations in every effectiveness study in which diverse groups are included.

### Need to Integrate Findings and Make Them Available

A mechanism is needed to integrate measurement findings and make them easily accessible to investigators. One mechanism could be for editors of various compendia of measures to include information on measurement properties across diverse populations. This would not only make the information readily available but would highlight the value of the information. Dedicated symposia to disseminate findings about how HRQOL measures work across groups would be useful. This would enable investigators to utilize an accumulating evidence base in selecting measures, as well as identify gaps needing further research.

### When Standard Measures Are Not Comparable

When particular studies find noncomparable or inadequate measures in relevant subgroups, no standard methods exist for addressing scoring and analytic options. Such methods could be developed through dialogue among measurement specialists. One approach is to develop alternative scoring procedures. For example, some have omitted items that don't work in subgroups and created scales with only reliable or nonbiased items in the subgroups.[58,120,140,166] For comparative studies, Bullinger and colleagues[84] created a set of scales from the FACT instrument that included only items that are unbiased across English and Spanish groups. They then created another set of "within-language" scales as additional outcomes that cannot be generalized across the groups. Thus, they have attempted to "strike a compromise" between obtaining a single set of items that are comparable across all cultures and creating culture-specific measures. Item response theory and test equating can be used to cocalibrate within-language measures that are constructed with a smaller set of unbiased items. The structural equation modeling literature suggests that it may be possible to have "group-specific" item loadings[96] to accommodate findings that items have different factor loadings across groups. The usefulness of combining etic and emic items needs to be further evaluated.

In studies examining whether treatments are equally effective for vulnerable subgroups, noncomparable measurement properties will make it hard to analyze the data and interpret the results. One analytic possibility is to stratify analyses within groups to ensure that findings are not attributable to measurement differences.[167] However, this can result in loss of power.

To the extent that evidence across several studies suggests substantial and consistent noncomparability of standard measures, it may be time to consider modifying original measures[161] or developing new ones that build on existing measures. The modification process could resemble the "parallel approach" from international studies.[84,168] This entails obtaining input from representatives of several diverse groups in selecting concepts and items to ensure their relevance, applicability, and meaning in each group. This could involve omitting items that are emic in nature or replacing them with more etic or pancultural items.[62] Although modifying standard instruments would certainly be controversial, it may nevertheless be the only direction to take if standard measures are found to be substantially and consistently biased in major subgroups.

If new measures are to be developed, researchers should follow the guidelines suggested above and attempt to define each concept fully from the perspective of many diverse groups using qualitative studies. To the extent that investigators begin a new measure by using items from prior measures, those items also need to be subjected to qualitative studies to ensure their appropriateness.

Given limited resources, it may be useful to begin this new measurement work in studies that aim primarily to improve HRQOL or that plan to test whether a given treatment is differentially effective across subgroups in terms of HRQOL. Another way to prioritize this work is to study subgroups that suffer a disproportionate burden of disease or are suspected to have poorer HRQOL outcomes. We could also begin with major ethnic groups, low-SES groups, and older persons who are the focus of current policy initiatives and thus are most likely to be included in studies of treatment effectiveness. The choice of HRQOL instruments might give priority to those most commonly used, such as the SF-36 and the CES-D. Another approach is to begin with studies of generic instruments that can be more widely applied and proceed to disease-specific measures.

### Recommendations

Recommendations for addressing the issues raised are summarized in Table 3. HRQOL measurement can best be advanced through systematic and coordinated efforts to improve the representation of diverse groups in treatment effectiveness research, expand the evidence base, coordinate measurement studies, optimize the universality of existing HRQOL measures, and address appropriate methodological strategies. These suggestions can serve to guide further efforts in this area.

## Conclusions

We advocate culturally sensitive research regarding HRQOL measures through continuous efforts to "mesh the process of inquiry with the cultural characteristics of the group being studied."[14] Research is made culturally sensitive by considering the culture of the groups involved throughout the entire research process: planning, recruitment, data collection, measurement, analysis, and interpretation.[161] Because examining the cultural equivalence of HRQOL measures within the United States is somewhat new, we have a unique opportunity to shape the direction of this work by developing and disseminating guidelines. This article represents a step toward generating dialogue on the issues and methods involved in ensuring the adequacy of HRQOL measures so that substantial portions of the US population are not overlooked in the development of treatment effectiveness policy.

## Acknowledgments

We would like to thank Dr Eliseo J. Pérez-Stable, Dr Steven Gregorich, Dr Martha Rangel-Lugo, and Deirdra Forté for their substantial assistance.

## References

1. US Census Bureau. Population projections of the United States by age, sex, race, and Hispanic origin: 1995 to 2050. Available at: http://www.census.gov/prod/1/pop/p25–1130/ Accessed June 6, 2000.

2. US Census Bureau. Table RDP-5: Percent of people in poverty, by definition of income: 1979 to 1998. Available at: http://www.census.gov/hhes/poverty/histpov/rdp05.html. Accessed September 12, 1999.

3. **Moss N, Krieger N.** Measuring social inequalities in health. Public Health Rep 1995;110:302–305.

4. **Williams DR, Collins C.** U.S. socioeconomic and racial differences in health: Patterns and explanations. Annu Rev Soc 1995;21:349–386.

5. **Pamuk E, Makuc D, Heck K, Reuben C, Lochner K.** Socioeconomic status and health chartbook: Health United States, 1998. Hyattsville, Md: National Center for Health Statistics; 1998.

6. **Swanson GM, Ward AJ.** Recruiting minorities into clinical trials: Toward a participant-friendly system. J Natl Cancer Inst 1995;87:1747–1759.

7. **Kaye JM, Lawton P, Kaye D.** Attitudes of elderly people about clinical research on aging. Gerontologist 1990;30:100–106.

8. **Vernon SW, Roberts RE, Lee ES.** Ethnic status and participation in longitudinal health surveys. Am J Epidemiol 1984;119:99–113.

9. **Areán PA, Gallagher-Thompson D.** Issues and recommendations for the recruitment and retention of older ethnic minority adults into clinical research. J Consult Clin Psychol 1996;64:875–880.

10. **McCarthy CR.** Historical background of clinical trials involving women and minorities. Acad Med 1994;69:695–698.

11. **Hughes D, DuMont K.** Using focus groups to facilitate culturally anchored research. Am J Community Psychol 1993;21:775–806.

12. **Marín G, Marín BV.** Research with Hispanic populations, volume 23. Newbury Park, Calif: Sage Publications; 1991.

13. **Pasick RJ.** Socioeconomic and cultural factors in the development and use of theory. In: Glanz K, Lewis FM, Rimer BK, eds. Health behavior and health education. San Francisco, Calif: Jossey-Bass Inc; 1997:425–440.

14. **Rogler LH.** The meaning of culturally sensitive research in mental health. Am J Psychiatry 1989;146:296–303.

15. **Cella D, Hernandez L, Bonomi AE, Corona M, Vaquero M, Shiomoto G, et al.** Spanish language translation and initial validation of the Functional Assessment of Cancer Therapy quality of life instrument. Med Care 1998;36:1407–1418.

16. **Kuyken W, Orley J, Hudelson P, Sartorius N.** Quality of life assessment across cultures. Int J Ment Health 1994;23:5–27.

17. **Hutchinson JF.** Quality of life in ethnic groups. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. Philadelphia, Pa: Lippincott-Raven; 1996:587–593.

18. **Weiss MG, Kleinman A.** Depression in cross-cultural perspective: Developing a culturally informed model. In: Dasen PR, Berry JW, Sartorius N, eds. Health and cross-cultural psychology: Toward applications. Newbury Park, Calif: Sage Publications Inc; 1988:179–206.

19. **Testa MA, Simonson DC.** Assessment of quality-of-life outcomes. N Engl J Med 1996;334:835–840.

20. **Flaskerud JH, Winslow BJ.** Conceptualizing vulnerable populations health-related research. Nurs Res 1998;47:69–78.

21. Agency for Health Care Policy and Research. AHCPR seeks proposals to develop quality of care measures for vulnerable populations. Available at: http://www.ahcpr.gov/news/press/vulnpr.htm. Accessed 1998.

22. **Russo J, Trujillo CA, Wingerson D, Decker K, Ries R, Wetzler H, et al.** The MOS 36-Item Short Form Health Survey: Reliability, validity, and preliminary findings in schizophrenic outpatients. Med Care 1998;36:752–756.

23. **Wu AW.** Quality-of-life assessment in clinical research: Application in diverse populations. Med Care. 2000;38(suppl II):II-130-II-135.

24. **Mueller KJ, Ortega ST, Parker K, Patil K, Askenazi A.** Health status and access to care among rural minorities. J Health Care Poor Underserved 1999;10:230–249.

25. **Krieger N, Rowley DL, Herman AA, Avery B, Phillips MT.** Racism, sexism, and social class: Implications for studies of health, disease, and well-being. Am J Prev Med 1993;9(suppl):82–122.

26. **Williams DR, Lavizzo-Mourey R, Warren RC.** The concept of race and health status in America. Public Health Rep 1994;109:26–41.

27. **Burns RB, McCarthy EP, Freund KM, Marwill SL, Shwartz M, Ash A, et al.** Black women receive less mammography even with similar use of primary care [see comments]. Ann Intern Med 1996;125:173–182.

28. **Carlisle DM, Leake BD, Shapiro MF.** Racial and ethnic disparities in the use of cardiovascular procedures: Associations with type of health insurance. Am J Public Health 1997;87:263–267.

29. Council on Ethical and Judicial Affairs. Black-white disparities in health care. JAMA 1990;263:2344–2346.

30. **Escarce JJ, Epstein KR, Colby DC, Schwartz JS.** Racial differences in the elderly's use of medical procedures and diagnostic tests. Am J Public Health 1993;83:948–954.

31. **Fuentes-Afflick E, Korenbrot CC, Greene J.** Ethnic disparity in the performance of prenatal nutrition risk assessment among Medicaid-eligible women. Public Health Rep 1995;110:764–773.

32. **Gornick ME, Eggers PW, Reilly TW, Mentnech RM, Fitterman LK, Kucken LE, et al.** Effects of race and income on mortality and use of services among Medicare beneficiaries. N Engl J Med 1996;335:791–799.

33. **Harlan L, Brawley O, Pommerenke F, Wali P, Kramer B.** Geographic, age, and racial variation in the treatment of local/regional carcinoma of the prostate [see comments]. J Clin Oncol 1995;13:93–100.

34. **Kuppermann M, Gates E, Washington AE.** Racial-ethnic differences in prenatal diagnostic test use and outcomes: Preferences, socioeconomics, or patient knowledge? Obstet Gynecol 1996;87:675–682.

35. **McWhorter WP, Mayer WJ.** Black/white differences in type of initial breast cancer treatment and implications for survival. Am J Public Health 1987;77:1515–1517.

36. **Todd KH, Samaroo N, Hoffman JR.** Ethnicity as a risk factor for inadequate emergency department analgesia. JAMA 1993;269:1537–1539.

37. **Wenneker MB, Epstein AM.** Racial inequalities in the use of procedures for patients with ischemic heart disease in Massachusetts. JAMA 1989;261:253–257.

38. **Leidy KN, Chan KS, Coughlin C.** Is the asthma quality of life questionnaire a useful measure for low-income asthmatics? Am J Respir Crit Care Med 1998;158:1082–1090.

39. **Cella DF, Lloyd SR, Wright BD.** Cross-cultural instrument equating: Current research and fu-

ture directions. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. Philadelphia, Pa: Lippincott-Raven; 1996:707–715.

40. **Burstin HR, Lipsitz SR, Brennan TA.** Socioeconomic status and risk for substandard medical care. JAMA 1992;268:2383–2387.

41. **Greenberg ER, Chute CG, Stukel T, Baron JA, Freeman DH, Yates J, et al.** Social and economic factors in the choice of lung cancer treatment. N Engl J Med 1988;318:612–617.

42. **Manson-Siddle CJ, Robinson MB.** Super profile: Analysis of socioeconomic variations in coronary investigation and revascularization rates. J Epidemiol Community Health 1998;52:507–512.

43. **Tobin JN, Wassertheil-Smoller S, Wexler JP, Steingart RM, Budner N, Lense L, et al.** Sex bias in considering coronary bypass surgery. Ann Intern Med 1987;107:19–25.

44. **Khan SS, Nessim S, Gray R, Czer LS, Chaux A, Matloff J.** Increased mortality of women in coronary artery bypass surgery: Evidence for referral bias. Ann Intern Med 1990;112:561–567.

45. **Ayanian JZ, Epstein AM.** Differences in the use of procedures between women and men hospitalized for coronary heart disease [see comments]. N Engl J Med 1991;325:221–225.

46. **Saha S, Stettin GD, Redberg RF.** Gender and willingness to undergo invasive cardiac procedures. J Gen Intern Med 1999;14:122–125.

47. **Safran DG, Rogers WH, Tarlov AR, McHorney CA, Ware JE Jr.** Gender differences in medical treatment: The case of physician-prescribed activity restrictions. Soc Sci Med 1997;45:711–722.

48. **Steingart RM, Packer M, Hamm P, Coglianese ME, Gersh B, Geltman EM, et al.** Sex differences in the management of coronary artery disease: Survival and Ventricular Enlargement Investigators [see comments]. N Engl J Med 1991;325:226–230.

49. **Shaw LJ, Miller DD, Romeis JC, Kargl D, Younis LT, Chaitman BR.** Gender differences in the noninvasive evaluation and management of patients with suspected coronary artery disease. Ann Intern Med 1994;120:559–566.

50. **Samet J, Hunt WC, Key C, Humble CG, Goodwin JS.** Choice of cancer therapy varies with age of patient. JAMA 1986;255:3385–3390.

51. **Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, et al.** The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med 1999;340:618–626.

52. **Schwartz LM, Woloshin S, Welch HG, for the VA Outcomes Group.** Sounding board: Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. N Engl J Med 1999;341:278–283.

53. **McHorney CA, Ware JE, Raczek AE.** The MOS 36-item short form health survey (SF-36), II: Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–263.

54. **Meredith LS, Siu AL.** Variation and quality of self-report health data: Asians and Pacific Islanders compared with other ethnic groups. Med Care 1995;33:1120–1131.

55. **Sherbourne CD, Meredith LS.** Quality of self-report data: A comparison of older and younger chronically ill patients. J Gerontol 1992;47:S204–S211.

56. **Markides KS, Stroup-Benham CA, Goodwin JS, Perkowski LC, Lichtenstein M, Ray LA.** The effect of medical conditions on the functional limitations of Mexican-American elderly. Ann Epidemiol 1996;6:386–391.

57. **Johnson RJ, Wolinsky FD.** Gender, race, and health: The structure of health status among older adults. Gerontologist 1994;34:24–35.

58. **Stump TE, Clark DO, Johnson RJ, Wolinsky FD.** The structure of health status among Hispanic, African American, and white older adults. J Gerontol B Psychol Sci Soc Sci 1997;52B(special issue):49–60.

59. **Berkanovic E, Telesky C.** Mexican-American, black American and white American differences in reporting illnesses, disability and physician visits for illnesses. Soc Sci Med 1985;20:567–577.

60. **Angel R, Thoits P.** The impact of culture on the cognitive structure of illness. Cult Med Psychiatry 1987;11:465–494.

61. **Raczynski JM, Taylor H, Cutter G, Hardin M, Rappaport N, Oberman A.** Diagnoses, symptoms, and attribution of symptoms among black and white inpatients admitted for coronary heart disease. Am J Public Health 1994;84:951–956.

62. **Johnson TP, O'Rourke D, Chavez N, Sudman S, Warnecke RB, Lacey L, et al.** Cultural variations in the interpretation of health survey questions. In: Warnecke R, ed. Health survey research methods. Hyattsville, Md: National Center for Health Statistics; 1996:57–62.

63. **Osmond DH, Vranizan K, Schillinger D, Stewart AL, Bindman AB.** Measuring the need for medical care in an ethnically diverse population. Health Serv Res 1996;31:551–571.

64. **Shetterly SM, Baxter J, Mason LD, Hamman RF.** Self-rated health among Hispanic vs non-Hispanic white adults: The San Luis Valley Health and Aging Study. Am J Public Health 1996;86:1798–1801.

65. **Ferketich S, Phillips L, Verran J.** Focus on psychometrics: Development and administration of a survey instrument for cross-cultural research. Res Nurs Health 1993;16:227–230.

66. **Barofsky I.** The role of cognitive equivalence in studies of health-related quality-of-life assessments. Med Care 2000;38(suppl II):II-125–II-129.

67. **Herdman M, Fox-Rushby J, Badia X.** A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. Qual Life Res 1998;7:323–335.

68. **Patrick DL, Wild DJ, Johnson ES, Wagner TH, Martin MA.** Cross-cultural validation of quality of life measures. In: Orley J, Kuyken W, eds. Quality of life assessment: International perspectives: Proceedings of the joint meeting organized by the World Health Organization and the Foundation IPSEN in Paris, July 2–3, 1993. Berlin, Germany: Springer-Verlag; 1994:19–32.

69. **Berry JW.** On cross-cultural comparability. Int J Psychol 1969;4:119–128.

70. **Triandis HC, Marín G.** Etic plus emic versus pseudoetic: A test of a basic assumption of contemporary cross-cultural psychology. J Cross-Cult Psychol 1983;14:489–500.

71. **Geisinger KF.** Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychol Assess 1994;6:304–312.

72. **Guyatt GH.** The philosophy of health-related quality of life translation. Qual Life Res 1993;2:461–465.

73. **Hunt SM.** Cross-cultural issues in the use of quality of life measures in randomized controlled clinical trials. In: Staquet MJ, Hays RD, Fayers PM, eds. Quality of life assessment in clinical trials. New York, NY: Oxford University Press; 1998:51–67.

74. **Guillemin F, Bombardier C, Beaton D.** Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417–1432.

75. **Sartorius N, Kuyken W.** Translation of health status instruments. In: Orley J, Kuyken W, eds. Quality of life assessment: International perspectives. Heidelberg, Germany: Springer-Verlag; 1994:3–18.

76. **Flaherty JA.** Appropriate and inappropriate research methodologies for Hispanic mental health. In: Gaviria M, Arana JD, eds. Health and behavior: Research agenda for Hispanics. Chicago, Ill: University of Illinois; 1987:177–186.

77. **Flaherty JA, Gaviria FM, Pathak D, Mitchell T, Wintrob R, Richman JA, et al.** Developing instruments for cross-cultural psychiatric research. J Nerv Ment Dis 1988;176:257–263.

78. **Hays RD, Anderson R, Revicki D.** Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993;2:441–449.

79. **Bullinger M.** Ensuring international equivalence of quality of life measures: Problems and approaches to solutions. In: Orley J, Kuyken W, eds. Quality of life assessment: International perspectives: Proceedings of the joint meeting organized by the World Health Organization and the Foundation IPSEN in Paris, July 2–3, 1993. Berlin, Germany: Springer-Verlag; 1994:33–40.

80. **Hui CH, Triandis HC.** Multistrategy approach to cross-cultural research: The case of locus of control. J Cross-Cult Psychol 1983;14:65–83.

81. **Hui CH, Triandis HC.** Measurement in cross-cultural psychology: A review and comparison of strategies. J Cross-Cult Psychol 1985;16:131–152.

82. **Anderson RT, McFarlane M, Naughton MJ, Shumaker SA.** Conceptual issues and considerations in cross-cultural validation of generic health-related quality of life instruments. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia, Pa: Lippincott-Raven; 1996:605–612.

83. **Ware JE, Gandek BL, Keller SD, for the IQOLA Project Group.** Evaluating instruments used cross-nationally: Methods from the IQOLA project. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia, Pa: Lippincott-Raven; 1996:681–692.

84. **Bullinger M, Anderson R, Cella D, Aaronson N.** Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual Life Res 1993;2:451–459.

85. **Cole NS, Moss PA.** Bias in test use. In: Linn RL, ed. Educational measurement. 3rd ed. New York, NY: American Council on Education and Macmillan Publishing Co; 1989:201–219.

86. **Hambleton RK.** Principles and selected applications of item response theory. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: American Council on Education and Macmillan Publishing Co; 1989:147–200.

87. **Nunnally JC, Bernstein IH.** Psychometric theory. 3rd ed. New York, NY: McGraw-Hill, Inc; 1994.

88. **González-Calvo J, González VM, Lorig K.** Cultural diversity issues in the development of valid and reliable measures of health status. Arthritis Care Res 1997;19:448–456.

89. **Szabo S, Orley J, Saxena S.** An approach to response scale development for cross-cultural questionnaires. Eur Psychol 1997;2:270–276.

90. **Hazuda H.** Minority issues in Alzheimer disease outcomes research. Alzheimer Dis Assoc Disord 1997;11:156–161.

91. **Osgood CE, Suci GJ, Tannenbaum PH.** The measurement of meaning. Urbana, Ill: University of Illinois Press; 1947.

92. **Keith KD, Heal LW, Schalock RL.** Cross-cultural measurement of critical quality of life concepts. J Intellect Dev Disabil 1996;21:273–293.

93. **Patrick DL, Sittampalam Y, Somerville SM, Carter WB, Bergner M.** A cross-cultural comparison of health status values. Am J Public Health 1985;75:1402–1407.

94. **Stewart AL, Hays RD, Ware JE.** Methods of constructing health measures. In: Stewart AL, Ware JE, eds. Measuring functioning and well-being: The Medical Outcomes Study approach. Durham, NC: Duke University Press; 1992:67–85.

95. **Ware JE, Harris WJ, Gandek B, Rogers BW, Resse PR.** MAP-R for Windows: Multitrait/multi-item analysis program: Revised user's guide. Boston, Mass: Health Assessment Lab; 1997.

96. **Byrne BM, Shavelson RJ, Muthen B.** Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychol Bull 1989;105:456–466.

97. **Meredith W.** Notes on factorial invariance. Psychometrika 1964;29:177–185.

98. **Meredith W.** Measurement invariance, factor analysis and factorial invariance. Psychometrika 1993;58:525–543.

99. **Dancer LS, Anderson AJ, Derlin RL.** Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. J Consult Clin Psychol 1994;62:710–717.

100. **Pope C, Mays N.** Reaching the parts other methods cannot reach: An introduction to qualitative methods in health and health services research. Br Med J 1995;311:42–45.

101. **Sudman S, Bradburn NM, Schwarz N.** Thinking about answers: The application of cognitive processes to survey methodology. San Francisco, Calif: Jossey-Bass Inc; 1996.

102. **Sudman S, Warnecke R, Johnson T, O'Rourke D, Davis AM.** Cognitive aspects of reporting cancer prevention examinations and tests. Hyattsville, Md: Public Health Service, National Center for Health Statistics; Vital and Health Statistics series 6, No. 7; 1994.

103. **Harris-Kojetin LD, Fowler FJ, Jr, Brown JA, Schnaier JA, Sweeny SF.** The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. Med Care 1999;37(suppl):MS10–MS21.

104. **Oksenberg L, Cannell C, Kalton G.** New strategies for pretesting survey questions. J Official Stat 1991;7:349–365.

105. **Morgan DL.** Focus groups as qualitative research: Volume 16, Qualitative research methods. Newbury Park, Calif: Sage Publications; 1988.

106. **Wagner L.** Focus groups: When and how to use them: A practical guide. 2nd ed. Stanford, Calif: Health Promotion Resource Center, Stanford Center for Research in Disease Prevention; 1992.

107. **Morse J.** Approaches to qualitative-quantitative methodological triangulation. Nurs Res 1991;40:253–270.

108. **Messick S.** Validity. In: Linn RL, ed. Educational measurement. 3rd ed. New York, NY: American Council on Education and Macmillan Publishing Co; 1989:13–103.

109. **McHorney CA.** Measuring and monitoring general health status in elderly persons: Practical and methodological issues in using the SF-36 health survey. Gerontologist 1996;36:571–583.

110. **Callahan CM, Wolinsky FD.** The effect of gender and race on the measurement properties of the CES-D in older adults. Med Care 1994;32:341–356.

111. **Roberts RE.** Reliability of the CES-D scale in different ethnic contexts. Psychiatry Res 1980;2:125–134.

112. **McHorney CA, Ware JE, Lu R, Sherbourne CD.** The MOS 36-item Short-Form Health Survey (SF-36), III: Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994;32:40–66.

113. **Colsher PL, Wallace RB.** Data quality and age: Health and psychobehavioral correlates of item nonresponse and inconsistent responses. J Gerontol 1989;44:P45–P52.

114. **Teresi JA, Holmes D.** Overview of methodological issues in gerontological and geriatric measurement. In: Lawton MP, Teresi JA, eds. Annual review of gerontology and geriatrics. New York, NY: Springer Publishing Co; 1994;14:1–22.

115. **Andresen EM, Patrick DL, Carter WB, Malmgren JA.** Comparing the performance of health status measures for healthy older adults. J Am Geriatr Soc 1995;43:1030–1034.

116. **Ren XS, Chang K.** Evaluating health status of elderly Chinese in Boston. J Clin Epidemiol 1998;51:429–435.

117. **Reuben DB, Valle LA, Hays RD, Siu AL.** Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. J Am Geriatr Soc 1995;43:17–23.

118. **Harada N, Chiu V, Tsuneishi C, Fukuhara S, Makinodan T.** Cross-cultural adaptation of the SF-36 health survey for Japanese-American elderly. J Aging Ethnicity 1998;1:59–80.

119. **Johnson PA, Goldman L, Orav EJ, Garcia T, Pearson SD, Lee TH.** Comparison of the Medical Outcomes Study Short-Form 36-Item Health Survey in black patients and white patients with acute chest pain. Med Care 1995;33:145–160.

120. **González VM, Stewart A, Ritter PL, Lorig K.** Translation and validation of arthritis outcome measures into Spanish. Arthritis Rheum 1995;38:1429–1446.

121. **Clark VA, Aneshensel CS, Frerichs RR, Morgan TM.** Analysis of effects of sex and age in response to items on the CES-D scale. Psychiatry Res 1981;5:171–181.

122. **Dunnigan T, McNall M, Mortimer JT.** The problem of metaphorical nonequivalence in cross-cultural survey research: Comparing the mental health statuses of Hmong refugee and general population adolescents. J Cross-Cult Psychol 1993;24:344–365.

123. **Golding JM, Aneshensel CS, Hough RL.** Responses to Depression Scale items among Mexican-Americans and non-Hispanic whites. J Clin Psychol 1991;47:61–75.

124. **Johnson JA, Nowatzki TE, Coons SJ.** Health-related quality of life of diabetic Pima Indians. Med Care 1996;34:97–102.

125. **Bollen KA.** Structural equations with latent variables. New York, NY: Wiley; 1989.

126. **Jöreskog KG.** A general method for estimating a linear structural equation system. In: Goldberger AS, Duncan OD, eds. Structural equation models in the social sciences. New York, NY: Seminar Press/Harcourt Brace; 1973:85–112 .

127. **Wolinsky FD, Stump TE.** A measurement model of the Medical Outcomes Study 36-Item Short Form Health Survey in a clinical sample of disadvantaged, older, black, and white men and women. Med Care 1996;34:537–548.

128. **Liang J, Van Tran T, Krause N, Markides KS.** Generational differences in the structure of the CES-D scale in Mexican Americans. J Gerontol 1989;44:S110–S120.

129. **Hertzog C, Van Alstine J, Usala PD, Hultsch DF, Dixon R.** Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. Psychol Assess 1990;2:64–72.

130. **Devins GM, Beiser M, Dion R, Pelletier LG, Edwards RG.** Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the Affect Balance Scale. Am J Public Health 1997;87:794–799.

131. **Guarnaccia PJ, Angel R, Worobey JL.** The factor structure of the CES-D in the Hispanic Health and Nutrition Examination Survey: The influences of ethnicity, gender and language. Soc Sci Med 1989;29:85–94.

132. **Dibble SL, Padilla GV, Dodd MJ, Miaskowski C.** Gender differences in the dimensions of quality of life. Oncol Nurs Forum 1998;25:577–583.

133. **Azen SP, Palmer JM, Carlson M, Mandel D, Cherry BJ, Fanchiang S, et al.** Psychometric properties of a Chinese translation of the SF-36 Health Survey questionnaire in the Well Elderly Study. J Aging Health 1999;11:240–251.

134. **Miller TQ, Markides KS, Black SA.** The factor structure of the CES-D in two surveys of elderly Mexican Americans. J Gerontol B Psychol Sci Soc Sci 1997;52:S259–S269.

135. **Dupuy HJ.** The Psychological General Well-Being (PGWB) Index. In: Wenger NK, Mattson ME, Furberg CD, Elinson J, eds. Assessment of quality of life in clinical trials of cardiovascular therapies. New York, NY: LeJacq; 1984:170–183.

136. **Whitelaw NA, Liang J.** The structure of the OARS physical health measures. Med Care 1991;29:332–347.

137. **Liang J, Bennett J, Whitelaw N, Maeda D.** The structure of self-reported physical health among the aged in the United States and Japan. Med Care 1991;29:1161–1180.

138. **Kessler RC, Mroczek DK.** Measuring the effects of medical interventions. Med Care 1995;33 (suppl):AS109–AS119.

139. **Teresi JA, Cross PS, Golden RR.** Some applications of latent trait analysis to the measurement of ADL. J Gerontol 1989;44:S196–S204.

140. **Teresi JA, Golden R.** Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. Alzheimer Dis Assoc Disord 1994;8(suppl):S291–S298.

141. **Kaplan RM, Bush JW, Berry CC.** The reliability, stability, and generalizability of a health status index. In: Proceedings of the Social Statistics Section, American Statistical Association. Alexandria, VA: The American Statistical Association; 1978:704–709.

142. **Angel R, Guarnaccia PJ.** Mind, body, and culture: Somatization among Hispanics. Soc Sci Med 1989;28:1229–1238.

143. **Arocho R, McMillan CA.** Discriminant and criterion validation of the US-Spanish version of the SF-36 Health Survey in a Cuban-American population with benign prostatic hyperplasia. Med Care 1998;36:766–772.

144. **Arnold BR, Cuellar I, Guzman N.** Statistical and clinical evaluation of the Mattis Dementia Rating Scale, Spanish adaptation: an initial investigation. J Gerontol B Psychol Sci Soc Sci 1998;53:364–369.

145. **Poston WS II, Olvera NE, Yanez C, Haddock CK, Dunn JK, Hanis CL, et al.** Evaluation of the factor structure and psychometric characteristics of the

General Well-Being Schedule (GWB) with Mexican American women. Women Health 1998;27:51–64.

146. **Krongrad A, Perczek RE, Burke MA, Granville LJ, Lai H, Lai S.** Reliability of Spanish translations of select urological quality of life instruments. J Urol 1997;158:493–496.

147. **Aday L, Chiu GY, Andersen R.** Methodological issues in health care surveys of the Spanish language population. Am J Public Health 1980;70:367–374.

148. **Carr LG, Krause N.** Social status, psychiatric symptomatology, and response bias. J Health Soc Behav 1978;19:86–91.

149. **Hays R, Hadorn D.** Responsiveness to change: An aspect of validity, not a separate dimension. Qual Life Res 1992;1:73–75.

150. **Kusek JW, Lee JY, Smith DE, Milligan S, Faulkner M, Cornell CE, et al.** Effect of blood pressure control and antihypertensive drug regimen on quality of life: The African American Study of Kidney Disease and Hypertension (AASK) Pilot Study. Control Clin Trials 1996;17(suppl):40S–46S.

151. **Croog SH, Kong BW, Levine S, Weir MR, Baume RM, Saunders E.** Hypertensive black men and women: Quality of life and effects of antihypertensive medications: Black Hypertension Quality of Life Multicenter Trial Group. Arch Intern Med 1990;150:1733–1741.

152. **Radloff LS.** The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;1:385–401.

153. **Hays RD, Morales LS, Reise SP.** Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38(suppl II):II-28–II-42.

154. **McHorney CA, Cohen AS.** Equating health status measures with item response theory: Illustrations with functional status items. Med Care 2000;38(suppl II):II-43–II-59.

155. **Liang MH.** Longitudinal construct validity: Establishment of clinical meaning in patient evaluative instruments. Med Care 2000;38(suppl II):II-84–II-90.

156. **Sherbourne CD, Meredith LS.** Quality of self-report data: A comparison of older and younger chronically ill patients. J Gerontol 1992;47:S204–S211.

157. **McGraw SA, McKinlay JB, Crawford SA, Costa LA, Cohen DL.** Health survey methods with minority populations: Some lessons from recent experiences. In: Becker DM, Hill DR, Jackson JS, Levine DM, Stillman FA, Weiss SM, eds. Health behavior research in minority populations: Access, design, and implementation. Washington, DC: US Dept of Health and Human Services; 1992:149–167.

158. **Ettl MK, Hays RD, Cunningham WE, Shapiro MF, Beck CK.** Assessing health-related quality of life in disadvantaged and very ill populations. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia, Pa: Lippincott-Raven; 1996:595–604.

159. **Cunningham WE, Bozzette SA, Hays RD, Kanouse DE, Shapiro MF.** Comparison of health-related quality of life in clinical trial and nonclinical trial human immunodeficiency virus-infected cohorts. Med Care 1995;33:AS15–AS25.

160. **Levine DM, Becker DM, Bone LR, Stillman FA, Tuggle MB, Prentice M, et al.** A partnership with minority populations: A community model of effectiveness research. Ethn Dis 1992;2:296–305.

161. **Rogler LH**. Methodological sources of cultural insensitivity in mental health research. Am Psychol 1999;54:424–433.

162. **Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, et al.** International Quality of Life Assessment (IQOLA) project. Qual Life Res 1992;1:349–351.

163. WHOQOL Group. The World Health Organization quality of life assessment (WHOQOL): Position paper from the World Health Organization. Soc Sci Med 1995;41:1402–1409.

164. AHCPR seeks proposals to develop quality of care measures for vulnerable populations [press release]. Rockville, Md: Agency for Health Care Policy and Research; December 22, 1998. Available at: http://www.ahcpr.gov/news/press/vulnpr.htm

165. National Institutes of Health. PA-98-031: Methodology and measurement in the behavioral and social sciences. Bethesda, Md: National Institutes of Health; 1998.

166. **Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R.** Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). Psychiatry Res 1993;49:239–250.

167. **Deyo RA.** Pitfalls in measuring the health status of Mexican Americans: Comparative validity of the English and Spanish Sickness Impact Profile. Am J Public Health 1984;74:569–573.

168. **Bullinger M, Power MJ, Aaronson NK, Cella DF, Anderson RT.** Creating and evaluating cross-cultural instruments. In: Spilker B, ed. Quality of life and pharmacoeconomics in clinical trials. Philadelphia, Pa: Lippincott-Raven Publishers; 1996:659–668.