Qual Life Res (2009) 18:889–900
DOI 10.1007/s11136-009-9506-y

# Readability estimates for commonly used health-related quality of life surveys

**Sylvia H. Paz · Honghu Liu · Marie N. Fongwa · Leo S. Morales · Ron D. Hays**

Accepted: 14 June 2009 / Published online: 10 July 2009


**Abstract**
*Purpose* To estimate readability of seven commonly used health-related quality of life instruments: SF-36, HUI, EQ-5D, QWB-SA, HALex, Minnesota Living with Heart Failure Questionnaire (MLHFQ), and the NEI-VFQ-25.
*Methods* The Flesch–Kincaid (F–K) and Flesch Reading Ease (FRE) formulae were used to estimate readability for every item in each measure.
*Results* The percentage of items that require more than 5 years of formal schooling according to F–K was 50 for the EQ-5D, 53 for the SF-36, 80 for the VFQ-25, 85 for the QWB-SA, 100 for the HUI, HALex, and the MLHFQ. The percentage of items deemed harder than "easy" according to FRE was 50 for the SF-36, 67 for the EQ-5D, 79 for the QWB-SA, 80 for the VFQ-25, 100 for the HUI, HALex, and the MLHFQ.
*Conclusions* All seven surveys have a substantial number of items with high readability levels that may not be appropriate for the general population.

**Keywords** Health-related quality of life · Survey research · Readability · Health literacy

S. H. Paz (✉)
Department of Health Services, UCLA School of Public Health, P.O. Box 951772, Los Angeles, CA 90095-1772, USA
e-mail: shpaz@ucla.edu

H. Liu · L. S. Morales · R. D. Hays
UCLA Department of Medicine, UCLA Division of General Internal Medicine and Health Services Research, 911 Broxton Plaza, Los Angeles, CA 90095-1736, USA

M. N. Fongwa
UCLA School of Nursing, 700 Tiverton Ave., 3-238 Factor Building, P.O. Box 956917, Los Angeles, CA 90095-6917, USA

## Introduction

A number of generic and disease-targeted health-related quality of life (HRQOL) instruments have been developed. Survey measurement of HRQOL assumes comprehension of the questions by respondents. Several studies have been conducted to evaluate national levels of literacy. For example, in the United Kingdom, a government report showed that 56% of a randomly selected sample of adults had literacy skills at the lowest level of ability [1]. In addition, Smith et al. report that 22% of the working population in the United Kingdom have a low level of literacy [2, 3]. Analysis from the 2003 National Assessment of Adult Literary Survey in the United States indicated that 44% of adults had basic or below basic literacy level [4]. This report also found that 36% of the adult US population had basic or below basic health literacy [4]. Gazmararian et al. [5] specifically examined functional health literacy in a US national sample of Medicare enrollees in a managed care organization and found that more than one-third of respondents had inadequate or marginal health literacy.

Low literacy is associated with lower socioeconomic levels and poor health. In the United States, it disproportionally affects ethnic minorities including those immigrants who often arrive with low levels of education, socioeconomic status, English proficiency, and discrepant cultural models with regard to disease and disease prevention compared to US models [6]. Discrepancies between the readability of health information and the literacy skills of patients have been extensively reported since the onset of health-related readability evaluation in the 1980s [2, 7–17].

Studies that have evaluated patient literacy have found that patient educational level is not always consistent with their literacy level. Davis et al. [18] reported that among

adult patients with a fifth to tenth grade education, 60% were reading at least three grades below their grade level. Similar results have been reported in other studies which report up to six grade reading levels below the highest grade completed [19].

US norms recommend that surveys do not include items that require more than 8 or 9 years of formal schooling for the general population; and more than 5 years of formal schooling for vulnerable populations [12, 13]. Likewise, in the United Kingdom, it is recommended that health literature is written so that no more than 5 years of education are needed to completely understand the passage [17]. Therefore, it seems appropriate to suggest that health materials be written assuming a maximum of 5 years of formal education to assure comprehension by the widest population possible [16, 17, 20]. Items that are not easily understood will have higher rates of non-response and the data may become unreliable due to items being incomprehensible to subjects with low literacy levels.

Reading ease evaluation has become increasingly important since research has shown that comprehension is higher when texts are easily read. The concept of readability refers to the ease of a piece of text to be read and understood. Most health-related readability studies have focused on educational materials, consent forms, and more recently some internet-based health information studies have also been done [16, 21, 22]. By contrast, relatively few studies have been conducted to evaluate the readability of health surveys. Furthermore, only a few of these studies evaluated readability of each item separately [16, 21, 22]. This is important since computerized methods calculate a weighted average of text readability, when the instrument is evaluated as a whole, and this average readability score only reflects the mean level of the readability of the whole instrument. But in a survey, the average readability score of the whole instrument tells only a part of the story because the subject needs to have an adequate literacy level to understand each item independently. In addition, mean readability scores are insufficient parameters to describe the real reading level that participants face in a survey as the variation of item reading levels may be high, and therefore the full range of scores would not be captured. Thus, before collecting survey data, assessing readability scores at the item level is an important contribution to the literature that will help close the gap between survey research and what is truly understood by the general population.

The Health Measurement Research Group conducted a multisite study to evaluate extensively used HRQOL instruments [23, 24]. Five of these are generic instruments: the Short-Form Health Survey-36 item (SF-36v2), Health Utilities Index (HUI), European Quality of Life-5-Dimensional (EQ-5D), Quality of Well-Being Scale-Self-Administered (QWB-SA), and the Health and Activities Limitations Index (HALex). In addition, two disease-targeted instruments were included to learn how health assessments function differently in subjects with specific conditions: The Minnesota Living with Heart Failure Questionnaire (MLHFQ) and the National Eye Institute Visual Functioning Questionnaire-25 item (VFQ-25). These latter instruments were selected because they focus on patients for whom the study data were collected, those with heart disease and cataracts. In addition, these disease-targeted instruments are considered to be legacy measures for these conditions [25]. The generic instruments used are among the most widely used measures. The purpose of this article is to assess the readability of these seven commonly used HRQOL instruments at the item level.

## Methods

### Instruments to be evaluated

A brief description of each of the instruments evaluated in this study is included in the following paragraphs. Additionally, the number of items in each domain for each survey is shown in parenthesis.

#### *Generic profile measures*

1. Short Forms-36 (SF-36)

The SF-36 was developed at the Research ANd Development (RAND) Corporation in the 1980s as part of the Medical Outcomes Study. The SF-36v2, a newer version of the SF-36 with improved instructions, item wording, response choices, and increased scoring range, was used in this study [26, 27]. This instrument is composed of eight scales: Physical functioning (10), role limitations due to physical problems (4), bodily pain (2), general health perceptions (5), energy/vitality (4), social functioning (2), role limitations due to emotional problems (3), and mental health/emotional distress (5). In addition, the instrument also includes another item that measures change in perceived health (1).

2. Health Utilities Index (HUI)

Developed at the McMaster University in Canada, the HUI measure contains a health status classification system and a preference-based scoring formula. The HUI combines two systems that have been developed, HUI2 and HUI3. The HUI2 has seven dimensions: sensation (vision, hearing, and speech) (6), mobility (2), emotion (1), cognition or mental health (2), self-care (1), pain (2), and fertility, which was not included in this study. The HUI3

measures eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. A utility function translates categorical health status measures from HUI2 and HUI3 into interval-scale utility scores and HRQOL utility scores [28, 29].

3. European Quality of Life-5-Dimensional (EQ-5D)

With the goal of having a standardized generic instrument that could be used across Europe to measure HRQOL, the European Quality of Life group developed the EQ-5D. Providing a single index value for general health status, the EQ-5D preference-based measure is composed of five dimensions: mobility (1), self-care (1), usual activities (1), pain/discomfort (1), and anxiety/depression (1). Each dimension is measured using a single item with three response choices: no problems, some or moderate problems, and extreme problems. The single value index is obtained by the combination of responses from each of the five dimensions [30].

4. Quality of Well-Being Scale-Self-Administered (QWB-SA)

The QWB-SA was developed by the Health Outcomes Assessment Program at the University of California, San Diego, to improve the original QWB, which was lengthy, and difficult to administer [31, 32]. Retaining the psychometric properties of the original version, the QWB-SA has been used extensively to evaluate patients with diverse health conditions. With a total of 75 items, the QWB-SA comprised five sections: presence or absence of 18 chronic and 25 acute physical symptoms (43), mental health symptoms (14), self-care (2), mobility (3), physical activity (9), social activity/role expectations (3), and general health (1). The QWB-SA reflects a societal perspective on the value of the subject's functioning and well-being. It combines preference-weighted values (the value that society places on different health states) for symptoms and functioning. Items directly ask about different symptoms or conditions [33].

5. Health and Activities Limitations Index (HALex)

Even though the original HALex was developed by the National Center for Health Statistics in the 1980s and 1990s for use in the National Health Interview Survey, the version used in this study was adapted to the one used by the Behavioral Risk Factor Surveillance Survey from the US Center for Disease Control and Prevention. The HALex assesses HRQOL based on the person's perceived health status as well as activity limitation. This instrument focuses on evaluating how health problems affect a person's daily activities. The seven items used in the HALex are part of the National Health Interview Survey, and correspond to the physical role limitations domain. With few questions, this instrument is easy to complete. The HALex single score index reflects the total impact of a specific health state on a person's overall HRQOL [34–36].

*Disease-targeted measures*

1. Minnesota Living with Heart Failure Questionnaire (MLHFQ)

Specifically designed to assess the effects of heart failure and its treatments on quality of life, the 21-item MLHFQ was developed in 1984 by Rector et al. [37, 38]. Items are representative of the key dimensions of quality of life that are affected by heart failure; physical (8), emotional (5), and heart-specific overall quality of life (8). Using a 6-point categorical response scale to ask the subject how much his/her life is affected by each dimension, the questionnaire produces a global score along with scores for each one of the previously mentioned dimensions [37–39].

2. National Eye Institute Visual Functioning Questionnaire-25 item (VFQ-25)

The VFQ-25 was developed with funding by the US National Eye Institute to measure self-reported, vision-targeted health status. Hence, the VFQ reflects the influence of visual disabilities and visual symptoms on generic health domains as well as task-oriented domains that are related to visual functioning. A 25-item vision-targeted measure of HRQOL, the VFQ-25, produces a single overall visual function score that rates the subject's perceived visual functioning. The 12 subscales include: General Health (1), General Vision (1), Near Vision Activities (3), Distance Vision Activities (3), Ocular Pain (2), Vision-Specific Social Function (2), Vision-Specific Role Difficulties (2), Vision-Specific Mental Health (4), Vision-Specific Dependency (3), Driving Difficulties (2), Color Vision (1), and Peripheral Vision (1). The NEI-VFQ-25 is scored using standard algorithms [40, 41].

## Readability assessment

There are a number of manual and computerized formulae that can be used to evaluate the readability of written text. These formulae are based on the number of syllables per word and the number of words per sentence; two components that have been found to be good predictors of readability [42]. These formulae provide an estimate of the reading level necessary to read and comprehend given text. Two commonly used formulae are the Flesch–Kincaid (F–K) and the Flesch Reading Ease (FRE) [16, 43]. Even though both methods are based on measuring word length and sentence length, their results are different because they

use different weighting factors. The F–K method produces a corresponding grade level, which is needed to read the material. Scores generated by the F–K method are highly correlated with the scores calculated by other formulae [16]. A previous limitation of this method, a 12th grade ceiling, has been resolved leaving no ceiling value for the readability calculation using this method. The FRE method rates text based on a 100-point scale so that 100 represents the easiest text and 0 the hardest. Table 1 shows the ratings that accompany the scores used for each of the two methods [16]. The formulae used to calculate the FRE and F–K scores are as follows:

$$\text{FRE score} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}), \text{ and}$$

$$\text{F - K reading grade level score} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59$$

where ASL is the average sentence length (number of words divided by number of sentences) and ASW is the average number of syllables per word (number of syllables divided by number of words).

The readability of all items was estimated using Microsoft Word. Computerized calculations are advantageous as they decrease the possibility of human error. However, they are challenging when calculating survey items since many of these have fragmented formatting and thus do not conform to the necessary structure of complete sentences or questions. Many items have a common stem which is followed by multiple questions. To overcome these obstacles, we completed each item with the corresponding stem. In general, each question that called for a response from the subject was completed if necessary.

Some questions have several response choices that are usually included as part of the question, or as a second phrase in the same question. This latter situation usually occurs when the survey is administered by an interviewer. This makes each item longer and usually presents a higher readability score. As an example of this situation, items in the VFQ were scored using both methods. First each question alone was scored and then a second score was obtained for the question including response choices. Both scores are graphically presented in Fig. 8.

Finally, overall survey readability scores were also calculated using the F–K and FRE readability formulae to contrast the difference between rating a survey as a whole and rating a survey at the individual item level.

**Table 1** Ratings of Flesch reading ease and Flesch–Kincaid grade level scores [16]

| Reading difficulty | Flesch reading ease score | Flesch–Kincaid grade level score |
|---|---|---|
| Very easy | 90–100 | 5th |
| Easy | 80–90 | 6th |
| Fairly easy | 70–80 | 7th |
| Standard | 60–70 | 8th–9th |
| Fairly difficult | 50–60 | 10th–12th |
| Difficult | 30–50 | 13th–16th |
| Very difficult | 0–30 | ≥College graduate |

**Table 2** Mean, median, standard deviation, and range of item readability scores

| | Mean (95% CI) | SD | Median | Range |
|---|---|---|---|---|
| **HUI** | | | | |
| **F–K** | 9.6 (8.5,10.7) | 2.2 | 9.0 | 6.1–12.4 |
| **FRE** | 62.9 (58.2,67.6) | 9.2 | 60.5 | 45.2–79.5 |
| **SF-36** | | | | |
| **F–K** | 7.1 (5.4,8.8) | 5.3 | 4.5 | 0.6–16.2 |
| **FRE** | 74.6 (67.9,81.3) | 20.4 | 79.6 | 31.7–100.0 |
| **EQ-5D** | | | | |
| **F–K** | 6.7 (3.9,9.5) | 3.5 | 6.1 | 3.0–12.0 |
| **FRE** | 66.6 (53.8,79.4) | 16.0 | 69.9 | 45.1–84.6 |
| **QWB-SA** | | | | |
| **F–K** | 8.6 (8.5,9.6) | 4.3 | 8.7 | 0.5–21.5 |
| **FRE** | 66.2 (62.0,70.4) | 18.7 | 66.3 | 0.0–100.0 |
| **HALex** | | | | |
| **F–K** | 10.9 (8.1,13.7) | 3.8 | 9.9 | 7.5–18.2 |
| **FRE** | 55.0 (43.1,66.9) | 16.0 | 59.6 | 26.2–68.9 |
| **MLHFQ** | | | | |
| **F–K** | 8.1 (7.5,8.7) | 1.3 | 9.9 | 7.7–12.4 |
| **FRE** | 65.4 (62.4,68.4) | 7.1 | 66.1 | 52.0–76.7 |
| **VFQ-25** | | | | |
| **F–K** | 8.1 (7.1,9.1) | 2.6 | 8.3 | 3.7–13.9 |
| **FRE** | 67.9 (63.4,72.4) | 11.6 | 66.3 | 41.8–87.9 |

## Statistical analysis

Means with 95% confidence intervals, medians, standard deviations, and ranges of F–K and FRE readability scores across items for each survey were calculated and are presented in Table 2. The F–K scores are also presented graphically in Figs. 1, 2, 3, 4, 5, 6, 7, and 8. Microsoft Office Excel, version 2007, was used for all analyses.

## Results (data available upon request from S. Paz)

1. Short Forms-36 (SF-36)

The mean and median F–K grade level scores judged this instrument to be at a "fairly easy" and "very easy" level of readability (See Tables 1, 2). Nineteen items (53%) scored above the recommended 5 years of schooling (See Fig. 1). In addition, nine items (25%) fell in the categories of "fairly difficult," "difficult," or "very difficult," and eight

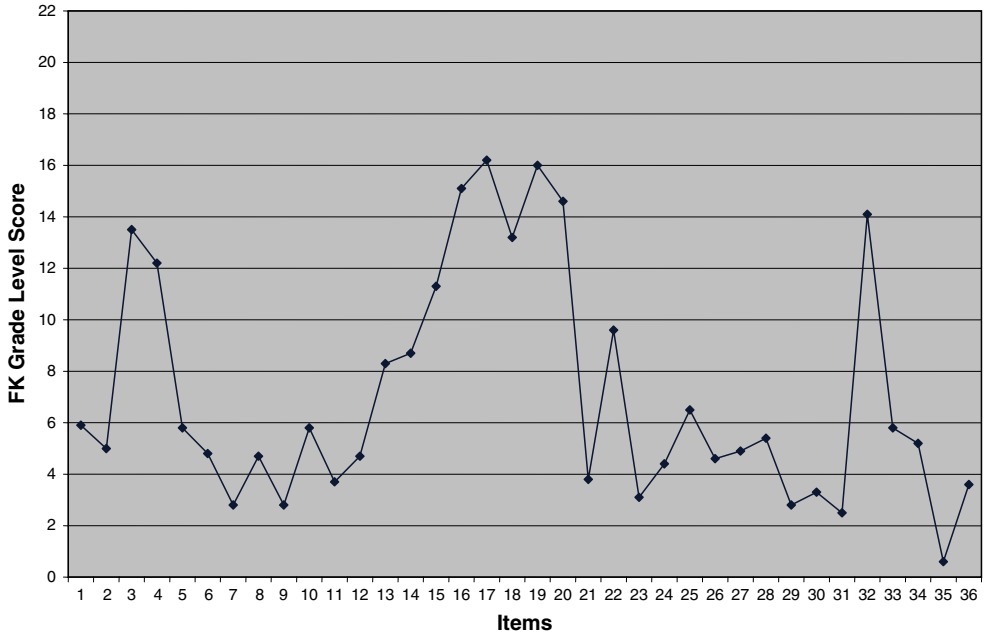



**Fig. 1** SF-36

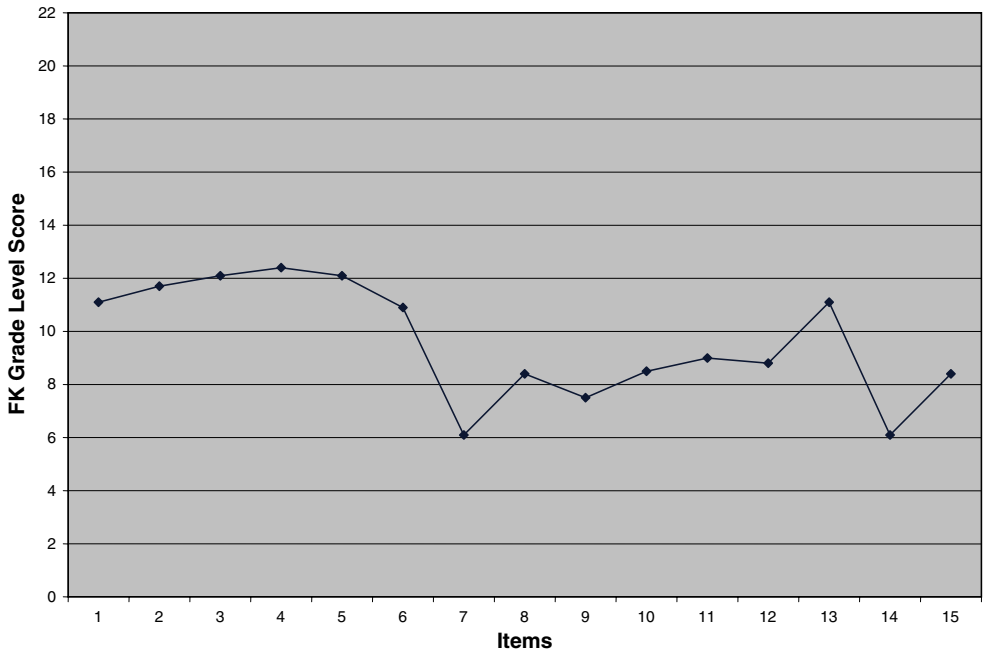



**Fig. 2** HUI

items (22%) require more than 12 years of formal schooling to be properly understood. The mean and median on the FRE readability index placed this survey at a "fairly easy" level of reading difficulty (see Tables 1, 2). Even though the mean and median values are "fairly easy," 18 items (50%) fell in the categories of "fairly easy," "standard," "fairly difficult," "difficult," or "very difficult" according to the FRE scoring method; i.e., 50% are harder than the recommended categories of "very easy" or "easy". Eight items (22%) scored "fairly difficult," "difficult," or "very difficult" according to both scoring methods. The readability scores for the SF-36 overall were 6.7 using the F–K grade level scoring and 70.3 using the FRE readability formula. These results set this survey at an "easy" and "fairly easy" level of readability, respectively, according to the classification presented in Table 1.

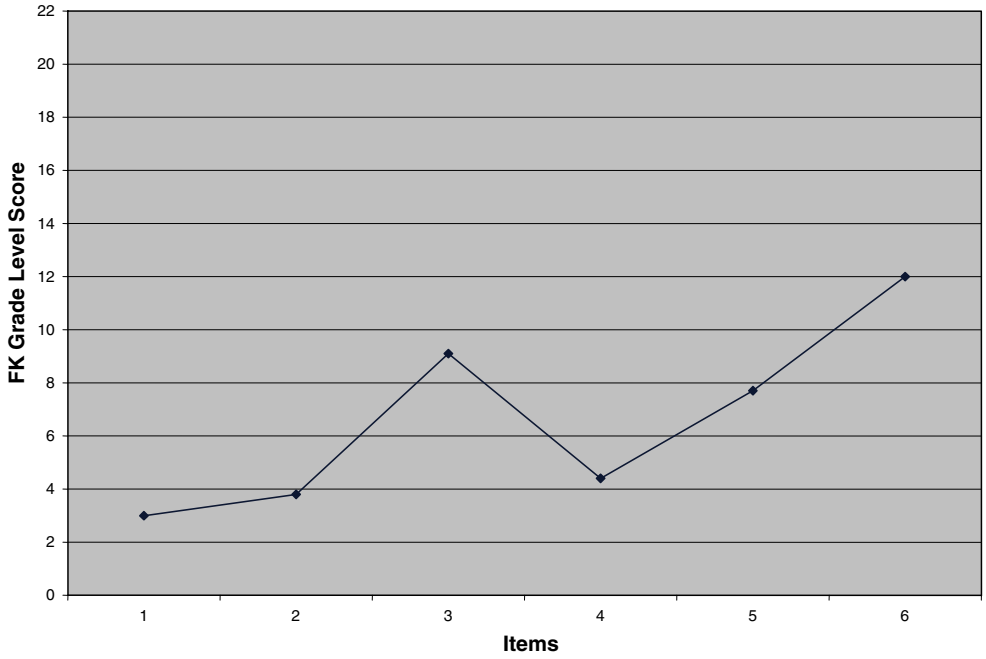



**Fig. 3** EQ-5D

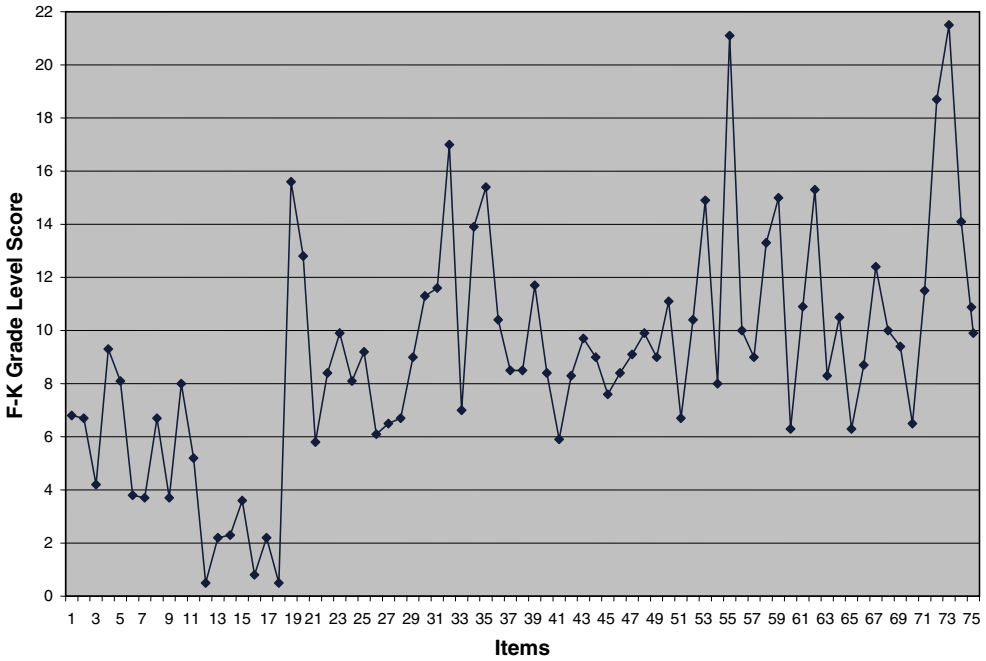



**Fig. 4** QWB-SA

2. Health Utilities Index (HUI)

The mean and median F–K grade level score for the HUI items were 9.6 and 9.0, respectively, setting this survey at a "standard" level of readability according to the classification presented in Tables 1 and 2. All 15 items (100%) scored above the recommended 5 years of formal schooling (see Fig. 2). Using the FRE readability formula, which does not depend on grade level score, the mean and median for this questionnaire's items also set the survey at "standard" level of readability on average (see Table 1, 2). Even though the mean and median values are at the "standard" level of readability, 100% of items (15/15) fell in the categories that are harder than "very easy" or "easy," and 40% of the items (6/15) fell in the categories of "fairly difficult," "difficult,"

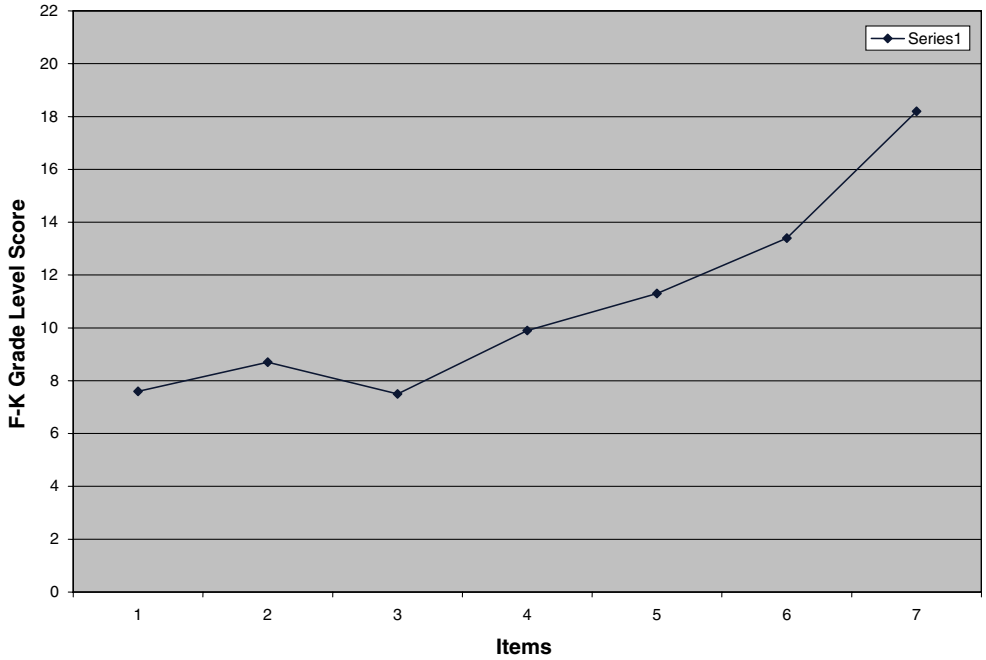



**Fig. 5** HALex

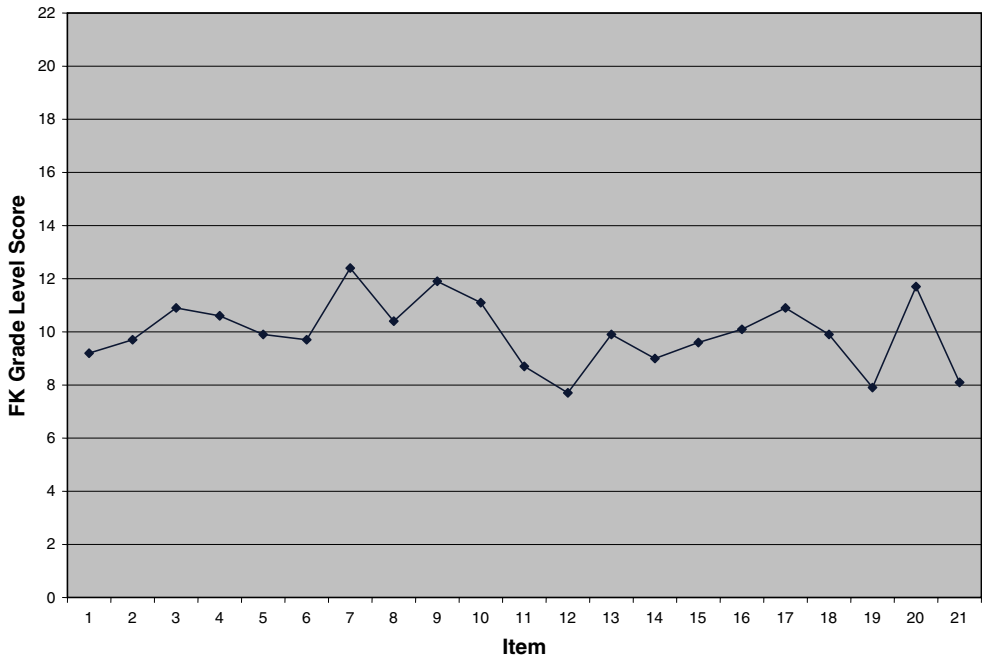



**Fig. 6** Living with Heart Failure Questionnaire

or "very difficult" according to both scoring methods (see Table 1). When calculated as a whole instrument, the overall readability scores for the HUI were 7.1 using the F–K grade level scoring and 65.7 using the FRE readability formula. These results would set this survey at a "fairly easy" and "standard" level of readability, respectively, according to the classification presented in Table 1.

3. European Quality of Life-5-Dimensional (EQ-5D)

The mean and median F–K grade level score for the EQ-5D items set this survey at the "easy" level of readability according to the classification given in Table 1. The standard deviation was 3.5 and the range of scores went from 3.0 to 12.0 (VAS item) (see Table 2). Three items (50%)

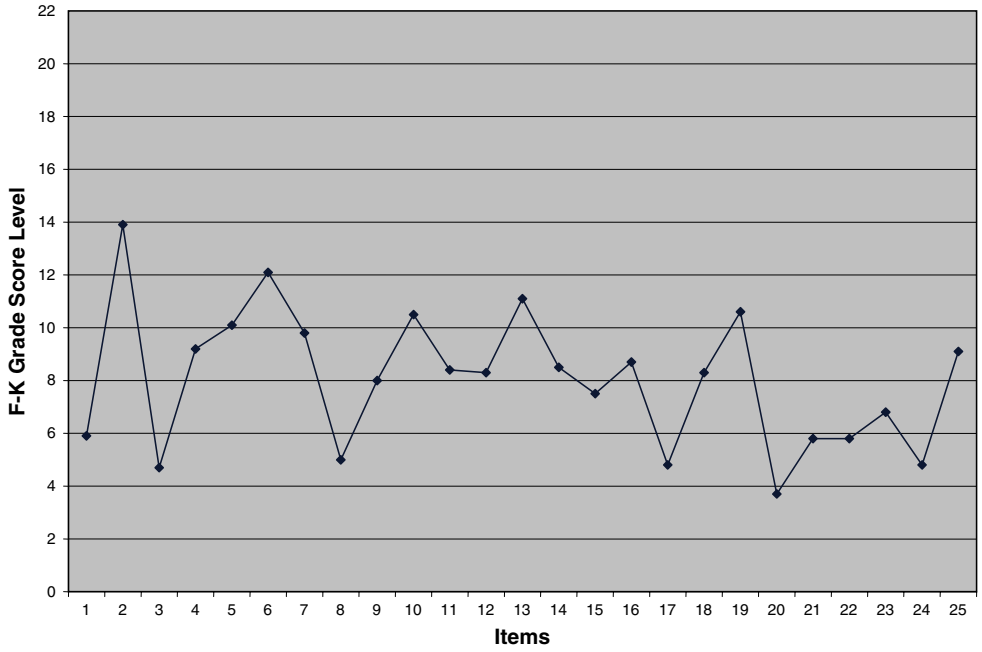
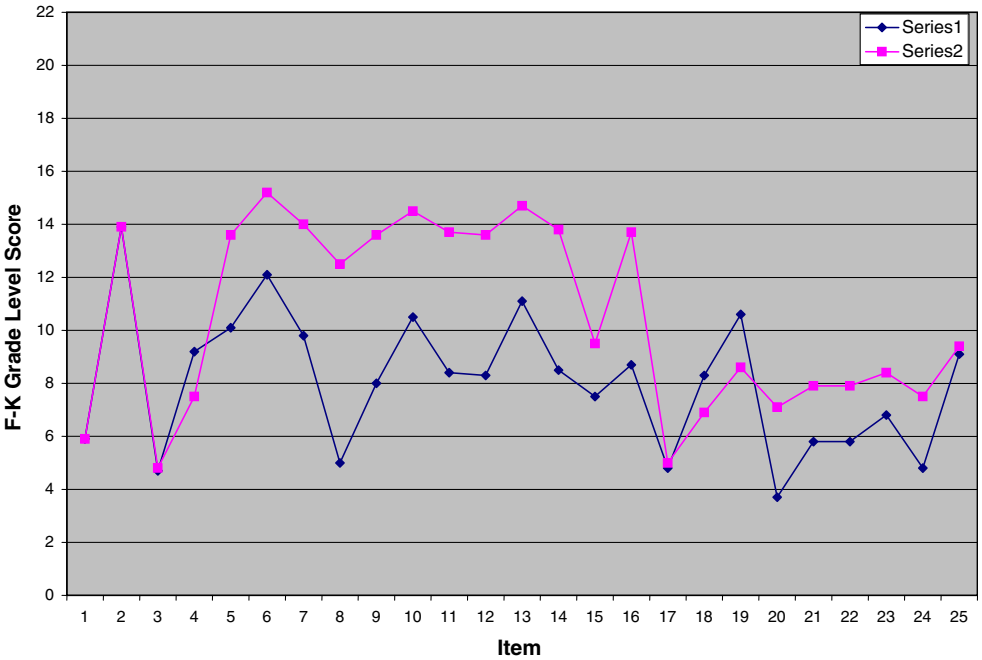



**Fig. 7** VFQ-25




* Item alone versus item in addition to response choices, for each item.

Series 1: Item alone

Series 2: Item in addition to response choices.

**Fig. 8** VFQ-25*

scored above the recommended 5 years of schooling (See Fig. 3). Using the FRE readability formula, the mean and median for the EQ-5D placed this survey at a "standard" level according to Table 1. Even though the mean and median values are standard, only 33% (2/6 items) fall in the categories of "very easy" or "easy". The VAS item, with a score of 12.0, the highest in the F–K scale and a rating of "fairly difficult," has a "standard" rating in the FRE scale,

thus not affecting as much the mean score using this latter method. Item 3, the hardest item using the FRE is the only "difficult" item in this survey according to this method. The overall readability scores for the EQ-5D were 4.2 using the F–K grade level scoring and 78.4 using the FRE readability formula. These results set this survey at a "very easy" and "fairly easy" level of readability respectively (see Table 1).

4. Quality of Well-Being Scale-Self-Administered (QWB-SA)

The mean and median F–K grade level score set this survey at a "standard" level of readability (see Tables 1, 2). Only 11 items scored at the recommended 5 years of formal schooling. This means that 85% (64/75) of the items in this survey may not be appropriately understood by individuals with less education (see Fig. 4). Furthermore, 14 items in this survey scored above 12.0 using the F–K method meaning that a college level education or higher is needed to appropriately comprehend 19% (14/75) of this survey. The FRE readability mean and median estimates for the QWB-SA placed this survey at a "standard" level according to Tables 1 and 2. The standard deviation was 18.7 and the range went from 0.0 to 100.0. Even though the mean and median values are "standard," only 21% (16/75 items) fell in the recommended categories of "very easy" or "easy" according to this method of evaluating readability. The overall readability scores for the QWB-SA were 3.1 using the F–K grade level scoring and 79.3 using the FRE readability formula, setting this survey at a "very easy" and "fairly easy" level of readability respectively (see Table 1).

5. Health and Activities Limitations Index (HALex)

The mean and median F–K grade level score for the HALex items set this survey at a "standard" and "fairly difficult" level of readability respectively (see Tables 1, 2). All seven items scored above the recommended 5 years of formal schooling meaning that 100% (7/7) of the items in this survey may not be appropriately understood by individuals with less education (see Fig. 5). Furthermore, one item (14%) requires completed 12 years of formal schooling and one item requires more than college level education to be properly understood. On the FRE readability formula, the mean and median placed this survey at a "fairly difficult" level according to Tables 1 and 2. As with the F–K formula, 100% (7/7 items) fell in the categories above the recommended "very easy" or "easy" categories using this scoring method. The overall readability scores for the HALex were 10.1 using the F–K grade level scoring and 55.4 using the FRE readability formula, setting this survey at a "fairly difficult" level of readability with both methods, according to the classification presented in Table 1.

6. Minnesota Living with Heart Failure Questionnaire (MLHFQ)

The mean and median F–K grade level score for the MLHFQ items set this survey at a "standard" level of readability when using the F–K scoring method (see Tables 1, 2). All 21 items (100%) scored above the recommended 5 years of formal schooling making this survey not appropriate for subjects with less education (see Fig. 6). The mean and median on the FRE readability estimates also placed this survey at a "standard" level according to Tables 1 and 2. Even though the mean and median values are standard, 100% (21/21 items) fell in the categories that are harder than the recommended "very easy" or "easy" using this scoring method. When calculated as a whole instrument, the overall readability scores for the MLHFQ were 5.5 using the F–K grade level scoring and 69.2 using the FRE readability formula. These results place this survey at a "very easy" and "standard" level of readability respectively (see Table 1).

7. National Eye Institute Visual Functioning Questionnaire-25 item (VFQ-25)

The mean and median F–K grade level score for the VFQ-25 questionnaire placed this survey at a "standard" level of readability (see Tables 1, 2). Twenty items (80%) scored above the recommended 5 years of schooling (see Fig. 7). Furthermore, two items require more than a High School level education to be properly understood. Using the FRE readability formula, the mean and median also placed this survey at a "standard" level according to Tables 1 and 2. Even though the mean and median values are standard, 80% (20/25) items did not fall in the recommended categories of "very easy" or "easy". The overall readability scores for the VFQ-25 were 8.9 using the F–K grade level scoring and 63.7 using the FRE readability formula. These results set this instrument at a "standard" level of readability using both calculation methods (see Table 1).

Figure 8 shows scores obtained by scoring the item alone, along with the item in addition to response choices, for each item of the VFQ-25. The graph shows that for two items the scores were identical, for 20 items including response choices had a higher score, and for three items not including response choices had a higher score. For these last three items, this occurred because adding a second sentence, which normally is longer but with a lower readability score, contributes with a higher weight to the average total item score. In these three items, the second sentence, which included the response choices was shorter than the first sentence, and therefore contributed less to the weighted average.

## Discussion

The results of this study reveal that current HRQOL measures may be inappropriate for general population surveys and in particular, they are inappropriate for populations with lower socio-economic status. Readability analysis for HRQOL surveys is important and furthermore analysis at the item level is essential. Mean scores for all of these widely used surveys required more than the recommended 5 years of formal schooling. Moreover, all surveys had a significant number of items with scores above the recommended threshold. These findings show that most readability studies, which report survey mean scores, are inadequate since a significant segment of the population will not have the literacy skills needed to comprehend and respond correctly to many items in the surveys. Furthermore, vulnerable populations will especially be affected with the administration of surveys, which are beyond their literary skills.

Ethnic minorities and underserved populations in the United States consistently show worse health outcomes, preventive screening rates, worse disease management, and lower survival rates [44]. Health literacy and limited reading skills are known to be important barriers to improving health outcomes. Meade et al. [44] reported on alarming low levels of literacy in the general population which happen to be disproportionately prevalent among vulnerable populations. There are multiple studies that report on health materials written at readability levels far above the recommended US national norms [20]. Although educational level is not always consistent with literacy level, before developing new measures of HRQOL, it behooves outcome researchers to consider the educational background of the target populations. A discrepancy between the readability level and the appropriate readability when including underserved populations was found in most surveys analyzed in this paper. In addition, data is at an even higher risk of poor quality when surveys are administered to populations who lack literacy levels necessary for full comprehension of items. This is exacerbated when immigrant populations who tend to have less education and English proficiency are included in the sample.

Readability formulae are useful in that they can assist with a quantifiable estimation of the reading ease of given text. However, they do not take into account other factors that are important in predicting survey comprehension. Content, layout, learning stimulation, and cultural appropriateness are some examples of additional factors that might influence the readability of surveys. Furthermore, they do not take into account complementarities of individual items which can also facilitate understanding when taken as a whole in a specific context. Other personal factors that have been studied and found to affect readability are previous experience, motivation, and interest. These formulae may underestimate the effect of new material with vocabulary not usually used by the general population.

Bailin and Grafstein [45] reported on a study documenting that reading ability is significantly determined by knowledge procedures involved in deriving significance from given text. An additional caveat of these formulae is that they rely solely on sentence and word length, and therefore score equally sentences with the same words but scrambled in a different order. Less useful in this context are other recommendations like design factors and other visuals that could accompany written text, and that have been found to increase readability. Even though most of these studies have been done on educational materials or web-based information, some extensively reported suggestions that might help with reading ease and that could be helpful when working with surveys are a font size of 12 or larger along with the use of black ink on white paper and the amount of white space in the page [46, 47].

An additional limitation of this study is that readability analyses were performed only in one language using methods used primarily for a US population. The use of other indices such as the SMOG (Simple Measure of Gobbledygook) index which estimates the years of education needed to appropriately understand a piece of text, and which is often used in the United Kingdom, would be an important contribution to the literature. In addition, future studies could estimate the readability in other languages. For example, it would be of interest to use the Fernandez Huerta formula to estimate the readability of the SF-36 in Spanish or the Kandel and Moles formula to estimate the readability in French.

Despite these limitations, readability formulae provide a fast and efficient measurement tool that is readily available in commonly used computer software. The use of these formulae when developing surveys could help investigators select simpler vocabulary and sentence structure. Both scoring algorithms used in this paper yield better results when using shorter and more commonly used words and shorter sentences. The methods used in this analysis may still be used as a helpful tool when developing new surveys and modifying existing ones focusing on reducing the discrepancy between survey readability and population skills. For example, in Part IV of the QWB-SA instrument, all nine items have the instruction "please fill in all days that apply" as part of the question. By removing this phrase and placing it at the beginning of the section as an instruction for all the following items, readability scores are reduced from 8.3 to 5.8 for item 1 and from 10.5 to 8.3 for item 2, using the F–K method.

An interesting finding of this study was the variation in the readability within surveys and between surveys. The

largest range within a survey was found in the QWB-SA with an item variation of 100.0 using the FRE algorithm and 21.0 using the F–K formula. The smallest range was seen in the MLHFQ with 24.7 using the FRE and 4.7 using the F–K algorithm. Both highest and lowest ranges were found in the same survey using both formulae. With regard to between surveys and considering the median value of each, the highest readability score with the FRE algorithm was seen in the HALex (59.6) and the lowest in the SF-36 (79.6). When using the F–K scoring algorithm, the highest median score was seen in both the HALex and MLHFQ, both 9.9, and the lowest was seen in the SF-36 (4.5). Not considering these extreme scores, the rest of the survey scores were all within the 60 s range using FRE algorithms, and showing more variability ranging in the 6th–9th grade level using F–K algorithm. Being a more stable statistic and less influenced by extreme values, the median was reported for this comparison.

No major differences were found between the generic and the disease-targeted instruments. Both disease-targeted instruments had means and medians above the recommended scores, as did most of the generic instruments. Of interest, both disease-targeted instruments had the same mean score using the F–K algorithm, but the median was lower in the VFQ-25. And as Fig. 7 confirms, this instrument has more items within the recommended range.

As seen in Fig. 8, most items have higher readability scores when including all response choices within the question. When surveys are administered by professional interviewers, most probably all response choices are read. The items may be longer literally, but the interviewer could be helpful in explaining items that are not clear to the subject, or emphasizing the item's important part; both options not being available with self-administered surveys. In addition, Krosnick and Alwin's study found that the order of response choices affects the response selected and differs when the questionnaire is self-administered versus interviewer-administered. While the likelihood of choosing the first response choices increased when the survey was self-administered, the likelihood of selecting the last choices increased with interviewer-administered surveys. Furthermore, the authors also concluded from their study that subjects with lower levels of education were more likely to be influenced by changes in the order of response choices [48].

The validity of data collected from self-reported outcome measures depends upon the subject's ability to comprehend each item in the survey. The gap between survey readability levels and necessary reading skills for comprehension must be reduced. Working along with educators and editors, researchers working with survey data need to become more conscious of the population's low literacy levels. If the goal of outcome measurement is ultimately to improve HRQOL, sensitivity to an ever changing population is necessary when using existing measures and when creating new methods of evaluation. Surveys that are multicultural, multilingual, and literacy sensitive to a demographically continuously changing population are warranted.



## References

1. "http://www.dcsf.gov.uk/research/data/uploadfiles/RR490.pdf" Retrieved Jan. 5, 2009.
2. Smith, H., Gooding, S., et al. (1998). Evaluation of readability and accuracy of information leaflets in general practice for patients with asthma. *BMJ, 317*, 264–265.
3. Carey, S., Low, S., et al. (1997). *Adult literacy in Britain*. London: Office for National Statistics.
4. "http://nces.ed.gov/naal/index.asp" Retrieved Jan. 5, 2009.
5. Gazmararian, J. A., Baker, D. W., et al. (1999). Health literacy among medicare enrollees in a managed care organization. *The Journal of the American Medical Association, 281*(6), 545–551.
6. Rudd, R. E. (2007). Health literacy skills of US adults. *American Journal of Health Behavior, 31*(Suppl 1), S8–S18.
7. Morrow, G. R. (1980). How readable are subject consent forms? *The Journal of the American Medical Association, 244*(1), 56–58.
8. Tarnowski, K. J., Allen, D. M., et al. (1990). Readability of pediatric biomedical research informed consent forms. *Pediatrics, 85*(1), 58–62.
9. Meade, C. D., & Howser, D. M. (1992). Consent forms: How to determine and improve their readability. *Oncology Nursing Forum, 19*(10), 1523–1528.
10. Ott, B. B., & Hardie, T. L. (1995). Readability of written materials: Implications for critical care nurses. *Dimensions of Critical Care Nursing, 14*(6), 328–334.
11. Ott, B. B., & Hardie, T. L. (1997). Readability of advance directive documents. *Image—The Journal of Nursing Scholarship, 29*(1), 53–57.
12. Calderon, J. L., & Beltran, R. A. (2004). Pitfalls in health communication: Healthcare policy, institution, structure, and process. *Medscape General Medicine, 6*(1), 9.
13. Calderon, J. L., Zadshir, A., et al. (2004). A survey of kidney disease and risk-factor information on the world wide web. *Medscape General Medicine, 6*(4), 3.
14. Hunter, J. L. (2005). Cervical cancer educational pamphlets: Do they miss the mark for Mexican immigrant women's needs? *Cancer Control, 12*(Suppl 2), 42–50.
15. Meade, C. D. (2005). Cancer, culture and literacy: Critical next steps in improving care for diverse populations. *Cancer Control, 12*(Suppl 2), 4–5.
16. Calderon, J. L., Morales, L. S., et al. (2006). Variation in the readability of items within surveys. *American Journal of Medical Quality, 21*(1), 49–56.
17. Estey, A., Musseau, A., et al. (1991). Comprehension levels of patients reading health information. *Patient Education Counseling, 18*, 165–169. (1996), *30*, 205–208.
18. Davis, T. C., Crouch, M. A., et al. (1990). The gap between patient reading comprehension and the readability of patient education materials. *Journal of Family Practice, 31*(5), 533–538.

19. Miller, B., & Bodie, M. (1994). Determination of reading comprehension level for effective patient health-education materials. *Nursing Research, 43*(2), 118–119.
20. Freda, M. C., Damus, K., et al. (1999). Evaluation of the readability of ACOG patient education pamphlets. The American College of Obstetricians and Gynecologists. *Obstetrics and Gynecology, 93*(5 Pt 1), 771–774.
21. Rogers, E. S., Spalding, S. L., et al. (2008). Are patient-administered attention deficit hyperactivity disorder scales suitable for adults? *Journal of Attention Disorders*. doi:10.1177/1087054708323017
22. Wallace, L. S., Keenum, A. J., et al. (2007). Readability and cognitive complexity of self-administered opioid assessment screening tools. *Journal of Opioid Management, 3*(6), 338–344.
23. Fryback, D. G., Dunham, N. C., et al. (2007). US Norms for six generic health-related quality-of-life indexes from the National Health Measurement Study. *Medical Care, 45*, 1162–1170.
24. "http://healthmeasurement.org/" Retrieved Aug. 17, 2008.
25. Hays, R. D., Kim S., et al. (2009). Effects of mode and order of administration on generic health-related quality of life scores. 2009, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). 1098-3015/09/. Value in health.
26. Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.
27. Hays, R. D., Sherbourne, C. D., et al. (1993). The RAND 36-Item health survey 1.0. *Health Economics, 2*(3), 217–227.
28. Torrance, G. W., Feeny, D. H., et al. (1996). Multiattribute utility function for a comprehensive health status classification system. Health utilities index mark 2. *Medical Care, 34*(7), 702–722.
29. Horsman, J., Furlong, W., et al. (2003). The health utilities index (HUI): Concepts, measurement properties and applications. *Health and Quality of Life Outcomes, 1*, 54.
30. "http://www.euroqol.org." Retrieved April 1st, 2008.
31. Andresen, E. M., Rothenberg, B. M., et al. (1998). Performance of a self-administered mailed version of the quality of well-being (QWB-SA) questionnaire among older adults. *Medical Care, 36*(9), 1349–1360.
32. Kaplan, R. M., Ganiats, T. G., et al. (1998). The quality of well-being scale: Critical similarities and differences with SF-36. *International Journal for Quality in Health Care, 10*(6), 509–520.
33. Sieber, W., Groessl, E., et al. (2004). *Quality of well-being self-administered (QWB-SA) Scale. User's manual. Health outcomes assessment program.* San Diego: University of California.
34. Erickson, P., Wilson, R., et al. (1995). Years of healthy life. *Healthy People 2000 Statistical Notes, 7*, 1–15.
35. Livingston, E. H., & Ko, C. Y. (2002). Use of the health and activities limitation index as a measure of quality of life in obesity. *Obesity Research, 10*(8), 824–832.
36. Asada, Y. (2005). Assessment of the health of Americans: The average health-related quality of life and its inequality across individuals and groups. *Population Health Metrics, 3*, 7.
37. Rector, T. S., & Cohn, J. N. (1992). Assessment of patient outcome with the Minnesota living with heart failure questionnaire: Reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Pimobendan Multicenter Research Group. *American Heart Journal, 124*(4), 1017–1025.
38. Rector, T. S., Kubo, S. H., et al. (1993). Validity of the Minnesota living with heart failure questionnaire as a measure of therapeutic response to enalapril or placebo. *American Journal of Cardiology, 71*(12), 1106–1107.
39. Rector, T. S., Anand, I. S., et al. (2006). Relationships between clinical assessments and patients' perceptions of the effects of heart failure on their quality of life. *Journal of Cardiac Failure, 12*(2), 87–92.
40. Mangione, C. M., Lee, P. P., et al. (1998). Psychometric properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). NEI-VFQ Field Test Investigators. *Archives of Ophthalmology, 116*(11), 1496–1504.
41. Mangione, C. M., Lee, P. P., et al. (2001). Development of the 25-item National Eye Institute Visual Function Questionnaire. *Archives of Ophthalmology, 119*(7), 1050–1058.
42. Meade, C., & Smith, C. (1991). Readability formulas: Cautions and criteria. *Patient education and counseling, 17*, 153–158.
43. Friedman, D. B., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education and Behavior, 33*(3), 352–373.
44. Meade, C. D., Byrd, J. C., et al. (1989). Improving patient comprehension of literature on smoking. *American Journal of Public Health, 79*(10), 1411–1412.
45. Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language and Communication, 21*, 285–301.
46. Bradley, B., Singleton, M., et al. (1994). Readability of patient information leaflets on over-the-counter (OTC) medicines. *Journal of Clinical Pharmacy and Therapeutics, 19*(1), 7–15.
47. Doak, L. G., Doak, C. C., et al. (1996). Strategies to improve cancer education materials. *Oncology Nursing Forum, 23*(8), 1305–1312.
48. Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201–219.