World Journal of Preventive Medicine, 2015, Vol. 3, No. 2, 28-39
Available online at http://pubs.sciepub.com/jpm/3/2/3
© Science and Education Publishing
DOI:10.12691/jpm-3-2-3

Systematic Review of Health-Related Quality of Life Assessments in Physical Activity Research

Peter D. Hart[1,*], Minsoo Kang[2], Norman L. Weatherby[3], Yun Soo Lee[3], Tom M. Brinthaupt[4]

[1]Health Promotion Program, Montana State University - Northern, & Health Demographics, Havre, MT 59501,USA
[2]Kinesmetrics Laboratory, Middle Tennessee State University, Murfreesboro, TN 37132, USA
[3]Department of Health and Human Performance, Middle Tennessee State University, Murfreesboro, TN 37132, USA
[4]Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132, USA
*Corresponding author: peter.hart@msun.edu

Received April 21, 2015; Revised April 27, 2015; Accepted April 29, 2015

Abstract In physical activity research, health-related quality of life (HRQOL) is an outcome variable of growing importance. Physical activity is directly associated with HRQOL and intervention-type studies seek to show improvements in HRQOL based on treatment effects. As interest grows in using HRQOL as an outcome measure in physical activity research, the need to investigate the measurement properties of HRQOL assessments increases in importance. The objective of this study was to systematically explore HRQOL assessments used in physical activity research by examining their instrument characteristics (items, dimensions, scoring, etc.) and their published psychometric properties. Results of this study showed that 10 HRQOL assessments are currently used in physical activity research. Recommendations were made relative to different study designs and goals.

Keywords: *health-related quality of life, systematic review, physical activity, measurement, item-response theory*

Cite This Article: Peter D. Hart, Minsoo Kang, Norman L. Weatherby, Yun Soo Lee, and Tom M. Brinthaupt, "Systematic Review of Health-Related Quality of Life Assessments in Physical Activity Research." *World Journal of Preventive Medicine*, vol. 3, no. 2 (2015): 28-39. doi: 10.12691/jpm-3-2-3.

1. Introduction

Physical activity is a strong predictor of health status and should be adopted by individuals of all ages [1]. Many studies have shown the positive effects of regular physical activity on specific health outcomes. Such physical activity-related health outcomes have included all-cause mortality [2,3], cause-specific mortality [2,4], premature chronic disease [5,6,7], obesity [8,9], and mental health [10,11,12].

Health-related quality of life (HRQOL) is a more personal health outcome that has seen a growing interest in physical activity research. HRQOL is a broad latent construct that includes both subjective and objective indicators of people's lives that affect their physical and/or mental health status [13]. Due to its ability to capture overall perceived health, HRQOL has become a standard outcome measure in public health and medical research [13]. In addition, because HRQOL usually includes a component of perceived functional status, it has been considered a measure as important to research outcomes as other more objective indicators [14]. In addition, HRQOL has been shown to be a strong predictor of physician visits, hospitalization, and mortality [15].

Participating in a physically active lifestyle is linked to greater HRQOL [16]. Specifically, meeting recommended levels of physical activity has shown to be related to superior levels of HRQOL [17]. Physical activity has been used as a predictor of HRQOL in both prospective [18,19,20] and cross-sectional [16,21] observational studies. Physical activity has also seen major impacts in clinical interventions with links to positive changes in HRQOL [22,23]. In fact, increasing HRQOL has been described as the most important goal in physical activity interventions [24].

Despite the empirical evidence confirming the positive effects of physical activity on health, the majority of Americans remain physically inactive [1]. The Healthy People 2020 publication states that more than 80% of adults do not meet the current guidelines for physical activity and has reported several national objectives aimed at increasing the percentage of adults who engage in physical activity [25]. With physical activity interventions serving as the primary tool for achieving a more physically active population, the need to assess HRQOL will be in even greater demand. Furthermore, with the overwhelming interest in HRQOL as an outcome measure in physical activity research, there is a strong need for a better understanding of the measurement properties of HRQOL assessments commonly used in physical activity research.

There are currently no studies that review the scale characteristics, score determination, feasibility issues, and psychometric properties of the common HRQOL instruments used in physical activity research. Therefore, the purpose of this study was to review the most common instruments used to measure generic HRQOL in physical activity research in adults by summarizing the characteristics and scoring options of each instrument as well as the psychometric properties of each HRQOL scale. The overall usefulness of each HRQOL instrument relative to different study designs will be discussed. The study will serve as a HRQOL resource guide for those conducting physical activity research.

2. Materials and Methods

2.1. Search Strategy

PubMed.gov was systematically searched for published articles of physical activity research containing measures of HRQOL. The following search terms were used: ("physical activity" OR exercise) AND ("health-related quality of life" OR "quality of life"). After pertinent articles were identified, their reference lists were searched for more relevant studies. After all HRQOL instruments used in physical activity research were identified, the assessments were each investigated to determine if they were appropriate for inclusion.

2.2. Inclusion and Exclusion Criteria

An article was included in the study if it 1) was published in English, 2) was available in full text, 3) had a primary objective of evaluating the effects of physical activity on HRQOL as an outcome measure, 4) used a measure of HRQOL assessed via a questionnaire, 5) used adults as participants, and 6) was published on or after January 1, 2000. An HRQOL assessment was excluded from this study if it 1) was not specifically health-related in nature (e.g., life satisfaction), 2) measured a construct other than generic HRQOL (e.g., living with heart failure), 3) did not consist of a set of items measuring the HRQOL construct (e.g., single item or proxy variable), or 4) completely lacked empirical measurement evidence (e.g., researcher developed questions).

2.3. Instrument Characteristics and Properties

For each identified HRQOL assessment tool included in the study, the following characteristics were retrieved: 1) mode of administration, 2) number of items contained in the assessment tool, 3) type of rating scale(s) used, 4) number and types of domains and sub-dimensions, 5) alternate forms, 6) target populations, 7) adopted languages, and 8) scoring methods. The psychometric properties retrieved from each HRQOL assessment were categorized into three domains: validity, reliability, and item response theory.

The validity properties included in this study were: 1) content validity, (2) criterion validity, 3) construct validity, and 4) responsiveness. *Content validity* is the extent to which an assessment tool measures the construct of interest [26]. Appropriate scale construction should include content validity methods such as literature review, expert panel advice and/or ratings, and theme saturation [27]. *Criterion validity* is the extent to which measurements from an assessment tool adequately reflect an agreed upon gold standard measurement. HRQOL may have no known gold standard and therefore criterion validity may have limited to no impact on this study. *Construct validity* is the ability of an assessment tool to measure the trait or construct that it was intended to measure [26]. Types of construct validity evidence include known group difference testing, assessment of uni- or multi-dimensionality of scales, and correlation with other measures of hypothesized direction (i.e., convergent or divergent validity). *Responsiveness* is the ability of an assessment tool to detect clinically important changes in the construct of interest [28]. Measures of responsiveness should include effect size measures or statistics from receiver operating characteristic (ROC) curves.

The reliability properties included in this study were: 1) internal consistency reliability and 2) test-retest. *Internal consistency* refers to the extent to which items in an assessment tool are inter-correlated [29]. If such an inter-correlation exists, the items of the scale are said to measure a unidimensional construct. *Test-retest* reliability measures the stability of measurements over repeated trials [26]. Measures of test-retest include limits of agreement, Pearson correlations, and intra-class correlations (ICCs).

Item response theory methods to be reviewed include: 1) item analysis, 2) model data fit, 3) rating scale assessment, 4) scoring, 5) test equating, and 6) differential item functioning (DIF). Item response theory stems from modern psychometric theory and incorporates various scale item and person ability parameters into a statistical probability model [30]. That is, the probability of a person's response to an item is a function of the person's trait being measured (i.e., HRQOL) and the characteristics of the item (i.e., difficulty, discrimination, etc.). Through the use of item response theory models, HRQOL assessment tools can be evaluated based on their item's usefulness, the scale's unidimensionality, and the functioning of the chosen rating scale [30]. Item response theory can also be used to create an interval level measurement of the construct and equate scores from different assessment tools, as well as determine whether model bias exists across population subgroups [31].

3. Results

A total of 8,263 articles were found using the search terms. After reviewing titles and abstracts, 2,556 articles were identified as meeting inclusion criteria, of which, 1,209 articles were dropped due to exclusion criteria. A total of 1,347 articles were included in the final sample and were examined for their HRQOL assessment. Table 1 displays the characteristics of 10 HRQOL assessments arranged according to their frequency of use in physical activity research. The majority of physical activity studies used the Short Form Health Survey (SF-36) or one of its variants. The next most commonly used HRQOL assessment was the Sickness Impact Profile (SIP) followed by the Euroqol assessment (EQ-5D). Other HRQOL assessments identified (from most common to less common) were the Nottingham Health Profile (NHP), WHO Quality of Life (WHOQOL-BREF), Quality of Well-Being Scale (QWB), Health Utilities Index 3 (HUI3), CDC's Healthy Days Core (HRQOL-4), Assessment of Quality of Life (AQoL), and the Duke Health Profile (DHP).

Table 1. Characteristics of generic HRQOL assessments in adult physical activity research

Instrument	Mode	Items	Scale	Scoring	Dimensions	Forms	Languages
Short-Form Health Survey (SF-36)	Self-Administered Computer Interviewer Telephone	36	Categorical Rating 3 to 6-point	1) Summated Scoring 2) Norm-Based T-scoring 8-dimensions 2-domains	Vitality Physical functioning Bodily pain General health Physical role functioning Emotional role functioning Social role functioning Mental health	SF-36v1* SF-36v2* SF-12v1 SF-12v2 SF-8 VF-36	Multiple
Sickness Impact Profile (SIP)	Self-Administered Interview	136	Yes/No	1) Standardized Weighted 2) Overall 3) 12-dimensions 4) 2-domains	Sleep and rest Emotional behavior Body care and movement Home management Mobility Social interaction Ambulation Alertness behavior Communication Work Recreation and pastimes Eating	SIP-136* SIP-68 SIP-66 SIP-30 SIP-24 SIP-82	Multiple
Euroqol (EQ-5D)	Self-Administered Interview Telephone	6	Categorical Rating 3-point VAS 0-100	1) Descriptive Profile (11111 to 33333) 2) Health Index Score (-0.11 to 1) 3) Self-Reported Health Status (0 to 100)	Mobility Self-Care Usual Activities Pain/Discomfort Anxiety/Depression	EQ-5D-3L* EQ-5D-5L	Multiple
Nottingham Health Profile (NHP)	Self-Administered Interview	45	Yes/No	Scaled Weights (0 to 100)	Physical mobility Pain Social isolation Emotional reactions Energy Sleep	NHP*	Multiple
WHO Quality of Life Assessment (WHOQOL-BREF)	Self-Administered Interviewer	26	Categorical Scale 5-point	1) 4 Domain Scores 2) 2 Descriptive Items	Physical Health Psychological Social Relationships Environment	WHOQOL-BREF* WHOQOL-100	Multiple
Quality of Well-being Scale (QWB)	Self-Administered Computer Interviewer Telephone	76	Categorical Scale 2 to 5-point	1) 4 Domain Scores 2) Health Index Score (0 to 1)	Symptoms Mobility Physical Activity Social Activity	QWB QWB-SA*	Multiple
Health Utilities Index Mark 3 (HUI3)	Self-Administered Computer Interviewer Telephone	8	Categorical Scale 5 to 6-point	1) Descriptive Profile 2) Health Index Score (0 to 1)	Emotion Pain Vision Hearing Speech Ambulation Dexterity Cognition	HUI1 HUI2 HUI3*	Multiple
CDC Healthy Days (CDC HRQOL)	Interview Telephone	4	1 Categorical Rating Scale 3 Continuous Measures	1) Descriptive Score 2) Summary Index	Physical Mental	HRQOL-4* HRQOL-9 HRQOL-12	English
Assessment of QoL (AQoL)	Self-Administered Interview Mail Telephone	15	Categorical Rating 4-point	1) Overall Score 2) 5-Dimension Scores 3) Utility Score (1 to 0)	Illness Independent Living Social Relationships Physical Senses Psychological Wellbeing	AQoL I AQoL II* AQoL-8	English
Duke Health Profile (DHP)	Self-Administered	17	Categorical Rating 3-point	1) 10 Dimension Scores 2) 1 Summary General Health Score	Physical Mental Social General Perceived Health Self-Esteem Anxiety Depression Pain Disability	DHP* DUHP	English

Note. * indicates the common form used in physical activity research.

3.1. Short-Form Health Survey (SF-36)

Characteristics. The SF-36 is the most widely used HRQOL instrument in physical activity research. The appeal of the SF-36 is that it is a relatively efficient scale with numerous published sources detailing its psychometric properties. The SF-36 was developed from the Medical Outcomes Study (MOS) conducted by RAND [32]. The SF-36 is a multi-dimensional scale consisting of 36 items, 8 health-related dimensions, and two domains. The dimensions include: 1) vitality, 2) physical functioning, 3) bodily pain, 4) general health, 5) physical role functioning, 6) emotional role functioning, 7) social role functioning, and 8) mental health. The physical domain consists of the physical functioning, bodily pain, general health, and physical role functioning dimensions and the mental domain consists of the vitality, emotional role functioning, social role functioning, and mental health dimensions [33]. The latest version (v2) of the SF-36 has 3 different rating scale categories, ranging from 3-point to 6-point.

The SF-36 is intended to measure HRQOL in adults and can be self-administered, administered via computer, with aid of an interviewer, or by telephone. The instrument can be modified to include either a (standard) 4-week recall or a 1-week recall and has been incorporated into both observational and intervention-type studies. The SF-36, due to advances in measurement theory, has made several transformations, and is now referred to as the 2nd version (SF-36v2). The newer version made changes to item wording, item layout, and number of response categories to certain items [33]. Three other alternate forms of the SF-36 are available. The SF-12, SF-12v2, and SF-8 are shorter forms of the original that, however, maintain the measurement of all 8 dimensions as well as the two domain-specific summary scores [34].

The scoring of the SF-36 is relatively simple, relying on the assumption that item scores are linearly related to the underlying construct with the scales summated according to the Likert approach [32]. The updated version of SF-36 (SF-36v2) allows scores to be normalized to allow for easy comparisons [33]. The normalizing process used national data to allow for standardization of summated scores, followed by T-score conversion.

Psychometric properties. The SF-36 was constructed from a pool of items retrieved from existing instruments used for measuring physical limitations, role functioning, mental health, and perceived general health [32]. The larger pool of 245-items was part of the Medical Outcomes Study (MOS), of which the 36-items of the SF-36 were a subset. Participants in the MOS who completed the lengthy survey and took the follow-up health examination (within 1-month) were used for the validity study. The health examination was required to allow for clinical diagnoses for the construction of contrasting groups (clinical tests of validity).

Two types of criteria were used for the initial validation of the SF-36. Psychometric criteria were considered by the use of principal components analysis (construct validity) and inspection of correlations among the eight scales. Clinical criteria were considered by comparing the specific scale scores between four distinct groups of subjects: 1) minor chronic conditions only, 2) serious chronic medical conditions only, 3) psychiatric conditions only, and 4) both serious medical and psychiatric conditions [35].

As hypothesized, the principal components showed high loadings of physical functioning, physical role, and bodily pain on the physical domain. Also, high loadings were seen of mental health, emotional role, and social functioning on the mental domain. Vitality and general health had cross-loaded on both domains. These results provided evidence of both convergent (i.e., physical functioning loading on physical domain) and divergent (i.e., physical functioning not loading on mental domain) validity. As well, results of the contrasting groups analysis provided acceptable validity evidence for the SF-36 scales [35].

Initial reliability was estimated for the SF-36 using corrected item-test correlations as well as Cronbach's alpha for each scale [36]. Using acceptable cut-point criteria, all eight scales had a 100% success rate. Average item-test correlations ranged from .42 to .74. As well, Cronbach's alpha ranged from .78 to .93.

Since its inception, the SF-36 has undergone hundreds of psychometric-related testings. Some of these studies have focused on validating the SF-36 instrument on different language speaking populations and/or cultures [37] or demographic-specific populations [38]. Other studies have focused on validating the scales on disease-specific populations [39] or condition-specific populations [40].

The widespread popularity and use of the SF-36 has drawn the attention of a few investigators trained in item response theory. Specifically, Rasch analysis has been used to compare its measurement results with the traditional Likert summation and [41]. Results showed that SF-36 scores from the Rasch analysis displayed stronger relative validity evidence as compared to the traditional summation approach. Rasch measurement has also been used to compare the two methods in relative precision [42] and confirm the unidimensionality and reproducibility of the instrument [43].

3.2. Sickness Impact Profile (SIP)

Characteristics. The SIP is another instrument used to measure HRQOL. The SIP was designed specifically as a measure of behavioral dysfunction in usual daily activities [44]. The final version consisted of 136-items of 12 categories: 1) sleep and rest, 2) emotional behavior, 3) body care and movement, 4) home management, 5) mobility, 6) social interaction, 7) ambulation, 8) alertness behavior, 9) communication, 10) work, 11) recreation and pastimes, and 12) eating [45].

The SIP is designed to be self-administered or given by face-to-face interview. The instrument is intended for generally healthy adults as well as adults with specific health conditions. The SIP is a relatively long instrument, as compared to SF-36 and EQ-5D. Like other popular HRQOL instruments, SIP has been used internationally and therefore has been translated into several language specific versions [46].

Several different scores can be obtained through SIP use: overall score, 12 different dimension scores, and 2 domain scores (physical and psychosocial). The response scale is a simple dichotomous yes or no type and the scoring is derived from a standardized weighting scheme [45].

Psychometric properties. The development of the SIP was driven by strong content validity [44]. The investigators developing the SIP began with an open-ended request form to elicit statements from individuals describing sickness-related changes in behavior. This procedure produced 1,250 statements of sickness-related behavior, which then resulted in 312 unique statements comprising 14 different dimensions. Using 25 judges and their ratings of the 312 items, content validity was affirmed by showing the correlations of each judge's rating of an item with the mean of the 25 judge's ratings.

Over the course of a few years, other psychometric data appeared regarding the SIP [47]. Construct validity, in experimental (clinical) format, was demonstrated using the differing health status and severity approach [48]. Reliability was also tested extensively for the SIP [49]. Test-retest reliability, internal consistency, and inter-rater reliability were tested on 119 respondents. Also, tests were carried out with two different forms (long and short), two different modes of administration (interviewer and self-administration), and with the sample stratified by disease severity. Overall, the reliability of the SIP was moderate to high in all circumstances.

There has only been one published study utilizing item response theory on the SIP [50]. The extended Rasch model was used to calibrate the SIP items, assess item bias, and create a shorter form via test equating. Results showed that 82 items fit the Rasch model. Item bias was seen in age, gender, and diagnosis groups, and the Rasch calibrated shorter 82-item form showed a moderate correlation with the SIP full form. Several shorter forms of the SIP have been developed (see Table 1), but their use in physical activity research is sparse.

3.3. Euroqol (EQ-5D)

Characteristics. The EQ-5D questionnaire is a standardized instrument used to measure HRQOL [51]. The EQ-5D is a very simple and short instrument that has two distinct parts [52]. The first part is a set of five items, each serving as a separate dimension: 1) mobility, 2) self-care, 3) usual activities, 4) pain/discomfort, and 5) anxiety/depression. Each item has a 3-category response: 1 = *no problems*, 2 = *some problems* and 3 = *extreme problems*. The second part is a visual analog scale (VAS) representing self-assessed health status. The scale ranges from *Best imaginable health state* (100) to the *Worst imaginable health state* (0). Respondents mark the vertical scale (which resembles a thermometer) at their perceived level of health.

The EQ-5D is designed to be self-administered or given by face-to-face interview. The instrument is intended for the general adult population as well as adults with specific health conditions. The EQ-5D is most efficiently used in large population-based surveys [53] but has also been widely used in clinical settings [54]. Like the SF-36 instrument, the EQ-5D has been used internationally and therefore has been adapted to several language specific versions [53]. Also, with advances in psychometric theory, researchers have found a possible benefit of having a 5-category response scale as opposed to the 3-category response scale [55]. These psychometric-based changes have resulted in separate EQ-5D-3L (3-level) and EQ-5D-5L (5-level) forms.

The EQ-5D has three different scoring methods [52]. The first is just the simple scoring profile of the five items (i.e., 32124). There are 243 possible combinations of these five components and therefore this scoring method yields 243 different *health states*. The second method is a population preference-weighted index score based on the five items. The index ranges from -0.11 (if scores are all 3s for each item) to 1.0 (if scores are all 1s for each item). Given that a score of 0.0 equates to death and a score of 1.0 equates to perfect health, it can be seen that the index can assume a quality of health worse than death itself. The last scoring method simply comes straight from the VAS instrument and serves as a measure of self-reported health.

Psychometric properties. The EQ-5D was developed by a multi-disciplinary group of researchers to measure HRQOL [51]. The most notable psychometric data on the EQ-5D are from a performance study of its construct, convergent, and divergent validity [51]. Construct validity was tested according to hypothesized relationships between special groups of people and the noted difference in their scoring profile. For example, older adults, women, professional workers, recent users of health services, and those diagnosed with a chronic health condition were hypothesized (and shown) to have lower scoring profiles, compared to their respective counterparts. Convergent and divergent validity were tested by comparing EQ-5D scores to dimensional scores of the SF-36. Convergent validity was established by showing that the EQ-5D anxiety/depression dimension was highly correlated with the mental health dimension of the SF-36. Likewise, divergent validity was established by showing that the EQ-5D anxiety/depression dimension was not highly correlated with the physical functioning dimension of the SF-36.

Another useful study that provided psychometric data for the EQ-5D investigated its construct validity and discriminant ability [56]. The EQ-5D reliability (internal consistency) could not be assessed because each dimension (scale) had only a single item. The construct validity was tested using polychoric correlation coefficients (PCCs) between its scales and those of the COOP/WONCA instrument. PCC results showed strong correlations with like scales and low correlations with unlike scales. Construct validity was further tested using common factor analysis. Results showed a two-factor model: mental and physical health. Also, as suspected, the anxiety/depression scale loaded on the "mental" factor and mobility, self-care, usual activities, and pain/discomfort scales loaded on the "physical" factor, providing adequate construct validity. The discriminant ability of the EQ-5D was tested using receiver operating characteristic (ROC) curves. The grouping variables in the study were migraine headache status and reporting an absence from work due to illness. EQ-5D successfully and significantly distinguished between both grouping variables, providing evidence for discriminant ability.

Item response theory has been used, in a state-of-the-art fashion, to determine the equivalency of EQ-5D measures between the 3-level response form and the 5-level response form [57]. Another study, using a Rasch measurement model, showed the benefits of having a 5-

category response scale as opposed to the 3-category response scale [55].

3.4. Nottingham Health Profile (NHP)

Characteristics. The NHP is a generic HRQOL instrument with physical, emotional, and social domains of health [58]. The NHP has a total of 45 items, all of which are dichotomous response. Two parts make up the instrument. The first part contains six different health dimensions: 1) physical mobility, 2) pain, 3) social isolation, 4) emotional reactions, 5) energy, and 6) sleep. The second part includes specific health status questions.

The NHP is designed for adults (16+ years) and to be self-administered or interviewer-administered. It was originally designed as an instrument for epidemiological research [46] but has since been used in several different arenas. No alternate forms (to date) have been found in the published literature. However, translated versions have been created in several languages.

Only the first part of NHP is considered in its scoring [59]. Scores can be obtained using weights associated with subject responses and yield a single value ranging from 0 (no health problems) to 100 (severe health problems). If all weights are summed, in part I, a score of 100 will occur.

Psychometric properties. The NHP was developed using methods of content validity, beginning with a large pool of statements (over 2,200) from approximately 700 people regarding their typical feelings about poor health [58]. The resulting instrument took on 45 items, 38 of which were part of the overall scoring profile, and six dimensions [58]. Construct validity evidence was published on the NHP by testing the instrument's ability to distinguish between different levels of pain severity [60]. Results successfully showed that the pain, energy, and sleep dimensions were highly correlated with pain severity (convergent validity evidence), whereas the other three dimensions were not highly correlated with pain severity (divergent validity).

Another study providing psychometric data for the NHP, investigated its construct validity and discriminant ability [56]. The NHP reliability (internal consistency) was determined by Cronbach's alpha. The construct validity was tested using intraclass correlation coefficients (ICCs) between its scales and those of the SF-36 instrument. Reliability results showed good internal consistency. ICC results showed strong correlations with like scales and low correlations with unlike scales. Construct validity was further tested using common factor analysis. Results showed a two-factor model: mental and physical health. Also, as suspected, the energy, emotional reactions, and social isolation scales loaded on the "mental" factor and energy, pain, and physical mobility scales loaded on the "physical" factor, providing adequate construct validity. The discriminative ability of the NHP was tested using ROC curves. Groups were formed by migraine headache status and reporting an absence from work due to illness. NHP significantly distinguished between both grouping variables, providing evidence for discriminant ability.

Item response theory has been used to assess the psychometric properties of the NHP [61]. Results showed adequate fit to the model, however, differential item functioning (DIF) was found in age and gender groups. A Rasch study was performed to reduce the number of items in the NHP from 38 to 22, while maintaining the new scale's validity [62]. Finally, the Rasch model was used to assess unidimensionality and item-fit of the Brazilian version of the NHP [63]. Despite adequate fit, some items were found to be too easy for the population under study.

3.5. WHO Quality of Life Assessment (WHOQOL-BREF)

Characteristics. The WHOQOL-BREF is a generic HRQOL tool developed from the larger WHOQOL-100 [64]. The assessment consists of 26 items which make up four HRQOL domains: 1) physical health, 2) psychological, 3) social relationships, and 4) environment. Additionally, two of the items are included to assess self-perceived general health and are for descriptive purposes. The WHOQOL-BREF was developed for adult use and has been designed to be an international HRQOL assessment. The scoring of the WHOQOL-BREF results in a single score for each domain.

Psychometric properties. The development of the WHOQOL-BREF stemmed from the larger WHOQOL version (WHOQOL-100). The initial WHOQOL-BREF project showed adequate validity with strong correlations between the WHOQOL-BREF and WHOQOL-100 domain scores [65]. The same project also showed strong evidence for content validity, discriminant validity, internal consistency, and stability in the WHOQOL-BREF scales. A more recent validation study of the WHOQOL-BREF used a large sample of participants from 23 countries. Participants were very diverse, consisting of people of various health ranges and diseases, and various sociodemographic characteristics. Results of the study showed strong evidence of internal consistency, discriminant validity, and construct validity [66].

The construct validity was evaluated in the WHOQOL-BREF using item response theory among a general population of adults [67]. Using a mail survey format and a random sample of Danish adults, the WHOQOL-BREF was administered to 1,101 respondents. Results indicated that each of the four domains of the WHOQOL-BREF fit a 2-parameter item response model. However, the total scale did not fit either a 2-parameter model or a Rasch model. The conclusion was that domain specific scores should be used when administering the WHOQOL-BREF and that the total scores of the WHOQOL-BREF may not be sufficiently valid.

3.6. Quality of Well-Being Scale (QWB)

Characteristics. The QWB is a generic HRQOL assessment tool that measures 3 different dimensions [46]. A recently updated version of the QWB scale has been developed specifically for participant self-administration (QWB-SA). The QWB-SA does not require a trained interviewer, as does the QWB, and therefore is easier and less expensive to use in research and practice. The different dimensions attempt to assess HRQOL in relation to daily functioning with scales in 1) mobility, 2) physical activity, and 3) social activity. The functioning scales ask questions about certain activities and ask respondents to respond using the past 3 days as their reference. A second component of health problems is assessed by asking questions about 26 (QWB) or 58 symptoms (QWB-SA).

Four different domain scores can be generated which can also be combined to form a total utility score representing HRQOL, ranging from 0 to 1 (death to optimal health, respectively).

Psychometric properties. Test-retest reliability was provided for both forms (QWB & QWB-SA) of the QWB scale [68]. English speaking primary care patients were used for the reliability study. Participants were randomized to receive either the QWB or the QWB-SA and were administered their respective forms twice with a one month interval. Results showed that the two forms were equivalent in terms of HRQOL scores. Also, results indicated that both QWB forms were stable in assessing HRQOL over time.

Construct validity has been established for the QWB scale by showing strong relationships between its HRQOL scores and various health outcomes among patients with chronic obstructive pulmonary disease [69]. Further validity evidence was presented when QWB scores were found to be associated with four health outcomes among HIV-infected adults [70]. Finally, evidence was also provided for the QWB's construct validity when HRQOL scores were significantly related to dementia ratings and behavioral problems among patients with and without Alzheimer's disease [70].

Although not specifically evaluated for its functioning, the QWB has been analyzed using item response theory, in comparison to four other HRQOL assessments [71]. As part of the National Health Measurement Study, the QWB was administered to 3,844 adults along with the EQ-5D, HUI2, HUI3, and SF-6D assessments. Findings showed that the five assessments combined contributed to 3 domains consisting of physical, psychosocial, and pain. However, the QWB only contributed to 2 of these domains, physical and psychosocial.

3.7. Health Utility Index Mark 3 (HUI3)

Characteristics. The HUI3 is another tool used often in economics research. Its development was driven by the need to describe 1) experiences of medical patients, 2) outcomes associated with therapy and disease, 3) the effectiveness of medical and health-related interventions, and 4) health status in large population studies [72]. The HUI3 is the most recent version of the Health Utility Index series, starting with HUI1 and then HUI2 [73]. The HUI3 consists of 8 attributes: 1) vision, 2) hearing, 3) speech, 4) ambulation, 5) dexterity, 6) emotion, 7) cognition, and 8) pain. With these attributes and a multi-attribute utility algorithm, the HUI3 can yield HRQOL values covering over 900,000 unique health states [46]. Scores can also be computed using a different set of algorithms to yield either single-attribute or multi-attribute values ranging from -0.36 (*worse than dead*) to 0.00 (*dead*) to 1.00 (*perfect health*).

Psychometric properties. The HUI3 is a third generation HRQOL instrument that began from earlier work with the HUI1 [74]. The rationale for the original items and dimensions came from a population perspective of health outcomes [46]. The evolution of the index to the HUI3 was driven by the desire to make the assessment practical for both clinical use as well as population studies. Each dimension of the HUI3 is assessed with a single item; therefore, internal consistency reliability has not been examined with this assessment. Stability has been evaluated in the HUI3 using the Kappa statistic of agreement. The HUI3 was administered to a large sample at two different time periods, one month apart. Results showed that 6 of the 8 dimensions had acceptable reliability [75].

Construct validity was evaluated in the HUI3 by comparing HRQOL scores between groups with known differences. Participants were used from the 1990 Ontario Health Survey. Adequate validity was shown as participants with stroke, arthritis, and both stroke and arthritis had significantly lower HUI3 scores [73]. Convergent validity evidence was evaluated on the HUI3 by comparing the scoring patterns between the HUI3, EQ-5D, and SF-36 HRQOL assessments [76]. Participants for this validity study were outpatients with rheumatic disease. Results provided adequate evidence for convergent validity. Those patients with higher SF-36 scores also had significantly higher EQ-5D and HUI3 scores. Total scores on EQ-5D and HUI3 were not significantly different from each other.

To compare 5 different HRQOL assessments for their interrelationships, item response theory was used on the HUI3 [77]. As part of the National Health Measurement Study, the HUI3 was administered to 3,844 adults along with the EQ-5D, HUI2, QWB-SA, and SF-6D. Results indicated, that the HUI3 was linearly related to the EQ-5D and the HUI2 scales. Although a linear relationship was shown, it was stated that the relationship was simplistic and that the different scales were in actuality measuring different aspects of generic HRQOL.

3.8. CDC Health-Related Quality of Life (HRQOL-4) Scale

Characteristics. The HRQOL-4 Scale consists of four items and was developed as a surveillance tool to be used in the U.S. Behavioral Risk Factor Surveillance System (BRFSS) [78]. The four items were created through the CDC's definition of HRQOL which includes perceived physical and mental health over time. The first item asks participants to rate their own general health on a 5-point scale starting with *excellent* and ending with *poor*. The second and third questions were specifically geared toward *physicalhealth* (physical illness and injury) and *mentalhealth* (stress, depression, and emotional problems), respectively. These questions ask respondents to report the number of days (out of the previous 30 days) that their physical (or mental) health was not good. The last question specifically addresses the amounts of *usualactivity* (self-care, work, or recreation) influenced by physical and/or mental health. Respondents are asked to report the number of days (out of the previous 30 days) that poor physical or mental health kept them from their usual activities [79].

The scoring methods for the CDC HRQOL-4 scale are twofold. The first option is a descriptive scoring method. This can be done by creating dichotomous categories for each item [17]. For example, for the first item, those reporting either *fair* or *poor* general health can be considered to exhibit poor HRQOL and those reporting *excellent*, *very good*, or *good* general health can be considered to exhibit good HRQOL. For the second and third items, those reporting 14 days of poor health or more can be considered to exhibit poor physical (or mental) health. For the fourth item, those reporting 14 days or more can be considered to be inactive due to poor health. The second scoring option is to create a summary index of unhealthy (or healthy) days. The index can be constructed from the physical and mental health items and used to assess the overall number of unhealthy days due to physical and/or mental health, not to exceed 30 days [78].

Psychometric properties. The CDC HRQOL-4 scale was developed using a strong conceptual framework [79]. Items were specifically constructed to be 1) individual-oriented, 2) subjective in nature, 3) non disease-specific, 4) sensible to the general public, 5) non-biased toward various ethnic groups, and 6) practical. The time frame was also considered to capture an adequate reflection of an individual's health. Developers of the CDC HRQOL-4 scale used early BRFSS data to test the scale's validity. This was accomplished first by showing the relationship between the first core HRQOL item (perceived general health rating) and the second core item (number of days respondents said their physical health was not good). The relationship provided convergent validity evidence as those reporting better general health had significantly fewer days of poor physical health and those reporting poor general health reported significantly more days of poor physical health. This relationship was also found between the first item and the third (number of days poor mental health) and fourth (number of days limited by physical and/or mental health) items.

The retest reliability was assessed for the CDC HRQOL-4 using a random sample of BRFSS respondents approximately two weeks after their initial survey [80]. The Kappa coefficient and proportion of agreement were used for the first core item and the intra-class correlation coefficient was used for the other three items as well as the healthy days index. Reliability coefficients for the general health item and the healthy days index were both acceptable. Reliability was moderate for the other three (number of days) items.

Another validation study of the CDC HRQOL-4 was conducted with a sample of Dutch adults [81]. First, reliability was evaluated by computing the Cronbach alpha on the three number of days core items. The reliability of the three items was deemed acceptable (alpha = .77). Second, criterion validity was assessed by comparing the HRQOL-4 items with three other well-respected HRQOL assessments: SF-36, WHOQoL-BREF, and GHQ-12. Spearman correlations confirmed that HRQOL-4 items of similar domain were highly related across instruments. As well, HRQOL-4 items of different domains were not correlated across instruments. Finally, construct validity was examined by comparing HRQOL-4 scores between groups of adults with known differences in health status. Those respondents reporting a chronic condition, depression, use of prescription drugs, and visiting a doctor, had significantly lower scores of HRQOL as assessed by the CDC HRQOL-4. To date, no item response theory studies have been published on the CDC HRQOL-4 assessment.

3.9. Assessment of Quality of Life (AQoL)

Characteristics. The AQoL instrument is a generic HRQOL assessment developed by Australian researchers [82]. The AQoL consists of five dimensions covering the HRQOL construct and contains questions specifically targeted for economic evaluation. The AQoL has a total of 15 items each measured on a four point categorical scale ranging from A (*Good HRQOL*) to D (*poor HRQOL*). The five dimensions consist of: 1) illness, 2) independent living, 3) social relationships, 4) physical senses, and 5) psychological wellbeing.

The AQoL is designed to be administered by self, interviewer, mail, or telephone. It was designed as a multi-attribute health utility index, however, it is also used as a health states assessment. The original AQoL was replaced by its developers and referred to as AQoL-II. A shorter version has been developed [83] consisting of only eight items (AQoL-8); however, this version has not been used in physical activity research.

There are three scoring options for AQoL users [82]. The first is an overall HRQOL score. This is computed by assigning a zero to an 'A' response, a one to a 'B' response, a two to a 'C' response, and a three to a 'D' response. Therefore, a low overall score of zero is possible and a high score of 45 is possible. The second option is to sum the same scale by subdomains. Therefore, each subdomain can range in score from zero to nine. Finally, an algorithm can be used to transform the raw AQoL scores to preference weighted utility scores ranging from -0.04 (*worse than death*) to zero (*death*) to one (*complete health*).

Psychometric properties. The AQoL was developed using a content validation procedure [82]. The development began using a strong conceptual framework based on the World Health Organization's (WHO) definition of health. With this framework in mind, researchers and professionals constructed appropriate items, reviewed the items for clarity and simplicity, and administered the selected items to both hospital patients and community members. After data collection, an item analysis was performed using 100% range criteria (all categories of an item being selected) for item sensitivity and a standard deviation of .50 cutoff as the criterion for item discrimination. Items surviving the preliminary analysis were further tested for construct validity. First, principal components analysis was run, dropping items that did not load on the underlying HRQOL construct. Second, exploratory factor analysis was performed, dropping items that failed to load on a single factor only. Results indicated a five factor structure and measures of internal consistency confirmed its reliability. Re-analysis of each factor separately by principal components analysis provided evidence of the unidimensionality of each factor. Finally, structural equation modeling was performed to assess the explanatory power of the AQoL in providing HRQOL information.

Another study providing psychometric information for the AQoL investigated its stability across different methods of administration. The developer of the instrument showed that administering the AQoL via mail or telephone resulted in statistically equivalent HRQOL scores [84]. Furthermore, the stability of scores was maintained as well for each set of subscale scores.

Item response theory has been used on the AQoL with a specific purpose to find the most parsimonious scale [83]. First, subscale unidimensionality was determined followed by full scale unidimensionality, using Mokken item response theory. Items which were not considered unidimensional (homogenous) were candidates for deletion. Second, a Rasch partial credit model was used to determine each item's set of category thresholds. Items with disordered thresholds (i.e., persons with low HRQOL endorsing categories representing higher HRQOL levels) were also candidates for deletion. The goal of the study was to reduce each of the four subscales by one item resulting in an 8-item AQoL scale using the two criteria of unidimensionality and ordered category thresholds. The resulting AQoL-8 correlated well (intraclass correlation coefficient = .95) with the full AQoL scale and showed 97% of a validation sample within +/- 2 SD limit of agreement and was therefore considered a more parsimonious measure of HRQOL.

3.10. Duke Health Profile (DHP)

Characteristics. The DHP is a generic self-report HRQOL assessment tool that contains 10 different measures of health [85]. Six of the measures are considered positive health measures (physical, mental, social, general, perceived health, and self-esteem) and the other four are considered measures of dysfunction (anxiety, depression, pain, and disability). The scale consists of only 17 items, each measured on a 3-point categorical rating scale. Scoring for the DHP is relatively simple, summing each separate dimension and multiplying (or dividing for the general health score) by a constant. Each dimension has a score range from 0 to 100 where 100 represents the best health for the positive health measures and the worst health for the measures of dysfunction.

Another unique characteristic of this assessment is that the general health dimension is a composite of the physical, mental, and social dimensions. Combining these three major dimensions of HRQOL allows for a more realistic measure of general health. The DHP was developed for adults but has been revised and validated for adolescents [86]. The DHP HRQOL assessment is primarily used in English speaking countries but has recently been validated in France [87].

Psychometric properties. The DHP was developed from a slightly larger (63-item) Duke-UNC Health Profile (DUHP) assessment [85]. Items were selected from the larger pool of items using content validity (or face validity) and item-remainder (item score and dimension score with item removed) correlations. Cronbach alphas provided evidence of internal consistency reliability with multi-item dimensions having alphas ranging from .55 to .78. Test-retest provided evidence of stability with all dimensions having reliability greater than .50, except disability and pain. Spearman correlations were used for item-convergent and item-divergent evidence against three other assessment tools: DUHP, SIP, and the Tennessee Self-Concept Scale. Validity was established for the DHP with strong positive correlations among similar constructs and strong negative correlations among different constructs. Finally, mean comparisons were used to provide construct validity evidence by showing DHP score differences between groups with known health problems. Validity was established when results showed that groups with lower levels of health had significantly lower DHP scores compared to groups with better health [85].

4. Discussion

The purpose of this study was to systematically review assessments used to measure generic HRQOL in physical activity research in adults. The review included summarizing the characteristics, scoring options, and psychometric properties of each HRQOL assessment. A total of 10 instruments were found and examined. By far, the SF-36 along with its variants was the most commonly used HRQOL assessment in physical activity research. Table 2 displays the recommendation for each assessment based on whether a researcher's reason for selecting it was its psychometric properties, amount of HRQOL information (scores from dimensions), or its length.

In terms of participant burden, the CDC HRQOL-4 and the EQ-5D both provide a valid HRQOL score given they contain only 4 and 6 items, respectively. The AQoL, DHP, and WHOQL-BREF, however, also allow for low participant strain (15, 17, and 26 items, respectively) and provide slightly more information. The AQoL provides a single HRQOL score along with 5 subdomain scores (illness, independent living, social relationships, physical senses, and psychological wellbeing). The DHP provides 10 dimensional scores (physical, mental, social, general perceived health, self-esteem, anxiety depression, pain, and disability), one of which is a general health score. The WHOQOL-BREF measures HRQOL with 4 separate dimensions (physical health, psychological, social relationships, and environment). These 3 mid-sized assessments may be useful to physical activity researchers who seek to investigate very specific HRQOL changes (i.e., pain or social relations) without overwhelming their subjects with several items or forms. In terms of psychometric properties, the SF-36 leads in both amounts and quality of supporting information. The evidence backing the SF-36's validity includes both classical test theory as well as modern test theory. The other 9 assessments all have several studies validating their scales using both psychometric approaches, with the exception of the CDC HRQOL-4 and DHP which have no published data (to date) using item response theory.

Table 2. Recommendation for HRQOL assessment based on psychometric properties, amount of HRQOL information, and length

Form	Psychometric	Information	Short Length
SF-36	Definitely	Definitely	SF-12/8
SIP	Yes	Yes	No
EQ-5D	Yes	Maybe	Definitely
NHP	Yes	No	No
WHOQOL	Yes	Maybe	Maybe
QWB	Maybe	Maybe	No
HUI3	Yes	Yes	Yes
HRQOL-4	Maybe	No	Definitely
AQoL	Maybe	Yes	Yes
DHP	Maybe	Definitely	Yes

5. Conclusion

In conclusion, 10 HRQOL assessments were found to be used in physical activity research. The SF-36, the most commonly used and validated assessment, provides the most information given its size. Other HRQOL assessments with good potential include AQoL, DHP, and WHOQOL-BREF. If time is the most important factor, the EQ-5D and CDC HRQOL-4 are useful and valid scales.

References

[1] U.S. Department of Health and Human Services. Physical Activity Guidelines Advisory Committee Report. Washington (DC): U.S. Department of Health and Human Services; 2008. p. A-7.

[2] Kampert, J. B., Blair, S. N., Barlow, C. E., & Kohl, H. W. (1996). Physical activity, physical fitness, and all-cause and cancer mortality: A prospective study of men and women. Annals of Epidemiology, 6, 452-457.

[3] Lee, D. C., Sui, X., Ortega, F. B., Kim, Y. S., Church, T. S., Winett, U., Ekelund, U., Katzmarzyk, P. T., & Blair, S. N. (2010). Comparisons of leisure-time physical activity and cardiorespiratory fitness as predictors of all-cause mortality in men and women. British Journal of Sports Medicine, 45(6), 504-510.

[4] Tanasescu, M., Leitzmann, M. F., Rimm, E. B., & Hu, F. B. (2003). Physical activity in relation to cardiovascular disease and total mortality among men with type 2 diabetes. Circulation, 107(19), 2435-2439.

[5] Durand, G., Tsismenakis, A. J., Jahnke, S. A., Baur, D. M., Christophi, C. A., & Kales, S. N. (2011). Firefighters' physical activity: Relation to fitness and cardiovascular disease risk. Medicine and Science in Sports and Exercise, In Print.

[6] Franco, O. H., de Laet, C., Peeters, A., Jonker, J., Mackenbach, J., & Nusselder, W. (2005). Effects of physical activity on life expectancy with cardiovascular disease. Archives of Internal Medicine, 165(20), 2355-2360.

[7] Sesso, H. D., Paffenbarger, R. S., Ha, T., & Lee, I. M. (1999). Physical activity and cardiovascular disease risk in middle-aged and older women. American Journal of Epidemiology, 150(4), 408-416.

[8] Brien, S. E., Katzmarzyk, P. T., Craig, C. L., & Gauvin, L. (2007). Physical activity, cardiorespiratory fitness and body mass index as predictors of substantial weight gain and obesity: The Canadian physical activity longitudinal study. Canadian Journal of Public Health, 98(2), 121-124.

[9] Buchowski, M. S., Cohen, S. S., Matthews, C. E., Schlundt, D. G., Signorello, L. B., Hargreaves, M. K., & Blot, W. J. (2010). Physical activity and obesity gap between black and white women in the southeastern U.S.*American World Journal of Preventive Medicine*, *39*(2), 140-147.

[10] Backmand, H., Kaprio, J., Kuiala, U., & Sarna, S. (2003). Influence of physical activity on depression and anxiety of former elite athletes.*International Journal of Sports Medicine*, *24*(8), 609-619.

[11] Strawbridge, W. J., Deleger, S., Roberts, R. E., & Kaplan, G. A. (2002). Physical activity reduces the risk of subsequent depression for older adults.*American Journal of Epidemiology*, *156*(4), 328-334.

[12] Vallance, J. K., Winkler, E. A., Gardiner, P. A., Healy, G. N., Lynch, B. M., & Owen, N. (2011). Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005-2006).*Preventive Medicine*, In Press.

[13] Centers for Disease Control and Prevention. *Measuring healthy days: Population assessment of health-related quality of life*. Centers for Disease Control and Prevention, Atlanta, Georgia 2000.

[14] Dominick, K. L., Ahern, F. M., Gold, C. H.,&Heller, D. A.(2004). Health-related quality of life among older adults with arthritis. *Health and Quality of Life Outcomes*, *13*(2), 5-12.

[15] Dominick, K. L., Ahern, F. M., Gold, C. H., &Heller, D. A.(2002). Relationship of health-related quality to health care utilization and mortality among older adults. *Aging Clinical and Experimental Research*, *14*(6), 499-508.

[16] Heath, G. W., & Brown, D. W. (2009). Recommended levels of physical activity and health-related quality of life among overweight and obese adults in the United States, 2005. *Journal of Physical Activity and Health, 6*(4), 403-411.

[17] Brown, D. W., Balluz, L. S., Heath, G. W., Moriarty, D. G., Ford, E. S., Giles, W. H., &Mokdad, A. H. (2003). Associations between recommended levels of physical activity and health-related quality of life. Findings from the 2001 Behavioral Risk Factor Surveillance System (BRFSS) survey.*Preventive Medicine*, *37*(5), 520-528.

[18] Aoyagi, Y., Park, H., Park, S., &Shephard, R. J. (2010). Habitual physical activity and health-related quality of life in older adults: Interactions between the amount and intensity of activity (the Nakanojo Study). *Quality of Life Research*, *19*(3), 333-338.

[19] Balboa-Castillo, T., León-Muñoz, L.M., Graciani, A., Rodríguez-Artalejo, F., & Guallar-Castillón, P., (2011). Longitudinal association of physical activity and sedentary behavior during leisure time with health-related quality of life in community-dwelling older adults.*Health and Quality of Life Outcomes*, *27*(9), 47.

[20] Tessier, S., Vuillemin, A., Bertrais, S., Boini, S., Le Bihan, E., Oppert, J. M., Hercberg, S., Guillemin, F., &Briançon, S. (2007). Association between leisure-time physical activity and health-related quality of life changes over time. *Preventive Medicine*, *44*(3), 202-208.

[21] Luncheon, C., & Zack, M. (2011). Health-related quality of life and the physical activity levels of middle-aged women, California Health Interview Survey, 2005. *Prevention of Chronic Disease*, *8*(2), A36.

[22] Courneya, K.S., Tamburrini, A.L., Woolcott, C.G., McNeely, M.L., Karvinen, K.H., Campbell, K.L., McTiernan, A., &Friedenreich, C.M. (2011). The Alberta Physical Activity and Breast Cancer Prevention Trial: Quality of life outcomes. *Preventive Medicine*, *52*(1), 26-32.

[23] Sørensen, J., Sørensen, J. B., Skovgaard, T., Bredahl, T., &Puggaard, L. (2011). Exercise on prescription: Changes in physical activity and health-related quality of life in five Danish programmes.*European Journal of Public Health*, *21*(1), 56-62.

[24] Bertheussen, G. F., Romundstad, P. R., Landmark, T., Kaasa, S., Dale, O., &Helbostad, J. L. (2011). Associations between physical activity and physical and mental health--a HUNT 3 study. *Medicine and Science in Sports and Exercise*, *43*(7), 1220-1228.

[25] U. S. Department of Health and Human Services. Healthy people 2020. Available at: http://www.healthypeople.gov/2020/topicsobjectives2020/default.aspx. Accessed June 7, 2011.

[26] Allen, M.J., & Yen, W. M. (2002). Introduction to Measurement Theory. Long Grove, IL: Waveland Press.

[27] Kline, T. (2009). *Psychological testing: A practical approach to design and evaluation*. (3rd Ed.). New Deli, India: Vistaar Publications.

[28] Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trial*, *12*(4 Suppl), 142S-158S.

[29] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

[30] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. New Deli, India: Sage Publications.

[31] Wood, T. M., Zhu, W. (2006*). Measurement theory and practice in kinesiology*. Champaign, IL: Human Kinetics.

[32] Ware, J.E., 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.

[33] Ware, J.E. (2004). SF-36 Health Survey Update. Retrieved August 15, 2011, from http://www.sf-36.org/announcements/Updated_SF36_bookChapter_Sept04.pdf

[34] QualityMetric. (2011). SF Health Surveys. Retrieved August 15, 2011, from http://www.qualitymetric.com/WhatWeDo/GenericHealthSurveys/tabid/184/Default.aspx.

[35] McHorney, C.A., Ware, J.E., Sherbourne, C.D.(1993). The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, *31*(3), 247.

[36] McHorney, C.A., Ware, J.E., Sherbourne, C.D.(1994). The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, *32*(1), 40.

[37] Laguardia, J., Campos. M,R,, Travassos. C.M., Najar, A.L., Anjos, L.A., Vasconcellos, M.M. (2011). Psychometric evaluation of the SF-36 (v.2) questionnaire in a probability sample of household: Results of the survey Pesquisa Dimensoes Sociais das Desigualdades (PDSD), Brazil, 2008. Health and Quality of Life Outcomes, 9(1), 61.

[38] Mishra, G.D., Gale, C.R., Sayer, A.A., Cooper, C., Dennison, E.M., Whalley, L.J., Craig, L., Kuh, D., Deary, I.J. (2011). How useful are the SF-36 sub-scales in older people? Mokken scaling of data from the HALCyon programme.*Quality of life Research, 20*(7), 1005.

[39] Laosanguanek, N., Wiroteurairuang, T., Siritho, S., Prayoonwiwat, N. (2011). Reliability of the Thai version of SF-36 questionnaire for an evaluation of quality of life in multiple sclerosis patients in multiple sclerosis clinic at Siriraj Hospital.*Journal of the Medical Association of Thailand*, *94* Suppl 1, S84.

[40] Freidheim, O. M. S., Borchgrevin, P. C., Saltnes, T., & Kaasa, S. (2007). Validation and comparison of the health related quality of life instrument EORTC QLQ C30 and SF36 in assessment of patients with non-malignant pain. *Journal of Pain and Symptom Management*, *34*(6), 657.

[41] Raczek AE, Ware JE, Bjorner JB Gandek B, Haley SM, Aaronson NK, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology,51*(11), 1203.

[42] McHorney CA, Haley SM, Ware JE Jr. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. Journal of Clinical Epidemiology, 50(4), 451.

[43] Haley, S. M., McHorney, C. A., & Ware, J. E., Jr. (1994). Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. [Research Support, Non-U.S. Gov't]. J Clin Epidemiol, 47(6), 671-684.

[44] Gilson BS, Gilson JS, Bergner M, Bobbit RA, Kressel S, Pollard WE, Vesselago M. (1975). The sickness impact profile. Development of an outcome measure of health care. American Journal of Public Health, 65(12), 1304-1310.

[45] Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). Sickness Impact Profile: Development and final revision of a health status measure. Medical Care, 19(8), 787.

[46] Coons, S. J., Rao, S., Keininger, D. L., & Hays, R. D. (2000). A comparative review of generic quality-of-life instruments. Pharmacoeconomics, 17(1), 13-35.

[47] Bergner, M, Bobbitt, R. A., Pollard, W. E., Martin, D. P., &Gilson, B.S. (1976). The sickness impact profile: Validation of a health status measure.*Medical Care*, *14*(1), 57.

[48] Bergner, M., Bobbitt, R. A., Kressel, S., Pollard, W. E., Gilson, B. S., &Morris, J. R.(1976). The sickness impact profile: Conceptual formulation and methodology for the development of a health status measure.*International Journal of Health Services*, *6*(3), 393.

[49] Pollard WE, Bobbitt RA, Bergner M, Martin DP, Gilson BS. (1976). The Sickness Impact Profile: Reliability of a health status measure. *Medical Care*, *14*(2), 146.

[50] Lindeboom R, Holman R, Dijkgraaf MG, Sprangers MA, Buskens E, Diederiks JP, De Haan RJ. (2004). Scaling the sickness impact profile using item response theory: An exploration of linearity, adaptive use, and patient driven item weights.*Journal of Clinical Epidemiology*, *57*(1), 66.

[51] Brazier, J., Jones, N., & Kind, P. (1993). Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. Quality of Life Research, 2(3), 169.

[52] Agency for Healthcare Research and Quality (AHRQ). (2005). Calculating the U.S. Population-based EQ-5D Index Score. Retrieved August 15, 2011, from http://www.ahrq.gov/rice/EQ5Dscore.htm.

[53] Rabin R, Oemar M, Oppe M, on behalf of the EuroQoL Group: EQ-5D-3L user guide. 4th edition. Rotterdam: EuroQoL Group; 2011.

[54] Vestergaard, S., Kronborg, C., & Puggaard, L. (2008). Home-based video exercise intervention for community-dwelling frail older women: a randomized controlled trial. Aging Clin Exp Res, 20(5), 479-486.

[55] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, In Press.

[56] Essink-Bot, M. L., Krabbe, P. F., Bonsel, G. J., Aaronson, N. K. (1997). An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument.*Medical Care*, *35*(5), 522-537.

[57] Pickard AS, Kohlmann T, Janssen MF, Bonsel G, Rosenbloom S, Cella D. (2007). Evaluating equivalency between response systems: Application of the Rasch Model to a 3-level and 5-level EQ-5D. *Medical Care*, *45*(9), 812.

[58] Hunt, S. M., McEwan, J., & McKenna, S. P. (1985). Measuring health status: A new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners*, 35(273), 185.

[59] Hunt, S. M., McKenna, S. P., McEwen, J., Williams, J., & Papp, E.(1981). The NottinghamHealth Profile: Subjective health status and medical consultations.*Social Science Medicine*, 15(3), 221.

[60] Mauskopf J, Austin R, Dix L, Berzon R. (1994). The Nottingham Health Profile as a measure of quality of life in zoster patients: Convergent and discriminant validity.*Quality of Life Research*, *3*(6), 43.

[61] Hagell P, Whalley D, McKenna SP, Lindvall O. (2003). Health status measurement in Parkinson's disease: Validity of the PDQ-39 and Nottingham Health Profile.*Movement Disorders*, *18*(7), 773-783.

[62] Prieto, L., Alonso J, Lamarca R, Wright BD, (1998). Rasch Measurement for Reducing the Items of the Nottingham Health Profile. *Journal of Outcome Measurement*, *2*(4), 285.

[63] Teixeira-Salmela LF, Magalhães Lde C, Souza AC, Lima Mde C, Lima RC, Goulart F. (2004). Adaptation of the Nottingham Health Profile: A simple measure to assess quality of life. Cadernos de saúde pública, 20(4), 905.

[64] Saxena, S., Carlson, D., & Billington, R. (2001). The WHO quality of life assessment instrument (WHOQOL-Bref): The importance of its items for cross-cultural research. *Quality of Life Research, 10*(8), 711-721.

[65] WHOQoL Group. "Development of the World Health Organization WHOQOL-BREF quality of life assessment." Psychological medicine 28.03 (1998): 551-558.

[66] Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL group. Quality of Life Research, 13(2), 299-310.

[67] Noerholm, V., Groenvold, M., Watt, T., Bjorner, J. B., Rasmussen, N. A., & Bech, P. (2004). Quality of life in the Danish general population--normative data and validity of WHOQOL-BREF using Rasch and item response theory models.*Quality of Life Research*, *13*(2), 531-540.

[68] Kaplan, R. M., Sieber, W. J., & Ganiats, T. G. (1997). The quality of well-being scale: Comparison of the interviewer-administered version with a self-administered questionnaire. Psychology & Health, 12(6), 783-791.

[69] Kaplan, R. M., Atkins, C. J., & Timms, R. (1984). Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *Journal of Chronic Disease, 37*(2), 85-95.

[70] Hughes, T. E., Kaplan, R. M., Coons, S. J., Draugalis, J. R., Johnson, J. A., & Patterson, T. L. (1997). Construct validities of the Quality of Well-Being Scale and the MOS-HIV-34 Health Survey for HIV-infected patients. *Medical Decision Making, 17*(4), 439-446.

[71] Cherepanov, D., Palta, M., & Fryback, D. G. (2010). Underlying dimensions of the five health-related quality-of-life measures used in utility assessment: Evidence from the National Health Measurement Study. *Medical Care, 48*(8), 718-725.

[72] Horsman, J., Furlong, W., Feeny, D., & Torrance, G. (2003). The Health Utilities Index (HUI): Concepts, measurement properties and applications. Health and Quality of Life Outcomes, 1, 54.

[73] Grootendorst, P., Feeny, D., & Furlong, W. (2000). Health Utilities Index Mark 3: Evidence of construct validity for stroke and arthritis in a population health survey. Medical Care, 38(3), 290-299.

[74] Torrance, G. W., Furlong, W., Feeny, D., & Boyle, M. (1995). Multi-attribute preference functions. Health Utilities Index. *Pharmacoeconomics, 7*(6), 503-520.

[75] Boyle, M. H., Furlong, W., Feeny, D., Torrance, G. W., & Hatcher, J. (1995). Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Quality of Life Research, 4*(3), 249-257.

[76] Luo, N., Chew, L. H., Fong, K. Y., Koh, D. R., Ng, S. C., Yoon, K. H., . . . Thumboo, J. (2003). A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. Journal of Rheumatology, 30(10), 2268-2274.

[77] Fryback, D. G., Palta, M., Cherepanov, D., Bolt, D., & Kim, J. S. (2010). Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Medical Decision Making, 30*(1), 5-15.

[78] Taylor, V. R., & National Center for Chronic Disease Prevention and Health Promotion (U.S.). Division of Adult and Community Health. (2000). Measuring healthy days : Population assessment of health-related quality of life. Atlanta, Ga.: U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Adult and Community Health.

[79] Hennessy, C. H., Moriarty, D. G., Zack, M. M., Scherr, P. A., & Brackbill, R. (1994). Measuring health-related quality of life for public health surveillance. Public Health Reports, 109(5), 665-672.

[80] Andresen, E. M., Catlin, T. K., Wyrwich, K. W., & Jackson-Thompson, J. (2003). Retest reliability of surveillance questions on health related quality of life. Journal of Epidemiology and Community Health, 57(5), 339-343.

[81] Toet, J., Raat, H., & van Ameijden, E. J. (2006). Validation of the Dutch version of the CDC core healthy days measures in a community sample. Quality of Life Research, 15(1), 179-184.

[82] Hawthorne, G., Richardson, J., & Osborne, R. (1999). The Assessment of Quality of Life (AQoL) instrument: A psychometric measure of health-related quality of life. Quality of Life Research, 8(3), 209-224.

[83] Hawthorne, G. (2009). Assessing utility where short measures are required: Development of the short Assessment of Quality of Life-8 (AQoL-8) instrument. Value Health, 12(6), 948-957.

[84] Hawthorne, G. (2003). The effect of different methods of collecting data: Mail, telephone and filter data collection issues in utility measurement. Quality of Life Research, 12(8), 1081-1088.

[85] Parkerson, G. R., Jr., Broadhead, W. E., & Tse, C. K. (1990). The Duke Health Profile. A 17-item measure of health and dysfunction. Medical Care, 28(11), 1056-1072.

[86] Vo, T. X., Guillemin, F., & Deschamps, J. P. (2005). Psychometric properties of the DUKE Health Profile-adolescent version (DHP-A): A generic instrument for adolescents. Quality of Life Research, 14(10), 2229-2234.

[87] Baumann, C., Erpelding, M. L., Perret-Guillaume, C., Gautier, A., Regat, S., Collin, J. F., & Briancon, S. (2011). Health-related quality of life in French adolescents and adults: Norms for the DUKE Health Profile. BMC Public Health, 27(11), 401.