EQUIVALENCE OF PAPER AND TOUCH SCREEN VERSIONS OF THE EQ-5D VISUAL ANALOG SCALE (EQ-VAS)

By

Sulabha Ramachandran

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHARMACEUTICAL SCIENCES

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2005

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Sulabha Ramachandran

entitled Equivalence of Paper and Touch Screen Versions of the EQ-5D Visual Analog Scale (EQ-VAS)

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

Date: 09-02-05
Stephen Joel Coons

Date: 09-02-05
Ronald J. Vogel

Date: 09-02-05
Benjamin M. Craig

Date: 09-02-05
Merrie L. Brucks

Date: 09-02-05
Alan J. Malter

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 09-02-05
Dissertation Director: Stephen Joel Coons

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements of an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: Sulabha Ramachandran

ACKNOWLEDGEMENTS

I would like to thank a number of individuals who have supported and guided me during the process of conducting my research and writing my dissertation as well as throughout the course of my graduate work.

I have been incredibly fortunate to have worked with Dr. Stephen Joel Coons, my dissertation chairperson and doctoral program advisor. I am thankful for his support, patience, and encouragement throughout my graduate education. His technical and editorial input has been essential to the completion of this dissertation. His friendship and advice, I value immensely. Thank you for guiding me to the finish line!

I would like to acknowledge the assistance and support of my dissertation committee, Drs. Ronald Vogel, Merrie Brucks, Alan Malter, and Benjamin Craig. I have benefited greatly from their instruction and advice. Their incisive comments and suggestions helped me tremendously in refining my dissertation.

I am extremely grateful to Theron Taber at Assist Technologies, whose financial contributions and technology (i.e., touch screen devices) enabled the completion of this dissertation research. In addition, I would like to thank his wonderful staff, who were always available to answer any questions or provide technical support.

I would like to acknowledge Denise deGeus for her valuable help in coordinating activities related to the data collection process. She has been a joy to interact with and I am incredibly thankful for her support and friendship.

I would like to thank present and former graduate students for sharing ideas, experiences and friendships. A special thanks to Yuda Chongpison, who took the time out of her busy schedule to help me during the data collection process.

This dissertation would not have been possible without the love, support, and encouragement of my family members.

I am sincerely grateful to my husband, Durgesh Bhandary, for being the best friend I could ask for. I am so grateful for his valuable assistance with my data collection, analysis, and dissertation formatting. Thank you for the love, encouragement, and emotional support during this somewhat arduous process.

To the Bhandary family, for always being supportive and welcoming me into their lives. Thank you to my sister, Sushma for her constant encouragement and endless faith in my abilities.

To my wonderful parents, who have always trusted my judgment and allowed me to think for myself. Thank you for making anything and everything possible!

TABLE OF CONTENTS

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

TABLE OF CONTENTS - Continued

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The EQ-VAS, a measure of self-reported health status, has been operationalized in ways that depart from the original format. The primary purpose of the study was to examine the equivalence of the original paper-based vertical format with a touch screen-based horizontal format. Non-probability sampling was used to recruit 314 subjects intended to reflect the primary socio-demographic characteristics of the general adult population. A two-part questionnaire completed roughly 10 minutes apart was administered in a randomized crossover design. One part was the original paper-based 20cm vertical EQ-VAS; the other part was touch screen computer-based (designed by Assist Technologies) and included, among other items/scales, a horizontal EQ-VAS, the SF-36, and socio-demographic items. A mean difference of ± eight points between the two versions was specified as the minimally important difference (MID). Almost a third (30.1%) of the respondents reported identical scores on both formats and 80.1% of the respondents had difference scores within ± 8 points. The 95% confidence intervals for both samples indicated that the difference in scores was relatively small and below our equivalence threshold. In addition, data collected via touch screen may be more reliable since 22% of subjects did not complete the EQ-VAS paper format as instructed. These results provided evidence for the measurement equivalence of the touch screen EQ-VAS with the original paper format.

A secondary purpose was to examine the psychometric properties of an electronic version of the SF-36. Floor and ceiling effects were comparable to that

observed in other studies using the paper SF-36 in the general population. All reliability coefficients exceeded 0.70; the range was from 0.75 to 0.93. There was support for the construct validity of the touch screen SF-36, as the direction and strength of the correlations between the SF-36 scales and the EQ-5D domains were as hypothesized. Overall, there was a high level of correspondence between the touch screen SF-36 scores and previously reported paper based SF-36 scores in the general population. The comparable psychometric properties and low level of missing data make touch screen questionnaires a very viable alternative to their paper-based formats.

CHAPTER 1

INTRODUCTION

I. Background

The rapid evolution of outcomes research during the last two decades has led to an increased emphasis on health-related quality of life (HRQoL) assessments. Traditionally, mortality data was used as an outcome measure to assess the effectiveness of health care interventions (Sullivan, 1966). However, in recent years, the paradigm for health care assessment has shifted by emphasizing outcome measures that incorporate the notion of not just the quantity of life but also the quality of life (Kongpatanakul and Strom, 1992) This viewpoint had also been espoused by the World Health Organization (WHO), which in 1948 defined health as "not only an absence of infirmity and disease but also a state of complete physical, mental and social well-being" (WHO, 1948). Hence, the interest in measuring HRQoL as an important treatment outcome has been growing (Spilker, 1992). Additionally, the US Food and Drug Administration (FDA) has begun to allow quality of life promotional claims for pharmaceuticals that demonstrate sufficient evidence of acceptable methodology in measurement and a clear relationship between therapy and any observed change in HRQoL (Leidy et al., 1999a,1999b).

HRQoL refers to how health impacts an individual's ability to function and his or her perceived well-being in physical, mental and social domains of life (Coons et al., 2000). It is a multidimensional construct and, in general, the most commonly measured

HRQoL dimensions include physical functioning, social/role functioning, and mental health/emotional well-being.

In randomized clinical trials, results obtained from HRQoL measures are increasingly being used to complement safety data, survival rates and traditional indicators of clinical efficacy as a measure of the value of medical interventions (McHorney, 1997). HRQoL indicators have been used to compare treatments where efficacy is similar but side effect profiles or effects on functional abilities differ. HRQoL data can also be used to describe and monitor the health of large populations, to establish priorities and allocate resources (program/policy evaluation), and in clinical practice by helping to select between treatment options and monitor patient outcomes (McHorney, 1999).

There are a wide range of HRQoL instruments being used currently. These include both generic measures as well as those designed specifically to measure the health status of persons with a particular illness or condition. These measures differ greatly in content, breadth and depth of measurement and therefore their appropriateness for different applications; populations and settings also vary considerably. When selecting between these different HRQOL measures, a variety of factors must be addressed, including conceptual and methodological issues. In addition, careful thought must also be given to the mode of administration.

II. Statement of the Problem

Traditionally, health related quality of life data have been collected via patient self-report using paper-based questionnaires. Responses are then manually entered into an electronic database. However, with improvements in technology and the development of personal computers, a more efficient alternative method of data collection has become available. Questionnaires used for health status assessments can be converted to an electronic format, allowing the respondent to answer questions using user-friendly computer interfaces, such as a touch screen monitor. Electronic completion of self-reported questionnaires has many potential advantages. It allows for instant data entry and processing, facilitates easy storage and retrieval of data and is well suited to skip patterns and branching logic. Response errors can also be eliminated by presenting questions individually on the screen thereby minimizing cognitive errors, which may result in missing data points. Data obtained using the touch screen is transferred electronically to a database thereby eliminating the labor intensive and error prone process of secondary data entry. Write-in responses can also be avoided and data can be analyzed fairly quickly. Previous studies have demonstrated the acceptability of using computers to administer HRQoL instruments (Buxton et al.,1998; Pouwer et al., 1998; Taenzer et al., 1997). Drummond et al (1995) compared completion of a paper-based quality of life questionnaire with a computer version and found that data collected electronically was more complete and easier to handle. Also, several studies have shown that patients completed the electronic questionnaires quicker than their respective paper versions (Taenzer et al., 1997; Velikova et al., 1999b).

However, before the widespread application of the automated methods of data collection can be recommended, it is essential to ensure that the data collected using this form of computer technology is comparable to data collected via traditional pencil and paper questionnaires. If an instrument is to be administered in a manner other than the one in which it was originally developed and tested, it is important to ensure that the alternate forms of the instrument demonstrate measurement equivalence. An example of one such instrument for which an electronic version has been developed is the EQ-5D.

The EQ-5D, a multi-attribute, preference-based HRQoL instrument, is receiving increasing attention in the US, particularly since it has been included as part of the Medical Expenditure Panel Survey (MEPS) conducted by the US Agency for Healthcare Research and Quality (AHRQ). The EQ-5D has two parts. The first is a descriptive system that classifies respondents into one of 243 distinct health states. The current descriptive system consists of five dimensions. Each dimension has three levels, reflecting "no problems," "some problems," and "extreme problems." A scoring function assigns a value to self-reported health states from a set of preference weights that have been empirically derived.

The second part of the EQ-5D, which is the focus of this study, is a 20 cm vertical visual analog scale (VAS) that has endpoints labeled "best imaginable health state" and "worst imaginable health state" anchored at 100 and 0, respectively. Respondents are asked to indicate how they rate their own health by drawing a line from an anchor box to that point on the VAS (i.e., EQ-VAS) which best represents their own health on that day. The EQ-VAS has been operationalized in ways that depart from the original version. For example, although still paper-based, the EQ-VAS in the MEPS 2000 self-administered questionnaire (SAQ) was modified from the original 20cm vertical VAS to an 11.2cm horizontal VAS. In addition, that approximate length and horizontal orientation has been adopted by Assist Technologies in its touch screen adaptation of the EQ-5D. However, there exists no prior research that has examined the measurement equivalence of the horizontal and vertical orientations of the EQ-VAS.

III. Purpose of the Research

The primary purpose of this research was to test the equivalence of the original paper-based vertical version of the EQ-VAS with the touch screen-based horizontal version. The study will also examine the scaling assumptions, data quality, reliability and validity of the electronic version of the SF-36.

IV. Research Objectives

This investigation had the following nine research objectives:

<u>Objective 1</u>

To determine the minimally important difference (MIID) on the EQ-VAS, which, in this situation, is the maximum difference between within-subject scores on the two versions (vertical vs. horizontal) that would still allow the two approaches to be considered equivalent.

Objective 2

To evaluate the measurement equivalence between the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal version.

Objective 3

To determine if measurement equivalence between the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal version exists across the EQ-VAS scale (i.e., for every level of health status).

Objective 4

To examine the effect of order of administration of the two versions of the EQ-VAS on measurement equivalence between the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal version.

Objective 5

To examine the effect of time (first vs. second administration) on measurement equivalence between the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal version.

Objective 6

To examine the inter-relationships between demographic variables (e.g., age, education level, sex, race/ethnicity), health status variables, prior computer experience and mode of administration of HRQoL assessments.

Objective 7

To examine the data quality and scaling assumptions of the touch screen version of the SF-36.

Objective 8

To determine the item and scale reliabilities of the touch screen version of the SF-36 and compare it to the original paper version of the SF-36.

Objective 9

To examine the construct validity of the touch screen version of the SF-36 and compare it to the original paper version of the SF-36.

V. Research Hypotheses

The following null hypothesis will be tested in this research:

Hypothesis One (H_{01})

There is no difference in scores obtained on the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal EQ-VAS version.

Hypothesis Two (H_{02})

There is no difference in reliability estimates obtained for the touch screen version of the SF-36 when compared with those obtained for the general US population.

Hypothesis Three (H_{03})

There is no difference in the validity of the touch screen version of the SF-36 when compared with that of the originally validated paper version.

CHAPTER 2

LITERATURE REVIEW

This chapter is a review of the literature relevant to this research. The literature review covers four broad areas:

I. An overview of health-related quality of life assessment

II. Review of visual analog scales and their role in HRQoL assessment

III. The EQ-5D instrument and its measurement properties

IV. Modes of administration of HRQoL measures

V. Measurement equivalence between alternate modes of administration

I. Health-related Quality of Life (HRQoL) Assessment

Importance of Measuring HRQoL

The field of outcomes research has progressed substantially over the past two decades, resulting in some fundamental changes in health care research, policy, and practice. One of these changes is reflected in the measurement of health, which has changed over time. Traditionally, measures of survival, clinical endpoints, and disease specific symptoms and problems were primarily used as outcome measures to assess the effectiveness of health care interventions (Sullivan, 1966). However, in recent years, it has been recognized that determination of the value of health care services requires an expansion of metrics beyond the traditional measures of mortality and major morbidity.

Many health care interventions significantly affect patient functioning and well-being while having little measurable impact on risk of death or disability. This is particularly true in the case of chronic illnesses that involve palliative therapies or therapies that may prevent further deterioration; the relevant measure of health status should include other domains of health other than length of life (Bergner and Rothman, 1987). Therefore, a broader range of patient outcomes need to be measured in order to understand the true benefits and risks of health care interventions.

These ideas however are not new. In 1948, the World Health Organization (WHO) defined health as a state of complete physical, mental, and social well-being and not merely the absence of disease and infirmity. This definition continues to serve as the broad conceptual framework for health. Those who argue that the goal of medical care is not just to prolong life and prevent disability but also to relieve distress and preserve function and well-being are now endorsing these objectives. Hence, other supplemental measures have been identified to provide a more multidimensional profile of health. Central to the idea of expanded measures is the concept of "Health-Related Quality of Life" (HRQoL).

Health-related quality of life (HRQoL) measures provide a patient perspective on the impact of disease and its treatment on the patient's daily life and functioning. These measures complement and extend information provided by clinical endpoints and add relevant outcome data not captured by them. Evaluating HRQoL helps translate clinical improvement into patient-centered outcomes. Additionally, the US Food and Drug Administration (FDA) has begun to allow quality of life promotional claims for pharmaceuticals that demonstrate sufficient evidence of acceptable methodology in measurement and a clear relationship between therapy and any observed change in HRQoL (Leidy et al., 1999b). Hence, the interest in measuring health-related quality of life (HRQoL) as an important treatment outcome has been growing (Spilker, 1992).

Conceptualizing HRQoL

Numerous definitions of HRQoL have been proposed in the literature. Conceptions relevant to health and quality of life are diverse, are scattered through many disciplines, and use many different labels (Deyo et al., 1989). The terms functional status, health status, quality of life and health-related quality of life are often used interchangeably but these terms have subtle but important differences with respect to dimensionality, perspective and scope (Bergner, 1989).

The term quality of life represents a broader range of human experiences related to one's overall well-being. It is an all-inclusive concept and incorporates all factors (e.g., physical, physiological, social, job satisfaction, spirituality, housing, economic and political) that impact an individual's life (Revicki et al., 2000). It has meaning beyond an individual's health. Therefore, outcomes of disease and treatment generally are evaluated under the more restricted concept and term "health-related quality of life" (HRQoL).

HRQoL represents those parts of quality of life that directly relate to an individual's health (Spilker, 1992). HRQoL refers to how health impacts an individual's ability to function and his or her perceived well-being in physical, mental and social domains of life (Coons et al., 2000). It is patient-oriented, multidimensional and variable over time (Schipper, 1990). Since, the purpose of HRQoL evaluations is to examine how disease and medical interventions impact a person's daily functioning, there is nothing more relevant than basing these decisions on the patient's own HRQoL assessment.

The actual dimensions of HRQoL vary from one instrument to another and are dependent on the characteristics of the specific study. Definitions of HRQoL frequently must be deduced from the dimensions that are assessed (Bergner, 1989). However, in general, the most commonly measured HRQoL dimensions include physical functioning, social/role functioning, and mental health/emotional well-being. Global perceptions of function and general well-being have also been included in many HRQoL measures (Shumaker and Berzon, 1995).

Another feature of HRQoL measurement is that it usually refers to an individual's health status at any given point in time. HRQoL is a point in time measure and therefore may change in the next day, month or year. The variable nature of health-related quality of life affects the methodology involved in its assessment. The timing of measurement can significantly influence whether a health care intervention will be shown to affect a person's health related quality of life.

Applications of HRQoL

HRQoL is increasingly being recognized as an important outcome in studies evaluating pharmaceutical agents, medical technology, and the organization of health care. In randomized clinical trials, results obtained from HRQoL measures are increasingly being used to complement safety data, survival rates and traditional indicators of clinical efficacy as a measure of the value of medical interventions (McHorney, 1997). It is the outcome of choice for conditions in which there are no physical, physiological, or biochemical markers of disease activity, or for situations in which treatment effectiveness is based primarily on patient perception (e.g., pain related conditions). HRQoL indicators have also been used to compare treatments where efficacy is similar but side effect profiles or effects on functional abilities differ (Leidy et al., 1999a, 1999b).

HRQoL data can also be used to describe and monitor the health of large populations, to establish priorities and allocate resources (program/policy evaluation), and in clinical practice by helping to select between treatment options and monitor patient outcomes (McHorney, 1999). Efforts are increasing to incorporate HRQoL as an indicator of quality of care and clinical effectiveness, and payers are beginning to consider HRQoL in reimbursement decisions (Weinberger et al., 1996).

HRQoL Measures

The rapid evolution of outcomes research during the last two decades has led to an increased emphasis on HRQoL assessments. There are a wide range of HRQoL instruments being used currently. These include both generic measures as well as those designed specifically to measure the health status of persons with a particular illness or condition. These measures differ greatly in content, breadth and depth of measurement and therefore their appropriateness for different applications; populations and settings also vary considerably.

Generic measures are non-disease specific and therefore broadly applicable. They can be used across patient populations, irrespective of the underlying condition and therefore allow for a broad comparison of the relative impact of various health care programs. Generic measures may, however be less responsive to changes in specific conditions. Generic measures are further categorized into health profiles and utility measures (Cramer and Spilker, 1998). Health profiles refer to instruments in which the dimensions are not aggregated. Items within a dimension as well as the dimension itself are given equal weight. On the other hand, utility measures have their basis in economic and decision theory, and reflect the preferences of patients for different health states. Preference weights are derived by using utility elicitation techniques and HRQoL is summarized as a single utility-based score.

Disease-specific measures are those that were developed specifically for any particular disease state. These measures may include only those dimensions of HRQoL that are relevant to the area of study. These measures may be more responsive than generic measures in disease states for which they were developed but cannot be used to compare across populations.

The choice of an HRQoL measure depends on a variety of factors including conceptual and methodological issues. The selection of the HRQoL instrument from the pool of potential measures is based on an evaluation of each instrument's psychometric characteristics. In general, a combination of generic and disease-specific instruments is used depending on the objective of the study. In addition, careful thought must also be given to the mode of administration.

II. Visual Analog Scales

Description

Visual analog scales (VAS) are a type of rating scale that have a long history in psychometric research and in the measurement of health. The VAS began to appear in the literature with increasing frequency in the 1960s. Although originally conceptualized as a method to be used for the evaluation of individuals by raters, the use was extended to the rating of subjective phenomena by individuals experiencing the phenomena of interest (Wewers and Lowe, 1990). The VAS has been used to measure a variety of subjective phenomena including mood, anxiety, alertness, craving for cigarettes, quality of sleep, functional abilities, nausea, fatigue, dyspnea and pain. In recent years, the visual analog scale has been incorporated in HRQoL instruments such as the EQ-5D in order to measure general health status. They have been used commonly as a means to measure both individuals' rating of their own health and their preference for other, hypothetical health states.

It is usually depicted as a straight line of specified length with verbal descriptors at each end. The VAS has traditionally been used as a horizontally oriented scale, generally 100mm (or 10 cm) in length without markings and with the endpoints being anchored by a word or short phrase to indicate the extremes of the variable being measured (Gift, 1989). While there is no clear scientific reason for this standard length, there is historical precedent (Freyd, 1923) and this length allows for easy quantitative reading. There have however, been numerous variations to this approach including modifying the length of the line, altering the labels at the ends of the line, addition of scale marking which may or may not have numeric/verbal descriptors and/or changing the orientation of the scale.

A VAS which has verbal descriptors (e.g., mild, moderate, severe) placed at intervals along the line is known as a graphic rating scale (GRS) whereas in numeric rating scales, the VAS is calibrated with numbers ranging from 1 to 20. The low end of the scale is to the left in a horizontally oriented scale and at the bottom of a vertically oriented scale. Subjects respond to the VAS by placing a mark on the line, between the two extremes, at a point that best represents their current perception of a given phenomenon. The scale is scored by measuring the distance from the low end of the scale to the subject's mark, on a predetermined measurement interval. The most commonly chosen interval is millimeters with a 10 centimeter line, producing a 100-point scale. The VAS has been used in a manner comparable to a Likert scale, with numbers replacing verbal descriptors. It has also been used to derive preference weights, as an alternative to methods such as paired comparisons, time-trade off (TTO) and standard gamble (SG).

Characteristics of VAS and Issues Regarding its Use

A VAS, because of the apparent simplicity of the technique and its adaptability to a wide range of research settings, is an attractive measurement options. The VAS is simple and quick to construct, easy to administer and score, relatively simple to understand and lends itself to self-completion thereby making it appropriate for a variety of subjects (Little and McPhail, 1973; Luria, 1975; Morrison, 1983; Rampling and Williams, 1977). Those with sight impairment, such as dilated pupils, can see it easily, especially if the anchors are printed with large type. In some cases, the anchors can also be read to the subject. Even in cases where patients have been medicated for pain, they have demonstrated sufficient manual dexterity required to use a VAS (Revill et al., 1976). Proponents of the VAS have also claimed that it is suitable for frequent and repeated use, requires little motivation for completion by subjects and is suitable for use by untrained staff (Folstein and Luria, 1973; Rampling and Williams, 1977). Visual analog scales also have very low usage of words and therefore the vocabulary level of the respondents is not a very important consideration.

Although ease of use is a frequently cited advantage, Carlsson (1983) questioned this assumption pointing out that the VAS requires an ability to transform a complex subjective experience into a visual-spatial display, involving perceptual judgment and accuracy. Some people find it difficult to comprehend the idea of converting a subjective sensation to a straight line. Huskisson (1974) found that seven percent of subjects were unable to use the VAS after a single explanation as compared to only three percent of subjects who needed more than one explanation to use a graphic rating scale. Researchers have attempted to overcome this limitation by either teaching subjects to use the scale or by providing detailed written instructions at the top of the scale (Guyatt et al., 1987).

Kremer and colleagues (1981) reported failure rates of 11% and claimed that the completion of the VAS is difficult for many patients in spite of careful instructions and practice, diminishing the validity of the method. Age and the loss of ability to think abstractly, mental disorganization and confusion, and perceptual skills and memory have all been suggested as factors which may contribute to respondent error (Carlsson, 1983; Hornblow and Kidson, 1976; Kremer et al., 1981). Hornblow and Kidson (1976) suggest that the ability to use the VAS for assessment of anxiety as a continuum decreases as the psychiatric symptoms of subjects become more severe. In contrast, Carlsson (1983) explored the impact of learning, memory and perceptual judgment on the reliability and validity of the VAS and found no relationship between these variables and the subject's ability to make reliable pain assessments on the VAS. It has also been recommended to present VAS in a horizontal rather than a vertical format, as Scott and Huskisson (1976) found a seven percent failure rate for the vertical VAS, which was reduced by changing to the horizontal mode.

Other pragmatic concerns include accurate reproduction of the instrument itself. For example, when visual analog scales are administered via paper format, distortion to the scales has been reported during photocopying (Lorig, 1983). Some machines have been known to decrease its size while others, after warming, increase its size; therefore, printing is recommended. Similarly, the angle at which the subject views the VAS may alter placement of the mark due to perceived variations in the length of the line. This may be particularly true for hospitalized patients who complete a VAS in bed. The VAS has been primarily designed to be used with a seated subject marking the scale, but not all subjects may be able to use the scale in this manner. It remains to be determined whether using the scale in the supine position influences the results obtained (Gift, 1989).

Additionally, other perceptual/motor issues include the effects of medications on comprehension, eye-hand coordination, and other limitations caused by the subject's clinical status and/or treatment (Wewers and Lowe, 1990).

The VAS provides a continuous scale for magnitude estimation and subjects who perceive their ratings to fall between categories (i.e., those specified in a Likert scale or a graphic rating scale) prefer the continuous nature of the VAS (Aitken, 1969). End aversion is a well-known measurement bias that affects scales scored on a continuum such as visual analog scales. It refers to the reluctance of some respondents to use the extreme categories of a category scale or to use the portion of the scale near the ends of a continuous scale (Torrance et al., 2001).

On the VAS, the anchors usually represent the extremes of the characteristic under study but in many cases, researchers do not agree on what constitutes the extremes. Even when the continuum of the characteristic has been agreed upon, the exact wording of the anchors is a source of considerable debate. For example, "no pain" is easily understood by subjects, whereas "pain as bad as it can be" has no absolute value and could be argued to be immeasurable. Therefore, any mark along the line between these two extremes is totally dependent upon the subject's unique interpretation of the maximal value and based on the subject's experience to date. This can be problematic particularly in repeated measures design, where intra-subject differences on repeated observations may reflect subjective changes in the perceived range of the phenomena represented by the VAS rather than "real" differences between the experience of the phenomena at given observation points. In other words, the metric of the VAS for the individual subject has changed due to the experience (Wewers and Lowe, 1990).

While most visual analog scales described in the literature are single-item scales, a number of multi-item visual analog scale-based measures have also been developed. Some of these measures have been used in exploratory research to define and weight relevant variables and to explore relationships between different dimensions of subjective experience (McCormack et al., 1988). Well-known examples of this approach include the use of two visual analog scales to capture the intensity and distress dimensions of pain and Padilla's multidimensional measure of quality of life (Padilla et al., 1983). In some cases, this can be problematic. When multiple horizontal scales have been used, subjects had a tendency to mark all the scales down the middle. Also, since the dimensions may have shared variances, it can impact statistical tests (Cella and Perry, 1986).

Psychometric Properties of the VAS

Validity

Validity for the VAS has been established in a variety of studies, the majority of which have focused on visual analog scales developed to measure mood or pain. Examinations have included assessment of construct, discriminant and criterion-related validity. However, the majority of investigators who have evaluated the VAS have utilized a criterion-related approach. In most of the studies, an established instrument was used as the criterion. The McGill pain questionnaire has been used to establish the validity of the VAS as a measure of pain, while the Beck Depression Inventory and the depression scale of the Symptom Checklist-90 have been used to establish its validity as a measure of depression (Davies et al., 1975; Melzack, 1975). Several other correlation studies have also been reported in the literature

Some studies tested the ability of the VAS to discriminate between subject groups based on objective criteria, as a means of validating the VAS. Gift et al. (1986) established the validity of the VAS as a measure of dyspnea by examining the degree of airway obstruction in patients with obstructive lung disease. The study found the VAS to have good discriminant validity. Similar results were also obtained when comparing non-patient volunteers with oncology in-patients receiving chemotherapy (Padilla et al., 1983). Joyce et al. (1975) found that VAS-based pain scores differentiated between responses to two analgesics, and between high and low doses of each of the drugs. Price et al. (1983) also reported the validity of a VAS for pain when used to discriminate different intensities of experimentally induced pain. Robinson et al (1975) used measures of food intake to validate a VAS for hunger for use with both normal and psychiatric patient populations. High correlations have been reported between the VAS and verbal/ numerical rating scales and 4-point scales (Ohnhaus and Adler, 1975; Scott and Huskisson, 1976; Woodforde and Merskey, 1972; Downie et al., 1978; Elton et al., 1979; Kremer et al., 1981).

Reliability

The reliability of a linear analog as a method of assessment depends, first, upon visual and motor co-ordination, that is, the ability of the patient to place a mark where he or she intends to put it. Second, there is a presumption that the same rating will be given for the same perception of health status on more than one occasion. Using the test-retest method, subjects have been shown to be reliably reporting measures of subjective sensation (Brazier et al., 1999; Luria, 1975).

The time interval between initial testing and the retest has shown to have an impact on reliability estimates. Ratings that were repeated after an hour had higher reliability than those repeated after two weeks (Carlsson, 1983; Padilla et al., 1983). Test-retest reliability has been assessed for the VAS on the basis of scores obtained at a variety of intervals, including six months apart, three weeks apart, and two hours apart (Hornblow and Kidson, 1976; Luria, 1975; Morrison, 1983). The studies that did recognize the continuously changing quality of subjective experience employed experimental designs that attempted to maintain the relevant subjective experience at a constant level during the test-retest procedure (Robinson et al., 1975). Revill et al. (1976) evaluated the reliability of a VAS for pain, designed to measure pain in childbirth, by recording repeated measures of feelings about remembered events. There were very high correlations between initial rating and the 5 minute repeat rating ($r = 0.95$) and 24 hour repeat rating ($r = 0.97$). There are several other factors that influence the reproducibility of the VAS including the length of the VAS line. Revill et al. (1976) reported that the VAS for measuring pain with 10, 15 and 20 cm lines demonstrated significantly lower variability than a 5 cm line. The smallest line length, 5cm, had the largest error and the authors recommended using one of the other lengths (i.e., 10, 15 or 20 cm). Another factor found to influence the reproducibility of the VAS is the intensity of the sensation being experienced. Dixon and Bird (1981) investigated a vertical 10 cm VAS and noted that the reproducibility was found to be variable along its length. Subjects were found to be better able to reproduce marks along the extremes or in the center of the scale, while the region +/- 2 cm of the midpoint was the least reproducible.

In a study to compare vertical and horizontal VAS for the measurement of chronic pain, no significant difference was identified between the distributions of scores on the two scales although scores on the horizontal VAS tended to be slightly lower. The horizontal VAS also produced less variability and more closely approximated a normal distribution (Scott and Huskisson, 1976; Sriwatanakul et al., 1983). In addition, subjects have expressed preferences for horizontal as compared to vertical scales (Sriwatanakul et al., 1983). Inter-rater reliability estimates have also been reported when the VAS has been used by more than one rater (Little and McPhail, 1973). Several studies have compared scores on patient-rated and physician-rated visual analog scales for depression to establish inter-rater reliability (Little and McPhail, 1973; Malpas et al., 1974).

Sensitivity

In addition to reliability and validity, the sensitivity of a scale (i.e., discriminating capacity) is an important consideration in measurement. Small increments of change in phenomena have been demonstrated with the VAS in several studies (Miller and Ferris, 1993). The VAS has been found to be a more sensitive measure of subjective sensations than a graphic rating scale or a four-point rating scale (Joyce et al., 1975; Ohnhaus and Adler, 1975; Scott and Huskisson, 1976). Some studies have indicated that the vertically oriented VAS is more sensitive, produces higher scores, and is easier for subjects to use than the horizontal scale (Scott and Huskisson, 1979; Gift et al., 1986; Sriwatanakul et al., 1983). The use of gradations on the scale has been shown to reduce its sensitivity, a situation that could be clinically significant for some phenomena (Gift, 1989). When numbers or verbal labels are used to define intermediate points, it results in clustering of scores around a preferred digit or around the verbal descriptors (Aitken, 1969; Huskisson, 1974; Scott and Huskisson, 1976).

Ceiling effects have been reported on the VAS in some subjects when measuring change over time (Little and McPhail, 1973). Little and McPhail (1973) noted that on several occasions while using a VAS for depression as a repeated measure, they found that patients scored the VAS at the maximum level of depression before the completion of the study, thus leaving themselves no room to record a higher score should their mood worsen. This issue was addressed in a study by Borg (1982) in which the upper end of the scale was opened and subjects were allowed to indicate amounts higher than the maximum.

Analysis of VAS Data

The choice of statistical procedures for use with the visual analog scales is influenced by whether the VAS is treated as an ordinal, interval or ratio scale. There is no disagreement in the literature as to whether the VAS provides ordinal scores that can be ranked and analyzed using non-parametric, rank based methods (McCormack et al., 1988). However, for the most part VAS scores have usually been treated as interval level data. The arguments advanced in favor of treating VAS scores as interval level data suitable for group comparisons are, first, that the assumptions underlying the acceptance of VAS as interval scales are shared with many of the accepted forms of psychological measurement and, second, when used in this manner they are able to discriminate between groups in the expected direction and to provide replicable results (McCormack et al., 1988).

A few studies conducted by Price and colleagues (1983b, 1987, 1994) have claimed that the VAS in addition to being a valid and reliable measure yields numerical values on a ratio scale. The authors demonstrated that the stimulus-response function derived from using a VAS was well described by a power function and this power function was consistent with separate judgments of ratios of pain intensity. Both of these separate but interrelated results served as evidence that visual analog scales have ratio scale properties. In recent years, Ludington and Dexter (1998) have suggested that VAS scores are ratio data, where zero represents a true absence of the phenomena of interest and that the scale has linear scale properties. Myles and colleagues (1999) provided additional support for the ratio scale properties of the VAS in post-operative patients with mild to moderate pain.

The choice of statistical test to analyze VAS data depends largely on the distributional assumptions about the data. Nonparametric approaches based on the rank order of data have an advantage. They do not incorrectly suggest that a difference exists among groups more often than specified by the nominal false-positive or type 1 error rate, regardless of whether the statistical distribution of the VAS measurement is normal or not (Dexter and Chestnut, 1995). Scott and Huskisson (1976) have recommended that nonparametric methods be used to analyze VAS measurements. Scores obtained with the VAS have been shown to be skewed (Aitken, 1969). A variety of techniques have been used to correct this, however, scores obtained with or without transformation show no differences in sensitivity or results obtained.

On the other hand, parametric procedures that assume that VAS measurements follow a normal distribution also have an advantage. If the distribution of the data were normal, the tests would have the highest power (i.e., ability) to detect differences among groups. However, even if the distribution were not normal, parametric tests usually give false-positive rates that are close to the nominal value (Fleiss, 1986). Therefore, Philip (1990) recommended that parametric statistics be used to analyze VAS data. Aitken (1969) recommended using parametric methods with VAS scores only after the data had been transformed to achieve a normal distribution. In a literature review of papers that analyzed VAS measurements in anesthesia, 54% of the studies used a nonparametric test (Mann-Whitney or Kruskal-Wallis), and 34% used a parametric test (t-test or ANOVA) (Mantha et al., 1993).

Dexter and Chestnut (1995) in a study examining the choice of statistical tests to compare VAS measurements found that no statistical test, either parametric or nonparametric incorrectly suggested that a difference existed among groups, when there was no difference, more often than the expected rate. The authors concluded that parametric tests without transformation were good choices to compare VAS measurements among groups since parametric methods, when compared to nonparametric tests, have greater power and more theory to guide experimental design and post hoc analysis.

III. The Development and Use of the EQ-5D

Background

In 1987, the EuroQol Group first met to determine the feasibility of jointly developing a standardized non-disease-specific instrument for describing and valuing health-related quality of life. The purpose was to develop an instrument that would complement other forms of quality of life measurement and would facilitate the collection of a common data set for reference purposes. The EuroQol Group has a global focus, hence of importance to the group was the capacity to generate cross-national comparisons of health state valuations (Brooks, 1996).

The EQ-5D instrument has been in the public domain since 1990 and is one of a small number of available generic index measures. Empirical work using the original six-dimensional format has been published in a number of countries. However, as a consequence of further development work, in 1991, the instrument was modified to a standard five-dimensional format. It has been cross-culturally translated for use in a variety of populations.

The EQ-5D Instrument

The EQ-5D is a generic HRQoL instrument and was designed for self-completion by the respondent. It consists of two main parts: a health status descriptive system and a visual analog scale (EQ-VAS). The descriptive system records self-reported problems on each of five domains: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. These dimensions/domains were selected following a detailed examination of the descriptive content of other existing health status measures including the Quality of Well-Being Scale, Sickness Impact Profile, Nottingham Health Profile, and the Rosser Index (The EuroQol Group, 1990). For each domain there are three response levels corresponding to no problems, moderate problems, and extreme problems. The resultant descriptive system is presented in Table 2.1.

Data from the EQ-5D descriptive system can be presented in two distinct forms. Raw scores on the five domains can be used in combination (as a health profile) and defined by a five-digit descriptor. For example, state 21132 would indicate some problems with mobility, no problems with self-care or usual activities, extreme pain or discomfort, and moderate anxiety or depression (Rabin et al., 2001). Theoretically, the design of the descriptive system identifies as many as 243 (3^5) unique health states, although a small number of these health states are not plausible in reality. This information is primarily used for descriptive purposes in clinical studies (Dolan, 1997).

Another way to present the data is in the form of a health index, which on a scale where 0 is equal to death and 1 equals full health. Scoring methods have been developed to assign a preference value or weight to each health state generated by the descriptive system (Dolan et al., 1996; Dolan and Roberts, 2002). These preference weights have been obtained in a number of countries using either VAS or time trade off (TTO) preference-elicitation methods, generally from samples from the general population (Dolan, 1997). Preference-based EQ-5D scores can be used to calculate quality-adjusted life-years (QALYs) for cost-utility analysis of health interventions (Drummond et al., 1997).

Table 2.1: EQ-5D Descriptive System

By placing a check mark in one box in each group below, please indicate which statements best describe your own health state today.	
Mobility	
I have no problems in walking about	❑
I have some problems in walking about	❑
I am confined to bed	❑
Self-Care	
I have no problems with self-care	❑
I have some problems washing or dressing myself	❑
I am unable to wash or dress myself	❑
Usual Activities (*e.g. work, study, housework, family or leisure activities)*	
I have no problems with performing my usual activities	❑
I have some problems with performing my usual activities	❑
I am unable to perform my usual activities	❑
Pain/Discomfort	
I have no pain or discomfort	❑
I have moderate pain or discomfort	❑
I have extreme pain or discomfort	❑
Anxiety/Depression	
I am not anxious or depressed	❑
I am moderately anxious or depressed	❑
I am extremely anxious or depressed	❑

The second part of the EQ-5D instrument is the visual analog scale (EQ-VAS). The EQ-VAS is a vertical, graduated (0–100 points) 20 cm 'thermometer', with 100 at the top representing 'best imaginable health state' and 0 at the bottom representing 'worst imaginable health state.' The VAS approach was chosen due to its relative simplicity for scaling purposes and it is intended to be a generic measure with interval scale properties. The EQ-VAS allows respondents to rate their health on the day of administration of the instrument (Brooks, 1996; Rabin and de Charro, 2001). The visual analog scale score can be interpreted directly as a quantitative measure of the patient's valuation of their own general health status (Hurst et al. 1997). The EQ-VAS score can be used as a measure of clinical outcome, using individual respondent's own judgment (Gudex et al., 1996)

Use of the EQ-5D

As noted before, the EQ-5D instrument is generic or general. It is not aimed at one particular disease or treatment but rather designed for use across a wide range of health interventions. The EQ-5D instrument provides a convenient way of collecting data about a respondent's own rating of his or her current health status. The descriptive system can be used as a profile, and allows for comparisons in a cross-sectional study with other patients or the general population. It can also be used to examine the health of a cohort of individuals over time with or without a health care intervention as part of a longitudinal study. Also, the self-rated health status information based on EQ-VAS scores can be used longitudinally to compare changes in health status of an individual patient or groups of patients. The EQ-5D instrument can be used to generate a tariff of individual or social valuations of health states. Using preferences elicited from the general population, a value can be assigned to any classified health state and then (with or without combining life years) used for comparison among individuals or groups. The descriptive information and/or the assigned valuations can be analyzed by demographic variables such as age, sex, or education.

Measurement Properties of the EQ-5D

Validity

The construct validity of the EQ-5D has been tested by examining the correlation between the EQ-5D and other established general health status and condition-specific instruments. Brazier et al. (1993) compared the EQ-5D instrument with the SF-36 and found abroad ordinal agreement between the two measures. Both sections of the EQ-5D have demonstrated moderate correlations with each other (r = 0.64) and with the physical functioning component score of the SF-12 (r= 0.65). These relationships are as expected since the majority of the EQ-5D is concerned with physical health. Small to moderate correlations have been observed for the EQ-5D index score (r = 0.54) and EQ-VAS (r = 0.44) with the mental component score of the SF-12 (Haywood et al., 2002a).

In a study of rheumatoid arthritis patients aimed at-testing the construct validity of the EQ-5D, Hurst et al. (1994) concluded that the EQ-5D is simple to use, showed clinically relevant correlations with other condition-specific instruments, and seems to measure both the current clinical symptoms of rheumatoid arthritis patients and changes in some of these symptoms over time. Johnson and Coons (1998) provided evidence for the convergent and discriminant validity of the EQ-5D by examining the relationship between EQ-5D scores and SF-12 component scores. Relationships between the EQ-5D dimensions and the PCS-12 and MCS-12 were as hypothesized. The EQ-VAS had a stronger relationship with the PCS-12, indicating that people give more consideration to their physical health in providing a global health rating. Wu et al (2002) examined the validity of the EQ-5D in a sample of subjects with advanced HIV disease. Each dimension of the EQ-5D was most highly correlated with the similar dimension measured concurrently using the MOS-HIV, which provides evidence for construct validity. All the hypothesized relationships of at least moderate correlations (0.3-0.6) between EQ-5D items and MOS-HIV scales were generally confirmed (Wu et al., 2002a).

The validity of the EQ-5D has also been examined in relationship to severity of illness and measures of socioeconomic status (Badia et al., 1996; Essink-Bot et al., 1995; Hurst et al., 1994; Richards and Irving, 1997). When EQ-5D items have been analyzed independently, they have been shown to discriminate between patients with and without disease (Essink-Bot et al., 1997). The EQ-5D index score has demonstrated predicted relationships with functional class and socioeconomic status with higher values being associated with higher functional class, employment, higher socioeconomic status and greater independence (Hurst et al., 1994). Respondents with higher levels of education and household income have been found to report better health (Brazier et al., 1993). Patients with significant chronic health problems report poorer health status than subjects without, at least in the functional dimensions. Higher scores on the EQ-VAS have been significantly associated with living independently or being employed. Compared with those unable to work due to ill health, patients in work have significantly better levels of health on the EQ-5D (Haywood et al., 2002b).

Reliability

When assessing the performance of a measure, it is important to know the extent to which any change reflected by the measure is due to 'real' change and not just due to measurement error. This characteristic was assessed by calculating test-retest reliability estimates, which measure the consistency of the instrument over time. It refers to the ability of the measuring instrument to produce the same result on two or more occasions, assuming that the characteristic under study has remained unchanged. The time interval is one of the most important considerations when assessing the test-retest reliability of a measure. If the interval is too short, respondents may remember what they answered on the first occasion, while if it is too long, the characteristic that is being measured might have changed.

The test-retest reliability is summarized by calculating a reliability coefficient, which estimates the proportion of total variance in the scores that is due to 'true' variance. The reliability coefficient ranges from 0 to 1.0, where 0 indicates no reliability and 1.0 indicates perfect reliability (no measurement error). There are a number of methods to calculate such coefficients but the intraclass correlation coefficient (ICC) is usually considered the most appropriate for continuous data and the Cohen's kappa or weighted kappa coefficient is considered best for nominal data (Cohen, 1960; Cohen, 1968; Fleiss and Cohen, 1973).

The test-retest performance (reliability and responsiveness) of the EQ-5D descriptive system and EQ-VAS has been examined in both the general population as well as in clinical samples. The same group of individuals are administered the EQ-5D on two occasions separated by a time period. In the 1993 MVH survey, a representative sample of 221 individuals completed the EQ-5D on two occasions, three months apart (Dolan, 1994). In terms of reliability, on the EQ-5D descriptive system, the mobility dimension was the most reliable (kappa coefficient = 0.75 [SE = 0.06]) and the pain dimension was the least reliable (kappa coefficient = 0.57 [SE = 0.07]). Eighty-four individuals (38% of the sample) changed their response on at least 1 dimension between time 1 and time 2, with 132 changes in total. The majority of cases who changed reported improved health at time 2 (61%) and also only changed on 1 dimension (63%). No individuals changed on all five dimensions. On the EQ-VAS, the mean score improved from 81.2 (SD = 17) to 83.8 (SD =17). Seventy-six percent of the sample showed a change in their EQ-VAS score; 50% of cases improved and 26% of cases worsened. Most of the changes were within ± 10 points. The EQ-VAS showed good reliability with an ICC of 0.84 (95% CI = 0.65 – 0.78). High reliability was associated with better self-rated health status.

Hurst et al (1997) analyzed the performance of the EQ-5D in terms of its validity, responsiveness and test-retest reliability in a sample of 224 out-patients with rheumatoid arthritis. The participants completed the EQ-5D at two time points, 3 months apart. A sub-sample also completed the EQ-5D again over a shorter time period of two weeks. With respect to test-retest reliability, the EQ-5D performed less well for this clinical sample than it had for the general population sample. Reliability estimates were based on participants who reported no change in their arthritis condition over both the three month and two week period. Over the three month test period, the study reported reliability estimates of 0.73 (95% CI = 0.63-0.83) for the EQ-5D index score and 0.70 (95% CI = 0.60-0.80) for the EQ-VAS. When tested over two weeks, reliability estimates were slightly better. The authors reported ICCs of 0.78 (95% CI = 0.60-0.96) for the EQ-5D index score and 0.85 (95% CI = 0.67-1.03) for the EQ-VAS. Haywood et al (2002) assessed test-retest reliability estimates of the EQ-5D for those patients indicating that their general health had remained the same at two weeks on a general health transition question. The intraclass correlation coefficients were 0.80 (95% CI = 0.74-0.85) for the EQ-VAS and 0.83 (95% CI = 0.78-0.87) for the EQ-5D index score (Haywood et al., 2002a).

Sensitivity

In addition to reporting the reliability, the responsiveness of a health measure (i.e., the ability to detect important clinical changes over time) is also an important characteristic (Deyo et al., 1991). The EQ-5D index, EQ-VAS and individual items have been shown to be responsive to changes in health (Badia et al., 1996; James et al., 1996). In the study by Hurst et al. (1997), patient self-reports as to whether their arthritis improved, worsened or remained unchanged over the three month test-retest period was used as the external criterion of change against which to examine the responsiveness of the EQ-5D. Regression models confirmed that change in the EQ-5D index score was predicted by change in disability, mood, pain, patient-assessed disease activity and self-reported drug side-effects, providing evidence that the EQ-5D is measuring clinically relevant change. Also, patients who reported that their arthritis had changed (positively or negatively) experienced significant changes in their EQ-VAS scores compared to those who said their arthritis was unchanged. Furthermore, the EQ-VAS seemed more responsive in patients who reported that their arthritis had improved than those who reported that it had worsened. Overall, the data suggested that both the EQ-5D index score and the EQ-VAS had good responsiveness to self-reported improvements in rheumatoid arthritis.

Other studies have also provided evidence for the responsiveness of the EQ-5D to changes in general health (Haywood et al., 2002b). However, the sensitivity of the EQ-5D might be hampered due to certain limitations. The EQ-5D dimensions have only three response categories/levels and are consequently prone to a considerable ceiling effect. Studies have reported considerably higher (> 85%) ceiling effects for the EQ-5D instrument compared to that (37% to 72%) for the SF-36 (Brazier et al., 1993; Johnson and Coons, 1998; Norum et al., 1996). The EQ-VAS on the other hand has a greater number of possible responses and is slightly more sensitive to change (Wu et al., 2002b). Therefore, the EQ-5D has been suggested to be more suitable for measuring the health of populations with more morbidity, where the distribution of the total score would be less skewed (Brazier et al., 1993; Johnson and Coons, 1998).

IV. Modes of Administration of HRQoL Measures

The primary sources of HRQoL information are standardized questionnaires that usually rely on patient self-report. Standardized questionnaires provide valid and reproducible data, and are comparatively inexpensive. HRQoL instruments may be administered by an interviewer either face-to-face or via the telephone or self-administered. Each method has its own advantages and disadvantages. The former method is resource intensive but optimizes compliance and minimizes errors and missing items. On the other hand, self-administered questionnaires enable uniform administration and unbiased scoring and quantification of data in a cost-effective fashion. Self-administered instruments can, in turn, be either paper-based or computerized.

Electronic Forms of Data Capture (EDC)

The advent of the microcomputer in the 1960s inspired psychologists and medical practitioners to explore the possibility of using computer-assisted patient interviews (Maultsby and Slack, 1971; Slack et al., 1966, 1967; Slack and Van Cura, 1968; Van Cura et al., 1971). HRQoL data have traditionally been collected using paper-based questionnaires and the data is then entered manually into a computer database for analysis. But, in the past 20 years, methods of testing and data capture in the health care industry have experienced vast changes. Computer technology and its application to patient care has been gaining attention within the medical community. The enormous advances of the last two decades with respect to computing power, innovations in mass storage of data, development of data analysis software and wireless communications have all combined to make routine use of computers in the health care industry inevitable. In spite of this, the health-care industry was relatively slower to adopt and incorporate computerized information technology effectively into the work environment. Researchers and practitioners alike have broken away from traditional data collection techniques, involving paper-based tests and surveys, and have begun using computerized methods to collect data. Computers have also been introduced into a number of clinical settings. They have been used for patients to self-report history of alcohol, drug and tobacco use in addiction treatment centers, for preventive health screening in general practitioners' offices, reporting of cancer needs and quality of life in oncology settings, and providing information on prenatal tests (Skinner and Allen, 1983; Campbell et al., 1997; Buxton et al., 1998a; Graham et al., 2000). One of the more frequently reported applications of computer technology in direct patient care is in the presentation of HRQoL questionnaires (Skinner and Allen, 1983; Newell et al., 1997).

Electronic data capture is being used in clinical trials and may soon become the standard because it represents a reliable, high-quality and cost-effective tool for capturing HRQoL data. Portable electronic devices have been developed that allow patients to self-report clinical and HRQoL data using validated instruments delivered on notebook computers. It is essential that the electronic device be suitable for use by a wide range of patients including those who are older, have significant disability, or who have no computer experience. An example of one such design is a touch-screen computer monitor fitted with sensors that allow the computer to be manipulated by the touch of a finger rather than a keyboard or mouse. Touch-screens are one of the most user-friendly interfaces to use, making it the system of choice for a wide variety of applications. They can be less intimidating and easier to use than other input devices, especially for novice users.

Not only do these devices eliminate the usual steps of abstracting patient interview data or computerizing marked responses, but the interface also permits customized, confidential, and private assistance in completely answering queries. HRQoL questionnaires have been converted to an electronic format, allowing the patient to answer questions using the touch-screen rather than the traditional paper-based methods. The computer is able to calculate scores and immediately produce a report.

Advantages and Disadvantages of EDC

Electronic data collection for self-reported questionnaires completed by patients has many potential advantages. Convenience, efficiency, ease of scoring and analyzing responses, decreased administration costs, decreased errors, fewer missing responses, and increased versatility of question presentation are just a few of the advantages cited in the literature. Electronic data collection can enhance completion rates and decrease the level of item non-response. This is particularly valuable in the case of HRQoL questionnaires where data analysis with missing information can be challenging and problematic.

Several studies have been conducted to examine the benefits of moving from paper-based questionnaires to electronic data capture. Drummond and colleagues (1995) compared completion of a paper-based quality of life questionnaire with a computer version and found that the data collected electronically were more complete and easier to handle. Hanscom et al (2002) reported that the overall missing rate for the computer-based questionnaires was approximately half the missing value rate for paper-based questionnaires. Patients using paper questionnaires may skip questions because of small font size and/or other format issues whereas in most cases, computer-based questionnaires are presented in a sizable font, with large, clearly delineated buttons for each response, making questions easier to read for patients with visual impairments. Also, some authors assert that displaying questions one at a time on a computer screen, as opposed to printing numerous questions on a single page of paper, allows respondents to focus more easily on individual questions and reduces the chance that they might skip items or become confused (Honaker and Robbins, 1988).

Another advantage of collecting data electronically is the speedier handling of data. It eliminates the need for the labor-intensive and error-prone process of manually entering primary data into an electronic database. The data is electronically transferred and stored in a database for statistical analysis. This allows for faster time to analysis and data integrity is also enhanced. In addition, on paper questionnaires, respondents can easily mark more than one response to a question, or make a mark halfway between two response options, which can be misinterpreted by data entry technicians or recorded as missing. Using computerized questionnaires, the respondent serves, in effect, as the data entry technician, and is forced to mark only one response per question.

Electronic questionnaires can also be customized and are better suited for instruments with skip patterns and branching logic. An electronic version of a questionnaire can minimize the extent to which responses are questionable or missing by restricting the respondent to a single answer and enforcing a response before allowing the respondent to proceed to the next question. Write-in responses can also be eliminated (Caro et al., 2001).

Further, as respondents enter their responses directly into the database, a greater degree of anonymity is provided, which elicits more complete responses to sensitive questions (Bloom, 1998). Tourangeau and Smith (1996) found that respondents tended to answer in a more "socially desirable way" during face-to-face interviews compared with computer-assisted self-administered questionnaires.

Several studies have shown that it is quicker for patients to complete electronic questionnaires as compared to paper-based versions. Velikova and colleagues (1999) found that it took an average of 9.6 minutes for patients to complete paper versions of the European Organization for the Research and Treatment of Cancer quality of life questionnaire (EORTC QLQ-C30) and Hospital Anxiety and Depression Scale (HADS), whereas it took an average of only 8.3 minutes for them to complete the electronic version. Similarly, Taenzer and colleagues (1997) found that a group of 36 women with breast cancer took an average of 8.6 minutes to complete an electronic version of the EORTC QLQ-C30, whereas it took a reference group an average of 11 minutes to complete the paper version. Significant differences in time to complete have also been found for different versions of the SF-36 (Caro et al., 2001).

Despite these potential advantages, electronic systems also have some disadvantages, which require consideration. The screen size on many electronic devices can be very small and difficult to read. The smaller devices also require the use of a stylus, making data entry potentially more tedious. A larger screen, similar to the new tablet personal computers, that could be manipulated with a touch of a finger would be ideal and speed data entry. In a study by Buxton et al. (1998), respondents had problems using the screen due to a lack of appropriate monitor responsiveness to touch. Wilkie and colleagues (2003) observed some difficulties with the touch screen device used in their study because of calibration issues. The device required calibration each time it was started and was less sensitive to a user's touch from an angle different from its calibration stage. This issue presented a difficulty for left versus right-handed users unless the screen was recalibrated for a new angle.

Nyholm et al. (2004) reported problems with battery level, which can be short and unpredictable with electronic diaries. Battery problems can result in sudden loss and corruption of data resulting from power loss and software problems. Although battery life can be an issue, some of the newer devices can run for many days on a single charge or set of batteries. If the electronic device being used is attached to a cord, requires a fixed table, and/or interferes with the clinical workflow, it can be another source of respondent burden. Wright et al. (2003) reported problems with the computer program being corrupted by patients not switching the computers off properly or switching off mid-program. This created error messages, not allowing patients to use the system temporarily. However, the technologies that allow electronic data collection are evolving rapidly and the procedures for collecting data are becoming more mobile and easy to use in both research and clinical settings.

Furthermore, it may not be important or cost effective to use computer-based questionnaires in all circumstances. The initial development costs of computer hardware, software and technical support and subsequent costs in terms of maintenance of hardware and updating information can be quite high (Allenby et al., 2002). The cost effectiveness of implementing a computerized survey depends on a variety of factors, including number of subjects, project scale, and system reliability. The added gains in data quality may not be sufficient to justify investing in a computer and an electronic questionnaire for a small study. However, a computer-based questionnaire intended for a large-scale clinical trial may be cost effective and offer substantial data quality and data management advantages. Despite the initial high cost of computer equipment and programming time, routine collection of data using electronic methods has proved to less expensive in the long term (Velikova et al., 1999b). In studies comparing paper-based questionnaires with electronic data capture, data collected electronically has been found to be cost-effective (Johannes et al., 2000; Kleinman et al., 2001).

Acceptability and Feasibility of EDC

A few studies have been conducted that have evaluated the acceptability of computerized technology to collect HRQoL data (Allenby et al., 2002; Wilkie et al., 2003; Mullen et al., 2004). Concern has been expressed about the feasibility of obtaining information via a computer especially from vulnerable populations such as older adults, less educated individuals, and persons who are economically disadvantaged. Hanscom et al. (1992) noted that older patients, women, and patients with less than a high school education were more likely to have missing and inconsistent data when it was collected via computerized questionnaires. Allenby et al. (2002) found that factors such as no prior computer experience, older age, poorer physical condition and lower educational levels increased the time it took for patients to complete the touch screen. Each increasing decade of age was associated with an estimated average 3% increase in the duration of time taken to complete the questionnaires.

However, several studies have revealed a high level of acceptability, even among these subgroups. Contrary to fears of researchers and health care providers, patients do not appear intimidated by computers (Kahn, 1993). Kohlmeier et al .(1997) have shown that persons over 65 years of age are willing and able to become computer literate, as long as ergonomic adaptations are made. Yarnold et al. (1996) reported that geriatric patients were quite accepting of computer-aided assessment. Low income participants were also able to use a computer-based demonstration program on their own. In a study by Hahn and Cella (2003), feedback was obtained from a diverse group of cancer patients regarding the acceptability of a touch-screen technology. Patients in this study were classified as either low (below seventh grade reading comprehension) or high literate and fifty-seven percent of the participants had no prior computer experience. In spite of this, nearly all patients (98%) felt the touch-screen was either easy or very easy to use. Furthermore, multimedia capabilities of computers and browsers will allow the development of audio cues to “read” the questions and allow illiterate patients to complete HRQoL instruments, a limitation that traditional paper versions of questionnaires will never overcome (Hahn and Cella, 2004). Touch-screen computer education has been especially effective when used to collect sensitive, private, anxiety-provoking or embarrassing information (Kahn, 1993).

Buxton et al. (1998) analyzed questionnaire responses of patients who completed the EORTC QLQ-C30 using a computerized program with a touch-sensitive video monitor (TSVM). The patients undertook a structured interview designed to assess the difficulties encountered with this process and the majority of them found all aspects of the TSVM system very easy to use, with only two percent ranking the process "somewhat difficult." There were no unanswered items on the questionnaire and all patients found the amount of time spent on answering the questions acceptable. Also, 96% of the patients stated they were willing to complete a similar questionnaire via TSVM on a future occasion or visit.

Furthermore, studies have also been conducted in various patient populations to demonstrate the feasibility and acceptability of touch-screen technology for administering HRQoL assessments (Taenzer et al., 1996; Newell et al., 1997; Wilkie et al., 2003). Several studies have found that patients readily accept and prefer to use electronic diaries rather than paper diaries (Drummond et al., 1995; Rabin et al., 1996). Patients' age, sex, and comfort/familiarity with technology were not associated with their electronic diary preference. In a group of asthma patients, Caro et al. (2001) found that 77% of respondents expressed a preference for the electronic implementation of the SF-36 and the vast majority (96%) found the electronic version easier to use. In a cancer pain clinic, participants completed a survey that assessed patients' impressions of a computerized HRQoL assessment. The authors concluded that patients found the computerized assessment easy to use, understandable, enjoyable, helpful, and a good use of waiting room time (Carlson et al., 2001). Taenzer and colleagues (1997) reported that the patient's attitudes toward computer surveys improve after they complete a computer-based questionnaire for the first time.

As with the introduction of any new technology, there is a learning curve during which its limitations become increasingly apparent. Certain minor changes in questionnaire presentation could be associated with differences in results. The reasons for these differences and their significance are still unclear but raise important questions about possible interactions between the context of the questions, method of presentation, and the cognitive processes used to formulate responses.

V. Measurement Equivalence Between Different Modes of Administration

Theoretical Framework for Measurement Equivalence

While the advantages of computerized electronic data capture for HRQoL measures are not in question, issues surrounding the measurement properties of this alternate administration mode require further examination. The primary concern that arises when a questionnaire is administered via two different modes is that results might differ based on the mode. Therefore, before widespread application of electronic data capture methods can be recommended, it is essential to establish that paper-based and electronic versions of the questionnaires exhibit measurement equivalence.

Measurement equivalence, also referred to as metric equivalence, differs from conceptual and psychometric equivalence (Anderson et al., 1996). Conceptual equivalence refers to similarity in meaning of items across language versions of an instrument, and aims to ensure that different language versions measure the same construct. Psychometric equivalence refers to similarity in psychometric properties such as floor and ceiling effects, reliability and construct validity (Lohr, 2000). Conceptual equivalence and psychometric equivalence are prerequisites for measurement equivalence but do not necessarily ensure measurement equivalence.

For agreement between alternate modes of administration for an instrument or scale (e.g., self-administered vs. interview administered, in-person interview vs. telephone interview, and paper form vs. touch screen), the critical question is whether these alternate modes can be used interchangeably. For two measurements made by the same method, it is important to know how variable measurements from a subject can be if the true value of the quantity does not change and what this measurement tells us about the subject's true or average value.

Customarily, a new measure or method is evaluated in terms of precision and accuracy. Precision refers to random errors whereas accuracy refers to systematic errors. Because error studies need to be extensive to demonstrate accuracy, it is generally more efficient to evaluate a method by comparison with a method already well characterized by error studies. Subjects are usually administered both methods and the comparison data are subjected to statistical analysis. The statistical calculations provide specific estimates of the type and magnitude of errors and not necessarily yes or no answers on the acceptability of a new measure/method. Ideal comparison data would have exactly the same values for both measures and all points would fall on a line making a 45 degree angle when graphed.

An instrument's alternate versions or administration modes can be tested for equivalence by examining the difference between means. However, when there are large samples (high power), even small differences in mean scores can be declared as "statistically significant," even though they may be of little significance to the subject. Equivalence is not necessarily equality and instead needs to be defined *a priori* by means of an equivalence region. The hypothesis test postulates that the difference between means lies within this equivalence region. This equivalence region has also been termed as the minimal important difference (MID).

In this context, MID refers to the maximum difference between within-subject scores on the alternate versions of the HRQoL instrument that would still allow the different approaches to be considered equivalent. The confidence interval around the observed difference is used to determine whether the interval includes the predetermined MID value. If not, we can conclude that that the two means are equivalent. This technique has been used in several studies and a similar application has been adopted by the US Food and Drug Administration for testing pharmacologic bioequivalence in the evaluation of generic drug applications (Makuch and Simon, 1978; Fuller et al., 1995; Schuirmann, 1987).

Several methods have been used to determine minimal important differences, and a 'gold standard' approach has not yet been agreed upon (Jacobson and Truax, 1991;

Lydick and Epstein, 1993; Juniper, 1998; Wyrwich et al., 1999; Hays and Woolley, 2000; Crosby et al., 2003). The MID can be set from an absolute difference in means or medians, a proportionate difference in results, a ratio of two group results, or an "effect size." The use of different perspectives, criteria, or methods can yield different results, and hence it seems prudent to use multiple methods to estimate the MID (Kelleher et al., 2004). In the area of HRQoL, two methods have been commonly used to calculate the minimal important difference : anchor-based and distribution based approaches (Guyatt et al., 2002; Sloan et al., 2002).

The anchor-based approach uses an external criterion or 'anchor' with which to compare improvements in health-related quality of life scores. This method defines MID as the within-patient change score irrespective of the direction of change or the between-patient score differences for matched patients who differ only on the severity of the condition. This anchor is frequently a global or overall question measuring well-being or treatment effect. The anchor-based approach can be criticized for a variety of reasons. Change scores may be biased because patients underestimate their HRQoL at a prior visit or provide retrospective change estimates that are highly correlated with their HRQoL at the current visit (Wyrwich et al., 1999, 2000). It is also possible that the number of response options on the anchor-based question, its comprehensiveness, and how conceptually similar it is to other such questions will affect MID estimates. Also, this approach only looks at change scores and inappropriately assumes that getting better is the same as getting worse. Also, using unreliable single-item anchors is another shortcoming.

On the other hand, distribution-based approaches rely on relating the difference between scores to some measure of variability (Norman et al., 2001). Although several methods are in use, the most common distribution-based method is to compare two subgroups at a point in time or measure the change over time in one group to the standard deviation at baseline (i.e. concept of effect size) (Juniper, 1998; Hays and Woolley, 2000). The effect size is entirely dependent on the within-group standard deviation at baseline, and therefore may vary widely with different samples taken from the same population (Wyrwich et al., 2000).

Although the effect size method is a subject of debate as to whether it is more of a measure of responsiveness than a minimal important difference measure, it is used quite often because it is a common, well-known measure and familiar to clinicians (Samsa et al., 1999). Conventional benchmarks established by Cohen (1969) are used to interpret group effect sizes. Cohen's recommendation for small (0.2-0.5), medium (0.51-0.8) and large (>0.8) effects are one attempt at interpretation that has been suggested as appropriate for HRQoL (Kazis et al., 1989). However, effect sizes are sensitive to the homogeneity of distribution of the sample from which the study findings were derived.

As concluded by Norman et al. (2003), empirical evidence demonstrates that under many circumstances (e.g., different chronic diseases), when patients identify minimal meaningful change on HRQoL scale scores, the estimate tends to be remarkably close to a half a standard deviation. Juniper and colleagues (1994) used an anchor-based approach to determine the magnitude of change corresponding to a minimally important difference on the Asthma Quality of Life Questionnaire (AQLQ). The authors determined that a 0.5 point within-patient change on the AQLQ's scales was the smallest amount of change perceived important by patients. In addition, as pointed out by Sloan et al. (2003) in their discussion of the distribution-based effect size approach, a 0.5 point difference on the AQLQ scales represents roughly a half a standard deviation, which lends additional support for it as an estimate of the MID. Determining the MID using anchor and distribution based approaches has resulted in similar conclusions.

Although there are limitations associated with using a threshold for minimally important differences in HRQoL assessment a general guideline can be very useful in interpreting HRQoL findings in the absence of other relevant evidence (Hays and Woolley, 2000; Beaton, 2003; Wright et al., 2003). Therefore, a reasonable assumption is to use a half of a standard deviation, which indicates a medium effect as the threshold to test the equivalence of different modes of administration of HRQoL instruments.

Review of Past Studies

Studies comparing electronic questionnaires with conventional paper questionnaires have shown no noteworthy differences between the two modes of data collection. Several generic and disease-specific HRQoL measures have exhibited stable psychometric properties across mode of administration. The SF-36 has been shown to be reliable and valid when administered face-to-face, by means of a telephone interview or by self-administration (Weinberger et al., 1996). Scores on the SF-36 gathered through computer administration and self-administration have also shown to be comparable (Mohr et al., 2005).

Several studies have also examined the psychometric properties of the EORTC QLQ-C30 administered via a computer. Taenzer and colleagues (1997) found the internal consistency of the measure similar for paper-and-pencil versus computer administration. Velikova et al. (1999) compared the computer touch-screen version of the questionnaire with optically scanned paper version in a randomized cross-over trial of 149 patients with cancer. At the group level, the differences between scores obtained with the two modes of administration were small, suggesting equivalence for most of the scales. The electronic questionnaire had good test-retest reliability. Correlation coefficients for the sub scales and items between the two administrations ranged from 0.78 to 0.95 and kappa coefficients of agreement from 0.55 to 0.90. The authors concluded that computer touch-screen questionnaires are well accepted by patients with cancer, with good data quality and reliability.

Drummond and colleagues (1995) also compared the use of pen-based electronic questionnaires and conventional paper questionnaires in a randomized cross-over study of patients with gastrointestinal disorders. All patients initially completed a paper HRQoL questionnaire and then on two subsequent visits completed electronic and paper questionnaires in random order. Study results indicated comparability in scores obtained on the two versions. An analysis of the psychometric characteristics of Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire suggests that the measure performs equally well with either paper-based or touch-screen computer versions. The QOLRAD exhibited evidence of internal consistency reliability, reproducibility and validity regardless of method of administration (Kleinman et al., 2001).

Bushnell et al. (2003), compared paper versions of the Asthma Quality of Life Questionnaire (AQLQ), Pediatric Asthma Quality of Life Questionnaire (PAQLQ), and the Pediatric Asthma Caregiver's Quality of Life Questionnaire (PACQLQ) with corresponding touch-screen versions designed by Assist Technologies. The results demonstrated no statistically significant differences between the within-patient scale scores on the three measures based on paired t-tests. Likewise, the intraclass correlation coefficients and the Pearson correlation coefficients were very high, indicating high levels of concordance. Again, the authors did not-test explicitly for meaningful differences, but the differences between the mean scale scores ranged from 0.02 to 0.06 points, which were well below the MID threshold of 0.5 points.

Several other studies have been conducted in various patient populations. Concordance between scores obtained on paper-based and electronic versions have been reported in patients having cancer, diabetes, and in psychiatry (Buxton et al., 1998a; Pouwer et al., 1998; Schmitz et al., 2000; Boyes et al., 2002). Most of the studies utilized a randomized cross-over design, wherein subjects were randomized to ensure that half the group first completed the paper-based version followed by the electronic version and the other half completed the questionnaires in reverse order. The time duration between completion of the two versions varied across studies. In a study by Velikova et al. (1999), one version was completed in the morning whereas the other version was completed in the afternoon with a target interval of three hours between the two administrations. The order in which the questionnaire versions were administered did not significantly affect the data obtained.

Results from studies on the extent to which socio-demographic factors influence respondent comfort and capability to complete computer-administered questionnaires are equivocal. Drummond et al. (1995) found no association between socio-demographic characteristics and preference for computer versus paper-based questionnaires. Similar results were also obtained by Kleinman and colleagues (2001), who reported that age, education level, and prior computer experience did not affect score stability across method of administration. Allenby et al. (2002) reported that older patients, on average did take longer to complete the questionnaire compared to younger patients and each increasing decade of age was associated with an estimated 3% increase in the duration of time taken to complete the questionnaires. Buxton et al. (1998) indicated that age, education level and previous computer use was associated with completion time. Age was also a predictor of data completeness in a study by Bliven and colleagues (2001). Older patients were significantly less likely to successfully complete the paper version of the questionnaires; however, age and education level did not predict completeness of questionnaire data collected via the computerized version. Bliven et al. (2001) also reported no significant relationship between computer literacy and successful completion of the computerized questionnaire. This was in contrast with another study in which it was found that prior computer experience was negatively correlated with the number of times subjects needed assistance (Taenzer et al. 1997).

The EQ-5D is an example of a generic HRQoL instrument for which computer touch-screen versions have been developed. To our knowledge, no published study has compared paper-based and electronic versions of the EQ-VAS.

CHAPTER 3

METHODOLOGY

The primary purpose of this research was to test the equivalence of the original paper-based vertical version of the EQ-VAS with the touch screen-based horizontal version. The study also examined the psychometric properties of an electronic version of the SF-36. This research was conducted under the auspices of the University of Arizona Human Subjects Protection Program.

I. Hypothesis Tests

As stated in Chapter 1, the hypotheses tested in this investigation include those listed below.

Hypothesis One (H_{01})

There is no difference in scores obtained on the original paper-based vertical version of the EQ-VAS and the touch screen-based horizontal EQ-VAS version.

Hypothesis Two (H_{02})

There is no difference in reliability estimates obtained for the touch screen version of the SF-36 when compared with those obtained for the general US population

Hypothesis Three (H_{03})

There is no difference in the validity of the touch screen version of the SF-36 when compared with that of the originally validated paper version.

II. Research Design

Subjects

Due to the nature of this research, it was not deemed essential that a random, US population-based sample be obtained. The intent was to collect data from a broad range of adults from the general population. Hence, a non-probability (i.e., convenience) sample was obtained from the Tucson (Arizona) Metropolitan Area,.

Sample Size Considerations

The sample size necessary for this study was based on several factors. In order to compute the sample size estimate, we specified the type I error rate (α), type II error rate (β), the effect size (*d*) and the variability in the outcome variable. A type I error occurs when a true null hypothesis is incorrectly rejected and a type II error occurs when a false null hypothesis is failed to be rejected. Power is the probability of avoiding a type II error ($1-\beta$). Conventional practice is to determine the sample size that gives 80% power at the 0.05 level of significance (two-sided). The effect size was defined as the maximum difference between within-subject scores on the two versions of the EQ-VAS scale that would still allow the two approaches to be considered equivalent. A moderate effect size of a half of a standard deviation (i.e., 8 points on a 100 point scale) was used as the threshold. The within-patient standard deviation (i.e. the standard deviation of repeated observations in the same individual) of the outcome variable was 16.67. The approximate sample size necessary to attain statistical significance was estimated using statistical power analysis tables developed by Cohen (1988). To obtain a statistically significant difference (two- tailed, alpha = .05) between scores obtained on the two versions of the EQ-VAS scale using a paired t-test , it was projected that a minimum of 64 participants were required for the study when power was specified at 80% and 85 participants when power was specified at 90%. In order to detect a smaller effect (d = 0.30), 175 and 234 participants were required when power was specified at 80% and 90% respectively.

Sampling Technique

Sampling techniques are usually classified as either probability or non-probability. In probability sampling, each member of the population has a known non-zero probability of being selected whereas in non-probability sampling, members are selected from the population arbitrarily in some non-random manner. Hence, in non-probability sampling, the degree to which the sample differs from the population (i.e., sampling variability) remains unknown. The only way to address data quality is to compare some of the study results with available information about the population and even then, there is no assurance that the estimates will meet an acceptable level of error.

Despite these drawbacks, non-probability sampling methods are useful and have several advantages, particularly when there is no intent to generalize beyond the sample itself. Non-probability sampling is relatively inexpensive and convenient to conduct. Also, in some circumstances, it is unfeasible or impractical to conduct probability sampling. Examples of non probability sampling technique include convenience sampling, judgment sampling, quota sampling, and snowball sampling. A convenience sample is one of the most common forms of non-probability sampling. The sample comprises subjects who are readily available to the researcher. There is no randomness and the likelihood of bias in terms of generalization to a broader population is high. This method is often the only feasible option due to restricted time and resources, and can legitimately be used provided its limitations are clearly understood and stated.

In this study, a convenience sample was obtained that was intended to reflect the primary socio-demographic characteristics of the general US adult population. The sampling was purposive in that subjects were selected by the researcher subjectively in an attempt to obtain a sample representative of the general adult population in the Tucson Metropolitan Area with respect to age, sex, and race/ethnicity. According to the 2000 US census data, the distribution of people by age, race, and sex in the Tucson Metropolitan Area is as follows:

Table 3.1: Tucson Population Demographic Characteristics

Characteristic	%
Age (median = 32.1 years)	
18-24 years	13.8
25-34 years	15.7
35-44 years	14.9
45-54 years	11.8
55-64 years	7.3
≥ 65 years	11.9
Sex	
Male	49.0
Female	51.0
Race/Ethnicity	
White, not Hispanic	54.2
Black, not Hispanic	4.1
Hispanic	35.7
Asian/Pacific Islander	2.5
Other	3.5

Source: Census 2000

A total of 314 persons from the Tucson Metropolitan Area agreed to participate in the study. Subjects were non-randomly asked to participate based on the recruiter's discretion and, hence, this can lead to sampling bias. Sampling was done until the total number of respondents required for the study was obtained.

Inclusion Criteria

Subjects were considered eligible to participate in the study based on the following inclusion criteria: (1) at least 18 years of age; and (2) able to speak and read English. Since we wanted a representation of the general US adult population, persons 18 years of age and older were invited to participate in the study. All instructions were provided in English and no other language versions of the data collection instrument were available, hence only participants who could speak and read English fluently were invited to participate.

Recruitment

Study participants were recruited primarily at the regional office of the Arizona Department of Motor Vehicles (DMV) in Tucson. Potential participants were approached by study personnel outside the DMV office entrance while they were entering, leaving, or awaiting services. Prior approval from the site was obtained and a letter of authorization was furnished to the University of Arizona Human Subjects Protection Office. The DMV approval indicated that data could only be collected outside the building, not inside the main service area. A flyer (Appendix B) stating the purpose of the study and eligibility criteria were given to potential participants. After reading the flyer, persons indicating an interest in participating were provided a copy of the disclaimer form (Appendix C), which provided additional study information. Tables and chairs were set up for participants to sit and comfortably complete the study questionnaire outside the main entrance to the DMV building.

Subjects were also recruited by email via the College of Pharmacy listserv at the University of Arizona. The email content was approved by the University of Arizona's Human Subjects Protection Program. The email message, which contained the text of the flyer, was sent to faculty, students, and staff at the College of Pharmacy and information regarding the time and place for data collection was also provided. Interested participants were provided a copy of the disclaimer form as indicated previously. Participation in the study was strictly voluntary.

Study Measures

EQ-5D

The EQ-5D self-report questionnaire consists of (1) a descriptive system and (2) a visual analog scale (EQ VAS). The EQ-5D descriptive system is a five item questionnaire, which asks respondents to describe their own health with respect to the following five dimensions: mobility, self-care, usual activities, pain and anxiety/depression. For each item there are three response levels (i.e., no problems, moderate problems, and extreme problems). These responses are used to classify the respondent into one of 243 unique EQ-5D health states. In addition to the multidimensional descriptive system, the EQ-5D also uses a visual analog scale (EQ-VAS) or "feeling thermometer" to measure an individual's overall health. The original

EQ-VAS is a thermometer-like 20-cm vertical line with endpoints labeled 'worst imaginable health state' and 'best imaginable health state' anchored at 0 and 100, respectively. In addition, it is demarcated in units of one and labeled in units of ten. The EQ-VAS was presented in the original vertical 20-cm format on the paper-based survey and as a horizontal 15-cm scale in the touch screen version. The touch-screen device has a small screen size and hence space is at a premium. As a result, the 20-cm vertical EQ-VAS scale had to be condensed to a 15-cm horizontal scale to adapt it to the screen display.

Medical Outcomes Study SF-36

Health status was also measured using the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) (Ware et al., 1992). It is currently one of the most widely used health status assessment tools. This generic health status measure is a self-administered 36-item questionnaire comprised of eight multi-item scales: physical functioning (10 items), bodily pain (2 items), role limitations due to physical problems (4 items), role limitations due to emotional health problems (3 items), social functioning (2 items), mental health (5 items), energy/fatigue (4 items) and, general health (5 items). The SF-36 also includes a single item health transition scale that provides an indication of the perceived change in health over the past year. These nine SF-36 scale scores range from 0-100; higher scores indicate better functioning and well-being. These items in these scales are also combined to yield two composite measures: the mental (MCS) and physical component summary (PCS) scores. Based on the published scoring recommendations, the MCS and PCS are scored normatively, that is, they are transformed to reflect a mean of 50 and a standard deviation of 10 in the general adult US population. The SF-36 has been shown to provide a reliable and valid measurement of health-related functioning and well being in multiple settings, populations, and conditions.

Socio-Demographic Information/Descriptive Items

Respondents were asked to provide socio-demographic information including: (1) age; (2) sex; (3) race/ethnicity; (4) level of education; (5) annual household income; and (6) presence of chronic conditions. In addition to standard socio-demographic data, study subjects were questioned about prior computer use. This included a single question asking respondents “Have you used a computer before today?” An item was also included to assess left or right handedness by asking respondents “With which hand do you normally write?”

Additional Assessments

In order to determine the degree to which a person’s overall quality of life can be explained by a person’s health status, respondents were questioned about their level of satisfaction with various aspects of life. These various aspects included respondents’ perception of satisfaction with their job situation, spiritual life, relationship with family members, their spouse and friends, wealth and income, education, leisure time, housing situation, transportation, neighborhood, self-worth, and life as a whole. A single item also asked respondents to rate their overall quality of life on a five-point scale ranging from excellent to poor.

III. Data Collection

Electronic Data Collection Instrument

The device and software for collecting the electronic data used in this study was designed by Assist Technologies of Scottsdale, Arizona. The hardware was a network-ready fully integrated liquid crystal display touch-screen workstation. The touch-screen computer display is fitted with sensors that allow the computer to be manipulated by the touch of the finger or a stylus. A copy of the data collection instrument was provided to Assist Technologies and the questions were programmed into the touch-screen computer device. The design and format of the data collection instrument were modified when adapting the paper questionnaire to the touch-screen version. Instructions on how to complete the instrument were changed to correspond with the new layout and design of the touch-screen instrument. Questions were presented individually on the screen and participants were asked to provide their responses by touching the corresponding button on the screen. Participants were not able to move on to the next question without answering the previous one. After a response was entered, the next item appeared automatically. Participants had the option of skipping a question by selecting the response, “Choose not to answer.” In addition, they were able to go back and review and change responses to previous questions. The computer electronically recorded the time taken to complete the questionnaire.

Questionnaire Administration

The questionnaire (Appendix C) was administered to study subjects in two parts and required approximately 10 to 15 minutes to complete. One part was paper-based and the other was touch screen computer-based. The paper-based part only contained the original 20cm vertical EQ-VAS. The touch screen-based part included the horizontal EQ-VAS plus the EQ-5D descriptive system, the SF-36, socio-demographic characteristics, a list of chronic conditions, questions about prior computer use and handedness and items assessing satisfaction with domains of life. These additional electronic questions provided an intervening task, which required approximately 10 minutes to complete. This delay was incorporated into the study design to minimize the impact of memory on the EQ-VAS responses obtained on the two formats while avoiding excess response burden.

Each subject was required to complete both parts. The sequence of the administration, however, was alternated such that the first subject completed the electronic version followed by the paper version, while the next subject completed the paper-based version followed by the electronic version. Either way, one of the two EQ-VAS versions appears at the beginning and one at the end of the questionnaire. Hence, both versions of the EQ-VAS were administered to each participant approximately 10 minutes apart. At the end of the data collection process, half of the subjects had completed the paper-based part first and then the computer-based part and the other half of the subjects had completed the parts in the opposite order.

A brief explanation of how to use the touch-screen device was provided at the start of the data collection procedure. Also, subjects received $5 for their participation. Participants could choose to not answer specific questions and withdraw from the study at any time. In that case, a question was asked regarding the reason for withdrawal from the study.

IV. Data Integrity

Data Entry

Data obtained using the touch screen was stored locally on the computer's hard drive. Upon completion of the data collection procedure, the touch-screen devices were shipped back to Assist Technologies in Scottsdale, Arizona and transferred to an electronic database. Data collected using the paper-based survey was entered manually into a computer database. Responses obtained from both questionnaire formats were then integrated and available for analysis as an electronic data file.

Confidentiality

To insure subject confidentiality, responses were coded and sorted by an identification number and were not linked to any specific individuals. No information on personal identifiers (e.g., name, social security number) was collected as part of this study. Only the study personnel at the University of Arizona and data entry personnel at Assist Technologies of Scottsdale Arizona had access to the data. All data analyses were performed and reported on aggregate data.

V. Data Analysis

All statistical analyses were performed using SAS Version 8.02 for Windows (2001) and Intercooled Stata 8.0 for Windows (2003). Datasets were converted from SAS to Stata data files using DBMS/COPY for Windows (1997). All statistical tests were evaluated using a conventional two-tailed, alpha (α) level of 0.05.

Descriptive Statistics

Information obtained from the survey questionnaire was analyzed in this investigation. Descriptive statistics for demographic and other individual questionnaire items were performed. For interval-level data, summary statistics included measures of central tendency (i.e., means, medians), standard deviations, skewness, and ranges. For ordinal and categorical data, percent responses and/or frequency tables were reported.

Mean scores were reported for each dimension of health on the EQ-5D and the SF-36. Scoring of the SF-36 included computing raw scores and transforming scale scores to a 0-100 scale. SF-36 mean scale scores were calculated and then the summary scores (i.e., PCS and MCS) were computed and compared with population norms.

Comparisons of responses to paper and touch screen EQ-VAS include examination of score distributions and review of frequency distributions for within-subject difference scores. A scatter plot of the data was also created and the line of perfect agreement was superimposed on it, thereby enabling a visual inspection of the existence and strength, if any, of agreement.

Bivariate Statistics

Correlation Coefficient

In order to examine the strength of the linear relationship between scores on the two EQ-VAS versions, the Pearson product moment correlation coefficient (r) was calculated. The Pearson product-moment correlation coefficient (r) is a measure of how well a linear equation describes the relationship between two variables (i.e., paper and touch screen EQ-VAS) measured on the same subject (Bland et al., 1994;Bland et al., 1995;Bland et al., 1996). It is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{\sum z_x z_y}{N-1}$$

The result obtained is equivalent to dividing the covariance between the two variables by the product of their standard deviations. The correlation coefficient adds a sign (+, -) to define the direction of the relationship, either positive or negative. A positive correlation coefficient indicates that as the value of one variable increases, the value of the other variable increases; as one decreases the other decreases. A negative correlation coefficient indicates that as the value of one variable increases, the other decreases, and vice-versa. Also, the coefficient ranges from −1.0 to 1.0. A value of 1.0 indicates that a linear equation describes the relationship perfectly and positively whereas a score of −1.0 shows that all data points lie on a single line but move in the opposite direction. A value of 0 shows that a linear model is inappropriate (i.e., there is no linear relationship between the variables).

Pearson's product-moment correlation coefficient (r) is a parametric statistic. It assumes that the variables being assessed are normally distributed. If this assumption was violated, a non-parametric alternative such as Spearman's rank-order correlation coefficient Rho (ρ) was used instead to detect a linear relationship. Unlike the Pearson product-moment correlation coefficient, it does not require the assumption that the relationship between the variables be linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. The Spearman's Rho is simply a special case of the Pearson's product-moment coefficient in which the computations are done after the data are converted to ranks. The raw scores are converted to ranks, and the differences (D) between the ranks of each observation on the two variables are calculated. Rho (ρ) is then given by:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Where, D = the difference between ranks of corresponding values of the two variables, and N = the number of pairs of values.

Another measure of correlation is Kendall's tau, which requires the two variables to be paired observations and at least ordinal. Much like the Spearman's rank correlation; it is carried out on ranks of data. In common with other measures of correlation Kendall's tau takes values between -1.0 and +1.0, with a positive correlation indicating that the ranks of both variables increase together whilst a negative correlation indicates that as the rank of one variable increases the other one decreases. The Spearman's rank correlation has been more widely used because it is much easier to compute than Kendall's tau. However, the main advantages of using Kendall's tau are that the distribution of this statistic has slightly better statistical properties and there is a direct interpretation of Kendall's tau in terms of probabilities of observing concordant and discordant pairs. In almost all situations the values of Spearman's rank correlation and Kendall's tau are very close and would invariably lead to the same conclusions.

As classified by Cohen (1977), the relationships measured by correlation coefficients are interpreted as small (0.2-0.5), moderate (0.5-0.8) or large (> 0.8).. Correlation coefficients however, are measures of association and not a measure of agreement. This is because a correlation coefficient measures the strength of a relationship between two variables, not the agreement between them. Data that is in poor agreement has shown to produce quite high correlations. This is because correlations ignore any systematic bias between the two variables. Also, correlations depend on the range and distribution of the variables and hence on the way in which the sample was selected (Bland et al., 1986).

Intraclass Correlation

The intraclass correlation coefficient (ICC) was used as a measure of agreement or consistency between the two methods of measurement, in conjunction with a significance test of the difference between means. The ICC may be conceptualized as the ratio of between-groups variance to total variance (Shrout and Fleiss, 1979). It approaches 1.0 when the between-groups effect is very large relative to the within-groups effect. In this way ICC is a measure of homogeneity. ICC is 0 when within-groups variance equals between-groups variance, indicative of the grouping variable having no effect. Though less common, the ICC can be negative when the within-groups variance exceeds the between-groups variance.

Limits of Agreement

Altman and Bland (1983, 1986) have explained in detail an analysis technique to evaluate agreement between two methods of measurement on a continuous scale. It is known as the 95% limits of agreement method. According to this approach, the mean and standard deviation of the difference in EQ-VAS scores (paper - touch screen) on the two measures were first obtained. The 95% limits of agreement were defined as the mean difference ± 2 (or more precisely, 1.96) standard deviations. It is expected that 95% of differences between EQ-VAS scores by the two methods to lie between the limits. The decision of how small the limits of agreement should be to conclude that the methods agree sufficiently is a clinical and not statistical decision. The 95% limits of agreement depend on some assumptions about the data: that the mean and SD of the differences is constant throughout the range of measurements, and that these differences are from an approximately normal distribution.

To check for these assumptions, a scatter plot of the differences against the average of the two EQ-VAS measurements [(paper + touch screen)/2] as well as a histogram of the differences was constructed. From the scatter plot, it is easy to see if there is any tendency for the amount of variation to change with the magnitude of the measurements. The correlation coefficient was tested against the null hypothesis of $r = 0$ for a formal test of independence. If the within-subject scores were independent of the size of the measurement, then a one-way analysis of variance was performed and the residual standard deviation was an overall measure of repeatability, pooled across subjects. This coefficient of repeatability represents the value below which the differences between the two EQ-VAS scores are expected to lie. The mean of the differences estimates the bias (i.e., a tendency for one of the methods to give consistently higher readings than the other method). Confidence intervals for these limits of agreement can also be formed (Bland et al., 1999).

Paired t-test

We examined the extent to which one mode of administration provided consistently higher or lower responses (bias) by calculating directional difference scores. A paired t-test was used to control for experimental variability and to test the hypothesis of zero bias (i.e., the mean scores for the two versions of the EQ-VAS are equal). It calculates the difference between each set of pairs and analyzes that list of differences. By analyzing only the differences, a paired t-test corrects for other sources of experimental variability as each person serves as their own control. The paired t-test assumes that the difference follows a Gaussian distribution.

The t ratio for a paired t-test is the mean of the differences divided by the standard error of the differences. With widely discrepant values the test will not detect that the mean difference between the readings is significantly different from zero, although the methods are in poor agreement. Conversely if the two methods give similar readings, but there is a bias in the results the mean difference may be small, but detected to be significant. In addition to bias testing, we assessed nonsystematic differences between modes by examining absolute mean differences (i.e., the magnitude of the difference without regard to the direction of the difference) and presented their means and standard deviations.

Statistical significance however does not necessarily indicate that meaningful difference exists. A p-value of less than 5% does not necessarily provide an indication of clinical significance or meaningfulness. Before interpreting the p-value or confidence intervals, the average difference between the paired measurements that was considered to be meaningful to subjects was determined. In HRQoL measurement, minimally important difference has been determined on the basis of two main approaches; distribution-based effect size methods and anchor-based methods.

Distribution-based methods rely on expressing an effect in terms of the underlying distribution of the results. In this method, the amount of overlap between two distributions of scores is compared to determine if there are meaningful differences (Norman et al., 2001). The most popular approach uses Cohen's standardized effect size; the mean change divided by the standard deviation to serve as an "effect size index" that is suitable for sample size estimation. Empirical evidence demonstrates that under many circumstances (e.g., different chronic diseases), when patients are used to identify minimal meaningful change on HRQOL scale scores, the estimates tend to be remarkably close to a half a standard deviation (Norman et al., 2003).

Anchor-based methods use some external anchor, such as patient's judgment of change, which is then used to compute a minimally important difference (MID). Juniper et al. (1994) used an anchor-based approach to determine the magnitude of change corresponding to an MID on the Asthma Quality of Life Questionnaire (AQLQ). Juniper et al. (1994) determined that a 0.5 point within-patient change on the AQLQ's scales was the smallest amount of change perceived important by patients on the scale of 1 to 7. A 0.5 point difference on the AQLQ scales represents roughly a half a standard deviation and lends additional support for it as an estimate of the MID (Sloan et al., 2003).

Although there are limitations associated with using a threshold for minimally important differences in HRQOL assessment, a general guideline can be very useful in interpreting HRQOL findings in the absence of other relevant evidence (Beaton, 2003). For the purpose of this investigation, a difference of half a standard deviation (roughly eight points on the 100 point EQ-VAS) or more was used as the equivalence threshold. In other words, if the mean difference between the scores on the two formats was less than eight points, the formats were deemed equivalent. In addition, a form of sensitivity analysis was performed by lowering the MID threshold to a more conservative threshold (e.g., quarter of a standard deviation).

Variance Comparison Test

In order to determine if there was a difference in the variability of EQ-VAS scores obtained via the touch screen method as compared to the original paper method, a two sample F-test for variances was performed. Also, the t-test makes the assumption that the variances of the two sampled populations are equal and if this is not the case, any inferences derived from the t-test are suspect. The null hypothesis for the F-test is that the standard deviation for both variables is equal (or, equivalently, that their ratio is equal to 1). The distribution of the ratio of the two variances approximates an F distribution with n-1 numerator and n-1 denominator degrees of freedom.

Other Associations

Differences in EQ-VAS and SF-36 scores among various sociodemographic groups (e.g., sex, gender, race/ethnicity, education, employment level, number of chronic conditions) were examined using independent group t-tests or one-way analysis of variance (ANOVA). In case the data was not normally distributed, Mann-Whitney U or the Kruskal-Wallis test was used as the non parametric alternative to the t-test and the one-way ANOVA respectively. Chi-square (X^2) tests were conducted to test for associations between responses on the EQ-5D descriptive system and some important demographic variables. Also, percentage of respondents scoring at the lowest (floor) and highest (ceiling) possible EQ-VAS scores were compared between modes of administration using chi-square techniques. Comparison of scores by questionnaire order (paper VAS vs. touch-screen VAS first) was determined by two sample t-tests.

Multivariate Statistics

While there was no formal hypothesis stated regarding the determinants of difference in EQ-VAS scores, multivariate statistics were conducted to help understand factors influencing the difference in scores between the two versions. The method for performing this analysis is outlined below.

Dependent and Independent Variables

The dependent variable for the multivariate model performed in this study included difference in EQ-VAS scores. The independent variables included age, sex, race/ethnicity, education, income, time to complete the questionnaire, mean EQ-VAS score, and order of administration. Age and sex information was obtained directly from the questionnaire. Race/ethnicity was dichotomized to “white, non-Hispanic” and “other racial/ethnic groups” due to small sample sizes in the other groups. Income was categorized into three levels: less than $20,000, $20,000-$40,000, and greater than $40,000. Education level was also collapsed into five levels; not a high school graduate, high school graduate, some college, no degree, associate/bachelor’s degree, and graduate degree (master’s /doctorate). Mean EQ-VAS score was the mean of the two EQ-VAS scores for each person (i.e., [paper + touch screen EQ-VAS score]/2). It served as the best estimate of the true EQ-VAS score and represented self-reported health status. The time to complete the questionnaire was automatically recorded by the touch screen device.

Multiple Linear Regression

Ordinary least squares (OLS) regression analyses were conducted to assess the relationship between potential predictors (i.e., independent variables) and the dependent variable. The coefficient for each independent variable represents the predicted increase in the dependent variable per unit increase in each of the independent variables, with all other variables in the model held constant. The degree of dependence between the dependent variable and the corresponding sets of independent variables were measured by multiple correlation coefficients (R). The proportion of variance in the dependent variable that could be explained by the independent variables in each model was reported. This value is called the coefficient of determination or the adjusted R^2 value ($R^2_{adjusted}$). The $R^2_{adjusted}$ value is bounded by zero and one, with values closer to one indicating a better fitting model.

VI. Scaling Assumptions, Reliability, and Validity of the SF-36

A secondary purpose of the study was to test the psychometric properties of the touch screen version of the SF-36 and compare it to those reported for the original paper-based version.

Item-Level Characteristics

The first step was to examine item-level characteristics of the SF-36. The extent of missing and out-of range data was examined for all the items. The frequency distribution of individual items was examined to determine if all of the response choices were used and if there were patterns across scales. However, because the study sample was obtained from the ambulatory general population, it was anticipated that the participants were relatively healthy and that not all response options would be used, especially for questions measuring functional limitations.

Item means were also examined and under traditional Likert scaling criteria, item means are expected to be roughly equivalent within a scale. While an average item value will vary with the population sampled, the ordering of item means and the rough differences between them is not expected to vary. In addition to examining item means, item standard deviations were also examined. Standard deviations should be roughly equivalent and close to 1.0 (for five-choice response scales) (Gandek et al., 1998b).

Item internal consistency was tested by examining the correlation between an item and the scale score computed from all other items in that scale (i.e., item-scale correlation after correction for overlap). The assumption here was that an item should be substantially linearly related to the underlying concept being measured. Item internal consistency is considered substantial and satisfactory if an item correlates 0.40 or more with its hypothesized scale (after correction for item-scale overlap) and that standard was adopted for comparison purposes (Ware and Gandek, 1998). In addition to demonstrating substantial correlations with their hypothesized scale, all items within a scale must have roughly equal item-scale correlations. This second assumption underlying Likert's simple scoring method is that items in a scale should contribute roughly equal proportions of information to the total scale score.

Scale-Level Characteristics

After item-level analyses were conducted, SF-36 scales were scored and the properties of the scales were examined focusing on scale means and standard deviations and the proportion of respondents scoring at the highest (ceiling) and lowest (floor) level. Scale means and standard deviations indicate where along a scale continuum the majority of individuals within the sample are likely to be found. Additionally, the SF-36 profile of mean scale scores is expected to have roughly the same shape across general population samples. The five scales which primarily measure health-related disability (physical functioning, role-physical, bodily pain, social functioning, role-emotional) are expected to have the highest mean scale scores, while lower mean scores should be found for the three scales which extend measurement to the well-being range (general health, vitality, mental health) (Gandek et al., 1998b).

The full range of the measure should be used and the scale score distribution should show substantial variability even if the distribution is skewed. The percentage of respondents at the highest possible (ceiling) or the lowest (floor) score is also noted, since the ability of an instrument to detect change over time is constrained by the percentage of respondents at the ceiling or floor.

Reliability and Correlation between Scales

Reliability of measurement refers to the extent to which the measured variance in a score reflects the true score, rather than random error; that is, the extent to which measures give consistent or accurate results. A reliability coefficient is an estimate of the proportion of total variance that is true variance and can be expressed as:

$$\text{Reliability} = 1 - (V_e / V_t)$$

Where, V_e = error variance and V_t = total measured variance

Reliability can be estimated in a number of ways, including test-retest, alternate forms, and internal consistency reliability. Within this study, the reliability of scale scores was estimated using the internal consistency method (Cronbach's alpha coefficient). A minimum reliability level of 0.70 has been recommended for group level comparisons and for person-level comparisons, a much higher standard of reliability (0.90 or higher) is considered appropriate.

To evaluate how distinct each scale is from other scales in the same matrix, correlations among all scales were computed and compared with reliability estimates. A reliability coefficient is thought of as a correlation between a scale and itself. To the extent that the correlation between two scales is less than their reliability coefficients, there is evidence of unique reliable variance measured by each scale. When the correlation between two scales equals their reliability, the scales are in essence performing like alternate form measures of the same concept, and there is no evidence that they have unique reliable variance. Thus, evaluation of inter-scale correlations helps determine whether each scale is measuring a distinct concept or construct.

Validity

Validity is the extent to which a score measures what it is intended to measure. The validity of questionnaires in the health field has most often been evaluated by means of content, construct, and criterion validity. Content validity examines the extent to which a measure or questionnaire represents the universe of concepts or domains; that is, whether the measure offers an adequate sample of the content of a construct. The content validity of the SF-36 has been evaluated based on standards published by Ware in 1987, and since the adaptation of the SF-36 to an electronic format did not alter the content in any way, the content validity did not need to be re-established (Gandek et al., 1998a).

Construct validity is a process in which validity is evaluated in terms of the extent to which a measure correlates with variables in a manner consistent with theory. According to tests of "known groups" validity, the mean scale scores were compared across groups known to differ (i.e., persons with zero and one or more chronic conditions). It is hypothesized that persons with no chronic conditions will have higher scale scores compare to those with one or more chronic conditions. Construct validity was also evaluated through examination of correlations between SF-36 scales and the EQ-5D descriptive system dimensions. The SF-36 is expected to have stronger relationships with some EQ-5D dimensions and weaker relationships with others. Both convergent and discriminant validity can be tested in this manner. Convergent validity is supported when scales measuring similar concepts (e.g., SF-36 physical functioning and role-physical with the EQ-5D mobility dimension) have substantial correlations. On the other hand, to meet the requirements of discriminant validity, scales that measure dissimilar constructs (e.g., SF-36 physical functioning with the EQ-5D anxiety/depression dimension) should have low correlations.

CHAPTER 4

RESULTS AND DISCUSSION

I. Response Rate

A total of 314 subjects from the Tucson metropolitan area participated in the study. The two primary recruitment locations were the University of Arizona Health Sciences Center (UA) and the Arizona Motor Vehicle Division (MVD) facility in Tucson. Although it was intended that only 300 subjects participate in the study, an additional fourteen subjects were included to compensate for any potentially unusable questionnaires due to missing data. One subject only partially completed the survey questionnaire and was counted as a non-respondent. Although all participants were assured of confidentiality, this particular subject expressed concerns regarding anonymity and confidentiality and withdrew midway through the questionnaire. Another participant completed all study questions except for the touch-screen administered EQ-VAS. This participant was not included in the analyses because the touch-screen EQ-VAS was one of the primary variables of interest and comparison between scores on the paper and touch-screen EQ-VAS could not be performed. Therefore, the analytical sample was composed of a total of 312 responses; 194 participants from the MVD and 118 participants from UA.

II. Respondent Characteristics

The mean (SD) age for respondents was 35.50 (13.59) years and a little over half (52.5%) of the respondents were male (Table 4.1). The sample from the MVD was significantly older ($t = 3.39$, $p < 0.0008$) and had a greater proportion of females ($\chi^2 = 4.45$, $p < 0.03$) compared to the sample from the University. The majority (57.2%) of respondents were white, non-Hispanic; 17.0% were Hispanic; 10.9% were Asian; and 10.0% self-identified as another race/ethnicity.

Approximately 42% of respondents indicated either having received a bachelor's or graduate (i.e., master's or doctorate/professional) degree. About 28% of respondents had one or more years of college with no degree and 11% of respondents had received an associate's degree. There were fifteen (4.8%) respondents who indicated that they had not graduated from high school. The study sample from the University had higher levels of educational attainment than the sample from the MVD and the general Tucson population.

Respondents reported total annual household income predominantly in the ranges of \$20,001 to \$40,000 (31.5%) and \$40,001 to \$75,000 (21.2%). About 14% of respondents reported annual household incomes at or below \$10,000. Income information was not available for 7.6% of the sample. An overwhelming majority (94.9%) of respondents had some prior computer use experience. This was especially true and as expected for the sample obtained from the University. When asked about which hand they use to normally write, 89.4% of the respondents reported the right hand, twenty–seven respondents (8.7%) reported the left, and 6 (1.9%) respondents indicated that they used either hand to write.

Table 4.1. Respondent Characteristics

Characteristic	MVD Sample n = 194	UA Sample n = 118	Total Sample N = 312
		n (%)	
Age			
Mean (SD)	37.50 (14.57)	32.21 (11.09)	35.50 (13.59)
Median	34.00	27.50	31.00
Sex			
Male	111 (57.22)	53 (44.92)	164 (52.46)
Female	83 (42.78)	65 (55.08)	148 (47.44)
Race/Ethnicity			
White, non- Hispanic	102 (52.58)	76 (64.96)	178 (57.23)
Black, non-Hispanic	12 (6.19)	3 (2.56)	15 (4.82)
Hispanic/Latino	44 (22.68)	9 (7.69)	53 (17.04)
Asian/Pacific Islander	10 (5.15)	24 (20.51)	34 (10.93)
Other	26 (13.40)	5 (4.27)	31 (9.97)
Education			
Not a high school graduate	15 (7.77)	0 (0)	15 (4.82)
High school graduate (or GED)	41 (21.24)	3 (2.54)	44 (14.15)
One or more years of college, no degree	63 (32.64)	25 (21.19)	88 (28.30)
Associate degree (e.g., *AA, AS)*	26 (13.47)	9 (7.63)	35 (11.25)
Bachelor's degree (*e.g., BA, BS*)	31 (16.06)	50 (42.37)	81 (26.05)
Master's degree	13 (6.04)	18 (15.25)	31 (9.97)
Doctorate/Professional degree	4 (2.07)	13 (11.02)	17 (5.47)
Annual Household Income			
$5,000 or less	21 (11.35)	8 (6.84)	29 (9.60)
$5,001 to $10,000	12 (6.49)	2 (1.71)	14 (4.64)
$10,001 to $20,000	29 (15.68)	25 (21.37)	54 (17.88)
$20,001 to $40,000	61 (32.97)	34 (29.06)	95 (31.46)
$40,001 to $75,000	35 (18.92)	29 (24.79)	64 (21.19)
More than $75,000	11 (5.95)	12 (10.26)	23 (7.62)
"Don't Know"	16 (8.65)	7 (5.98)	23 (7.62)
Prior computer use			
Yes	179 (92.27)	117 (99.15)	296 (94.87)
No	15 (7.73)	1 (0.85)	16 (5.13)

III. Descriptive Analyses of Questionnaire Items

The following section illustrates the descriptive findings from the questionnaire including: (1) frequency of different chronic conditions in the study sample; (2) respondents' self-reported health status based on scores on the EQ-5D descriptive system and the SF-36; and (3) respondents' satisfaction with different domains of life.

Chronic Conditions in the Sample

The presence of various chronic conditions was assessed by asking respondents "Have you ever been told by a health professional that you have any of the following conditions?" A little over two-thirds (68.17%) of all respondents indicated they had been diagnosed with at least one chronic condition. One hundred and thirty two (42.14%) respondents had at least two chronic conditions and 4.18% (n = 13) of the respondents reported the presence of five or more. Depression (30.4%), arthritis (23.2%), and migraine headaches (22.7%) were the three most commonly reported conditions among respondents in the MVD sample whereas sinusitis (22.0%), migraine headaches (18.6%), and hay fever (18.6%) were the most commonly reported by the UA sample. Overall, the MVD sample had a higher percentage of respondents who reported at least one chronic condition compared to the UA sample (72.7% vs. 61%). For both samples, the least reported conditions were congestive heart failure, MI, stroke, emphysema, and weak or failing kidneys.

Table 4.2. Reported Chronic Conditions

Condition	MVD n = 194	UA n = 118	Total N = 312
		n (%)	
Depression	59 (30.41)	17 (14.41)	76 (24.36)
Migraine Headaches	44 (22.68)	22 (18.64)	66 (21.15)
Sinusitis	39 (20.10)	26 (22. 03)	65 (20.83)
Arthritis	45 (23.20)	12 (10.17)	57 (18.27)
Hay Fever	33 (17.01)	22 (18.64)	55 (17.63)
Asthma	40 (20.62)	11 (9.32)	51 (16.35)
Hypertension, also called high blood pressure	37 (19.17)	11 (9.32)	48 (15.23)
Ulcer, this could be stomach, duodenal, or peptic ulcer	25 (12.89)	7 (5.93)	32 (10.26)
Chronic bronchitis	15 (7.73)	5 (4.24)	20 (6.41)
Diabetes, also called sugar diabetes	14 (7.22)	2 (1.69)	16 (5.13)
Any kind of liver condition	11 (5.67)	5 (4.24)	16 (5.13)
Cancer or malignancy of any kind	11 (5.67)	4 (3.39)	15 (4.81)
Angina, also called angina pectoris	8 (4.12)	0 (0)	8 (2.56)
Congestive heart failure	5 (2.58)	0 (0)	5 (1.60)
Heart attack, also called myocardial infarction	5 (2.58)	0 (0)	5 (1.60)
Stroke	5 (2.58)	0 (0)	5 (1.60)
Emphysema	5 (2.58)	0 (0)	5 (1.60)
Weak or failing kidneys-Do not include kidney stones, bladder infections or incontinence	5 (2.58)	0 (0)	5 (1.60)
At least 1 chronic condition	141 (72.7)	72 (61.0)	213 (68.2)

[a]N = 312; Multiple responses were permitted

Self-Reported Health Status: EQ-5D Descriptive System

Across both samples, the large majority of respondents indicated no problems in mobility (86.2%), self-care (96.8%), and performing usual activities (87.5%) on the EQ-5D descriptive system. However, a few participants did indicate that they had some problems walking about (13.8%) and performing their usual activities (12%). Most of the respondents also reported having no pain/discomfort (55.8%) and no anxiety/depression (64.3%). The remaining reported moderate (40.4%) and extreme (3.9%) pain/discomfort and moderate (33.4%) and extreme (2.3%) anxiety/depression. The response patterns on all five dimensions of the EQ-5D descriptive system were similar across both samples.

Table 4.3. Self-Reported Health Status: EQ-5D Descriptive System

Category	MVD n = 194	UA n = 118	Total N = 312
		n (%)	
Mobility			
No problems walking about	163 (84.02)	106 (89.83)	269 (86.22)
Some problems walking about	31 (15.98)	12 (10.17)	43 (13.78)
Confined to bed	0 (0)	0 (0)	0 (0)
Self-Care			
No problems with self-care	184 (95.34)	117 (99.15)	301 (96.78)
Some problems washing or dressing myself	9 (4.66)	1 (0.85)	10 (3.22)
Unable to wash or dress myself	0 (0)	0 (0)	0 (0)
Usual Activities			
No problems with performing my usual activities	165 (85.49)	107 (90.68)	272 (87.46)
Some problems with performing my usual activities	26 (13.47)	11 (9.32)	37 (11.90)
Unable to perform my usual activities	2 (1.04)	0 (0)	2 (0.64)
Pain/Discomfort			
No pain or discomfort	96 (49.48)	78 (66.10)	174 (55.77)
Moderate pain or discomfort	87 (44.85)	39 (33.05)	126 (40.38)
Extreme pain or discomfort	11 (5.67)	1 (0.85)	12 (3.85)
Anxiety/Depression (n = 311)			
Not anxious or depressed	120 (62.18)	80 (67.80)	200 (64.31)
Moderately anxious or depressed	67 (34.72)	37 (31.36)	104 (33.44)
Extremely anxious or depressed	6 (3.11)	1 (0.85)	7 (2.25)

Self-Reported Health Status: SF-36

Self-reported health status was also measured using the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36). The SF-36 assesses eight multi-item health scales/concepts: physical functioning, bodily pain, role limitations due to physical problems (i.e., role-physical), role limitations due to emotional problems (i.e., role-emotional), social functioning, emotional well-being, energy/fatigue and general health perceptions. The SF-36 scale scores can range from 0 to100; higher scores indicate better health status. Two summary scores reflecting the physical and mental components of health status also were computed. They are called the physical (PCS) and mental component summary (MCS) scores.

For the combined sample, the respondents' mean scores ranged from 59.3 to 86.6 across the eight scales. Scores were lowest for the vitality scale and highest for the physical functioning scale. Compared to the MVD sample, UA sample respondents had significantly better physical health ($t= -2.635$, $p < 0.008$) as indicated by their physical component summary score. On the other hand, the MVD sample had higher scores compared to the UA sample with respect to the mental component summary score. This difference was however not statistically significant. Across both groups, the average score on the MCS (47.6) was significantly lower than that observed in the general population ($t = - 3.379$, $p < 0.001$).

Table 4.4. Self-Reported Health Status: SF-36

Measure	MVD n = 194	UA n = 118	Total N = 312
		Mean (SD)	
SF-36 Scale			
Physical Functioning	83.09 (23.71)	92.45 (11.52)	86.63 (20.48)
Role-Physical	78.09 (35.21)	82.41 (31.74)	79.73 (33.96)
Role-Emotional	74.05 (34.72)	71.18 (37.97)	72.97 (35.95)
Bodily Pain	77.29 (23.56)	82.22 (18.18)	79.16 (21.79)
Mental Health	73.48 (18.63)	74.0 (15.35)	73.68 (17.44)
Vitality	61.52 (20.28)	55.53 (18.62)	59.26 (19.86)
Social Functioning	76.35 (23.68)	82.62 (18.34)	78.73 (21.99)
General Health	70.69 (21.63)	72.11 (17.11)	71.23 (20.03)
SF-36 Summary Scores			
PCS	49.80 (10.53)	52.77 (8.05)	50.92 (9.76)
MCS	48.18 (10.83)	46.73 (10.48)	47.63 (10.71)

PCS=Physical component summary score; MCS=Mental component summary score

Satisfaction with Different Aspects of Life

Respondents were asked about their level of satisfaction with various aspects of life. The response choices ranged from one to five, where 1 = dissatisfied, 2 = somewhat dissatisfied, 3 = neither satisfied nor dissatisfied, 4 = somewhat satisfied, and 5 = satisfied. Therefore, higher scores indicate a greater level of satisfaction.

The majority of respondents were either satisfied or somewhat satisfied with regard to their job situation (66.0%), leisure time (67%), education (73.7%), neighborhood (76%), transportation (77.4%), housing situation (71.2%), and wealth (62.5%). Furthermore, respondents also indicated being satisfied or somewhat satisfied in their relationships with family members (81.1%), friends and colleagues (84.6%), and spouse or significant others (66.3%). With regard to health, a large majority (80.1%) of subjects were either satisfied or somewhat satisfied. Respondents also indicated high satisfaction with self-worth (81.7%), spiritual life (70.4%) and life as a whole (84%).

Table 4.5. Self-Reported Satisfaction with Life Domains

Category	Mean (SD)	n (%)
Job situation (n = 306)		
Dissatisfied		30 (9.80)
Somewhat dissatisfied		36 (11.76)
Neither satisfied nor dissatisfied	3.78 (1.37)	38 (12.42)
Somewhat satisfied		70 (22.88)
Satisfied		132 (43.14)
Spiritual life (n= 311)		
Dissatisfied		9 (2.89)
Somewhat dissatisfied		29 (9.32)
Neither satisfied nor dissatisfied	3.94 (1.09)	54 (17.36)
Somewhat satisfied		99 (31.83)
Satisfied		120 (38.59)
Relationship with family numbers (n = 312)		
Dissatisfied		10 (3.21)
Somewhat dissatisfied		29 (9.29)
Neither satisfied nor dissatisfied	4.15 (1.09)	20 (6.41)
Somewhat satisfied		97 (31.09)
Satisfied		156 (50.0)
Health (n =312)		
Dissatisfied		11 (3.53)
Somewhat dissatisfied		37 (11.86)
Neither satisfied nor dissatisfied	4.02 (1.12)	14 (4.49)
Somewhat satisfied		122 (39.10)
Satisfied		128 (41.03)
Possessions, wealth and income (n = 312)		
Dissatisfied		24 (7.69)
Somewhat dissatisfied		40 (12.82)
Neither satisfied nor dissatisfied	3.58 (1.20)	53 (16.99)
Somewhat satisfied		120 (38.46)
Satisfied		75 (24.04)
Education (n = 312)		
Dissatisfied		10 (3.21)
Somewhat dissatisfied		46 (14.74)
Neither satisfied nor dissatisfied	3.91 (1.16)	26 (8.33)
Somewhat satisfied		111 (35.58)
Satisfied		119 (38.14)

Table 4.5. Self-Reported Satisfaction with Life Domains *(continued)*

Category	Mean (SD)	n (%)
Relationship with friends and colleagues (n = 311)		
Dissatisfied		3 (0.96)
Somewhat dissatisfied		21 (6.75)
Neither satisfied nor dissatisfied	4.26 (0.93)	24 (7.72)
Somewhat satisfied		107 (34.41)
Satisfied		156 (50.16)
Leisure Time (n =312)		
Dissatisfied		11 (3.53)
Somewhat dissatisfied		43 (17.31)
Neither satisfied nor dissatisfied	3.73 (1.11)	49 (15.71)
Somewhat satisfied		125 (40.06)
Satisfied		84 (26.92)
Housing Situation (n = 312)		
Dissatisfied		15 (4.81)
Somewhat dissatisfied		42 (13.46)
Neither satisfied nor dissatisfied	3.88 (1.21)	33 (10.58)
Somewhat satisfied		99 (31.73)
Satisfied		123 (39.42)
Relationship with spouse/significant others (n = 300)		
Dissatisfied		26 (8.67)
Somewhat dissatisfied		27 (9.0)
Neither satisfied nor dissatisfied	3.81 (1.30)	48 (16.0)
Somewhat satisfied		76 (25.33)
Satisfied		123 (41.0)
Transportation		
Dissatisfied		19 (6.09)
Somewhat dissatisfied		33 (10.58)
Neither satisfied nor dissatisfied	4.06 (1.25)	19 (6.09)
Somewhat satisfied		79 (25.32)
Satisfied		162 (51.92)
Neighborhood (n = 312)		
Dissatisfied		12 (3.85)
Somewhat dissatisfied		30 (9.62)
Neither satisfied nor dissatisfied	4.05 (1.14)	33 (10.58)
Somewhat satisfied		92 (29.49)
Satisfied		145 (46.47)

Table 4.5. Self-Reported Satisfaction with Life Domains ***(continued)***

Category	Mean (SD)	n (%)
Self-Worth (n = 312)		
Dissatisfied		6 (1.92)
Somewhat dissatisfied		19 (6.09)
Neither satisfied nor dissatisfied	4.26 (1.00)	32 (10.26)
Somewhat satisfied		85(27.24)
Satisfied		170 (54.49)
Life as a whole (n = 312)		
Dissatisfied		6 (1.92)
Somewhat dissatisfied		22 (7.05)
Neither satisfied nor dissatisfied	4.27 (1.00)	22 (7.05)
Somewhat satisfied		93 (29.81)
Satisfied		169 (54.17)

IV. Other Survey Characteristics

Completion Time

Respondents took a mean of 9.75 (SD=3.40) minutes to complete the questionnaire. Respondents in the MVD sample took significantly longer to complete the questionnaire as compared to respondents in the UA sample ($t = 2.26$, $p < 0.024$). The difference in time to complete the questionnaire was compared between age, sex, race/ethnicity, education level and order of administration subgroups (Table 4.6.).

Table 4.6. Time to Complete the Survey (minutes)

Characteristic	n	Mean (SD)	Statistic[a]	p-value
Time to complete survey (minutes)				
Total sample	312	9.75 (3.40)		
MVD sample	194	10.09 (3.37)	t = 2.26	0.024
UA sample	118	9.19 (3.39)		
Age (n=312)			t = -4.69	0.0001
< 50	254	9.33 (3.10)		
50 or older	58	11.59 (4.04)		
Sex (n=312)			t = -0.018	0.986
Male	164	9.75 (3.41)		
Female	148	9.76 (3.42)		
Race/Ethnicity (n=311)			F = 4.88	0.0008
White, non- Hispanic	178	9.15 (2.49)		
Black, non-Hispanic	15	9.73 (2.60)		
Hispanic/Latino	53	10.49 (3.68)		
Asian/Pacific Islander	34	11.65 (5.97)		
Other	31	9.94 (3.23)		
Education (n=311)			F = 2.52	0.0297
Not a high school graduate	15	10.0 (2.54)		
High school graduate (or GED)	44	11.07 (4.07)		
1 or more years of college, no degree	88	9.66 (2.99)		
Associate degree (e.g., *AA, AS*)	35	9.14 (3.42)		
Bachelor's degree (*e.g., BA, BS*)	81	9.04 (3.63)		
Graduate degree	48	10.21 (2.95)		
Order of administration (n=312)			t = - 1.432	0.153
Paper EQ-VAS first	154	10.03 (3.33)		
Touch-Screen EQ-VAS first	158	9.48 (3.47)		

[a]Independent group t-test (t); One-way ANOVA (F)

There were statistically significant differences in time to complete the questionnaire by age, race/ethnicity, and education level. Overall, respondents 50 years of age and older took longer to complete the questionnaire as compared to those under 50 (11.59 minutes versus 9.33 minutes, $p < 0.0001$). Blacks, Hispanics, and Asians took longer to complete the questionnaire as compared to non-Hispanic whites. Differences in time to complete were also noted across education level. There were, however, no statistically significant differences in time to complete the questionnaire by sex or order of administration.

Data Quality and Completeness

The number of missing or problematic responses was counted for the paper version of the EQ-VAS. The instructions ask the respondent to indicate his or her current health state by drawing a line from the box marked "Your own health state today" to the appropriate point on the EQ VAS. Twenty-two percent of respondents did not complete the paper EQ-VAS as per those instructions. Although, the percent of problematic or ambiguous responses was higher in the MVD sample (23.2%) compared to the UA sample (19.5%), this difference was not statistically significant. There were three main kinds of problematic responses: (1) placing a mark (i.e., X) on or between two values on the scale; (2) circling multiple values on the scale; or (3) drawing a line from the lowest point on the scale (i.e., 0) to the intended scale score. There were no missing or ambiguous scores for the electronic version of the EQ-VAS because the software only permits the respondent to mark one score on the scale and although it allows for questions to be skipped, it does not allow for questions to be missed unintentionally.

V. Comparison of EQ-VAS Responses

Frequency Distribution of Responses

Histograms were constructed to provide a graphic representation of the frequency distribution of EQ-VAS scores. For each sample, the distributions of scores on the paper and touch screen versions of the EQ-VAS are presented in Figure 4.1 and Figure 4.2, respectively.

The distribution of responses across versions for both samples is very similar. The distributions were markedly negatively skewed with the participants scoring mostly toward the positive end of the scales (i.e., best imaginable health). This was expected since our sample consisted of individuals from the general population. The full range of scores (i.e., 0 to100) was not observed for either version.

Figure 4.1. Distribution of EQ-VAS Scores (paper version)

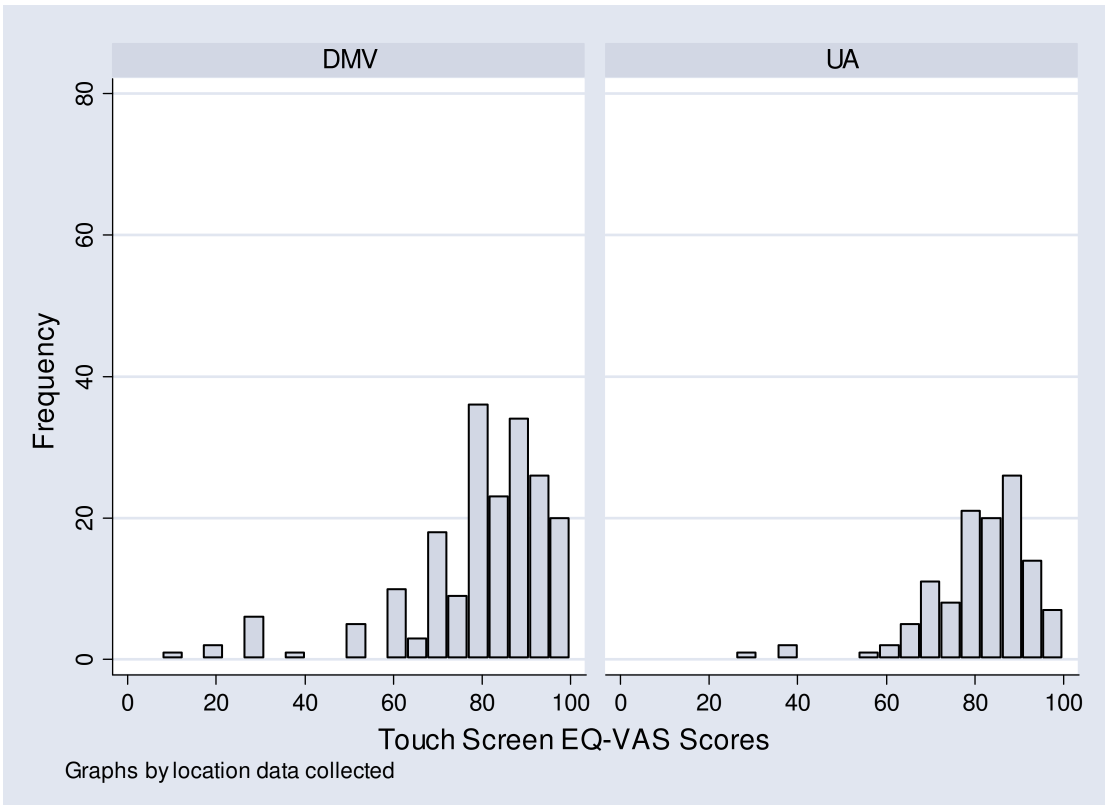


Figure 4.2. Distribution of EQ-VAS Scores (touch screen)

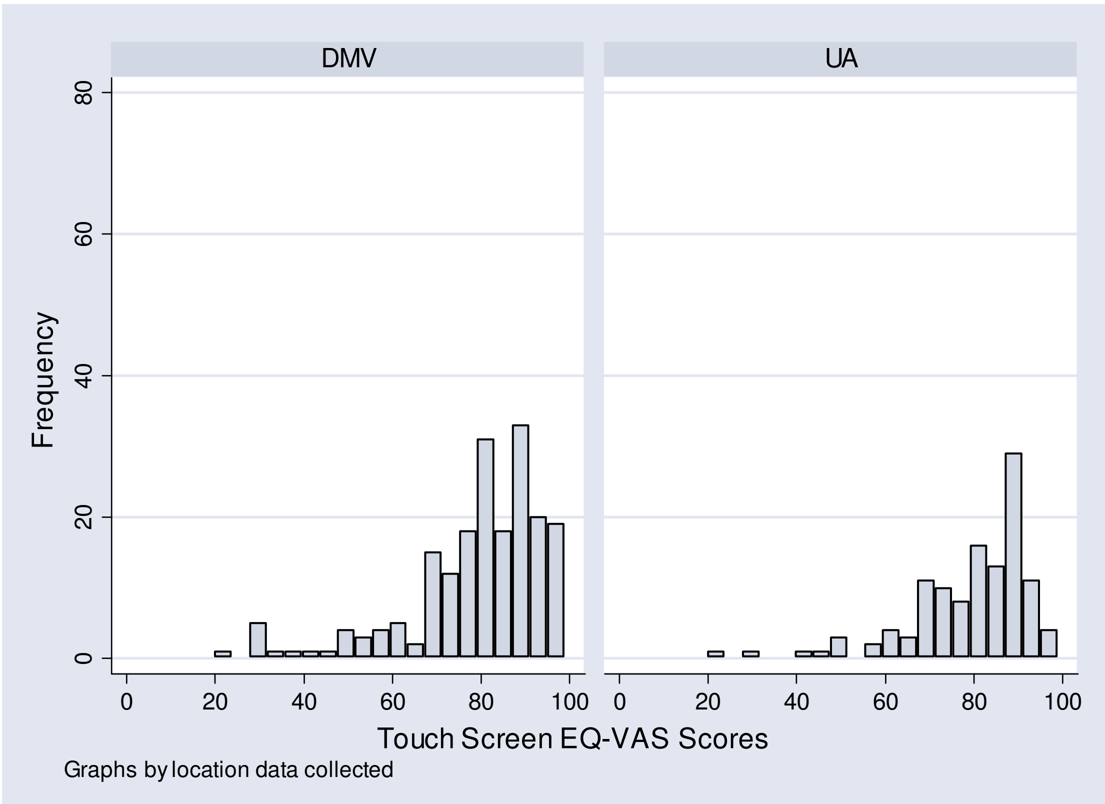
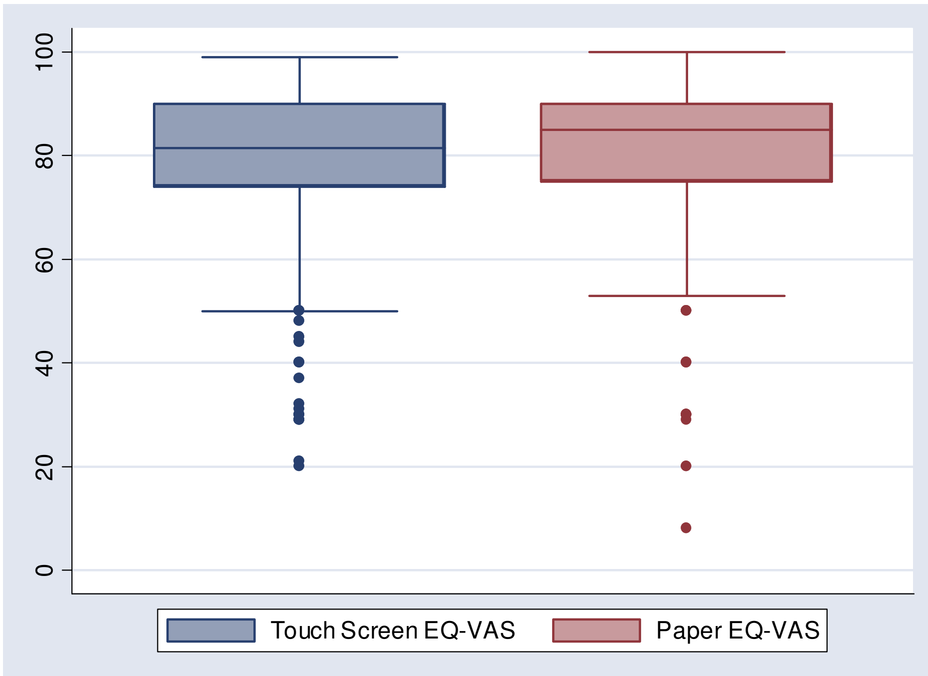


Figure 4.3. Box Plots of Paper and Touch Screen EQ-VAS Scores



Agreement refers to the extent of correspondence between the two methods for measuring the same outcome (i.e., general health status). A descriptive method to quantify agreement was to determine the percent of subjects showing exact agreement and the percent of subjects showing agreement within ± 8 points on the 100-point scale.

The percentage of participants who had identical scores on both versions was 36.4% and 26.3% for the UA sample and MVD sample, respectively (Table 4.7). Overall, a little over 80% of respondents had a difference score of eight points or less, which was the equivalence threshold. Only thirteen (3.85%) respondents had an absolute score difference of greater than 30 points and four participants (1.3%) had differences of 50 points or greater.

Table 4.7. Frequency Distribution of Difference between EQ-VAS Scores

Absolute Difference Score	MVD sample		UA sample		Total Sample	
	n (%)	Cum. %	n (%)	Cum. %	N (%)	Cum. %
0	51 (26.29)	26.29	43 (36.44)	36.44	94 (30.13)	30.13
1	34 (17.53)	43.81	25 (21.19)	57.63	59 (18.91)	49.04
2	20 (10.31)	54.12	3 (2.54)	60.17	23 (7.37)	56.41
3	15 (7.73)	61.86	6 (5.08)	65.25	21 (6.73)	63.14
4	8 (4.12)	65.98	5 (4.24)	69.49	13 (4.17)	67.31
5	5 (2.58)	68.56	10 (8.47)	77.97	15 (4.81)	72.12
6	7 (3.61)	72.16	1(0.85)	78.81	8 (2.56)	74.68
7	7 (3.61)	75.77	2 (1.69)	80.51	9 (2.88)	77.56
8	5 (2.58)	78.35	3 (2.54)	83.05	8 (2.56)	80.13
9	3 (1.55)	79.90	-	-	3 (0.96)	81.09
10	11 (5.67)	85.57	7 (5.93)	88.98	18 (5.77)	86.86
11	3 (1.55)	87.11	3 (2.54)	91.53	6 (1.92)	88.78
12	2 (1.03)	88.14	-		2 (0.64)	89.42
13	3 (1.55)	89.69	2 (1.69)	93.22	5 (1.60)	91.03
14	2 (1.03)	90.72	-	-	2(0.64)	91.67
15	-	-	1(0.85)	94.07	1(0.32)	91.99
16	2 (1.03)	91.75	1(0.85)	94.92	3 (0.96)	92.95
17	-	-	1(0.85)	95.76	1(0.32)	93.27
18	1(0.52)	92.27	-	-	1(0.32)	93.59
19	2 (1.03)	93.30	1(0.85)	96.61	3 (0.96)	94.55
20	-	-	1(0.85)	97.46	1(0.32)	94.87
21	1(0.52)	93.81	-	-	1(0.32)	95.19
23	1(0.52)	94.33	-	-	1(0.32)	95.51
25	1(0.52)	94.85	-	-	1(0.32)	95.83
31	1(0.52)	95.36	-	-	1(0.32)	96.15
32	1(0.52)	95.88	-	-	1(0.32)	96.47
33	1(0.52)	96.39	-	-	1(0.32)	96.79
35	-	-	1(0.85)	98.31	1(0.32)	97.12
37	2(1.03)	97.42	-	-	2(0.64)	97.76
40	-	-	1(0.85)	99.15	1(0.32)	98.08
45	-	-	1 (0.85)	100.00	1(0.32)	98.40
49	1(0.52)	97.94	-	-	1 (0.32)	98.72
50	3 (1.55)	99.48	-	-	3 (0.96)	99.68
72	1 (0.52)	100.00	-	-	1 (0.32)	100.00

Examining Outliers

Across both samples, 13 “outliers” had an absolute difference of 43.15 (SD = 11.32) points between the two versions of the EQ-VAS. When these 13 respondents (i.e., those who had difference scores > 30 points) were eliminated from the analytical sample, the Pearson correlation coefficient and the intraclass correlation coefficient increased to 0.91. The absolute mean (95% CI) difference dropped to 3.94 (3.21 - 4.66) for the MVD sample and 3.29 (2.44 - 4.15) for the UA sample.

In order to determine the association between sociodemographic characteristics and difference scores, these 13 respondents were compared with the rest of the sample on these characteristics (Table 4.8). There were no significant differences between these respondents and the rest of the sample with respect to age and sex. Due to the small cell sizes, for analytical purposes, the race/ethnicity variable was collapsed to only two categories; white, non Hispanic and “Others.”

In addition, there were no significant differences between the two groups with respect to race/ethnicity, education, household income, and handedness. Categories within education level, household income, and handedness were also collapsed for analytical purposes. For the education level variable, master’s degree and doctorate/professional degree were combined to form a new category called “Graduate degree.” Household income was categorized as “less than $20, 000,” “$20,001-$40, 000,” and “More than $40,000.”

General health status was also compared between those who had difference scores greater than or less than 30 points. Scores on the two versions of the visual analog scale were averaged to create a new variable called “mean EQ-VAS.” Respondents who had difference scores of greater than 30 points had significantly ($p = 0.0004$) lower mean EQ-VAS scores compared to those who had difference scores ≤ 30 points. They reported significantly lower scores on both the paper and the touch-screen version of the EQ-VAS. This group of respondents also scored lower on both the physical (PCS) and mental component summary (MCS) scores of the SF-36, with the difference between the two groups on the PCS being statistically significant ($p = 0.02$).

Table 4.8. Comparison of Respondents with Difference Scores ≤ 30 and > 30

Characteristic	Absolute difference score ≤ 30	Absolute difference score > 30	t-statistic	p-value
Age, mean (SD)[a]	35.57 (13.54)	33.76 (15.03)	0.426	0.677
Sex				
Male	157	7		
Female	142	6	0.008	1.000
Total	299	13		
Race/Ethnicity[b]				
White, non-Hispanic	173	5	1.913	0.252
Other	126	8		
Total	298	13		
Education[b]				
Not a high school graduate	14	1		
High school graduate/GED	41	3		
Some college, no degree	86	2	4.88	0.203
Associate/Bachelor's degree	109	7		
Graduate degree (Master's/Doctorate/Professional)	48	0		
Total	298	13		
Annual Household Income[b]				
$20,000 or less	90	7		
$20,001 to $40,000	92	3	4.69	0.100
More than $40,000	86	1		
Total	268	13		
Prior computer use[b]				
Yes	286	10	8.98	0.023
No	13	3		
Total	299	13		
Handedness[b]				
Right	267	12	0.730	1.000
Left/Either	32	1		
Total	299	13		
Health status, mean (SD)[a]				
Mean EQ-VAS	81.08 (13.72)	61.26 (14.72)		0.0004
Paper EQ-VAS	81.74 (14.24)	62.76 (27.24)		0.027
Touch Screen EQ-VAS	80.42 (13.84)	59.76 (27.55)		0.019
SF-36 PCS	51.29 (9.51)	42.46 (11.8)		0.020
SF-36 MCS	47.85 (10.6)	42.72 (12.45)		0.168
Total	299	13		

[a] t-test, [b] Fischer Exact-test

PCS = physical component summary score; MCS = mental component summary score

Digit Bias

Another characteristic observed in VAS scales in general is the tendency for respondents to cluster at certain digits. A greater proportion of respondents scored at multiples of five or ten on the paper version as compared to the touch screen version of the EQ-VAS (Table 4.9). This characteristic has also been noted in other studies collecting data using the EQ-VAS (Alderson and Wittenberg, 1995; Parkin et al., 2004) Rounding off to the closest multiple of five or ten might create an upward bias resulting in overall higher scores on the paper version of the EQ-VAS.

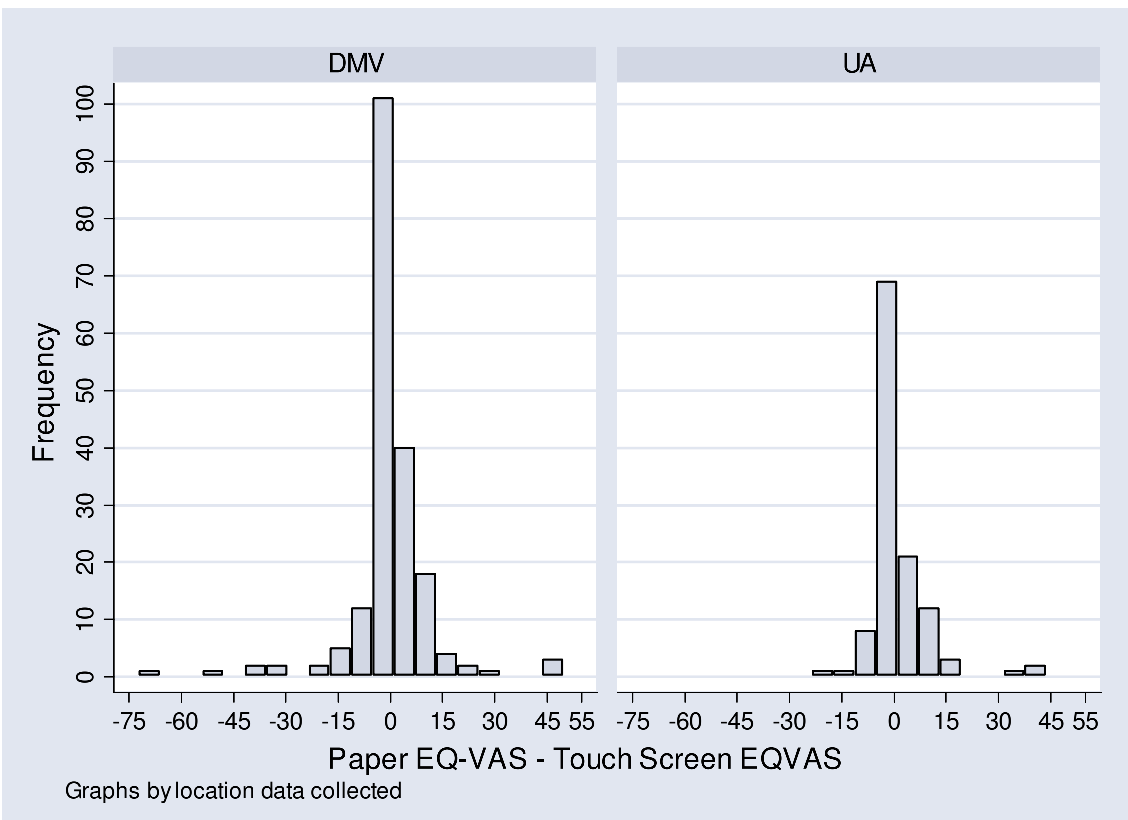
Table 4.9. Scoring at Multiples of 5 and 10 on the Paper and Touch Screen EQ-VAS Versions

Scale	Multiples of 5	Multiples of 10
	n (%)	
Paper EQ-VAS	244 (78.2)	165 (52.9)
Touch screen EQ-VAS	144 (46.2)	99 (31.7)

Testing Difference in Means: Estimate of Bias

A histogram of the differences (paper EQ-VAS - touch screen EQ-VAS) was constructed for both samples. As would be expected, all within subject differences were normally distributed and centered on zero in both samples.

Figure 4.4. Histogram of EQ-VAS Differences (paper -touch screen)



Mean differences were examined between the two administrations using a paired t-test (Table 4.9). Separate analyses were conducted for both samples (MVD and UA). The paired t-test tests the hypothesis that the mean difference between scores on the two versions is equal to zero. There are several assumptions for this test: 1) the observations in the sample were drawn at random from their respective populations; 2) the samples come from populations whose distributions are normal; and 3) the samples come from populations whose variances are equal. This final assumption was checked and results are presented in Table (4.13).

For the MVD sample, there was no statistically significant difference in scores on the two versions of the EQ-VAS whereas for the UA sample, the mean difference in scores on the two versions was significantly greater than zero. Subjects scored significantly higher on the paper EQ-VAS version compared to the touch screen version.

However, determining the difference that is acceptable is not a statistical decision but one based on meaningful differences. According to our definition of equivalence, a mean difference of within ± eight points between the two versions was considered to be less than a meaningfully important difference. A confidence interval for the mean difference was calculated to determine the limits within which the true difference is likely to lie.

The 95% confidence intervals for both samples indicate that the difference in scores was relatively small and below our equivalence threshold of ≤ 8 points. We can be 95% sure that the true mean difference between scores for the UA and MVD samples was between -0.98 and 2.44 points and 0.97 and 3.95, respectively. The mean absolute difference was higher for the MVD sample compared to the UA sample but the 95% confidence interval of the absolute difference for both samples was within the predetermined MID of eight points.

Table 4.10. Testing Mean Difference in EQ-VAS Scores

Measure	N	Mean (SD)	95% CI	t*	P-value
Paper EQ-VAS	312	80.95 (15.41)	79.22 – 82.57		
MVD sample	194	80.25 (17.11)	77.82 – 82.67		
UA sample	118	82.11 (12.10)	79.90 – 84.31		
Touch Screen EQ-VAS	312	79.56 (15.16)	77.95 – 81.21		
MVD sample	194	79.52 (15.87)	77.27 – 81.76		
UA sample	118	79.64 (14.00)	77.09 – 82.20		
Difference	312	1.38 (10.80)	0.18 – 2.59	2.26	0.024
MVD sample	194	0.73 (12.09)	-0.98 – 2.44	0.84	0.400
UA sample	118	2.46 (8.17)	0.97 – 3.95	3.27	0.001
Absolute Difference	312	5.33 (9.48)	4.27 – 6.39		
MVD sample	194	6.01 (10.51)	4.52 – 7.50		
UA sample	118	4.23 (7.41)	2.88 – 5.58		

* Paired t-test (t)

Effect of Order of Administration

In order to determine if the difference in EQ-VAS scores varied across the two administration orders for the questionnaire, further sub analysis were performed for both samples. The order of administration was alternated such that half of the respondents completed the paper EQ-VAS first and the other half completed the touch screen version first. Independent samples t-tests were performed to test the difference in bias between the two sub-groups.

The difference in scores based on the administration order was examined for paper as well as touch screen EQ-VAS in both samples. There were no statistically significant differences between paper EQ-VAS scores when completed first or second across both samples. Touch screen EQ-VAS scores were higher when completed second but this difference was statistically significant only for the sample recruited at the University.

Table 4.11. Difference in EQ-VAS Scores between First and Second Administrations of Paper and Touch Screen Versions

Measure	n	Mean (SD)	95% CI	t*	p-value
MVD Sample	194				
Paper First	95	80.51 (16.79)	77.10 – 83.93	-0.213	0.831
Paper Second	99	79.99 (17.49)	76.50 – 83.58		
Touch Screen First	99	78.07 (15.46)	74.99 – 81.16	-1.295	0.196
Touch Screen Second	95	81.02 (16.22)	77.72 – 84.32		
UA Sample	118				
Paper First	59	83.05 (12.87)	79.70 – 86.40	-0.843	0.400
Paper Second	59	81.17 (11.32)	78.22 – 84.11		
Touch Screen First	59	76.46 (14.55)	72.67 – 80.25	-2.527	0.012
Touch Screen Second	59	82.83 (12.78)	79.50 – 86.16		
Total Sample	312				
Paper First	154	81.49 (15.41)	79.03 – 83.94	-0.604	0.545
Paper Second	158	80.43 (15.45)	78.00 – 82.86		
Touch Screen First	158	77.47 (15.10)	75.10 – 79.84	-2.493	0.013
Touch Screen Second	154	81.71 (14.98)	79.33 – 84.10		

* Independent t-test (t)

In the UA sample, the mean difference in EQ-VAS scores was significantly greater for respondents' who completed the touch screen first followed by the paper version. However, there were no differences between the two groups for the sample obtained at the MVD (Table 4.12).

Table 4.12. Testing Differences in Bias by Order of Administration

Measure	N	Mean (SD) Difference (Paper-Touch Screen)	95% CI	t*	p-value
MVD Sample	194				
Paper First	95	-0.51 (10.83)	-2.71 – 1.70	1.398	0.163
Touch Screen First	99	1.92 (13.14)	-0.70 – 4.54		
UA Sample	118				
Paper First	59	0.22 (3.95)	-0.80 – 1.25	3.093	0.002
Touch Screen First	59	4.71 (10.43)	1.99 – 7.43		
Total Sample	312				
Paper First	154	-0.23 (8.84)	-1.64 – 1.18	2.640	0.008
Touch Screen First	158	2.96 (12.24)	1.04 – 4.88		

Variance Comparison Test

One of the assumptions of the t-test is that the samples come from populations whose variances are equal. In order to confirm this, a two-sample variance-comparison test was performed. It-tests the null hypothesis that the standard deviation in one variable is equal to the standard deviation in another.

To test the null hypothesis, an F value was calculated from the division of one sample variance by another sample variance. The ratio of the two sample variances is distributed as an F distribution. There are two degrees of freedom associated with the F distribution (i.e., one degree of freedom for the numerator variance and another for the denominator variance). The values of the F distribution vary from 0 to + infinity; however, the F tables only give values for alpha less than or equal to 0.50. Therefore, to insure that the calculated F value will be within the range of values given in the table, it is important that the largest of the two sample variances be in the numerator.

As evidenced by the results, the calculated F statistic for both samples (Table 4.13) was less than the critical F value, and hence we failed to reject the null hypothesis that the population variances are equal. Hence, we pooled the variances and performed the t-test.

Table 4.13. Variance Comparison Test

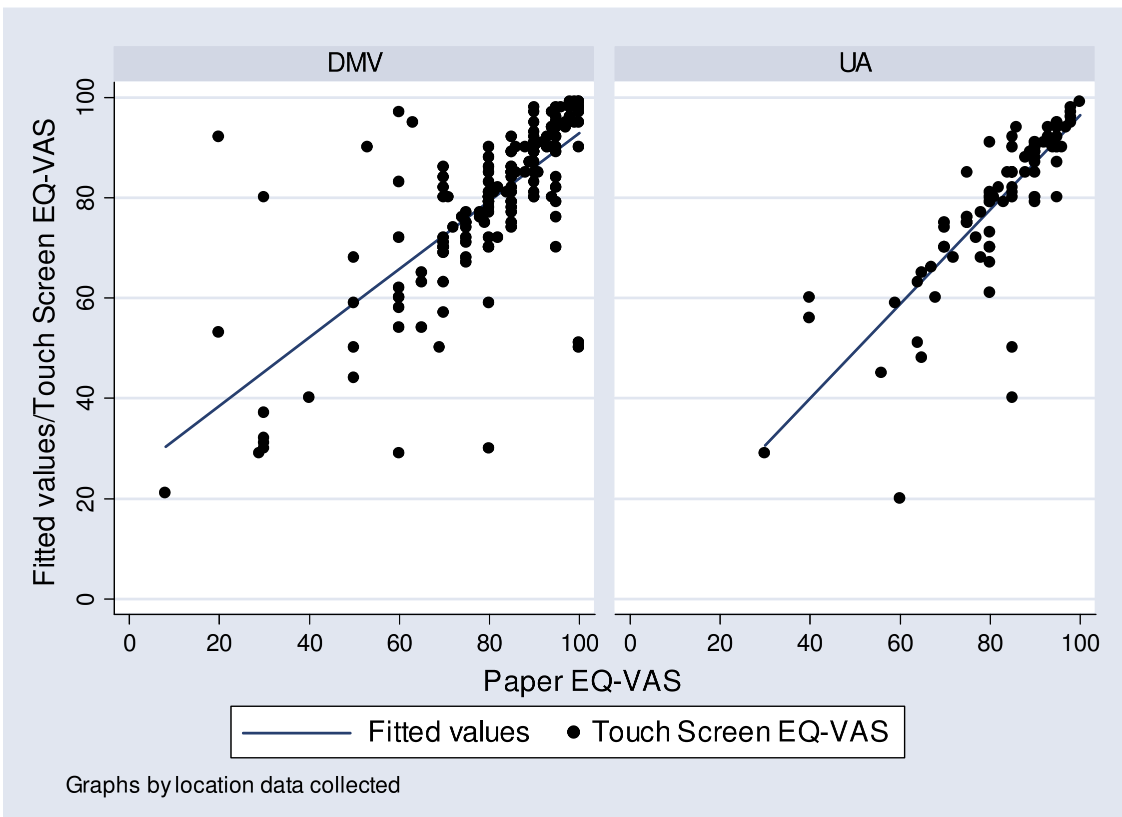
Measure	N	SD	F-statistic	p-value
MVD Sample				
Paper EQ-VAS	194	17.11	0.860	0.296
Touch Screen EQ-VAS	194	15.87		
UA Sample				
Paper EQ-VAS	118	12.10	1.339	0.115
Touch Screen EQ-VAS	118	14.00		
Total Sample				
Paper EQ-VAS	312	15.41	0.968	0.777
Touch Screen EQ-VAS	312	15.17		
Difference (Paper-Touch Screen)				
MVD Sample	194	12.10	1.35	0.246
UA Sample	118	8.17		

Test of Association: Correlation Coefficients

Correlations between the two sets of scores were examined for both samples. A bivariate scatter plot of the EQ-VAS scores from one method against those of the other with a regression line was constructed to examine the relationship.

The touch screen EQ-VAS score is marked on the vertical axis (i.e., y-axis) and the paper EQ-VAS score is marked on the horizontal axis (i.e. x-axis). Each point on the graph represents a pair of values. As shown in Figure 4.5, the line represents the predicted values of touch-screen by paper.

Figure 4.5. Two-Way Scatter Plot of EQ-VAS Scores



However, in such plots, the data points are usually clustered near the line and it is difficult to assess between-method differences. Correlation coefficients were calculated to examine the strength of the linear relationship between the two modes. The most common measure of correlation is the Pearson Product Moment Correlation.

For the data in Table 4.14, the Pearson correlation coefficient (*r*) between the two sets of scores was 0.73 ($p < 0.001$) for the MVD sample and slightly higher in the UA sample (0.81, $p < 0.0001$). Pearson's *r* is a parametric statistic and it assumes that the variables being assessed are normally distributed. As the distribution of both versions of the EQ-VAS were highly skewed, the Pearson correlation coefficient may be less useful if the underlying assumption of normality is violated. Therefore, non-parametric alternatives such as Spearman's rho (ρ) and Kendall's tau were also calculated (Table 4.14).

Spearman's rho assumes that the variables under consideration were measured on at least an ordinal (rank order) scale (i.e., the individual observations can be ranked into two ordered series). Spearman's rho can be thought of as the regular Pearson product moment correlation coefficient, except that Spearman's rho is computed from ranks.

Kendall's tau is equivalent to Spearman's rho with regard to the underlying assumptions. It is also comparable in terms of its statistical power. However, Spearman's rho and Kendall's tau are usually not identical in magnitude because their underlying logic as well as their computational formula's are very different.

The intraclass correlation coefficient was also computed as a measure of agreement or consistency between the two methods of measurement, in conjunction with the significance test of the difference between means (Table 4.14). The ICC is based on mean squares from ANOVA models and is conceptualized as the ratio of between-groups variance to total variance. It has been used in the literature as the preferred measure of strength of association because it corrects for lack of independence between the two sets of scores. The intraclass correlation was 0.747 ($p < 0.0001$), demonstrating a good level of agreement between the two sets of scores.

Past studies have provided reliability estimates in both the general population and patient samples. The EQ-VAS showed good reliability with an ICC of 0.84 (95% CI = 0.65 – 0.78) in a sample (n = 221) of the general population over a three month test period. In a patient sample (n = 224), Hurst et al. (1997) have reported reliability estimates of 0.70 (95% CI = 0.60-0.80) over a three month period and 0.85 (95% CI = 0.67-1.03) over a two week period. Other studies have reported an ICC of 0.80 (95% CI = 0.74-0.85) for the EQ-VAS (Haywood et al., 2002a). Results from the current study demonstrate ICC values falling right within the 95% CI estimated in previous studies (Table 4.14).

Table 4.14. Tests of Association: Correlation Coefficients

Measure	MVD n = 194		UA n = 118		Total N = 312	
	Statistic	95% CI	Statistic	95% CI	Statistic	95% CI
Pearson's *r*	0.733	0.66 - 0.79	0.814	0.74 - 0.87	0.750	0.70 - 0.79
Spearman's rho	0.742	0.67 - 0.80	0.872	0.82 - 0.91	0.787	0.74 - 0.83
Kendall's tau-b	0.644	0.55 - 0.72	0.756	0.67 - 0.82	0.683	0.62 - 0.74
Intraclass correlation (ICC)	0.731	0.66 - 0.79	0.790	0.71 - 0.85	0.747	0.69 - 0.79

However, a high correlation does not mean that the EQ-VAS scores from the two versions agree. The use of correlation can be misleading because the correlation between scores usually indicates association (linear trend) and not necessarily agreement. Data that seem to be in poor agreement can produce quite high correlations.

Although the ICC avoids the problem of a linear relationship being mistaken for agreement, it does not avoid other problems associated with correlation coefficients in general. Correlation coefficients depend on the range of the measurement in the sample. If this is wide, the correlation will be greater than if it is narrow. The number of measurements also affects the observed correlation.

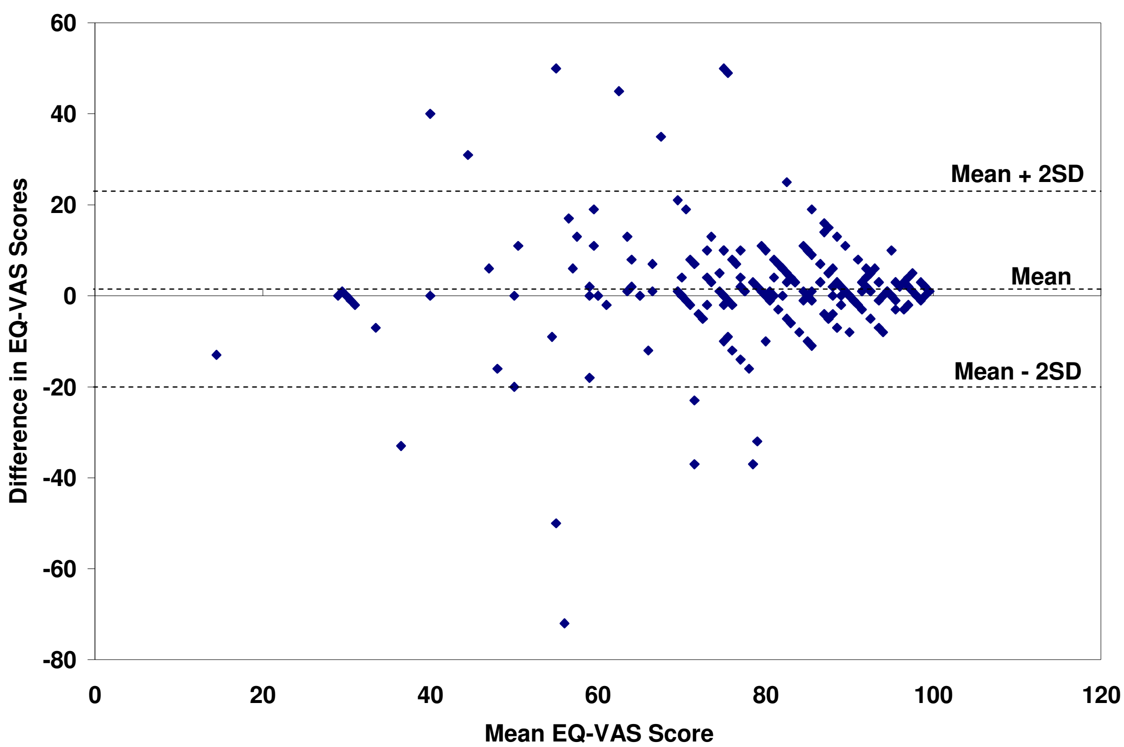
Limits of Agreement

Altman and Bland (1986; Bland and Altman 1999) proposed an alternative analysis technique to compare to measurement techniques. The Bland-Altman plot is a graphical method in which the differences (or alternatively the ratios) between the two measures are plotted against the averages of the two. They proposed examining the differences between the measurements for each subject and if the differences were exactly normally distributed, a multiplier of 1.96 would give the best estimate of the interval within which 95% of the differences would lie. However, in practice, an approximate figure of two is considered satisfactory, and does not rely on the assumption of exact normality. Approximately 95% of the differences will lie between two standard deviations of the mean (d-2SD to d+2SD), which is termed the limits of agreement. Furthermore, since we do not know the true value of the EQ-VAS, the mean of the two EQ-VAS scores (paper and touch screen) serves as the best estimate of it.

Extremely useful insight pertaining to the concordance of scores on the versions can be gained by visually examining the pattern of discord. This type of plot makes it much easier to assess the magnitude of disagreement (both random error and systematic bias), spot outliers, and observe trends. It allows for the separation of both the systematic and random error, which other methods combines into a single measure (Altman and Bland, 1986; Bland and Altman, 1986, 1994).

In Figure 4.6, the vertical axis on the graph shows the difference in scores between the paper and touch screen versions of the EQ-VAS (i.e., paper - touch screen) and the horizontal axis shows the mean EQ-VAS score (i.e., [paper + touch screen]/2) for each subject. According to the plot, the differences do not vary in any systematic way over the range of measurement. There are only 15 responses (i.e., 16/312 = 4.8%) of the responses outside the 2SD limits of agreements.

Figure 4.6. Bland–Altman Plot (Total Sample)



When the plots were constructed separately for each individual sample, there appears to some systematic bias for the UA sample (Fig 4.7). This was also reflected in the preceding analysis for the test for means. A greater proportion of respondents scored higher on the paper version of the EQ-VAS, particularly when they completed it last.

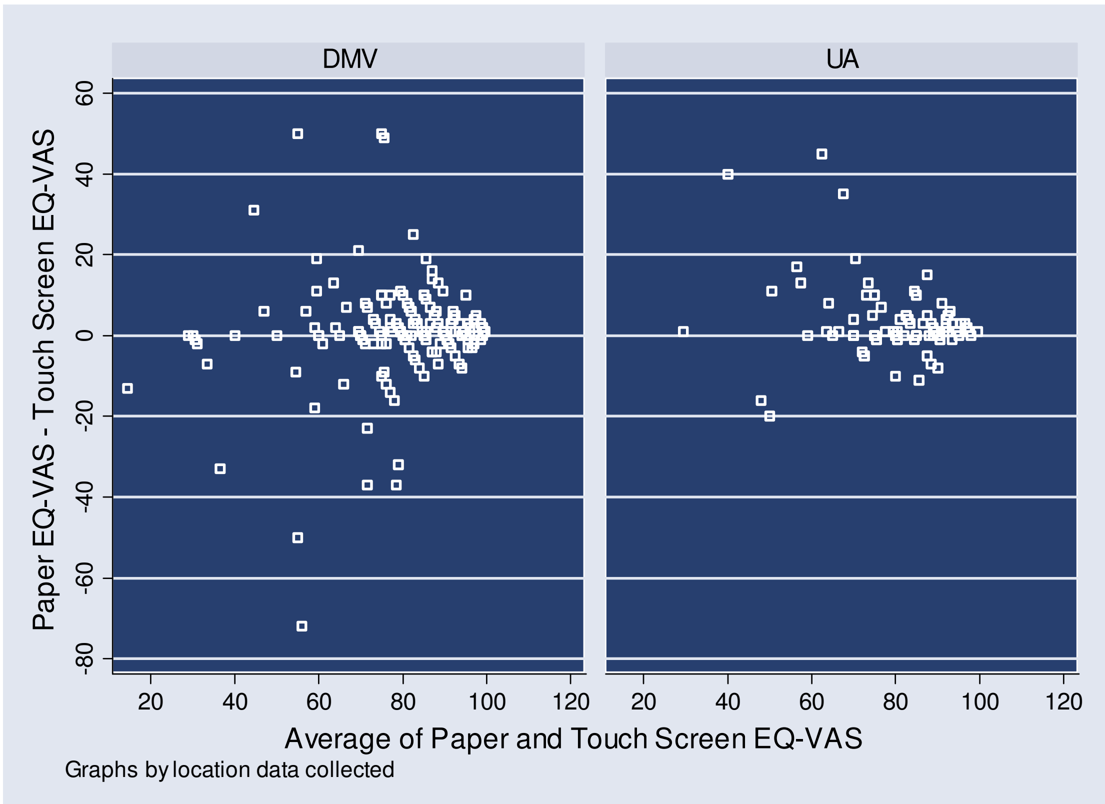
However, there was no association between the difference in EQ-VAS scores and the mean value for both samples (Table 4.15).

Although the 2SD limits of agreement reflect the observed nature of the differences, it is important to compare these observed limits of agreement with a priori ones. These observed limits are wider than the clinically acceptable margin of ± 8 points. This plot is used for method comparison studies that cover a narrower measurement range but when the standard deviation of the difference changes over the range of the measurement, it is recommended to use a percent (or log) difference (on the y-axis) plot.

Table 4.15. Relationship between Magnitude and the Difference

	Mean EQ-VAS Score	
Difference in EQ-VAS scores	MVD (n = 194)	UA (n = 118)
Spearman's rho	0.063	-0.070
p-value	0.383	0.454

Figure 4.7. Bland–Altman Plot for Each Sample



Additional Assessment of Relationship with Difference in EQ-VAS Scores

With the exception of order of administration, there were no differences in bias across various subgroups of sex, race, education, income, health status, and prior computer use.

Table 4.16. Testing Bias for Subgroups

Characteristic	n	Difference Mean (SD)	Statistic [a]	p-value
Age	312		r = - 0.14	0.014
Sex (n = 312)			t = 0.069	0.944
Male	164	1.43 (10.93)		
Female	148	1.35 (10.72)		
Race/Ethnicity (n = 312)			t = -0.151	0.879
White, non-Hispanic	178	1.30 (9.10)		
Other race/ethnicity	134	1.50 (12.75)		
Education (n = 311)			F = 0.79	0.531
Not a high school graduate	15	-2.06 (9.16)		
High school graduate/ GED	44	0.22 (11.96)		
≥ 1 year of college, no degree	88	1.71 (9.94)		
Associate/Bachelor's degree	116	2.31 (12.66)		
Graduate degree (Masters/Doctorate)	48	0.66 (5.70)		
Annual Household Income (n = 279)			F = 0.96	0.384
$20,000 or less	97	2.66 (14.33)		
$20,001 to $40,000	95	1.38 (9.52)		
More than $40,000	87	0.45 (7.25)		
Prior computer use (n = 312)			t = 1.627	0.123
Yes	296	1.73 (10.38)		
No	16	-4.87 (3.07)		
Handedness (n = 312)			t = 0.245	0.807
Right	279	1.44 (10.87)		
Left/Either	33	0.97 (10.30)		
Time to Complete Questionnaire	312		r = -0.133	0.018
Order of Administration (n = 312)			t = 2.64	0.008
Paper EQ-VAS First	154	-0.22 (8.84)		
Touch Screen EQ-VAS First	158	2.96 (12.24)		

Table 4.16. Testing Bias for Subgroups (*continued*)

Characteristic	n	Difference Mean (SD)	Statistic [a]	p-value
General Health Status (n = 312)			F = 1.47	0.210
Excellent	50	1.26 (7.31)		
Very Good	146	1.41 (10.26)		
Good	89	1.58 (11.65)		
Fair	22	3.23 (15.39)		
Poor	5	-9.6 (14.15)		

[a] Pearson correlation coefficient (*r*), Two-sample t-test (t), One-way ANOVA (F)

In addition to testing the measurement equivalence between the two EQ-VAS versions, sociodemographic variables were also evaluated for significant associations with absolute differences (Table 4.17). The absolute difference was significantly different across race/ethnicity group ($z = -2.54$, $p = 0.011$). White, non-Hispanic subjects had lower variability in EQ-VAS scores across the two versions as compared to other race/ethnicity groups. Income level was significantly associated with variability in EQ-VAS scores ($\chi^2 = 6.44$, $p = 0.04$). The mean absolute difference across the various income groups is illustrated in Table 4.17. When tested in post hoc analyses, there was a significant difference in the variability in EQ-VAS scores between the less than $20,000 group and greater than $40,000 group. The time to complete the questionnaire was also significantly associated with variability in EQ-VAS scores ($r_s = 0.120$, $p = 0.033$). Subjects who took longer to complete the questionnaire had a greater difference in scores between the two versions.

Table 4.17. Relationship between Absolute Difference, Health Status, Demographics and Other Variables

Characteristic	n	Absolute Difference Score Mean (SD)	Statistic [a]	p-value
Age	312		$r_s = 0.068$	0.233
Sex (n = 312)				
Male	164	5.19 (9.71)	t = -0.260	0.745
Female	148	5.47 (9.31)		
Race/Ethnicity (n = 312)				
White, non-Hispanic	178	4.47 (8.03)	z = -2.54	0.011
Other race/ethnicity	134	6.49 (11.07)		
Education (n = 311)				
Not a high school graduate	15	4.33 (8.27)	$\chi^2 = 3.11$	0.538
High school graduate/ GED	44	5.68 (10.49)		
≥ 1 year of college, no degree	88	5.42 (8.49)		
Associate/ Bachelor's degree	116	6.05 (11.35)		
Graduate degree (Masters/Doctorate Degree)	48	3.50 (4.52)		
Annual Household Income (n = 279)				
$20,000 or less	97	7.42 (12.52)	$\chi^2 = 6.44$	0.040
$20,001 to $40,000	95	4.43 (8.53)		
More than $40,000	87	4.03 (6.02)		
Prior computer use (n = 312)				
Yes	296	5.02 (9.24)	z = -2.83	0.004
No	16	11.13 (3.07)		
Handedness (n = 312)				
Right	279	5.41 (9.54)	z = 0.195	0.845
Left/Either	33	4.73 (9.17)		
Time to Complete Questionnaire	312		$r_s = 0.120$	0.033
Mean EQ-VAS Score	312		$r_s = -0.208$	0.0002

Table 4.17. Relationship between the Difference Score, Health Status, Demographics and Other Variables (*continued)*

Characteristic	n	Absolute Difference Score Mean (SD)	Statistic [a]	p-value
General Health Status (n = 312)				
Excellent	50	4.10 (6.16)		
Very Good	146	4.33 (9.40)		
Good	89	6.48 (9.78)	χ^2= 6.402	0.171
Fair	22	9.23 (12.59)		
Poor	5	9.6 (14.15)		
Order of Administration (n = 312)				
Paper EQ-VAS First	154	3.47 (8.13)	z = 4.352	< 0.00001
Touch Screen EQ-VAS First	158	7.15 (10.35)		

[a] Two-sample t-test (t), Two-sample Wilcoxon rank-sum (Mann-Whitney) test (z), Kruskal-Wallis Chi-square (χ^2),

Persons who had no prior computer experience had significantly greater difference in scores on the two versions compared to those who did have some prior computer experience (z = -2.83, p = 0.004). General health status represented by the mean EQ-VAS score was also significantly associated with a greater variability in scores across both versions (r_s = - 0.208, p = 0.0002). Subjects who were in poorer health reported more disparate scores on the two versions of the EQ-VAS. As reported earlier, the order of administration was significantly associated with greater difference scores (z = 4.352, p < 0.0001); subjects who completed the touch screen version first had significantly greater variability in scores between the two versions as compared to those who completed the paper EQ-VAS first.

The results of the multiple linear regression for variability in difference scores are presented in Table 4.18. The analysis was conducted to test for the effects of demographic characteristics (i.e., age, education, race/ethnicity, and income level), health status, prior computer use, time to complete the questionnaire, and order of administration on score consistency across the two EQ-VAS versions. Multiple regression analyses assess the impact of these multiple independent variables while controlling for the effect that each may have on the others. The mean absolute difference score served as the dependent variable. Age, sex, race/ethnicity (0 = white, 1 = other race/ethnicity), education level (0 = not a high school graduate, 1 = high school graduate, 2 = Some college, 3 = Associate/Bachelor's degree, 4 = graduate degree), income level (0 = less than \$20,000, 1 = \$20,000-\$40,000, 2 = greater than \$40,000), health status (mean of the two EQ-VAS scores), order of administration, and time to complete the questionnaire served as independent variables.

The model was statistically significant ($F = 5.22$; $p < 0.000001$) and accounted for 16.5% of the variability in EQ-VAS score difference on the two versions. Race/ethnicity, education, income, mean EQ-VAS, and order of administration were significant predictors of the score difference.

Table 4.18. Determinants of Absolute Difference

Independent Variable	Regression coefficient	Std. Error	t-statistic	p-value
Age	0.042	0.048	0.87	0.385
Sex (reference: female)	0.774	1.065	0.73	0.468
Race/ethnicity (reference: white, non-Hispanic)	2.152	1.114	1.93	0.055
Income (reference: < $20,000)				
$20,000-$40,000	-2.841	1.369	-2.08	0.039
> $40,0000	-3.111	1.517	-2.05	0.041
Education (reference: not a high school graduate)				
High School Graduate	5.134	2.991	1.72	0.087
≥ 1 year of college, no degree	5.736	2.809	2.04	0.042
Associate/Bachelor's degree	7.428	2.785	2.67	0.008
Graduate Degree	5.055	3.020	1.67	0.095
Mean EQ- VAS	-0.222	0.038	-5.80	0.000
Order of Administration (reference: Touch Screen first)	-3.351	1.091	-3.07	0.002
Prior Computer Use (reference: yes)	0.029	2.729	0.01	0.992
Time to complete	-0.049	0.172	-0.29	0.772

n = 279, Model Statistic: $F = 5.22$; $p < 0.000001$; Adjusted $R^2 = 0.1649$

VI. Testing Data Quality, Scaling Assumptions, Reliability and Validity of the Touch Screen SF-36

Item-Level Descriptive Statistics

Only one respondent had missing values and less than 1.0% (0.96%) of all responses were missing on the SF-36. The median percent missing for any given item was 0. All responses to the thirty-six items were in the range of possible answers and all response choices were used for all but one question concerning bodily pain (Table 4.19). The items exhibited response variability and although the response distributions were not perfectly symmetrical, they were not extremely skewed either.

Table 4.19. SF-36 Item Response Values Frequency

Item Label	Response Values Frequency 1	2	3	4	5	6
Physical Functioning (PF)						
PF01	43	111	158	_	_	_
PF02	21	38	253	_	_	_
PF03	12	37	263	_	_	_
PF04	16	83	213	_	_	_
PF05	6	32	274	_	_	_
PF06	14	89	209	_	_	_
PF07	18	54	240	_	_	_
PF08	15	42	255	_	_	_
PF09	9	15	288	_	_	_
PF10	9	7	296	_	_	_
Role- Physical (RP)						
RP1	51	261	_	_	_	_
RP2	77	234	_	_	_	_
RP3	62	250	_	_	_	_
RP4	63	249	_	_	_	_
Bodily Pain (BP)						
BP1	104	82	68	40	14	4
BP2	181	89	29	10	3	0
General Health (GH)						
GH1	50	146	89	22	5	_
GH2	4	24	23	102	159	_
GH3	88	125	57	23	19	_
GH4	10	28	88	69	117	_
GH5	69	161	31	31	20	_
Vitality (VT)						
VT1	20	77	85	71	39	19
VT2	18	90	77	80	32	15
VT3	8	24	42	73	117	48
VT4	8	32	55	89	107	21

Table 4.19: SF-36 Item Response Values Frequency (*continued*)

Social Functioning (SF)						
SF1	143	103	38	24	4	_
SF2	3	19	53	189	148	_
Role-Emotional (RE)						
RE1	76	236	_	_	_	_
RE2	108	204	_	_	_	_
RE3	69	243	_	_	_	_
Mental Health (MH)						
MH1	5	11	27	52	113	104
MH2	2	7	11	40	82	169
MH3	21	116	68	61	35	11
MH4	6	11	12	49	132	104
MH5	43	137	62	32	25	3
Health Transition (HT)	43	72	154	38	5	_

Within almost all scales, the item means were fairly comparable. For the bodily pain scale, the two items had a slighter larger discrepancy in mean scores. Although most item standard deviations did not approach 1.0, they were similar within scales (Table 4.20).

Table 4.20. SF-36 Item Descriptive Statistics

Item Label	Abbreviated Item Content	n	Mean	SD
Physical Functioning (PF)				
PF01	Vigorous activities	312	2.37	0.71
PF02	Moderate activities	312	2.74	0.57
PF03	Lifting/carrying groceries	312	2.80	0.48
PF04	Climbing several flights of stairs	312	2.63	0.58
PF05	Climbing one flight of stairs	312	2.86	0.40
PF06	Bending, kneeling, stooping	312	2.63	0.57
PF07	Walk more than a mile	312	2.71	0.57
PF08	Walking several blocks	312	2.77	0.52
PF09	Walking one block	312	2.89	0.39
PF10	Bathing or dressing	312	2.92	0.36
Role- Physical (RP)				
RP1	Cut down time spent on work	312	1.84	0.37
RP2	Accomplished less than would like	311	1.75	0.43
RP3	Limited in kind of work/activities	312	1.80	0.40
RP4	Difficulty performing work/activities	312	1.80	0.40
Bodily Pain (BP)				
BP1	Intensity of bodily pain	312	2.33	1.25
BP2	Extent pain interfered with work	312	1.61	0.86
General Health (GH)				
GH1	Rating of general health	312	2.31	0.88
GH2	I seem to get sick easier than others	312	4.24	0.97
GH3	I seem as healthy as anyone I know	312	2.23	1.12
GH4	I expect my health to get worse	312	3.82	1.13
GH5	My health is excellent	312	2.27	1.11
Vitality (VT)				
VT1	Feel full of pep	311	3.29	1.31
VT2	Have a lot of energy	312	3.20	1.26
VT3	Feel worn out	312	4.32	1.25
VT4	Feel tired	312	4.02	1.20

Table 4.20. SF-36 Item Descriptive Statistics (c*ontinued*)

Item Label	Abbreviated Item Content	n	Mean	SD
Social Functioning (SF)				
SF1	Extent health problems interfered	312	1.86	0.99
SF2	Frequency health problems interfered	312	4.15	0.98
Role-Emotional (RE)				
RE1	Cut down time spent on work	312	1.76	0.43
RE2	Accomplished less than would like	312	1.65	0.48
RE3	Work not done as carefully as usual	312	1.78	0.42
Mental health (MH)				
MH1	Been a nervous person	312	4.82	1.18
MH2	Felt down in the dumps	311	5.25	1.03
MH3	Felt calm and peaceful	312	3.02	1.27
MH4	Felt downhearted and blue	312	4.91	1.12
MH5	Been a happy person	312	2.54	1.15
Health Transition (HT)	Change in health from a year ago	312	2.65	0.92

Item internal consistency was tested by examining the correlation between an item and the scale score computed from all other items in that scale (item scale correlation after correction for overlap) (Table 4.21). Item internal consistency was considered substantial and satisfactory as all items correlated 0.40 or more with its hypothesized scale (after correction for item-scale overlap). Also all items within a scale had roughly equivalent item-scale correlations.

Table 4.21. Pearson Item-Scale Correlations Corrected for Overlap

Item Label	PF	RF	BP	GH	VT	SF	RE	MH
Physical Functioning (PF)								
PF01	**0.64**	_	_	_	_	_	_	_
PF02	**0.79**	_	_	_	_	_	_	_
PF03	**0.78**	_	_	_	_	_	_	_
PF04	**0.75**	_	_	_	_	_	_	_
PF05	**0.77**	_	_	_	_	_	_	_
PF06	**0.72**	_	_	_	_	_	_	_
PF07	**0.78**	_	_	_	_	_	_	_
PF08	**0.80**	_	_	_	_	_	_	_
PF09	**0.74**	_	_	_	_	_	_	_
PF10	**0.60**	_	_	_	_	_	_	_
Role- Physical (RP)								
RP1	_	**0.74**	_	_	_	_	_	_
RP2	_	**0.67**	_	_	_	_	_	_
RP3	_	**0.77**	_	_	_	_	_	_
RP4	_	**0.71**	_	_	_	_	_	_
General Health (GH)								
GH1	_	_	_	**0.70**	_	_	_	_
GH2	_	_	_	**0.50**	_	_	_	_
GH3	_	_	_	**0.69**	_	_	_	_
GH4	_	_	_	**0.48**	_	_	_	_
GH5	_	_	_	**0.74**	_	_	_	_
Vitality (VT)								
VT1	_	_	_	_	**0.53**	_	_	_
VT2	_	_	_	_	**0.64**	_	_	_
VT3	_	_	_	_	**0.63**	_	_	_
VT4	_	_	_	_	**0.67**	_	_	_
Role-Emotional (RE)								
RE1	_	_	_	_	_	_	**0.62**	_
RE2	_	_	_	_	_	_	**0.65**	_
RE3	_	_	_	_	_	_	**0.46**	_

Item Label	PF	RF	BP	GH	VT	SF	RE	MH
Mental health (MH)								
MH1	_	_	_	_	_	_	_	**0.53**
MH2	_	_	_	_	_	_	_	**0.69**
MH3	_	_	_	_	_	_	_	**0.55**
MH4	_	_	_	_	_	_	_	**0.64**
MH5	_	_	_	_	_	_	_	**0.61**

Scale-Level Descriptive Statistics

After item-level analyses were conducted, SF-36 scales were scored and properties of the scales were examined with a focus on scale means and standard deviations and the proportion of respondents scoring at the highest (ceiling) and lowest (floor) level (Table 4.22). The percentage of respondents at the floor for almost all scales was less than 10% and similar to that seen in the general US population (Ware et al. 1994). Floor and ceiling effects were lowest for scales measuring both disability and well-being (GH, VT, and MH). As seen in previous studies, floor effects were higher for the role-physical and role-emotional scales. The role functioning scales also had high ceiling effects, as expected for these coarse scales in general population samples. Ceiling effects were also fairly high for the physical functioning, bodily pain, and social functioning scales. In general, these floor and ceiling effects were very similar to that observed in the general US population.

Table 4.22. Comparison of Descriptive Statistics for SF-36 Scales between the Sample and the General US Population

SF-36 Scale	Mean (US)	Std. Dev (US)	Lowest score (observed/ possible)	Highest score (observed/ possible)	% at observed floor (US)	% at observed ceiling (US)
GH	71.23 (71.95)	20.03 (20.34)	10/0	100/100	0.96 (0)	8.01 (7.40)
PF	86.63 (84.15)	20.49 (23.28)	10/0	100/100	0.64 (0.84)	41.35 (38.79)
RP	79.73 (80.96)	33.96 (34.00)	10/0	100/100	9.94 (10.33)	66.99 (31.85)
RE	72.97 (81.26)	35.95 (33.04)	10/0	100/100	11.54 (9.61)	57.69 (71.01)
SF	78.73 (83.28)	22.00 (22.69)	10/0	100/100	0.32 (0.64)	35.90 (52.32)
BP	79.16 (75.15)	21.79 (23.69)	10/0	100/100	0.64 (0.58)	31.73 (31.85)
VT	59.26 (60.86)	19.86 (20.96)	5/0	100/100	1.60 (0.52)	1.92 (1.50)
MH	73.68 (74.74)	17.44 (18.05)	12/0	100/100	0.32 (0)	2.24 (3.91)

GH = general health perceptions; PF = physical functioning; RP = role limitations due to physical problems; RE = role limitations due to emotional problems; SF = social functioning; BP = bodily pain; VT = vitality; MH = mental health

The scale means and standard deviations for the sample were very similar to that seen in the general US population (Table 4.22). As expected, the five scales which primarily measure health-related disability (physical functioning, role-physical, bodily pain, social functioning, role-emotional) had the highest mean scale scores, while lower mean scores were found for the three scales that extended measurement to the well-being range (general health, vitality, mental health).

Reliability and Correlation between Scales

The reliability of scale scores was estimated using the internal consistency method (Cronbach's coefficient alpha [α]). A minimum reliability level of 0.70 has been suggested for group level comparisons (Nunnally and Bernstein, 1994). All of the reliability coefficients exceeded 0.70; the range was from 0.75 to 0.93 (Table 4.23). The reliability was greater than 0.90 for the physical functioning scale, the suggested minimum standard for individual comparisons.

Table 4.23. SF-36 Scale Internal Consistency: Reliability Coefficients (Cronbach's α) and Inter-Scale Correlations

Scale	GH	PF	RP	RE	SF	BP	VT	MH
GH	(0.82)							
PF	0.50	(0.93)						
RP	0.51	0.62	(0.87)					
RE	0.41	0.25	0.43	(0.75)				
SF	0.57	0.43	0.51	0.57	(0.75)			
BP	0.53	0.65	0.62	0.26	0.49	(0.85)		
VT	0.59	0.37	0.38	0.44	0.54	0.39	(0.80)	
MH	0.50	0.17	0.28	0.56	0.65	0.20	0.62	(0.81)

GH = general health perceptions; PF = physical functioning; RP = role limitations due to physical problems; RE = role limitations due to emotional problems; SF = social functioning; BP = bodily pain; VT = vitality; MH = mental health

To evaluate how distinct each scale is from other scales in the same matrix, correlations among all scales were computed and compared with reliability estimates. As would be expected, in all cases, the reliability coefficients exceeded the inter-scale correlations for the same scale (Table 4.23).

Validity

Content validity was not re-examined since the electronic version did not differ from the original questionnaire with respect to the content. The content validity of the SF-36 has been evaluated based on standards published by Ware (1987), prior to construction of the SF-36.

Construct validity is the extent to which a measure behaves as expected (e.g., the extent to which it correlates in the expected or hypothesized direction and magnitude with other measures/indicators). In tests of known groups' validity, mean scale scores were compared across groups known to differ, using criteria such as the presence of chronic conditions. Mean scale scores between respondents with no self-reported chronic conditions were compared to those with one or more chronic conditions. It was hypothesized that person's with no chronic illnesses would report higher scale scores (i.e., better health) than those with one or more chronic conditions. Respondents with one or more chronic conditions reported significantly lower scores on all eight scales of the SF-36 compared to those with no chronic conditions (Table 4.24).

Table 4.24. Difference in Mean SF-36 Scale and Summary Scores between Groups with Zero and ≥ 1 Chronic Conditions

Scale	No Chronic condition	≥ 1 chronic condition	t-statistic	p-value
	Mean (SD)			
Physical Functioning (PF)	92.47 (16.4)	83.92 (21.6)	3.86	0.0001
Role- Physical (RP)	89.14 (24.3)	75.35 (36.9)	3.93	0.0001
Bodily Pain (BP)	89.77 (12.1)	74.22 (23.5)	7.72	< 0.0001
General Health (GH)	79.85 (15.7)	67.23 (20.6)	5.96	< 0.0001
Vitality (VT)	64.70 (17.8)	56.73 (20.3)	3.52	0.0005
Social Functioning (SF)	86.49 (17.0)	75.12 (23.1)	4.88	< 0.0001
Role-Emotional (RE)	82.83 (30.6)	68.39 (37.4)	3.61	0.0004
Mental Health (MH)	78.75 (14.0)	71.32 (18.4)	3.93	0.0001
Physical Component Summary (PCS) Scale	54.40 (5.6)	49.32 (10.8)	5.47	< 0.0001
Mental Component Summary (MCS) Scale	50.39 (8.6)	46.36 (11.4)	3.46	0.0006

n = 312, t = Independent Samples t-test

Convergent and discriminant validity was also evaluated through examination of correlations between SF-36 scales and the EQ-5D dimensions. Scales measuring similar concepts (e.g., SF-36 physical functioning and role-physical with the EQ-5D mobility score) were expected to have substantial correlations, while scales that measure dissimilar constructs were expected to have low correlations (Table 4.25).

Table 4.25. Convergent and Discriminant Validity: Correlations between SF-36 Scale and Summary Scores and the EQ-5D Dimensions

	EQ-5D Dimension				
SF-36 Scale	**Mobility**	**Self-care**	**Usual Activities**	**Pain**	**Anxiety/ Depression**
Physical Functioning	-0.60	-0.43	-0.54	-0.55	-0.11
Role- Physical	-0.42	-0.31	-0.52	-0.42	-0.22
Bodily Pain	-0.53	-0.31	-0.52	-0.73	-0.22
General Health	-0.40	-0.27	-0.41	-0.42	-0.38
Vitality	-0.29	-0.23	-0.29	-0.31	-0.41
Social Functioning	-0.27	-0.26	-0.26	-0.33	-0.51
Role-Emotional	-0.17	-0.12	-0.18	-0.20	-0.51
Mental Health	-0.13	-0.07	-0.17	-0.18	-0.66
Physical Component Summary (PCS) Scale	-0.58	-0.40	-0.59	-0.62	-0.04
Mental Component Summary (MCS) Scale	-0.05	-0.05	-0.11	-0.09	-0.65

Scores on the physical functioning, bodily pain, and role-physical scales had high correlations with scores on the EQ-5D mobility and usual activities dimensions whereas scores on the mental health and role-emotional scales were highly associated with the EQ-5D anxiety/depression dimension. On the other hand, the scales on the SF-36 measuring physical health had very low correlations with the anxiety/depression dimension on the EQ-5D and the mental health, role-emotional, vitality and social functioning had poor correlations with measures of physical health on the EQ-5D.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

I. Summary

The use of paper questionnaires in research is pervasive and long standing. However, with advances in technology, paper questionnaires can be replaced by electronic versions, which offer numerous advantages. Many generic and disease specific health-related quality of life questionnaires have been adapted to electronic formats and, despite the many potential benefits of electronic data capture, it remains unclear whether electronic versions of these instruments yield similar results to their corresponding paper versions. One such questionnaire for which an electronic version has been developed is the EQ-5D, a measure of self-reported general health status.

The EQ-5D instrument is receiving increasing attention in the US, particularly since it has been included as part of the Medical Expenditure Panel Survey (MEPS) conducted by the US Agency for Healthcare Research and Quality (AHRQ). The EQ-5D has two parts; the first is a descriptive system that classifies respondents into one of 243 distinct health states and the second part is a 20 cm vertical visual analog scale (EQ-VAS). When the EQ-5D has been adapted to an electronic format, the EQ-VAS has been operationalized in ways that depart from the original version. For example, an application developed by Assist Technologies for use on a portable, touch screen device created an electronic version of the EQ-VAS that was modified from the original 20 cm vertical VAS to a 15 cm horizontal VAS.

While equivalence of paper and electronic versions of visual analog scales in general has been studied, no prior published reports of research were found that had examined the measurement equivalence of the horizontal and vertical orientations of the EQ-VAS. When an instrument is to be administered in a manner other than in which it was originally developed and tested, it is important to ensure that the alternative forms of the instrument demonstrate measurement equivalence to the original. The primary aim of this research was to investigate the measurement equivalence of the electronic touch screen EQ-VAS compared to the original paper-based EQ-VAS. A secondary objective was to examine the psychometric properties of the touch screen version of the SF-36.

The sample for this investigation was intended to reflect the general population in the Tucson metropolitan area. The sampling was purposive in that participants were selected subjectively to obtain a sample that generally represented the general population with respect to demographic characteristics, particularly age, sex, and race/ethnicity. Subjects were recruited from two locations; the Division of Motor Vehicles (DMV) regional office and the University of Arizona (UA) in Tucson. Potential participants were eligible for the study if they were 18 years of age and older and could speak and read English.

The study sample, overall, was comparable to the general population in Tucson with respect to age and sex. The race/ethnicity composition of the respondents recruited at the MVD office was more representative of the Tucson general population than the respondents recruited at the UA site. As a result, there were a disproportionately greater number of Asians/Pacific Islanders in the overall study sample due to the UA recruitment. Also, when compared to the general population, there were a greater number of respondents in the “Other” category, which for the sample also included persons who were Aleut/Eskimo or American Indian.

The sample was more educated than the general population in Tucson, largely due to the high number of respondents recruited at the UA with bachelors and graduate degrees. An overwhelming majority of respondents across both recruitment site samples had some prior experience using computers. As would be expected, both samples were in very good or good health as reflected by scores on the EQ-5D and the SF-36. While this may not be the case in most clinical trial settings, evidence from prior research has demonstrated that those patients find the computer mode easier to use and prefer it to paper-based formats.

Although a probability sampling technique would have produced a sample that was more representative of the general population, non-probability purposive sampling did result in a fairly broad cross-section of the population with respect to socio-demographic characteristics.

In order to meet the primary objective of the study, a questionnaire consisting of two parts was specifically designed. One part was paper-based and the other was touch screen computer-based. The device and software for collecting the electronic data was designed and developed by Assist Technologies of Scottsdale, Arizona. The paper-based part only contained the original 20 cm vertical EQ-VAS and the touch screen-based part included a modified EQ-VAS plus the EQ-5D descriptive system, the SF-36, socio-demographic characteristics, a list of chronic conditions, questions about prior computer use and handedness, and items assessing satisfaction with domains of life.

Instructions on how to use the touch screen device were provided and each subject was asked to complete both parts. The EQ-5D and SF-36 are standardized questionnaires, the paper versions of which are widely used and of proven acceptability to patients. Sociodemographic items and questions regarding chronic conditions were included to provide a more detailed description of the sample. Items on prior computer experience and right or left handedness were included to identify how these factors may be related to the respondent's use of the touch screen device. The additional items were also aimed at serving as a distraction by having participants answer other questions between the two versions of the EQ-VAS. Hence, the two versions of the EQ-VAS were administered to each participant approximately ten minutes apart.

The length of the time interval between completion of the two versions was a compromise between reducing memory effects on the one hand and reducing the possibility of extraneous events occurring during the intervening time on the other hand. It was stressed to each participant at the beginning that the questionnaire had two separate parts and they should approach the second part as if it were a separate exercise. It is still possible that some participants remembered and simply repeated the earlier EQ-VAS response. However, there are precedents in the literature of very short inter-version intervals (Maitland et al., 1994). In addition, increasing the interval time was not the preferred route because if the questionnaire were too long, potential subjects may have been reluctant to participate in the study and those who did could have found it burdensome. Pouwer and colleagues (1998) also reported possible contamination from events that occurred during a long inter-version interval. Therefore, for this study, a questionnaire that took approximately 10-15 minutes to complete was considered to be less burdensome yet sufficiently long enough to minimize bias based on memory effects.

Respondents took a mean of 9.75 (SD=3.40) minutes to complete the questionnaire however, there were statistically significant differences in time to complete the questionnaire by age, race/ethnicity, and education level. Respondents 50 years of age and older required more time to complete the questionnaire as compared to those under 50 (11.59 minutes versus 9.33 minutes, $p < 0.0001$). Although dated, past research has found that five percent of people older than 65 use computers, compared with about 45% of young and middle-aged adults (US Census, 1993). The present study also made no additional allowances for potential visual defects or reduced functional abilities among older subjects. The number of older adults with computer experience will increase in the coming decades with computer use becoming more pervasive. However, some consideration needs to be given to these issues when implementing either of these modes of administration in studies involving older respondents.

Non-Hispanic blacks, Hispanics, and Asians took longer to complete the questionnaire as compared to non-Hispanic whites. Evidence suggests that minorities with lower incomes are less likely to own a computer and consequently less likely to be familiar with their use. Difference in time to complete was also noted across education level. Although the vast majority of the sample indicated some prior experience using a computer, the degree of proficiency with computer use was not assessed. Also, respondents were not directly questioned regarding the ease of use of the touch screen device or their preference between the modes of administration. The learning curve associated with the use of the touch screen device may have differed across education and computer proficiency levels. Also, longer and more complex questionnaires need to be compared in a similar manner to fully explore these observed differences.

Respondents at the DMV took significantly longer to complete the questionnaire as compared to respondents in the UA sample (10.1 vs. 9.2 minutes, $p < 0.024$). Since data were collected at an outside location at the DMV, the mid-day sunshine made seeing the touch screen monitor difficult at certain angles. Also, the electronic device could not be plugged into an electrical outlet while respondents were completing the questionnaire and in a few instances, error messages indicating a low battery would pop up on the screen, not allowing respondents to use the system until the messages were deleted. Although battery life has been identified as a problem in past studies, the touch screen device used in this study can run for six days after being fully charged. However, despite the battery concern, no data were lost in the process as the system stores a back up at all times. The above factors could have been partly responsible for the longer time taken by MVD respondents to complete the questionnaire.

Data Quality and Completeness

A major advantage of electronic questionnaires is the ability to collect high quality data with fewer missing or problematic responses. It minimizes the extent to which responses are unusable by restricting the respondent to a single answer and enforcing a response before allowing the respondent to proceed to the next question. Write-in responses are eliminated. In general, all of these advantages were borne out in this investigation.

The computer program was designed to allow respondents to mark only one category, and in an unambiguous manner. Also, patients could alter a response on the touch screen by returning to the previous questions, but they could not move to the next screen without providing a response or explicitly indicating that they didn't wish to respond. However, response problems were detected in many of the paper questionnaires. About 22% of participants marked the paper EQ-VAS ambiguously by not following instructions correctly. Other studies have reported similar problems with paper questionnaires. Ryan et al. (1992) in a study comparing paper and electronic versions of the SF-36 reported that 44% of participants had at least one missing or problematic response (Ryan et al., 2002). Also, Streiner and Norman (1995) in their review found that 5% to 10% of returned paper questionnaires were reported as unusable because of omitted, illegible, or invalid responses.

After the data collection process was completed, the data from the touch screen device was transferred directly to a final computer database keeping data handling and processing to a minimum. On the other hand, for the paper questionnaire approximately three hours were spent entering and verifying entry into the database. This process however, would have been far more time consuming if the paper questionnaire was longer in length, which is generally the case. For this study, only the original EQ-VAS was in a paper format.

Although subjects were not directly asked about the ease of use of the touch screen device, it was evident from the results that respondents found the electronic version easy to use. During the data collection procedure, none of the participants expressed any difficulty following instructions or using the electronic device. Other studies have evaluated subjects' preference for paper and electronic questionnaires. Pouwer et al. (1998) reported that 99% of respondents found it easy to use the electronic diary whereas Drummond et al. (1995) found 57% of respondents preferred the electronic version and 30% had no preference. Murrelle and colleagues (1992) reported that 93% of subjects expressed that the electronic version appeared shorter. Overall, almost all studies have either reported the majority of subjects either preferring the electronic version or liking it as much as the paper version. The efficiency and reliability of computer assessment in comparison to paper instruments was as good or better.

Equivalence of EQ-VAS Scores

Histograms revealed similar patterns and distributions of responses to paper and electronic versions of the EQ-VAS. The equivalence of the touch screen EQ-VAS to the paper version was assessed at the group and individual level. At the group level, the equivalence was assessed by examining mean differences using a paired t-test. One of the assumptions of the t-test is that the samples come from populations whose variances are equal. In order to confirm this, a two-sample variance-comparison test was performed. The calculated F statistic for both samples was less than the critical F value, and hence we failed to reject the null hypothesis that the population variances were equal.

For the MVD sample, there was no statistically significant difference in scores on the two versions of the EQ-VAS whereas for the UA sample, subjects scored significantly higher on the paper EQ-VAS version compared to the touch screen version. A possible explanation for the difference is that people respond differently if they are asked by a machine than if presented with questions on paper. Tourangeau and Smith (1996) found that respondents tended to answer in a more “socially desirable way” to face-to-face interviews compared with computer-assisted self-administered questionnaires. The more positive reporting on the paper EQ-VAS (i.e., being socially-desirable) might be due to a subject’s perception that the information reported on paper questionnaires is less private compared to the touch screen questionnaire.

Another possible reason as to why participants rate their health higher on the paper EQ-VAS may be explained by the format of the paper scale. Participants tend to round their measurements up (in a majority of cases) on the paper version resulting in an upward bias. The reason some people overestimate and some do not, as well as the relationship between these tendencies and other variables, requires further study that is beyond the scope of this research. Also, drawing a line from a neutral box to the VAS on the paper version vs. touching a point on the VAS with the touch screen version may have an influence on the way people respond. However, to the best of our knowledge, we found no prior published literature that had addressed this issue and it remains an area that requires further exploration.

However, determining the difference that is acceptable is not purely a statistical decision but one based on minimally important differences (MID) between the two versions. Defining an MID involved specifying the magnitude of difference in scores between versions that is considered meaningful. A systematic review of the literature by Norman et al. (2003) reported that in all but six studies, the MID estimates were close to one half of a standard deviation (SD). An explanation for this consistency was that research in psychology has shown that the limit of people's ability to discriminate over a wide range of tasks is approximately 1 part in 7, which is very close to half a standard deviation. For a VAS ranging from 0 to 100, a half of a standard deviation is approximately eight points. Therefore, a mean difference of ± eight points between the two versions was specified as the MID, that is, the two formats were deemed equivalent if the score differences were less than or equal to eight points. However, to the best of our knowledge, the MID for the EQ-VAS has not been established. The MID used in this study was based on the best available literature and further confirmation of this MID is needed in studies specifically addressing the EQ-VAS scores.

The 95% confidence intervals for both samples indicated that the difference in scores was relatively small and below our equivalence threshold. Based on the results, we can be 95% sure that the true mean difference between scores for the UA and MVD samples were between -0.98 and 2.44 points and 0.97 and 3.95, respectively. Our findings are consistent with those of other randomized crossover comparisons of paper and touch screen questionnaires for health status assessment. Prior studies have shown uniform psychometric properties between modes of administration, with high correlations and no significant mean differences (Cook et al., 1993; Cook et al., 2004).

In order to determine if the difference in EQ-VAS scores varied across the two administration orders of the questionnaire, further sub-analyses were performed for both samples. There were no statistically significant differences between the group that completed the paper version first and the group that completed the touch screen version first with respect to age, race/ethnicity, sex, education, income, and health status. The difference in scores between first and second testing was examined for paper as well touch screen EQ-VAS in both samples. There was no statistically significant difference between paper EQ-VAS scores when completed first or second for both samples. However, touch screen EQ-VAS scores were consistently higher when completed second, but this difference was statistically significant only for the sample recruited at the University.

On testing the difference in bias by order of administration, the mean difference in EQ-VAS scores was higher for respondent's who completed the touch screen first followed by the paper version compared to those who completed it in the opposite order but this difference achieved statistical significance only for the UA sample. Using the current design, however, it is difficult to provide a precise explanation for this finding.

At the individual level, a little over 30% of the respondents had identical scores on both versions; 36.4% for the UA sample and 26.3% for the MVD sample. A little over 80 percent of respondents had a difference score of eight points or less, which was the equivalence threshold. For a minority of subjects (n = 13 [4.16%]) there were notable

($\geq$ 30 points) differences in paper and electronic versions of the EQ-VAS. It is important to consider whether these response differences were due to random measurement error, the mode of administration or other variables. This discrepancy may be the consequence of some participants having misunderstood the EQ-VAS or a lack of incentive to complete the questionnaire as per the instructions provided.

When these respondents were compared with the rest of the sample in terms of demographic characteristics, there were no significant differences with respect to age, sex, race/ethnicity, education, or income. These respondents, however, were in poorer health as compared to the rest of the sample. They had significantly lower mean EQ-VAS scores and reported significantly lower scores on both the physical and mental component summary scores of the SF-36. On excluding these respondents from the analysis, the absolute difference between the two versions across the two samples was within $\pm$ 4 points.

Correlation coefficients were calculated to examine the strength of the linear relationship between the two modes. The Pearson correlation coefficient between the two sets of scores was good ($r = 0.75$) but as the distribution of EQ-VAS was skewed, non-parametric alternatives such as the Spearman's rho and Kendall's tau were also calculated. The spearman's rho (0.79) and Kendall's tau (0.68) also demonstrated a strong association between the two sets of scores. All the correlation coefficients between the two sets of scores were higher in the UA sample as compared to the MVD sample. The intraclass correlation was 0.75, demonstrating a good level of agreement between the two sets of scores. Results from this study demonstrate ICC values within the 95% CI estimated in previous studies. However, correlations coefficients were given lower weight in the interpretation of results as they are highly influenced by the variability of responses between subjects, and as such are not the best indicators of agreement between measurement methods when variability is low (Bland and Altman, 1990).

Psychometric Properties of the SF-36

There is extensive evidence supporting the reliability, construct validity, and responsiveness of the SF-36 in the general population as well as in samples of patients with particular diseases. A secondary purpose of the study was to examine the psychometric properties of an electronic version of the SF-36. The SF-36 was completed between the two versions of the EQ-VAS. The study evaluated the data quality, scaling assumptions, reliability, and validity of the SF-36. The electronic SF-36 was easy to administer and instructions were modified to suit the touch screen device. Only one respondent had missing values and less than one percent of all SF-36 responses were missing. In the touch screen version of the SF-36, the cognitive burden on respondents was reduced as each item was presented one at a time. None of the respondents expressed any difficulty in completing the electronic SF-36.

The scaling properties of the touch screen SF-36 were very similar to that observed for paper versions. All the items of the SF-36 exhibited good response variability indicating a good spread and the response distributions were neither perfectly symmetrical nor were they extremely skewed. Within most scales, the item means and standard deviations were fairly comparable. Item internal consistency was substantial as all items correlated 0.40 or more with its hypothesized scale (after correction for item-scale overlap). Also all items within a scale had roughly equivalent item-scale correlations.

As seen in previous studies, floor and ceiling effects of the scales were similar to that seen in the general US population. Scale means and standard deviations were comparable to that observed in other studies. The reliability of all scale scores exceeded 0.70 (range = 0.75 to 0.93) and in all cases the reliability coefficients exceeded the inter-scale correlations for the same scale. There was support for the construct validity of the touch screen SF-36 as respondents with no chronic illnesses reported higher scale scores (i.e., better health) than those with one or more chronic conditions. Also, scales measuring similar concepts (i.e., SF-36 physical functioning and role-physical compared with the EQ-5D mobility score) demonstrated substantial correlations (convergent validity), whereas scales that measure dissimilar constructs had low correlations (discriminant validity). Overall, there was a high level of correspondence between the touch screen SF-36 scores and previously reported paper based SF-36 scores in the general population.

II. Limitations

Before placing the results in the context of other investigations of computer-based applications of HRQoL measures, the limitations of this study need to be addressed. A primary limitation of these data relate to the issue of external validity. The sample for this study came from a purposive sampling of the general population in the Tucson metropolitan area. Although purposive, a non-probability sample such as this limits the generalizability of our findings. The sample from the MVD as compared to that from the UA more closely approximated the Tucson general population with respect to age, sex, race/ethnicity, and education level.

With respect to health status, the sample mostly consisted of individuals who were relatively healthy thereby restricting the variance. Most of the EQ-VAS scores were at the high end of the scale and this ceiling effect raises the probability of a Type 2 error. The study sample was also restricted to the English-speaking population. Therefore, it is important, nonetheless to replicate this protocol in different patient populations and other geographic regions.

Second, we did not systematically ask about reasons for refusal to participate. As a result, the non participants are not well characterized. It is possible that only persons who were comfortable and not intimidated with the use of computers self-selected to be in the sample. However, our impression was that most of subjects who refused to participate seemed very busy and/or did not want to do additional work and were not refusing specifically to avoid using computers.

Third, prior computer use in this study was assessed based on a single item asking respondents “Have you used a computer before today?” There are however, varying levels of computer proficiency between individuals and unfamiliarity and lower proficiency for computer use could have accounted for response differences by some subjects.

Fourth, during the data completion process, the short delay between versions of the EQ-VAS might have produced memory effects in our results but this is likely to have been minimized by the intervening battery of questions. Also, other combinations of test-retest reliabilities (e.g., paper vs. paper) in this sample were not estimated and therefore are not available for comparison.

Finally, the specification of a minimally important difference for the EQ-VAS is arbitrary and requires further exploration. However, our study results revealed mean differences between scores on the two modes that were not just small enough to be within the predetermined MID of ± eight points, but also within limits of ± four points (i.e., a quarter of a standard deviation).

III. Conclusions and Recommendations

The purpose of the study was to test the equivalence of an electronic touch screen version of the EQ-VAS with the original 20 cm paper version. Overall, the results indicate measurement equivalence between the two versions of the EQ-VAS with less than an eight point difference between the two sets of scores. In addition, this special case of test-retest reliability (e.g., paper vs. touch screen) resulted in ICC similar to prior test-retest analyses using the paper-based EQ-VAS (Dolan, 1994; Hurst et al. 1997).

The touch screen EQ-VAS also generated higher quality data as no items were missed or marked ambiguously. Due to a lower proportion of end-digit bias, the touch screen EQ-VAS may also be more precise than the paper version, and has major advantages in terms of flexibility, data quality, and the speed of data handling.

Also, in the present study the electronic SF-36 demonstrated reliability and construct validity comparable to the original paper version. In addition, convergent and discriminant validity requirements were satisfied. Two further psychometric aspects were also considered in this study, missing values and ceiling and floor effects. The proportion of respondents at floor and ceiling were comparable to that seen in the general US population and the proportion of missing values was much lower due to the ease of use and simplicity of the touch screen device.

Such advantages, however, may be lost if the subjects are not comfortable with the electronic device or if the study personnel do not provide sufficient introduction to its use. It is therefore very important to ensure that the functioning of the device is tested beforehand and that both research personnel and respondents are properly instructed regarding its use. Training for subjects must include the most basic elements of using the device, as well as features that are specific to the software application.

Despite the initial higher cost of computer equipment and programming time, routine collection of data using electronic methods may ultimately prove to be less expensive than employing a staff member to enter and verify responses to paper questionnaires. These findings support the validity of electronic questionnaires and along with ease of use and high data quality, suggest strong potential for their expanded use in health status assessment.

IV. Recommendations for Future Research

Future comparative studies could be enhanced by sampling both paper and touch screen responses on repeated occasions within and between days, to provide multiple estimates of test-retest reliabilities. In this investigation, the effects of length (20 cm vs. 15 cm) of the EQ-VAS scale and orientation (vertical vs. horizontal) could not be isolated as each format was completed only once by each participant. This is an area that requires further exploration. Further investigation is also needed to evaluate the influence of longer inter-version intervals between the paper and electronic formats.

Also, computer proficiency could possibly be assessed through a more detailed standardized task assessing comprehension of on screen instructions. The potential effects of visual presentation style and novelty on acceptance and preference for electronic questionnaires could be assessed though comparison of responses to questions presented in different formats and/or color schemes.

The results from this study apply to the sample in which the methodology was tested. If these data collection devices are to be used in clinical trials, it will be imperative to assess their reliability and acceptability in the specific patient populations in which they may be used. More definitive testing should be conducted to determine if specific subgroups of the population provide less valid and reliable data via these devices. Future studies should be conducted to test the performance of this technology in more diverse populations incorporating variations in language, race/ethnicity, educational level, disease, and burden of disease.

Although references are made to the potential economic value of touch-screen technology, there is a lack of detailed and thorough economic studies in the published literature. Economic evaluations of this method of data collection in comparison with more traditional methods are necessary to provide further evidence of its cost-effectiveness.

As subjective end-points of health related quality of life are being included with increasing frequency alongside traditional endpoints as measures of treatment effectiveness, electronic self-reporting of such endpoints in clinical trials could potentially enhance the data collection process. In the coming years, it is likely that electronic data capture will become the primary mode of data collection in many areas of health care practice and research. If this is the case, strategies for ensuring that all respondents have a sufficient level of proficiency with the data collection task should be implemented.

APPENDIX A

HUMAN SUBJECTS APPROVAL LETTERS

Human Subjects Protection Program
http://www.irb.arizona.edu

THE UNIVERSITY OF
ARIZONA®
TUCSON ARIZONA

1350 N. Vine Avenue
P.O. Box 245137
Tucson, AZ 85724-5137
(520) 626-6721

15 October 2004

Stephen Coons, Ph.D.
Sulabha Ramachandran, M.S.
Department of Pharmacy Practice/Science
PO Box 210207

RE: USEFULNESS OF TOUCH SCREEN TECHNOLOGY FOR HEALTH STATUS ASSESSMENT

Dear Dr. Coons and Mr. Ramachandran:

We received documents concerning your above cited project. Regulations published by the U.S. Department of Health and Human Services [45 CFR Part 46.101(b)(2)] exempt this type of research from review by our Institutional Review Board. **Note: A copy of your disclaimer form, with IRB approval stamp affixed, is enclosed for duplication and use in enrolling subjects.**

Please be advised that **clearance from academic and/or other official authorities for site(s) where proposed research is to be conducted must be obtained prior to performance of this study. Evidence of this must be submitted to the Human Subjects Protection Program office.**

Exempt status is granted with the understanding that no further changes or additions will be made either to the procedures followed or to the consenting instrument used (copies of which we have on file) without the review and approval of the Human Subjects Committee and your College or Departmental Review Committee. Any research related physical or psychological harm to any subject must also be reported to each committee.

Thank you for informing us of your work. If you have any questions concerning the above, please contact this office.

Sincerely,

Rebecca Dahl, R.N., Ph.D.
Director
Human Subjects Protection Program

cc: Departmental/College Review Committee

APPENDIX B

STUDY FLYER

STUDY PARTICIPANTS NEEDED

The University of Arizona is recruiting 300 people in the Tucson metropolitan area to test the use of a touch-screen device for collecting health-related information. Your participation in this study will take about 15 minutes.

You must be:

- 18 years of age or older
- able to read and understand English

Upon completion, you will be compensated.

APPENDIX C

SUBJECT DISCLAIMER FORM

SUBJECT'S DISCLAIMER FORM

Project Title: **USEFULNESS OF TOUCH SCREEN TECHNOLOGY FOR HEALTH STATUS ASSESSMENT**

You are being invited to voluntarily participate in the above-titled research study. The purpose of the study is to determine the usefulness of a touch screen technology for health status assessment. You are eligible to participate because you are (1) 18 years of age or older and (2) able to read and understand English

If you agree to participate, you will be asked to complete a questionnaire regarding your health and satisfaction with other aspects of your life. Your participation will take approximately 15 minutes. By completing the questionnaire, you are giving permission for the investigator to use your information for research purposes only.

Your participation in the study is completely voluntary and any questions that you have will be answered. You may choose to not answer some or all of the questions and you may withdraw from the study at any time. There are no known risks associated with participating in this study and also no cost to you except for your time. In addition, no direct benefit from your participation is expected and on completion of the survey, you will receive $5 for your participation.

Only the principal investigator, Stephen Joel Coons, PhD, and his research collaborators will have access to your data. All information obtained will be kept in a secure place. Your name will not be recorded or collected in any way. Your responses will be kept confidential and will be used for research purposes only.

You can obtain further information from the principal investigator at (520) 626-4450. If you have questions concerning your rights as a research subject, you may call the University of Arizona Human Subjects Protection Program office at (520) 626-6721.

Thank you.

APPENDIX D

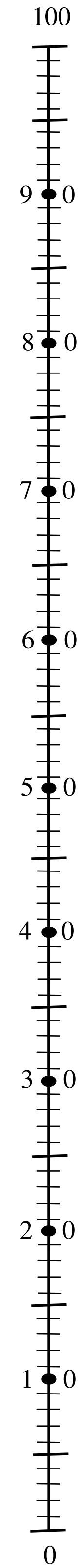
STUDY QUESTIONNAIRE

STUDY QUESTIONNAIRE

(PAPER-BASED PART*)

***NOTE: This one item is the original EQ-VAS**

Best imaginable health state



Worst imaginable health state

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your health state is today.

Your own health state today

STUDY QUESTIONNAIRE

(TOUCH-SCREEN BASED PART*)

***NOTE: Items 2–6 are from the EQ-5D. Items 7-42 represent the complete SF-36**

By touching one answer below, please indicate which statement best describes your own health state today.

2. **Mobility**

❑ I have no problems walking about	❑ I have some problems walking about	❑ I am confined to bed

3. **Self-Care**

❑ I have no problems with self-care	❑ I have some problems washing or dressing myself	❑ I am unable to wash or dress myself

4. **Usual Activities** (*e.g. work, study, housework, family or leisure activities)*

❑ I have no problems walking about	❑ I have some problems walking about	❑ I am confined to bed

5. **Pain/Discomfort**

❑ I have no problems with performing my usual activities	❑ I have some problems with performing my usual activities	❑ I am unable to perform my usual activities

6. **Anxiety/Depression**

❑ I am not anxious anxious or depressed	❑ I am moderately anxious or depressed	❑ I am extremely or depressed

7. In general, would you say your health is:

Excellent	Very Good	Good	Fair	Poor
☐	☐	☐	☐	☐

8. **Compared to one year ago**, how would you rate your health in general **now**?

Much better now than one year ago	Somewhat better now than one year ago	About the same as one year ago	Somewhat worse now than one year ago	Much worse now than one year ago
☐	☐	☐	☐	☐

The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
9. <u>Vigorous activities</u>, such as running, lifting heavy objects, participating in strenuous sports	☐	☐	☐
10. <u>Moderate activities</u>, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐	☐	☐
11. Lifting or carrying groceries	☐	☐	☐
12. Climbing <u>several</u> flights of stairs	☐	☐	☐
13. Climbing <u>one</u> flight of stairs	☐	☐	☐
14. Bending, kneeling, or stooping	☐	☐	☐

15. Walking more than a mile	☐	☐	☐
16. Walking several blocks	☐	☐	☐
17. Walking one block	☐	☐	☐
18. Bathing or dressing yourself	☐	☐	☐

During the **past 4 weeks,** have you had any of the following problems with your work or other regular daily activities **as a result of your physical health?**

	YES	NO
19. Cut down on the amount of time you spent on work or other activities	☐	☐
20. Accomplished less than you would like	☐	☐
21. Were limited in the kind of work or other activities	☐	☐
22. Had difficulty performing the work or other activities (for example, it took extra effort)	☐	☐

During the **past 4 weeks,** have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)?

	YES	NO
23. Cut down the amount of time you spent on work or other activities	☐	☐
24. Accomplished less than you would like	☐	☐

25. Didn't do work or other activities <u>as carefully as usual</u> ☐ ☐

26. During the **past 4 weeks,** to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

Not at all	Slightly	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

27. How much **bodily** pain have you had during the **past 4 weeks**?

None	Very mild	Mild	Moderate	Severe	Very severe
☐	☐	☐	☐	☐	☐

28. During the **past 4 weeks,** how much did **pain** interfere with your normal work (including both work outside the home and housework)?

Not at all	Slightly	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

These questions are about how you feel and how things have been with you **during the past 4 weeks.** For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the **past 4 weeks**. . .

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
29. Did you feel full of pep?	☐	☐	☐	☐	☐	☐

30. Have you been a very nervous person?	☐	☐	☐	☐	☐	☐
31. Have you felt so down in the dumps that nothing could cheer you up?	☐	☐	☐	☐	☐	☐
32. Have you felt calm and peaceful?	☐	☐	☐	☐	☐	☐
33. Did you have a lot of energy?	☐	☐	☐	☐	☐	☐
34. Have you felt downhearted and blue?	☐	☐	☐	☐	☐	☐
35. Did you feel worn out?	☐	☐	☐	☐	☐	☐
36. Have you been a happy person?	☐	☐	☐	☐	☐	☐
37. Did you feel tired?	☐	☐	☐	☐	☐	☐

38. During the **past 4 weeks,** how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

How TRUE or FALSE is <u>each</u> of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
39. I seem to get sick a little easier than other people.	☐	☐	☐	☐	☐
40. I am as healthy as anybody I know.	☐	☐	☐	☐	☐
41. I expect my health to get worse.	☐	☐	☐	☐	☐
42. My health is excellent.	☐	☐	☐	☐	☐

43. What is your current age? ________ (in years)

44. What is your sex?

❒ Male

❒ Female

45. What race/ethnicity do you consider yourself to be?

❒ American Indian/Native American

❒ Aleut/Eskimo

❒ Asian/Pacific Islander

❒ Black/African American (Non-Hispanic)

❒ White (Non- Hispanic)

❒ Hispanic/Latino

❒ Other

❒ Don't Know

46. What is the highest degree or level of school that you have ever completed? If currently enrolled, mark the previous grade or highest degree received.

❒ No schooling completed

❒ Not a high school graduate

❒ High school graduate (or GED)

❒ 1 or more years of college, no degree

❒ Associate degree (*for example: AA, AS*)

❒ Bachelor's degree (*for example: BA, AB, BS*)

❒ Master's degree

❒ Doctorate or professional degree

47. What is your total annual household income before deductions for taxes or anything else (e.g., social security, child support)

    ❒ $5,000 or less

    ❒ $5,001 to $10,000

    ❒ $10,001 to $20,000

    ❒ $20,001 to $40,000

    ❒ $40,001 to $75,000

    ❒ More than $75,000

    ❒ Don't Know

48. Have you used a computer before today?

    ❒ Yes

    ❒ No

49. With which hand do you normally write?

    ❒ Right

    ❒ Left

    ❒ Either

Have you ever been told by a health professional that you have any of the following conditions? Check all that apply.

	Yes	No
50. Hypertension, also called high blood pressure	❒	❒
51. Arthritis	❒	❒
52. Diabetes, also called sugar diabetes	❒	❒
53. Congestive heart failure	❒	❒
54. Angina, also called angina pectoris	❒	❒
55. Heart Attack. also called myocardial infarction	❒	❒
56. Stroke	❒	❒
57. Chronic bronchitis	❒	❒
58. Emphysema	❒	❒
59. Asthma	❒	❒
60. Depression	❒	❒
61. Migraine Headaches	❒	❒
62. Sinusitis	❒	❒
63. Hay Fever	❒	❒
64. Ulcer, this could be stomach, duodenal, or peptic ulcer	❒	❒
65. Any kind of liver condition	❒	❒
66. Weak or failing kidneys-Do not include kidney stones, bladder infections or incontinence	❒	❒
67. Cancer or malignancy of any kind	❒	❒

Given below are some words and phrases that people use to identify various aspects of their lives. For each of the words or phrases, please circle the number that best indicates how you feel about that particular aspect of your life.

		Dissatisfied	Somewhat dissatisfied	Neither satisfied nor dissatisfied	Somewhat satisfied	Satisfied
68.	**Job situation**	1	2	3	4	5
69.	**Spiritual life**	1	2	3	4	5
70.	**Relationship with family members**	1	2	3	4	5
71.	**Health**	1	2	3	4	5
72.	**Possessions, wealth, and income**	1	2	3	4	5
73.	**Education**	1	2	3	4	5
74.	**Relationship with friends and colleagues**	1	2	3	4	5
75.	**Leisure time**	1	2	3	4	5
76.	**Housing situation**	1	2	3	4	5
77.	**Relationship with spouse/significant other**	1	2	3	4	5
78.	**Transportation**	1	2	3	4	5
79.	**Neighborhood**	1	2	3	4	5
80.	**Self-worth**	1	2	3	4	5
81.	**Life as a whole**	1	2	3	4	5

82. How would you rate your overall quality of life?

❒ Excellent

❒ Very Good

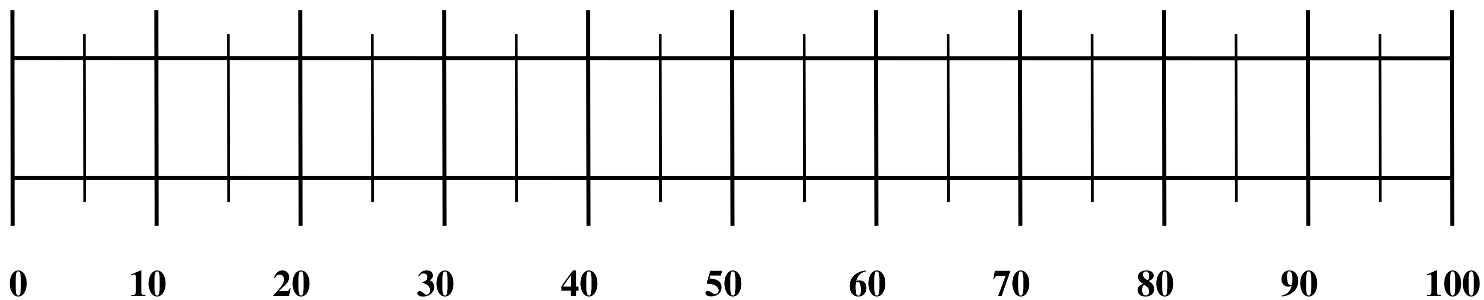
❒ Good

❒ Fair

❒ Poor

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by touching the scale below at whichever point on the scale indicates how good or bad your health state is today.

Worst Imaginable Health State



Best Imaginable Health State

REFERENCES

Aitken RC. Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine* 1969; 62: 989-993.

Allenby A, Matthews J, Beresford J, McLachlan SA. The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. *European Journal of Cancer Care* 2002; 11: 245-253.

Altman DG, Bland JM. Comparison of methods of measuring blood pressure. *Journal of Epidemiology and Community Health* 1986; 40: 274-277.

Anderson RT, Aaronson NK, Bullinger M, McBee WL. A review of the progress towards developing health-related quality-of-life instruments for international clinical studies and outcomes research. *Pharmacoeconomics.* 1996; *10(4):336-355*.

Badia X, Diaz-Prieto A, Rue M, Patrick DL. Measuring health and health state preferences among critically ill patients. *Intensive Care Medicine* 1996; 22: 1379-1384.

Beaton DE. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Medical Care* 2003; *41(5):593-596*.

Bergner M. Quality of life, health status, and clinical research. *Medical Care* 1989; 27: S148-S156.

Bergner M, Rothman ML. Health status measures: an overview and guide for selection. *Annual Review of Public Health* 1987; *8:191-210*.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307-310.

Bland JM, Altman DG. Correlation, regression, and repeated data. *British Medical Journal* 1994; 308: 896.

Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part 1--Correlation within subjects. *British Medical Journal* 1995; 310: 446.

Bland JM, Altman DG. Measurement error and correlation coefficients. *British Medical Journal* 1996; 313: 41-42.

Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; 8: 135-160.

Bliven BD, Kaufman SE, Spertus JA. Electronic collection of health-related quality of life data: Validity, time benefits, and patient preference. *Quality of Life Research* 2001; 10: 15-21.

Boyes A, Newell S, Girgis A. Rapid assessment of psychosocial well-being: are computers the way forward in a clinical setting? *Quality of Life Research* 2002; 11: 27-35.

Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. *Journal of Health Service Research Policy* 1999; 4: 174-184.

Brazier J, Jones N, Kind P. Testing the validity of the EuroQol and comparing it with the SF-36 health survey questionnaire. *Quality of Life Research* 1993; 2: 169-180.

Brooks R. EuroQol: the current state of play. *Health Policy* 1996; 37: 53-72.

Bushnell DM, Martin ML, Parasuraman B. Electronic versus paper questionnaires: A further comparison in persons with asthma. *Journal of Asthma* 2003; 40: 751-762.

Buxton J, White M, Osoba D. Patients' experiences using a computerized program with a touch-sensitive video monitor for the assessment of health-related quality of life. *Quality of Life Research* 1998; 7(6):513-519.

Campbell E, Peterkin D, Abbott R, Rogers J. Encouraging underscreened women to have cervical cancer screening: the effectiveness of a computer strategy. *Preventive Medicine* 1997; 26: 801-807.

Carlson LE, Speca M, Hagen N, Taenzer P. Computerized quality-of-life screening in a cancer pain clinic. *Journal of Palliative Care* 2001; 17: 46-52.

Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. *Pain* 1983; 16: 87-101.

Caro Sr JJ, Caro I, Caro J, Wouters F, Juniper EF. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research* 2001; 10: 683-691.

Cella DF, Perry SW. Reliability and concurrent validity of three visual-analogue mood scales. *Psychological Reports* 1986; 59: 827-833.

Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; 20: 37-46.

Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; 70: 213-220.

Cohen J. *Statistical Power Analysis for the Social Sciences*, Second edition. New York: Academic Press, 1977.

Cook AJ, Roberts DA, Henderson MD, Van Winkle LC, Chastain DC, Hamill-Ruth RJ. Electronic pain questionnaires: A randomized, crossover comparison with paper questionnaires for chronic pain assessment. *Pain* 2004; 110: 310-317.

Cook DJ, Guyatt GH, Juniper E, Griffith L, McIlroy W, Willan A, Jaeschke R, Epstein R. Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. *Journal of Clinical Epidemiology* 1993; *46(6):529-34.*

Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. *Pharmacoeconomics* 2000; 17: 13-35.

Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology* 2003; 56: 395-407.

Davies B, Burrows G, Poynton C. A comparative study of four depression rating scales. *Australian and New Zealand Journal of Psychiatry* 1975; 9: 21-24.

Dexter F, Chestnut DH. Analysis of statistical tests to compare visual analog scale measurements among groups. *Anesthesiology* 1995; 82: 896-902.

Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials* 1991; 12: 142S-158S.

Deyo RA, Patrick DL. Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care* 1989; 27: S254-S268.

Dixon JS, Bird HA. Reproducibility along a 10 cm vertical visual analogue scale. *Annals of Rheumatic Diseases* 1981; 40: 87-89.

Dolan P. Search for a critical appraisal of EuroQol: a response by the EuroQol group to Gafni and Birch. *Health Policy. 28(1):67-9,* 1994.

Dolan P, Gudex C, Kind P, Williams A. Valuing health states: A comparison of methods. *Journal of Health Economics* 1996; 15: 209-231.

Dolan PD. Modeling Valuations for EuroQol Health States. *Medical Care* 1997; 35: 1095-1108.

Dolan PD, Roberts JP. Modeling Valuations for EQ-5D Health States: An Alternative Model Using Differences in Valuations. *Medical Care* 2002; 40: 442-446.

Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Annals of the Rheumatic Diseases* 1978; 37: 378-381.

Drummond HE, Ghosh S, Ferguson A, Brackenridge D, Tiplady B. Electronic quality of life questionnaires: a comparison of pen-based electronic questionnaires with conventional paper in a gastrointestinal study. *Quality of Life Research* 1995; 4: 21-26.

Drummond M, Jonsson B, Rutten F. The role of economic evaluation in the pricing and reimbursement of medicines. *Health Policy* 1997; 40: 199-215.

Elton D, Burrows GD, Stanley GV. The relationship between psychophysical and perceptual variables and chronic pain. *British Journal of Clinical Psychology* 1979; 18: 425-430.

Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Medical Care* 1997; 35: 522-537.

Essink-Bot ML, van Royen L, Krabbe P, Bonsel GJ, Rutten FF. The impact of migraine on health status. *Headache* 1995; 35: 200-206.

Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability *Educational and Psychological Measurement* 1973; 33: 613-619.

Fleiss JL: *The Design and Analysis of Clinical Experiments*. New York, John Wiley and Sons, 1986, pp 62, 66, 76-77, 103, 369

Folstein MF, Luria R. Reliability, validity, and clinical application of the Visual Analogue Mood Scale. *Psychological Medicine* 1973; 3: 479-486.

Fuller RW, Hallett C, Dahl R. Assessing equivalence of inhaled drugs. *Respiratory Medicine* 1995; 89: 525-527.

Gandek B, Ware JE Jr. Methods for validating and norming translations of health status questionnaires: the IQOLA Project approach. International Quality of Life Assessment. *Journal of Clinical Epidemiology* 1998a; 51: 953-959.

Gandek B, Ware JE Jr., Aaronson NK, Alonso J, Apolone G, Bjorner J, Brazier J, Bullinger M, Fukuhara S, Kaasa S, Leplege A, Sullivan M. Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology* 1998b; 51: 1149-1158.

Gift AG. Visual analogue scales: measurement of subjective phenomena. *Nursing Research* 1989; 38: 286-288.

Gift AG, Plaut SM, Jacox A. Psychologic and physiologic factors related to dyspnea in subjects with chronic obstructive pulmonary disease. *Heart and Lung* 1986; 15: 595-601.

Graham W, Smith P, Kamal A, Fitzmaurice A, Smith N, Hamilton N. Randomised controlled trial comparing effectiveness of touch screen system with leaflet for providing women with information on prenatal tests. *British Medical Journal* 2000; 320: 155-160.

Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. *Quality of Life Research* 1996; 5: 521-531.

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings* 2002; 77(4):371-383.

Guyatt GH, Townsend M, Berman LB, Keller JL. A comparison of Likert and visual analogue scales for measuring change in function. *Journal of Chronic Diseases* 1987; 40: 1129-1133.

Hahn EA, Cella D. Health outcomes assessment in vulnerable populations: Measurement challenges and recommendations. *Archives of Physical Medicine and Rehabilitation* 2003; 84: S35-S42.

Hahn EA, Cella D, Dobrez D, Shiomoto G, Marcus E, Taylor SG, Vohra M, Chang CH, Wright BD, Linacre JM, Weiss BD, Valenzuela V, Chiang HL, Webster K. The talking touch screen: A new approach to outcomes assessment in low literacy. *Psycho-Oncology* 2004; 13: 86-95.

Hanscom B, Lurie JD, Homa K, Weinstein JN. Computerized questionnaires and the quality of survey data. *Spine* 2002; 27: 1797-1801.

Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 2000; 18(5):419-23.

Haywood KL, Garratt AM, Dziedzic K, Dawes PT. Generic measures of health-related quality of life in ankylosing spondylitis: reliability, validity and responsiveness. *Rheumatology* 2002; 41: 1380-1387.

Honaker CR, Robbins MF. Quality: the healthcare watchword. *Group Practice Journal* 1988; 37(3):10-14.

Hornblow AR, Kidson MA. The visual analogue scale for anxiety: a validation study. *Australian and New Zealand Journal of Psychiatry* 1976; 10: 339-341.

Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H. Validity of EuroQol--a generic health status instrument--in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *British Journal of Rheumatology* 1994; 33: 655-662.

Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *British Journal of Rheumatology* 1997; 36: 551-559.

Huskisson EC. Measurement of pain. *Lancet* 1974; 2: 1127-1131.

Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology* 1991*;* 59(1):12-19.

James M, Leger S, Rowsell KV. Prioritizing elective care: a cost utility analysis of orthopaedics in the north west of England. *Journal of Epidemiology and Community Health* 1996; 50: 182-189.

Johannes CB, Crawford SL, Woods J, Goldstein RB, Tran D, Mehrotra S, Johnson KB, Santoro N. An electronic menstrual cycle calendar: comparison of data quality with a paper version. *Menopause* 2000; 7(3):200-208.

Johnson JA, Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. *Quality of Life Research* 1998; 7: 155-166.

Joyce CRB, Zutshi DW, Hrubes V, Mason RM. Comparison of fixed interval and visual analogue scales for rating chronic pain. *European Journal of Clinical Pharmacology* 1975; 8: 415-420.

Juniper EF. Quality of life questionnaires: does statistically significant = clinically important? *Journal of Allergy & Clinical Immunology* 1998; 102(1):16-17.

Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *Journal of Clinical Epidemiology* 1994; 47(1):81-87.

Kahn G. Computer-based patient education: a progress report. *MD Computing* 1993; 10: 93-99.

Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Medical Care* 1989; 27(3 Suppl):S178-89.

Kelleher CJ, Pleil AM, Reese PR, Burgess SM, Brodish PH. How much is enough and who says so?. The case of the King's Health Questionnaire and overactive bladder. *British Journal of Obstetrics and Gynecology* 2004; 111: 605-612.

Kleinman L, Leidy NK, Crawley J, Bonomi A, Schoenfeld P. A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire. *Medical Care* 2001; 39: 181-189.

Kohlmeier L, Mendez M, McDuffie J, Miller M. Computer-assisted self-interviewing: a multimedia approach to dietary assessment. *American Journal of Clinical Nutrition* 1997*;* 65(4 Suppl):1275S-1281S.

Kongpatanakul S, Strom B. Quality of Life, health status and clinical drug research. *Pharmacoeconomics* 1992; 2: 8-14.

Kremer E, Atkinson JH, Ignelzi RJ. Measurement of pain: patient preference does not confound pain measurement. *Pain* 1981; 10: 241-248.

Leidy NK, Rentz AM, Zyczynski TM. Evaluating Health-related quality-of-life outcomes in patients with congestive heart failure. A review of recent randomized controlled trials. *Pharmacoeconomics.* 1999a; 15: 19-46.

Leidy NK, Revicki DA, Geneste B. Recommendations for Evaluating the Validity of Quality of Life Claims for Labeling and Promotion. *Value in Health* 1999b; 2: 113-127.

Little JC, McPhail NI. Measures of depressive mood at monthly intervals. *British Journal of Psychiatry* 1973; 122: 447-452.

Lohr KN. Health outcomes methodology symposium: summary and recommendations. *Medical Care* 2000; 38(9 Suppl): II194-208. .

Lorig K. Measurement of pain. *Nursing Research* 1983; 33: 376.

Ludington E, Dexter F. Statistical analysis of total labor pain using the visual analog scale and application to studies of analgesic effectiveness during childbirth. *Anesthesia and Analgesia* 1998; 87: 723-727.

Luria RE. The validity and reliability of the visual analogue mood scale. *Journal of Psychiatric Research* 1975; 12: 51-57.

Lydick E, Epstein RS. Interpretation of quality of life changes. *Quality of Life Research* 1993; 2(3):221-226.

Maitland ME, Mandel AR. A client-computer interface for questionnaire data. *Archives of Physical Medicine and Rehabilitation* 1994; 75: 639-642.

Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports* 1978; 62: 1037-1040.

Malpas A, Legg NJ, Scott DF. Effects of hypnotics on anxious patients. *British Journal of Psychiatry* 1974; 124: 482-484.

Mantha S, Thisted R, Foss J, Ellis JE, Roizen MF. A proposal to use confidence intervals for visual analog scale data for pain measurement to determine clinical significance. *Anesthesia and Analgesia* 1993; 77: 1041-1047.

Maultsby MC, Slack WV. A computer-based psychiatry history system. *Archives of General Psychiatry* 1971; 25: 570-572.

McCormack HM, Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. *Psychology and Medicine* 1988; 18: 1007-1019.

McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine* 1997; 127: 743-750.

McHorney CA. Health status assessment methods for adults: past accomplishments and future challenges. *Annual Review of Public Health* 1999; 20: 309-335.

Melzack R. The McGill Pain Questionnaire: Major properties and scoring methods. *Pain* 1975; 1: 277-299.

Miller MD, Ferris DG. Measurement of subjective phenomena in primary care research: the Visual Analogue Scale. *Family Practice Research Journal* 1993; 13: 15-24.

Mohr DC, Burke H, Beckner V, Merluzzi N. A preliminary report on a skills-based telephone-administered peer support program for patients with multiple sclerosis. *Multiple Sclerosis* 2005; 11(2):222-226.

Morrison DP. The Crichton Visual Analogue Scale for the assessment of behaviour in the elderly. *Acta Psychiatrica Scandinavica* 1983; 68: 408-413.

Mullen KH, Berry DL, Zierler BK. Computerized symptom and quality-of-life assessment for patients with cancer part II: acceptability and usability. *Oncology Nursing Forum* 2004; 31: E84-E89.

Murrelle L, Bulger JD, Ainsworth BE, Holliman SC, Bulger DW. Computerized mental health risk appraisal for college students: user acceptability and correlation with standard pencil-and-paper questionnaires. *American Journal of Health Promotion* 1992; 7: 90-92.

Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analog scale: is it linear or nonlinear? *Anesthesia and Analgesics* 1999; 89: 1517-1520.

Newell S, Girgis A, Sanson-Fisher RW, Stewart J. Are touch screen computer surveys acceptable to medical oncology patients? *Journal of Psychosocial Oncology* 1997; 15: 37-46.

Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 2003; 41(5):582-592.

Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care* 2001; 39(10):1039-1047.

Norum J, Angelsen V, Wist E, Olsen JA. Treatment costs in Hodgkin's disease: a cost-utility analysis. *European Journal of Cancer* 1996; 32A: 1510-1517.

Nyholm D, Kowalski J, Aquilonius SM. Wireless real-time electronic data capture for self-assessment of motor function and quality of life in Parkinson's disease. *Movement Disorders* 2004; 19: 446-451.

Ohnhaus EE, Adler R. Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. *Pain* 1975; 1: 379-384.

Padilla GV, Presant C, Grant MM, Metter G, Lipsett J, Heide F. Quality of life index for patients with cancer. *Research in Nursing and Health* 1983; 6: 117-126.

Parkin D, Rice N, Jacoby A, Doughty J. Use of a visual analogue scale in a daily patient diary: modelling cross-sectional time-series data on health-related quality of life. *Social Science and Medicine* 2004; 59: 351-360.

Philip BK. Parametric statistics for evaluation of the visual analog scale. *Anesthesia and Analgesia* 1990; 71: 710.

Pouwer F, Snoek FJ, van der P, Heine RJ, Brand AN. A comparison of the standard and the computerized versions of the Well-being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research* 1998; 7: 33-38.

Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.

Price DD, Harkins SW, Baker C. Sensory-affective relationships among different types of clinical and experimental pain. *Pain* 1987; 28: 297-307.

Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. *Pain* 1994; 56: 217-226.

Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain* 1983; 17: 45-56.

Rabin JM, McNett J, Badlani GH. "Compu-Void II": the computerized voiding diary. *Journal of Medical Systems* 1996; 20: 19-34.

Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Annals of Medicine* 2001; 33(5):337-343.

Rampling DJ, Williams RA. Evaluation of group processes using visual analogue scales. *Australia and New Zealand Journal of Psychiatry* 1977; 11: 189-191.

Revicki DA, Osoba D, Fairclough D, Barofsky I, Berzon R, Leidy NK, Rothman M. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Quality of Life Research* 2000; 9: 887-900.

Revill SI, Robinson JO, Rosen M, Hogg MI. The reliability of a linear analogue for evaluating pain. *Anesthesia* 1976; 31: 1191-1198.

Richards DM, Irving MH. Assessing the quality of life of patients with intestinal failure on home parenteral nutrition. *Gut* 1997; 40: 218-222.

Robinson RG, McHugh PR, Folstein MF. Measurement of appetite disturbances in psychiatric disorders. *Journal of Psychiatric Research* 1975; 12: 59-68.

Ryan JM, Corry JR, Attewell R, Smithson MJ. A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Quality of Life Research* 2002; 11: 19-26.

Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. *Pharmacoeconomics. 15(2):141-55,* 1999.

Schipper H. Guidelines and caveats for quality of life measurement in clinical practice and research. *Oncology* 1990; 4(5):51-57.

Schmitz N, Hartkamp N, Brinschwitz C, Michalek S, Tress W. Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. *Acta Psychiatrica. Scandinavica.* 2000; 102: 147-152.

Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; 15: 657-680.

Scott J, Huskisson EC. Graphic representation of pain. *Pain* 1976; 2: 175-184.

Scott J, Huskisson EC. Vertical or horizontal visual analogue scales. *Annals of the Rheumatic Diseases* 1979; 38: 560.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979;86:420-428.

Skinner HA, Allen BA. Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting and Clinical Psychology* 1983; 51: 267-275.

Slack WV, Hicks GP, Reed CE, Van Cura LJ. A computer-based medical-history system. *New England Journal of Medicine* 1966; 274: 194-198.

Slack WV, Peckham BM, Van Cura LJ, Carr WF. A computer-based physical examination system. *The Journal of the American Medical Association* 1967; 200: 224-228.

Slack WV, Van Cura LJ. Patient reaction to computer-based medical interviewing. *Computers and Biomedical Research* 1968; 1: 527-531.

Sloan JA, Cella D, Frost M, Guyatt GH, Sprangers M, Symonds T, Clinical Significance Consensus Meeting Group. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clinic Proceedings* 2002; 77(4):367-370.

Sloan JA, Cella D, Frost MH, Guyatt G, Osoba D. Quality of life III: translating the science of quality-of-life assessment into clinical practice-an example-driven approach for practicing clinicians and clinical researchers. *Clinical Therapeutics.* 2003; 25 Suppl D: D1-D5.

Spilker B. Standardization of quality of life trials: an industry perspective. *Pharmacoeconomics* 1992; 1: 73-75.

Sriwatanakul K, Kelvie W, Lasagna L, Calim JF, Weis OF, Mehta G. Studies with different types of visual analog scales for measurement of pain. *Clinical Pharmacology and Therapeutics* 1983; 34: 234-239.

Sullivan DF. A single index of mortality and morbidity. *HSMHA Health Report* 1966; 86: 347-354.

Taenzer PA, Sauve L, Burgess ED, Milkavich L, Whitmore B. The health interview: automated assessment in a multidisciplinary outpatient hypertension treatment program. *MD Computing* 1996; 13(5):423-426.

Taenzer PA, Speca M, Atkinson MJ, Bultz BD, Page S, Harasym P, Davis JL. Computerized quality-of-life screening in an oncology clinic. *Cancer Practice* 1997; 5: 168-175.

The EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy* 1990; 16: 199-208.

Torrance GW, Feeny D, Furlong W. Visual Analog Scales: Do They Have a Role in the Measurement of Preferences for Health States? *Medical Decision Making* 2001; 21: 329-334.

US Bureau of the Census. Use of computers at home, school, and work by persons 18 years and older. US Bureau of Census Washington, DC, 1993

Van Cura LJ, Slack WV, Frey SR. Elements of a computer medical interview system. *Biomedical Sciences Instrumentation* 1971; 8: 33-42.

Velikova G, Stark D, Selby P. Quality of life instruments in oncology. *European Journal of Cancer* 1999a; 35(11):1571-1580. .

Velikova G, Wright EP, Smith AB, Cull A, Gould A, Forman D, Perren T, Stead M, Brown J, Selby PJ. Automated collection of quality-of-life data: a comparison of paper and computer touch-screen questionnaires. *Journal of Clinical Oncology* 1999b; 17(3): 998-1007.

Ware JE Jr. Standards for validating health measures: definition and content. *Journal of Chronic Diseases* 1987; 40: 473-480.

Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey: Manual and Interpretation Guide*. Boston (MA): The Health Institute, New England Medical Center, 1994.

Ware JE, Jr., Gandek B. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. International Quality of Life Assessment. *Journal of Clinical Epidemiology* 1998; 51: 945-952.

Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care* 1992; 30: 473-483.

Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? *Journal of Clinical Epidemiology* 1996; 49: 135-140.

Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing and Health* 1990; 13: 227-236.

Wilkie DJ, Judge MK, Berry DL, Dell J, Zong S, Gilespie R. Usability of a computerized PAIN Report in the general public with pain and people with cancer pain. *Journal of Pain and Symptom Management* 2003; 25: 213-224.

Woodforde JM, Merskey H. Some relationships between subjective measures of pain. *Journal of Psychosomatic Research* 1972; 16: 173-178.

Wright EP, Selby PJ, Crawford M, Gillibrand A, Johnston C, Perren TJ, Rush R, Smith A, Velikova G, Watson K, Gould A, Cull A. Feasibility and compliance of automated measurement of quality of life in oncology practice. *Journal of Clinical Oncology* 2003; 21: 374-382.

Wu AW, Jacobson DL, Frick KD, Clark R, Revicki DA, Freedberg KA, Scott-Lennox J, Feinberg J. Validity and responsiveness of the EuroQol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial. *Quality of Life Research* 2002; 11: 273-282.

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care* 1999; 37(5): 469-78.

Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice* 2000; 6(1): 39-49.

Yarnold PR, Stewart MJ, Stille FC, Martin GJ. Assessing functional status of elderly adults via microcomputer. *Perceptual and Motor Skills* 1996; 82(2): 689-690.