NOT FOR QUOTATION WITHOUT PERMISSION

MULTIPLICATIVE MULTI-ATTRIBUTE UTILITY FUNCTION FOR THE HEALTH UTILITIES INDEX MARK 3 (HUI3) SYSTEM: A TECHNICAL REPORT

William Furlong
David Feeny
George W Torrance
Charles H Goldsmith
Sonja DePauw
Zenglong Zhu
Margaret Denton
Michael Boyle

Cite as: Furlong W, Feeny D, Torrance GW, Goldsmith CH, DePauw S, Zhu Z, Denton M, Boyle M: Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report, McMaster University Centre for Health Economics and Policy Analysis Working Paper 98-11, December 1998.

PREFACE

This technical report has been prepared as a reference document to record the details of the development of the multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3). The details of the study design and results presented here will be important for readers interested in understanding the development process, assessing the validity of the scoring functions or replicating the study.

Readers interested only in applying the HUI3 preference scoring functions are referred to Appendix B (beginning on page 95), which presents the HUI3 utility scoring systems concisely for use by data managers and analysts.

ACKNOWLEDGEMENTS

Funding from Ontario (Canada) Ministry of Health grant #04020; Natural Science and Engineering Research Council grant #10020; grants from Astra Pharma Inc., Bayer Inc. Boehringer-Ingelheim, Bristol-Myers Squibb, Burroughs Wellcome Inc., Eli Lilly, Glaxo Canada Inc., Janssen Pharmaceutica Research Foundation, Nordic Merrell Dow Research, Ortho-McNeil and CILAG, Pharmaceutical Manufacturers Association of Canada (Health Economics Committee), and Sandoz Canada Inc.

Assistance from Jean-Marie Berthelot, Stephen Birch, Nancy Bishop, Shelley Chambers, Paul Grootendorst, Ralph Keeney, Humaira Khan, Lynda Marsh, Roger Roberge, Vicki Rynard, Saba Safdar, Carol Siksay, David Streiner, Marilyn Swinton, Qinan Wang, Michael Wolfson, Regional Municipality of Hamilton-Wentworth staff for sampling, interviewers, and respondents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

GLOSSARY OF ABBREVIATIONS AND ACRONYMS

10% Trd Mean	-	10% trimmed mean (5% trimmed off each end of the distribution)
Ambln	-	Ambulation
ANOVA	-	analysis of variance
B12	-	block 1, state number 2: vector (V6,H1,S1,A6,D6,E1,C6,P1)
B38	-	block 3, state number 8: vector (V6,H1,S5,A1,D1,E1,C1,P1)
Cogtn	-	Cognition
col%	-	column percent
df	-	degrees of freedom
Dxtry	-	Dexterity
Emotn	-	Emotion
EOSB	-	end of scale bias
EOSBA	-	end of scale bias adjustment
EOSBA MA Val	-	10% trimmed mean value scores on multi-attribute scale adjusted for EOSB
EQ-5D	-	EuroQol 5-dimension instrument
F3/3	-	HUI2 health state with fertility level of 3 of 3 and all 6 other attributes at level 1
GSS	-	General Social Survey
Hearg	-	Hearing
HRQL	-	health-related quality of life
HUI	-	Health Utilities Index
HUI1	-	Health Utilities Index Mark 1
HUI2	-	Health Utilities Index Mark 2
HUI3	-	Health Utilities Index Mark 3
HUI3-D	-	HUI3 Direct Preference Survey
HUI3-M	-	HUI3 Modelling Preference Survey
HUI3-M&D	-	HUI3 Modelling and Direct Preference Surveys
ICC	-	intra-class correlation coefficient
I1	-	HUI2 intermediate state 1: vector (S1, M4, E2, C1, SC1, P1, F1)
I3	-	HUI2 intermediate state 3: vector (S3, M3, E2, C3, SC3, P2, F2)
M3/5	-	HUI2 health state with mobility level of 3 of 5 and all 6 other attributes at level 1
MA	-	methodological marker state A (V2, H1, S1, A1, D1, E1, C1, P3)
MAD	-	mean absolute difference
MADUF	-	multi-attribute disutility function

Glossary of Abbreviations and Acronyms (Cont'd)

MARS	-	multi-attribute risk-seeking
MAUF	-	multi-attribute utility function
MAUT	-	multi-attribute utility theory
Max	-	maximum
MB	-	methodological marker state B (V2, H1, S1, A3, D1, E2, C1, P3)
MC	-	methodological marker state C (V2, H1, S1, A1, D1, E2, C3, P5)
MD	-	mean difference
Min	-	minimum
n	-	number of measurements or respondents or subjects
n/a	-	not applicable
Neg	-	negative
NPHS	-	National Population Health Survey
OSD	-	overall standard deviation
OWCB	-	Ontario Workers Compensation Board
p	-	probability value
P2R2H4	-	HUI1 health state vector (P2, R2, SE1, H4)
P5R2H5	-	HUI1 health state vector (P5, R2, SE1, H5)
PH	-	Perfect Health health state: vector (V1, H1, S1, A1, D1, E1, C1, P1)
Pits	-	health state having lowest level on each of eight attributes: vector (V6, H6, S5, A6, D6, E5, C6, P5)
PLT	-	positive linear transformation
P-value	-	probability value
PWC	-	pair-wise comparison.
QALE	-	quality-adjusted life expectancy
QALYs	-	quality-adjusted life years
QWB	-	Quality of Well Being scale
Rescaled MA Val-		re-scaled EOSBA MA Val
SA	-	single-attribute
SAVF	-	single-attribute value function
SEM	-	standard error of the mean
SG	-	standard gamble
Spech	-	Speech
SV5	-	single-attribute vision level 5 health state
SV6	-	single-attribute vision level 6 health state

EXECUTIVE SUMMARY

This paper is of interest to analysts, policy makers and decision makers involved with descriptive clinical studies, clinical trials, programme evaluations, measuring population health, and planning assessments. It describes a recently developed system for measuring the overall health status and health-related quality of life of individuals, clinical groups and general populations.

The measurement system is the Health Utilities Index Mark 3 (HUI3) and it consists of two components: the health status classification system; and the preference-based scoring system. The health status classification system was first published in 1995 and this paper focuses on the scoring system. The HUI3 is the latest member of the Health Utilities Index family developed by researchers at McMaster University during the past 20 years.

HUI3 is a brief but comprehensive system for describing the health status of individuals and for assigning a preference score to that health status. The HUI3 is generic in the sense that it is designed to be applicable to all people. The scores are based on preference measures from a random sample of the general population. HUI3 is founded directly on multi-attribute utility theory. These scores are, therefore, referred to as utility scores and represent community preferences. Community preferences are considered an appropriate source of preferences for calculating quality-adjusted life years (QALYs) for use in cost-effectiveness or cost-utility analyses and for use in the measurement of population health. The HUI3 is also useful in clinical studies as a method of describing the health status of patients and tracking it over time.

The HUI has been included in studies being undertaken by more than one hundred investigative teams based in major centres around the world. The HUI3 has also been included in every major Canadian general population health survey since 1990. The early inclusion of the HUI3 in population health surveys has placed Canada in the forefront of regional and local surveillance of population health, including health-related quality of life (HRQL) and health-adjusted life expectancy (HALE) measures.

This technical report provides the first public release of the multiplicative multi-attribute utility function (MAUF) for the Health Utilities Index Mark 3 (HUI3). The report provides details of the study design, preference survey results, and modelling techniques. It also includes an appendix which presents the HUI3 utility scoring systems concisely for use by data managers and analysts.

The purpose of the HUI3 MAUF is to provide a formula for calculating scores for all of the 972,000 health states defined by the HUI3 health status classification system. This report presents a HUI3 utility function on the conventional Dead = 0.00 to Perfect Health = 1.00 scale. This is the most appropriate scale for calculating aggregated indices of morbidity and mortality such as quality-adjusted life years (QALYs).

Evidence to date indicates that the HUI3 measurement system is acceptable, reliable, valid, responsive and useful.

1.0 INTRODUCTION

The Health Utilities Index Mark 3 (HUI3) system is a brief but comprehensive system for describing the health status of individuals and for assigning a utility score to that health status. The results are useful in studies of health-related quality of life, in the measurement of population health, and in cost-effectiveness and cost-utility analyses.

The HUI3 consists of two components: the health status classification system; and the preference-based scoring system. The HUI3 classification system has been described previously (Feeny et al. 1995a). HUI3 provides a set of categorical variables to describe functional health status of individuals, and describes the comprehensive health state of each subject in terms of their capacity within eight attributes (i.e., dimensions of health status).

The HUI3 is generic in the sense that it is designed to be applicable to all people. The utility scores are founded directly on multi-attribute utility theory (Keeney and Raiffa 1976 and 1993; Torrance et al. 1996a), and are based on community preferences which are considered an appropriate source of preferences for economic evaluations (Ontario Ministry of Health 1994; Gold et al. 1996; Canadian Coordinating Office for Health Technology Assessment 1994 and 1997). The scores are appropriate for calculating quality-adjusted life years (QALYs) for use in cost-effectiveness or cost-utility analyses and for use in the measurement of population health (Patrick and Erickson 1993). The HUI3 is also useful in clinical studies as a method of describing the health status of patients and tracking it over time. As of August 1998 the HUI has been included, or proposed to be included, in studies being undertaken by approximately 200 investigative teams. These teams are based in major centres around the world: North and South America, Europe, Australia, and Asia.

The HUI3 has been included in every major Canadian general population health survey since 1990: the 1990 Ontario Health Survey (Berthelot et al. 1993; Furlong et al. 1989; Ontario Ministry of Health 1993; Roberge et al. 1995b); the 1991 Canadian General Social Survey (Roberge et al. 1995a); the 1994 and on-going National Population Health Survey (Hood et al. 1996; Roberge et al. 1996; Wolfson 1996; Statistics Canada 1998a; Statistics Canada 1998b); and the National Longitudinal Survey of Children and Youth (Statistics Canada and Human Resources Development Canada 1996a; Statistics Canada and Human Resources Development Canada 1996b). The early inclusion of the HUI3 in population health surveys has placed Canada in the forefront of regional and local surveillance of population health, including health-related quality of life (HRQL) and quality-adjusted life expectancy (QALE) measures (Hennessy et al. 1994).

To date, HRQL scores for health states defined by the HUI3 have been calculated using a provisional algorithm (Torrance et al. 1992a). Provisional scores were calculated using the multiplicative multi-attribute utility function estimated for an earlier system: the Health Utilities Index Mark 2 (HUI2). The absence of a final validated scoring function limited the usefulness of the HUI3. In spite of the limitations of the provisional HUI3 scoring system, the HUI3 health status classification system has been incorporated in numerous studies undertaken by a variety of investigators.

In 1991, the Ontario Ministry of Health funded a proposal by Torrance and colleagues (Torrance et al. 1994) which included the development of a sophisticated survey design and the surveying of approximately 500 members of the general population to obtain preference measures for the HUI3 health status classification system. The field work for the HUI3 preference survey was completed in November 1994. This report presents the multiplicative multi-attribute utility function (MAUF) for the HUI3 health status classification system.

The purpose of the HUI3 MAUF is to provide a formula for calculating scores for all of the 972,000 health states defined by the HUI3 health status classification system.

2.0 BACKGROUND

2.1 The Measurement of Health State Preferences

There is broad consensus that health is an important component as well as an indicator of human welfare. At the same time, it is recognized that little is known about the health status outcomes of clinical programmes or about the comprehensive health status of populations, in part because of the lack of appropriate tools (Hennessy et al. 1994). Traditionally, efforts to assess health have relied on simple dichotomies (eg., alive, dead) or trichotomies (eg., normal, disabled, dead). HUI3 provides a richer framework to describe the full range of impairments and captures more fully the burdens of morbidity.

There are two major types of HRQL measures: generic instruments; and specific instruments (Guyatt et al. 1993). The HUI3 provides a generic measure of health status and a preference-based scoring system. The HUI3 classification system was constructed to have content validity in regard to the specification of attributes and levels (Feeny et al. 1995a). Evidence of the reliability, validity, and responsiveness of HUI is accumulating rapidly (Barr et al. 1993; Barr et al. 1997; Boyle et al. 1995; Feeny et al. 1996; Furlong et al. 1995/6; Gemke and Bonsel 1996; Glaser et al. 1997; Mathias et al. 1997; Whitton et al. 1997; Torrance et al. 1998).

2.2 HUI3 Health Status Classification System

Multi-attribute approaches such as the HUI3 are the product of almost 30 years of research (Barr 1998). The HUI3 classification system (Table 1) includes eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort. Each attribute has 5 or 6 defined functional levels, and a comprehensive HUI3 health state is defined by a vector consisting of one level from each of the eight attributes. The HUI3 health status classification system has 972,000 unique health states. This descriptive richness is an important characteristic for both clinical and general population studies. For example, the 1994 National Population Health Survey (NPHS) reported 1,076 unique health states among 1,555 institutionalized subjects, 1,130 unique health states among the 16,920 non-institutionalized subjects, and 1,995 unique health states from the combined non-institutionalized and institutionalized samples (n=18,475). In addition, it was noted that all levels for each of the 8 attributes were reported for the subjects in the NPHS non-institutionalized survey (Roberge, 1996a).

The HUI has been applied in a wide variety of clinical studies, both in Canada (eg., Feeny et al. 1992; Barr et al. 1993; Barr et al. 1994; Barr et al. 1995; Saigal et al. 1994a and

1994b; Saigal et al. 1996; Whitton et al. 1994 and 1997) and abroad (eg., Billson and Walker 1994; Bosch et al. 1996; Costet et al. 1998; Feeny et al. 1993; Gemke et al. 1995; Gemke and Bonsel 1996; Grossman et al. 1998; Le Gales et al. 1997; Kanabar et al. 1995; Kiltie and Gattamaneni 1995).

2.3 Role of Multi-Attribute Utility Theory

Currently major multi-attribute health status systems include: the Quality of Well Being (QWB) originally developed in California in the early 1970s (Fanshel and Bush 1970; Bush et al. 1972; Kaplan et al. 1978; Kaplan and Bush 1982; Patrick et al. 1973a; Patrick et al. 1973b); the EQ-5D originally called the EuroQol and developed in Europe (Essink-Bot et al. 1993); and the Health Utilities Index (HUI). An important distinction is the fact that the HUI is the only one of these three that has a scoring function that is based on a well-specified and explicit theoretic foundation in utility theory. The HUI scoring function is based on multi-attribute utility theory (MAUT) (Keeney and Raiffa 1976 and 1993). This is an extension of expected utility theory (von Neumann and Morgenstern 1944 and 1947) to the situation where the outcomes to be valued are described by multiple attributes. Expected utility theory, in turn, is a well-established method to measure preferences on a cardinal scale under conditions of uncertainty.

MAUT is important for two major reasons. First, the theory indicates how to measure the preferences themselves. In this regard, MAUT specifies that preferences should be measured consistent with the axioms of expected utility theory. This, in turn, explains why the standard gamble is the preference elicitation method of choice. In the HUI, the preference scores are measured using the standard gamble (SG) as the gold standard. (Measures taken with a visual analogue scale (VAS), for efficiency, are converted to what they would have been had they been measured by SG.) In contrast, for the QWB the preferences were measured using a visual analog scale; and for the EQ-5D, using VAS and time trade-off techniques. Neither of these latter methods are founded directly on expected utility theory.

Second, MAUT indicates what functional forms (eg., types of models) are admissible for modelling preference utility scores in multi-attribute space. There are three basic forms available: additive, multiplicative and multilinear. (In addition, more complicated forms can be made by nesting these basic forms among themselves.) The HUI uses functional forms that are consistent with the underlying theory. In contrast, the EQ-5D has an implicit ad hoc functional form. The QWB has an additive structure, which is allowed by the theory. We have found, however, that the additive form never fits our data, and on introspection is not reasonable for the attributes that comprise HUI3 (although the additive form may well be reasonable for the structure of the QWB health-status classification system).

2.4 Evidence of Preference Interactions

There are three possible general types of preference interactions among attributes: none, synergistic; and antagonistic interactions. Fitting an additive preference scoring model would imply that there are no important preference interactions among attributes.

A synergistic preference interaction among attributes is when the total is greater than the sum of the effects of parts; an antagonistic preference interaction is when the total is less than the sum of the parts. In applying this thinking to HUI attributes, one can use either the utility scale or the disutility scale. (Disutility equals 1-utility.) An interaction that is synergistic on the utility scale is antagonistic on the disutility scale, and vice versa. In general, it is easier to think of the interactions in terms of the disutility scale. For example, a synergistic interaction in disutility would occur in a situation in which the sum of the disutility of two single impairments is less than the disutility of both impairments occurring simultaneously. For example, if a respondent reported being blind (ie., all other attributes at level 1) as having a utility of 0.70 (ie., a disutility of 0.30), reported the utility of being deaf as having a utility of 0.80 (ie., a disutility of 0.20), and reported the utility of being both blind and deaf as being 0.30 (ie., disutility of 0.70) the sum of the disutilities for each of blind and deaf (0.30 + 0.20 = 0.50) would be less than the disutility of the two impairments occurring together (0.70). Note that this synergistic interaction in disutility is also an antagonistic interaction in terms of utility. Specifically, starting out with being blind and deaf (utility = 0.30) and restoring sight adds 0.5 units of utility (utility = 0.80). Starting out with being blind and deaf (utility = 0.30) and restoring hearing adds 0.4 units of utility (utility = 0.70). However, starting out with being blind and deaf (utility = 0.30) and restoring both sight and hearing would add 0.7 units of utility (utility = 1.00) which is less than the sum of the two parts (0.5 + 0.4 = 0.9 and 0.7 < 0.9). When attributes interact this way, they are called preference substitutes. The term denotes that, to some extent, the attributes can substitute one for the other in terms of their impact on utility. For example, restoring function of either attribute will add a lot of utility, while restoring both does not add that much more.

An example of antagonistic, or negative, interaction in disutility would occur if the responses in the preceding example were unchanged except for the utility of blind = 0.50. Then the sum of the disutilities of each of blind and deaf (0.50 + 0.20 = 0.70) would be greater than the disutility of the two disabilities occurring together (0.60). Note, also, that this would be a synergistic, or positive, interaction in utility, as indicated by the fact that the sum of the utility added by restoring each of sight and hearing (0.40 + 0.10 = 0.50) is less than the utility added by restoring both (0.60). When attributes interact this way they are called preference complements. The term denotes that, to some extent, the attributes are complementary in their effect on utility.

That is, both attributes are needed to have a big impact on improving utility; either one alone does not add that much utility.

Previous work, for HUI1 (Torrance et al. 1982) and for HUI2 (Torrance et al. 1992b; Torrance et al. 1996a), indicates that the additive form does not fit the preference-measurement data for health states defined according to multi-attribute classification systems, and that the fitted multiplicative functions indicate that the preference interaction among attributes is strongly antagonistic in disutility terms (ie., strong negative disutility interaction and preference complementarity).

3.0 METHODS

3.1 Terminology, Abbreviations and Acronyms

Numerous technical terms, abbreviations and acronyms are used in this report. When introduced, terms are defined in detail. The terms describe various types of health state descriptions (eg., attribute-level states, corner states, Pits, Perfect Health, Marker States), types of respondent groups (ie., Group A respondents and Group B respondents), and summary statistics of preference measures for specific groups. For instance, Person-Mean(A) represents the mean score for Group A respondents and Person-Mean(B) represents the mean score of Group B respondents. Person-Mean is the weighted mean score of Person-Mean(A) and Person-Mean(B) scores. A glossary of abbreviations and acronyms is provided on pages xi and xii, for ease of reference.

3.2 Description of the Modelling Space

The conceptual space for modelling preference scores for the HUI3 system is eight-dimensional, and is defined by the underlying health state classification system (Table 1). The space is bounded by preference scores for two anchor states: Perfect Health with a utility of 1.00 (by definition); and Pits with a utility score of 0.00 (by definition). Perfect Health (PH) is described as full function on all of the eight HUI3 classification system attributes (ie., vector V1,H1,S1,A1,D1,E1,C1,P1), and Pits is described as having the lowest level on each of the eight attributes (ie., vector V6,H6,S5,A6,D6,E5,C6,P5).

Initial fitting of the multiplicative multi-attribute utility function is undertaken in the eight-dimensional space defined by the health state classification system and provides a formula for calculating utility scores for all of the 972,000 health states defined by the HUI3 health status classification system (Feeny et al. 1995a), on the Pits = 0.00 to Perfect Health = 1.00 scale (Pits/PH scale). Subsequently, the preference scores for dead are integrated into the scoring system. It is important to integrate preference scores for dead into the scale for at least three reasons. First, the preference scores for Dead are necessary for adding up outcomes in studies involving both morbidity and mortality. Second, the Dead = 0.00 to Perfect Health = 1.00 scale (Dead/PH scale) is the conventional scale reported in the literature and the scale necessary to calculate QALYs. Thus, to make HUI3 utility scores commensurate with scores reported from other studies and to be able to use HUI3 scores to calculate QALYs, it is necessary to develop a formula for calculating scores on this scale. The third reason for integrating dead into the utility scale is that the face validity of HUI3 scores will, at least in part, depend on the ability to compare the agreement between HUI3 utility scores and utility scores of hypothetical states collected in other studies, and reported on the conventional Dead/PH scale.

Two additional steps, after fitting the MAUF in the Pits/PH scale, are required to integrate the utility of dead into the preference scales. First, the Person-Mean utility score for dead is estimated on the Pits = 0.00 / Perfect Health = 1.00 scale. In theory, the score may be negative or positive, depending upon whether the mean utility score for dead is less than or greater than the score for Pits. The second step involves integrating the Person-Mean score for dead, on the Pits = 0.00 / Perfect Health = 1.00 scale, into the fitted MAUF which is also defined on the Dead = 0.00 / Perfect Health = 1.00 scale.

3.3 Types of Multi-Attribute Utility Theory Models

Multi-attribute utility theory (MAUT) describes the functional forms which are admissible for modelling preference utility scores in multi-attribute space. There are three basic forms available: additive, multiplicative and multilinear.

3.3.1 Additive versus multiplicative models

Previous applications of multi-attribute utility theory have used additive or multiplicative models because the measurements required for the multilinear models were considered too complex to be practical (Keeney and Raiffa 1976 and 1993; Torrance et al. 1982 and 1992b and 1996). The additive MAUT model is appropriate only if the interactions in preferences among attributes are not considered important. This constraint is intuitively difficult to accept, and in health status applications, the data have, in general, strongly rejected the additive model.

To appreciate the unreasonableness of additive utility independence, consider the eight HUI3 attributes and the conventional utility scale with healthy as 1.0 and dead as 0.0. Determine x_1, the disutility (reduction in utility) associated with being blind but otherwise healthy. For illustrative purposes, say that many people would assess it as having a disutility of about 0.4. Determine the disutility x_2 associated with being deaf but otherwise healthy. Again assume many would assess it as about 0.4. Continue this process for each of the eight attributes. Now consider having all eight attributes at their worst levels simultaneously (i.e., being blind and deaf and dumb, and ...) and determine the disutility associated with having all eight attributes at the lowest level, label this a_8. Now ask yourself: is $a_8 = x_1 + x_2 + ... + x_8$? The additive model is only appropriate if this equality holds. If a_8 < sum of x's (our usual finding) or if a_8 > sum of x's, the additive model does not hold. We invite readers to try this for themselves. This introspective exercise will demonstrate, for most people, that the additive model does not accurately reflect their preferences.

3.3.2 Multiplicative versus multilinear models

The multiplicative MAUT model is the least complex case in the family of non-additive models. In previous efforts to fit preference-based scoring functions for multi-attribute health state classification systems, the MAUT multiplicative model has been shown to fit the data better than the additive model (Torrance et al. 1982 and 1992a and 1995a and 1996). The results of the HUI3 preference measurement survey presented in this report again reject the additive form and favour the multiplicative function.

Data collected in earlier studies did not permit the fitting of the more complex set of multilinear models. The HUI3 preference measurement survey collected data based on a fractional factorial design that will permit the estimation of a complex set of multilinear models, but these analyses have not yet been completed and will be the subject of a future report.

In the multiplicative MAUT function, the interactions that can be represented are highly constrained. That is, the multiplicative model forces the interaction among attributes to be the same among all attributes. For example, consider the case described previously where the disutility of two disabilities together is less than the sum of the disutility of each disability separately. This type of preference structure is described as multi-attribute risk-seeking (MARS) and the attributes are complements (Keeney and Raiffa 1976 and 1993). MARS means that a person would prefer a 50/50 gamble between no deficits and two deficits, rather than a 50/50 gamble between one deficit and one other deficit; or in other words, they would prefer the riskier of the two gambles.

Attributes being complements means that each attribute alone (sight, hearing, speech, ambulation, ...) is not that valuable, but together they are quite valuable; i.e., to enjoy any one attribute you also need to have the other attributes. The alternative case, where the disutility of multiple disabilities is greater than the sum of the disutility of each disability separately, is described as multi-attribute risk-aversion and the attributes are preference substitutes; ie., either attribute alone is pretty good, because one can substitute for the other (Keeney and Raiffa 1976 and 1993).

Fitting a multiplicative model results in all attributes being treated as either complements (ie., having antagonistic disutility interactions) or all being treated as substitutes (ie., having synergistic disutility interactions). The form of the model does not allow some sub-set of attributes to be preference complements and the others to be preference substitutes. However, our pilot testing regarding interactions among attributes in the HUI3 suggested that most attributes

were seen as complements. The exception, if any, noted from the pretest was for the preference interaction of vision and hearing. Vision and hearing, two major ways in which a person can receive information from the outside world, were often seen as preference substitutes. This would be the case if a person said that it would be bad to go blind, and bad to go deaf, but it would be devastating to go both blind and deaf. That is, the disutility of being both blind and deaf exceeds the sum of the disutilities of being blind alone and being deaf alone. Further, in the pretest, emotion and cognition were often considered to be complements. Additional evidence for complementarity among some attributes has been reported for HUI2 in which the marginal disutility associated with bad outcomes on each attribute decreases as the levels on the other attributes drop (Torrance et al. 1992b and 1996).

The multilinear functional form is less restrictive: some pairs of attributes can be complements and other pairs can be substitutes. However, the enhanced richness of the model comes with a price: the estimation of a large number of model parameters is required. As the number of parameters to be estimated increases so does the number of different preference measurements required, and the practical demands for efficient survey techniques become more apparent. Few studies have estimated multilinear multi-attribute utility functions (eg., Cadman et al. 1986), especially in health.

3.4 Design of Preference Measurement Study and Instrumentation

The design of the HUI3 preference measurement study included two complementary surveys: a survey to collect measurements required for fitting HUI3 multi-attribute utility functions, the HUI3 Modelling survey (HUI3-M); and an associated survey which collected direct utility measures for 74 states reported as prevalent in the general population (Feeny et al. 1995b and Torrance et al. 1995b). The survey of preference scores for the 74 prevalent health states provides a valuable commensurate data set for assessing inter-survey or external agreement of HUI3 utility scores, and will be referred to as the HUI3-D (HUI3-Direct Measurement) survey. Respondents were randomly allocated to the HUI3-M (n=256) or HUI3-D (n=248) survey.

Value scores were measured using a newly developed prop for eliciting preference scores based on a visual analogue scale technique: the two-sided Feeling Thermometer (Figure 1). Standard gamble questions were administered using a modification of the original Chance Board prop (Furlong et al. 1990): the flip-card Chance Board (Figure 2). Figure 3 presents a schematic of the flip-card Chance Board question and response combinations used to determine specific utility scores.

The structures of the HUI3-M and HUI3-D survey interviews are summarized in Figures 4 and 5, respectively. The general format of the interviews was common to both surveys. However, the specifications for the health states to be evaluated and the mix of standard gamble and VAS measures differed across the two surveys.

3.4.1 HUI3 Modelling (HUI3-M) and Direct (HUI3-D) surveys

3.4.1.1 Sampling frame and sample size

A sample was obtained from the Planning Department of the Regional Municipality of Hamilton-Wentworth, Ontario, Canada. The sample consisted of a list of 3,000 randomly selected households from within the boundaries of the City of Hamilton. Households from this list were randomly assigned to either the HUI3-M or HUI3-D surveys. A schematic of the sampling process is presented in Figure 6.

The HUI3-M survey sample size of 256 respondents was determined on the basis of two major factors: a requirement for a balanced number of measurements for each health state specified by the 2^8 fractional factorial design (eg., $2^8 = 256$); and maintenance of the precision acheived in our previous work. The latter factor is of relevance to the topic of this report. Instrument and model precision estimates for the HUI3 preference measurement survey were based on results from a previous study (Torrance et al. 1992b and 1996). The mean of the standard deviations of directly measured VAS value scores for 10 states (states #2-11 from Table 3 of Torrance et al. 1992b and from Table 3 of Torrance et al. 1996a) was 0.22 and the mean of the standard deviations of directly measured standard gamble utility scores for 4 states (from Table 3 of Torrance et al. 1992b and from Table 3 of Torrance et al. 1996a) was 0.21. The mean of the standard deviations of utility scores calculated using VAS value scores and the power curve was estimated to be 0.275 (Torrance 1992). The precision of the multiplicative model was 0.10 based on n=200 for the multi-attribute disvalue function (Table 11 of Torrance et al. 1992b and Table 9 of Torrance et al. 1996a) and was 0.06 based on n=194 for the multi-attribute disutility function (Table 12 of Torrance et al. 1992b and Table 10 of Torrance et al. 1996a).

The HUI3-M preference survey collected value and utility measurements from 256 respondents who were English-speaking, non-institutionalized, members of the general public, age 16 years and older. Sets of health states were randomly allocated to respondents according to strata. Health state strata were defined as follows: scale anchor states (eg., Dead and Perfect Health); methodological marker states; single-attribute states; and block states (see section 3.4.1.4 for detail of these states). The number of respondents (n) providing value and utility measures

varied by health state strata and, therefore, the precision of the mean preference scores varies by strata. In general, anchor state scores are defined by the scale (eg., Dead = 0.00 and Perfect Health = 1.00 on the conventional health state utility scale) and, therefore, the precision of these scores was not an important consideration in the design phase. The number of measurements and expected precision of mean measured value and utility scores for the HUI3-M survey, based on HUI2 results, were as follows:

State Strata	Mean Values n	Mean Values SEM	Mean Utilities n	Mean Utilities SEM
Markers	256	0.014	256	0.014
Single-attribute	64	0.028	0	n/a
Blocks	16	0.055	0	n/a

(SEM - standard error of the mean)

The HUI3-D preference survey collected value and utility measurements from 248 respondents who were English-speaking, non-institutionalized, members of the general public, age 16 years and older. As in the HUI3-M survey, sets of health states were randomly allocated to HUI3-D respondents according to strata. The health state strata for HUI3-D survey were defined as follows (see section 3.4 and Figure 5 for details): scale anchor states (n = 3); methodological marker states (n = 3); most prevalent states (n =5); and less prevalent states (n = 65). As for the HUI3-M survey the number of respondents (n) providing value and utility measures varied by health state strata and, therefore, the precision of the mean preference scores varies by strata. Again, the precision of anchor state scores was not an important consideration. The number of measurements and expected precision of mean measured value and utility scores for the HUI3-D survey, based on HUI2 results, were as follows:

States Strata	Mean Values n	Mean Values SEM	Mean Utilities n	Mean Utilities SEM
Markers	247	0.014	247	0.014
Most prevalent	247	0.014	75	0.025
Less prevalent	19	0.050	14	0.061

(SEM - standard error of the mean)

3.4.1.2 Randomization and other techniques of controlling for extraneous survey design factors

The interview process used extensive randomization and blocking to control for potential interviewer effects, possible measurement order effects, and various other potentially confounding factors.

The first 3 characters of the postal code for each household, referred to as the Area Code by Canada Post, describes contiguous geographic areas within the City roughly the size of 25 carrier routes. These Area Codes were used as a stratification factor when randomizing the assignment of interviewers to households. Area code was used as a stratification factor to ensure that the geographic distribution of households assigned to each interviewer was reasonably balanced across interviewers, between attractive and less attractive districts of the city, and that the distribution was not determined by potentially confounding factors such as the proximity of specific households to interviewers residences.

Sets of health states were randomly assigned to respondents, and the order that sets of health states, and states within sets were introduced to the respondent was randomized. In the HUI3-M survey, two sets of single-attribute states and one set of "block states" (in two parts) were each randomly assigned to respondents. In the HUI3-D survey, 14 sets of 5 prevalent states were formed. Among the criteria used in creating sets of states was that each set be heterogeneous so that it would be more likely to contain health states that spanned the "entire" scale from highly disabled to very mildly disabled. Provisional Index utility scores (Torrance et al. 1992a) were used to "forecast" scores which were then used to allocate states such that each set had hetrogeneity of forecast scores. (Another criterion for the selection of health states for the Direct Survey was that all levels for each of the 8 attributes would be represented at least once.) The sets of 5 states were randomly assigned to respondents.

At first contact with the household, the interviewers enumerated all eligible household residents. Randomization labels, prepared by Statistics Canada, were used by interviewers to identify a target respondent to avoid self-selection of respondents within households. Refusal by the randomly selected target household respondent resulted in the household being declared ineligible.

3.4.1.3 Chronicity, duration of health states and exogenous factors for preference measurement

Health states were explicitly defined as being chronic and the duration of remaining life expectancy was provided by each respondent. Typically, the life expectancy was identified by each respondent directly in terms of the number of additional years they expected to live. For some respondents, perceived life expectancy was estimated by subtracting the respondent's present age from the age at which the respondent reported that they expected to die. For a few respondents who did not provide direct reports of life expectancy, standard life tables for Canada were used to estimate additional years of life expectancy. Life expectancy, estimated from life tables, was confirmed with respondents to ensure each respondent was comfortable with the duration before undertaking the health state preference measures.

Respondents were requested to focus only on the differences in health state descriptions and keep all other factors constant. The interview script read "When imagining yourself in these health states please remember that where you live, your income, your friends, and family would be the same as now."

3.4.1.4 Specification of multi-attribute health state descriptions for preference measurement

The HUI3-M survey design required that preference measurements be obtained for Dead, and 168 unique multi-attribute health state descriptions: 3 methodological marker states; 2 multi-attribute anchor states (Perfect Health and Pits); 37 other states for fitting the multiplicative MAUF and 126 other states for fitting the multilinear functional form. The 43 states used to fit the multiplicative MAUF are: 3 scaling anchor states (Pits, Dead and Perfect Health); 3 methodological "marker" states (MA, MB, MC; see Glossary of Abbreviations and Acronyms on page viii for detailed attribute vector notationof each state); 5 "vision" single-attribute level states (ie., all but the top level of vision; the top level of each attribute is already included as Perfect Health); 5 "hearing" single-attribute level states; 4 "speech" single-attribute level states; 5 "ambulation" single-attribute level states; 5 "dexterity" single-attribute level states; 4 "emotion" single-attribute level states; 5 "cognition" single-attribute level states; 4 "pain" single-attribute level states; and 126 "block" states as specified by the fractional factorial design plan 2.8.8 (the fractional factorial design plan actually specified 128 states but 2 of the states were Pits and Perfect Health, which have already been specified as scaling anchor states).

The 126 "block" states were not used in the fitting of the multiplicative MAUF and, therefore, the total number of states used to fit the multiplicative MAUF was 43 (43 = 168 - 126 "blocks" + Dead).

To facilitate estimating MAUFs, it is common to use disutility corner states (Torrance et al. 1982 and 1992b and 1996). (The disutility of a state is 1.00 - utility score for that state, and a disutility corner state has one attribute at the lowest level of function and all other attributes at the highest level of function. There are, therefore, 8 corner states of the HUI3 system.) This approach was adopted and leads to the multi-attribute preference function being fitted in "dis" terms.

Each "single-attribute level state" was described as an 8-attribute health state with one attribute at less than full function (ie., level 2 or worse) and all other attributes at level 1 function. For example, the 5 "vision" single-attribute level states were each described by eight phrases, one from each of the eight HUI3 attributes. The levels of vision varied across the 5 "vision" single-attribute level states such that one was described as level 2 vision, another as level 3 vision, a third as level 4 vision, a fourth as level 5 vision, and the remaining state as level 6 vision. The other 7 attributes were each described at the level 1 functional level. (Note that the "corner state" for each attribute is the single-attribute level state with that attribute at the lowest level.) Perfect Health, with each of the attributes described at level 1 function, represented the level 1 vision state (and the level 1 state for each of the other single-attribute level states).

3.4.1.5 Instrumentation to collect preference measures

The design of the HUI-M survey preference measurement interview included six major measurement tasks:

1) establishment of an appropriately anchored VAS scale for each respondent such that Perfect Health, the most desirable state, was assigned a value score of 100 and the least desirable state (ie., either Pits state or Dead as selected by each respondent) was assigned a value score of 0;
2) measurement of 3 methodological marker states and either Pits or Dead, whichever the respondent did not report as the least desirable state in step #1, on the VAS scale with conceptual anchors established in step #1;
3) measurement of 2 sets of single-attribute value functions (SAVF) on the VAS scale with conceptual anchors and marker states established in steps #1 and #2;
4) measurement of one set of "block" states (in two parts) from the 2^8 fractional factorial plan on the VAS scale with conceptual anchors and marker states established in steps #1 and #2;
5) standard gamble measurement of the 3 methodological marker states on the Pits = 0.00 / Perfect Health = 1.00 scale; and

6) standard gamble measurement of Dead on the Pits = 0.00 / Perfect Health = 1.00 scale, or standard gamble measurement of Pits on the Dead = 0.00 / Perfect Health = 1.00 scale (depending upon whether the respondent reported Pits or Dead respectively to be the least desirable state at the time of step #6).

In the SAVF section of the HUI3-M (step #3 above) each of the 256 respondents measured 2 SAVF's consisting of 4-5 states each (2 sets of 4-5 states). Each individual SAVF measurement was conducted using holistic states, that is, each state description not only described the level of the attribute being measured, but also described level 1 (the highest) for each of the other 7 attributes. The measurements from this section of the interview provide scores required for the multiplicative multi-attribute function: the single-attribute level scores including the multi-attribute "corner" state scores.

The second part of the HUI3-M survey measurement task asked each respondent to measure one block of 8 states (2 packets of 4) on the VAS. The block of 8 states were chosen based on the codes specified by the published 2^8 fractional factorial design: plan 2.8.8 (Statistical Engineering Laboratory 1957). It should be noted that these scores are not used in fitting the multiplicative MAUF, but are described here to provide readers with a complete description of the interview context within which the relevant measures were collected.

In the third part of the HUI3-M survey interview, all respondents were asked to provide SG utility measurements for the three "marker" living health states, on a scale bounded by Pits = 0.00 and Perfect Health = 1.00. A fourth standard gamble question elicited the score for Dead on the Pits/PH scale or the score for Pits on the Dead/PH scale, depending upon whether the respondent considered Dead to be more or less preferable than Pits. Mean SG scores for 4 states, paired with mean VAS scores for the same states on the same scale, were used to estimate the model for converting value scores to utility scores prior to fitting the HUI3 Person-Mean multiplicative MAUF.

The final part of the interview involved asking the respondent a variety of closed- and open-ended questions about factors the respondent used to decide on responses to questions on preferences for health states, and a series of questions asking the respondent their opinion about aspects of the interview process.

The initial part of the measurement task was common for the HUI-M and the HUI3-D: rating of scale anchoring states (ie., Perfect Health, Dead or Pits); and rating of 3 marker states

and Dead or Pits. The second part of the HUI3-D survey asked respondents to provide value scores for 2 sets of "prevalent" health states: a set consisting of the 5 "most prevalent" states (z-states); and a set of 5 states randomly selected from the 13 sets of "less prevalent states". The order that the states, within each set, were presented to the respondent was randomized. In the third part of the HUI3-D survey interview, all respondents were asked to provide SG utility measurements for the three "marker" living health states and for the 5 states from one of the "prevalent" health state sets (ie., either the set of z-states or the set of "less prevalent states") that they rated in the second part of the measurement task. As in the HUI3-M survey, each respondent was asked a fourth standard gamble question which elicited the score for Dead on the Pits/PH scale or the score for Pits on the Dead/PH scale, depending upon whether the respondent considered Dead to be more or less preferable than Pits.

3.4.2 Modelling survey details

3.4.2.1 Person-Mean approach to estimating multiplicative MAUF parameters

There are two basic approaches to estimating MAUFs: the Person-Mean MAUF approach; and the mean of individual MAUF approach. The testing of both approaches during the development of the HUI2 MAUF, resulted in the selection of the Person-Mean approach for use in the design of the HUI3 preference measurement study. Specifying the Person-Mean approach during the design phase facilitated collection of value measurements for many more health states than would have been possible using the same sample size and levels of precision if individual MAUFs were to be fitted. The Person-Mean approach required that preference scores be collected to estimate Person-Mean scores for 8 "corner" states defined on the Pits = 0.00 to Perfect Health = 1.00 scale; and to estimate eight Person-Mean single-attribute utility functions, each single-attribute utility function scale defined such that the lowest level for the attribute has a utility of 0.00 and the highest level (ie., level 1) has a utility of 1.00.

Preference measures for 43 states are required to fit the multiplicative MAUF. All respondents were asked to provide measures for a common set of 6 of these states: 3 scaling anchor states (Pits, Dead and Perfect Health); and 3 methodological marker states (MA, MB and MC). In addition, each individual respondent was asked to provide measures for two sets of states formed from the remaining 37 states required to fit the HUI3 multiplicative MAUF: the set of 5 "vision" states; the set of 5 "hearing" states; the set of 4 "speech" states; the set of 5 "ambulation" states; the set of 5 "dexterity" states; the set of 4 "emotion" states; the set of 5 "cognition" states; and the set of 4 "pain" states. All preference scores were measured using comprehensive health state descriptions and the VAS anchored by Pits or Dead = 0.00, and Perfect Health = 1.00.

Estimation of a single Person-Mean model is complicated by the need to aggregate across two commonly reported natural scales: a scale on which the Pits state has the lowest possible score; and a scale on which Dead has the lowest possible score. The Person-Mean approach selected for developing the HUI3 multiplicative MAUF involves the use of two parallel Person-Mean models until late in the modelling process when the parallel models are aggregated to form a unified model. The parallel models are defined by the lower anchor state reported by individual respondents when using the feeling thermometer: the Pits or Dead.

Eight sets of single-attribute preference functions were calculated by normalizing each set of Person-Mean single-attribute health state utility scores onto single-attribute utility scales using positive linear transformation, such that for each attribute the state with the lowest level has a utility score of 0.00 and the state with the highest level (represented by the Perfect Health state) has a utility score of 1.00. For example, the 5 "vision" states are those represented by the following vectors: (2,1,1,1,1,1,1,1); (3,1,1,1,1,1,1,1); (4,1,1,1,1,1,1,1); (5,1,1,1,1,1,1,1); and (6,1,1,1,1,1,1,1). The "corner" state of the vision attribute is the score for the last vector in this list: (6,1,1,1,1,1,1,1). The single-attribute utility function for the attribute vision is calculated by a linear re-scaling of the utility scores for each of the 5 vision vectors onto a scale where the score for the vector representing the lowest level of vision, (6,1,1,1,1,1,1,1), has a utility of 0.00 and the vector representing the highest level of vision (normal), (1,1,1,1,1,1,1,1), has a utility of 1.00.

3.4.2.2 Selection of appropriate functional form of MAUF: additive or multiplicative?

The design of the HUI3-M survey did not require that the functional form be specified in advance. Selection of the additive or multiplicative MAUF form was to be based on the parameter estimates obtained from the utility scores of the "corner" states. When fitting the multiplicative model, if the eight c_j estimates sum to 1, the solution yields the additive model. Otherwise, the multiplicative functional form was to be fitted.

3.4.3 Summary of HUI3-M and HUI3-D preference measurements

The HUI3-M survey interviews were conducted with 256 respondents randomly selected adults in the general population of Hamilton, Canada. All respondents (n=256) provided VAS value scores for each of PH, Pits, Dead, and three Marker states; and standard gamble utility scores for each of three Marker states on the Pits/PH scale, and for Dead on the Pits/PH scale or Pits on the Dead/PH scale (depending upon which of Pits and Dead each respondent reported as least preferable). In addition, one quarter of the respondents (n=64) provided VAS value scores for each of two sets of single-attribute states and one-sixteenth of the respondents

(n=16) provided VAS value scores for one set of eight block states defined according to the half-fraction 2^8 factorial design plan.

The HUI3-D survey interviewed 247 randomly selected adults in the general population. VAS value scores were reported by all respondents (n=247) for 11 states: PH, Pits, Dead, three Marker states, and the 5 most prevalent states reported in the General Social Survey conducted by Statistics Canada in 1991. Nineteen randomly allocated respondents provided VAS value scores for each set of five less prevalent states. Standard gamble utility scores were also collected from all HUI3-D respondents (n=247) for each of the three Marker states, and for Dead on the Pits/PH scale or Pits on the Dead/PH scale (depending upon which of Pits and Dead each respondent reported as least preferable). Sixty-five respondents provided SG scores for the 5 most prevalent states, and fourteen respondents provided SG scores for each of the sixty-five less prevalent states (n = 247 = 65 + 182; and 247 equals 65 prevalent state respondents plus 14 respondents for each of the 13 sets of five less prevalent states).

3.5 Analytical Strategies and Techniques

The overall analytical approach to estimating the multiplicative MAUF is illustrated in Figure 7. In brief, respondents were classified into two groups according to the state each respondent selected as the lowest anchor state when using the Feeling Thermometer (Group A respondents reported Pits to be equally or less preferable to Dead and Group B respondents reported Dead less preferable than Pits); Person-Mean value scores were calculated for each of Groups A [Person-Mean(A)] and B [Person-Mean(B)]; the Person-Mean(A) and Person-Mean(B) scores were corrected for end-of-scale-bias, or end-aversion bias (Streiner and Norman 1993), and converted to utility scores; overall Person-Mean utility scores were calculated as the weighted average of Person-Mean(A) and Person-Mean(B) scores; overall Person-Mean utility scores were converted to disutility scores (1-utility scores); the Person-Mean dis-utility scores were used to fit a multi-attribute disutility function (MADUF) with the scale defined such that Perfect Health = 0.00 and Pits = 1.00; the MADUF was converted into a multi-attribute utility function (MAUF) with the scale defined such that Pits = 0.00 and Perfect Health = 1.00; and the MAUF on the Pits/PH scale was then converted to a MAUF on the conventional Dead = 0.00 to PH = 1.00 scale.

In general, the overall approach was guided by four key concepts:

1) the function should be defined in terms of Person-Mean;
2) the function should facilitate calculation of utility scores on both the HUI3-specific Pits = 0.00 to Perfect Health = 1.00 scale, and the conventional Dead = 0.00 to Perfect Health = 1.00 scale;

3) that special analytical techniques should be accorded Dead because the state of Dead is undefined within the framework of the HUI3 health status classification system; and

4) that the opinions of reports of living-states being less desirable than dead should be weighted equally to reports of living states being more desirable than dead (in a sense, one person - one vote).

Additional analytical techniques were used to create commensurate scales among respondents to enable Person-Mean calculations, to adjust for end-of-scale bias, and to convert value scores into utility scores.

The first step was to normalize each respondent's natural scale such that the value score for the least desirable state is 0.00 and the value score for the most desirable state is 1.00. The least desirable state for each HUI3-M survey respondent was selected by the respondent from a choice of two states: either Pits, or Dead. Therefore, each respondent's choice of least desirable state defines whether or not they considered any of the HUI3 states to be worse than dead. Furthermore, the choice of least desirable state defines the lower anchor for each respondent's "natural" value scale: Pits or Dead. To ensure equal weighting of the opinions for respondents having each of these two natural scales, the Person-Mean approach was applied to each group separately and the data were aggregated across natural scales immediately before fitting the multiplicative MAUF.

3.5.1 Strategies common to both HUI3-M and HUI3-D surveys

3.5.1.1 Measure of central tendency for Person-Mean estimates

Direct preference measures, both values and utilities, are summarized using a variety of statistics: 10% trimmed mean (5% trimmed off of each end of the distribution), standard deviation, minimum, and maximum. The trimmed mean was selected, rather than the median or mode, to maintain most of the statistical properties associated with using mean-type estimates while reducing the effects of outlier scores on the estimates of central tendency for distributions of health state preference scores with skewed distributions. The Person-Mean score is defined as the trimmed mean for a specific health state.

3.5.1.2 End-of-scale bias adjustment (EOSBA)

End-of-scale bias, or end-aversion bias or central tendency bias, describes a tendency for people to avoid using intervals at the extremes of the scale (Streiner and Norman 1993). A preference-measurement interview survey undertaken by the Ontario Workers Compensation

Board (OWCB) provided evidence of end-of-scale bias for values reported at the most desirable end of a visual analogue scaling (VAS) prop (Torrance 1996). The OWCB survey first asked each of 125 respondents to provide value scores for a set of health states using a visual analogue scaling (VAS) prop similar to a feeling thermometer. After completing this task, three pairs of states were identified: the state rated closest to Perfect Health (ie., the state ranked second best in terms of desirability) and Perfect Health; the state rated closest to Dead (ie., the state ranked second worst in terms of desirability) and Dead; and two states ranked adjacent and rated in the centre of the scale.

Using a prop for eliciting preference difference scores using the pair-wise comparison (PWC) technique, the gold standard technique for measuring differences between pairs of states, the respondent was asked to scale the difference for each pair against the differences for the other pairs (the difference for a pair is the preference difference between the two states of the pair). The prop consisted of 3 vertical lines of equal length. The respondent was told to select the pair of states from the set of three pairs having the greatest preference difference and place one of these states at the top of line #1 and the other at the bottom of line #1, and to interpret the distance between the pair of states (ie., the length of the lines) as representing the largest preference difference among the set of health state pairs. The respondent was then asked to indicate the proportion of line #2 that would best represent the difference in preference between the two states having the next largest difference. Finally, the respondent was asked to indicate the length of the third line that would best represent the difference in preference between the pair of states having the smallest difference.

The results of comparisons of differences elicited using the VAS and PWC measurement techniques showed that the difference between the VAS score for Perfect Health and the VAS score for the state ranked closest to Perfect Health was 1.78 times that of the analogous difference in PWC scores, and that this finding was statistically significant ($p<0.0001$). The results of this study also suggested that there may be a small end-of-scale bias at the other end of the scale as well; the state ranked next better than dead was 1.12 times too far away from dead. However, this finding was not statistically significant, the magnitude of the bias was relatively small, and the anchor point in the OWCB study was dead, while our anchor point was generally Pits. For these reasons, we did not adjust for an end-of-scale bias at the lower end of the VAS scale in our study. The EOSBA for the top end of our VAS scale was operationalized using a minimalist approach.

This approach involved adjusting scores for a minimum number of states. Only mean value (not utility) scores of health states ranked second most desirable, and fairly close to the most desirable end of the VAS (ie., states having mean value scores greater than 0.75) were adjusted for end-of-scale bias. The EOSBA factor used was 1.78. To illustrate the approach we consider a simple example consisting of mean value scales for a set of 4 health states measured simultaneously on the feeling thermometer: mean score of State A (Perfect Health) = 100.0; mean score for State B = 90.0; mean score for State C = 70.0; and mean score for State D = 50.0. The difference between 90.0 (the mean score for the rank-2 state) and 100.0 (the end-of-scale value score) of 10.0 units is the subject of EOSBA, and is divided by the 1.78 factor to obtain an EOSBA difference of 5.6 units. The EOSBA difference of 5.6 units is then subtracted from 100.00 (the end-of-scale value score), to calculate the EOSBA mean score for the rank-2 state: 94.4 (ie, = 100.0 - 5.6). Subsequently, positive linear transformation (PLT) is used to rescale the mean value score for State C (representing in this example intermediately ranked states in the set of health states) to ensure that the relative size of the differences in mean value scores between states is retained in subsequent calculations. In this example, the PLT would ensure that that rescaled mean value score for State C would be at the mid-point of the interval defined by the mean score for State D (50.0) and the EOSBA mean score for State B (94.4): 72.2. (Note that the original mean value score of State C was 70.0, the mid-point between the original mean score of 90.0 for State B and the mean value of 50.0 for State D.)

The EOSBA was applied at the third stage of data analysis: after normalization of individual respondent value scores on one of two natural scales, and subsequent calculation of Person-Mean(A) and Person-Mean(B) for each of the 41 (43 minus PH minus Pits, or 43 minus PH minus Dead) states on each of the two natural scales. Value scores were adjusted for end-of-scale bias (EOSB) before estimation of value-utility power curves for each natural scale group. The natural scales of individual respondents are:

i) normalized value scale anchored by Pits = 0.00 and Perfect Health = 1.00, for Group A respondents (ie., respondents who reported Pits to be the least desirable state or both Pits and Dead to be tied as the least desirable states according to feeling thermometer ratings);

ii) normalized value scale anchored by Dead = 0.00 and Perfect Health = 1.00, for Group B respondents (ie., respondents who reported Dead to be the least desirable state according to feeling thermometer ratings).

3.5.1.3 Imputation of missing value and utility scores

The overall rate of missing preference measurement scores in the HUI3 preference study surveys was less than a quarter of one percent, 0.19 percent, of number of expected (ie., planned) measurements. The maximum missing rate of 0.3% was associated with two sets of measurements: values from the HUI3-M survey and utilities from the HUI3-D survey. The number of missing scores, expected scores and missing rates are presented in the following table by survey and type of preference measurement.

HUI3 Survey	Type of Measure	Number of Scores Missing	Number of Scores Expected	% Missing of Expected
HUI3-M	value	17	5920	0.29
	utility	1	1024	0.10
HUI3-D	value	0	3963	0.00
	utility	6	1979	0.30
HUI3-M&D	value	17	9883	0.17
	utility	7	3003	0.23
	both	24	12886	0.19

In general, the rate of missing scores was very low and imputation of scores was not necessary for the sole purpose of fitting Person-Mean MAUT models (two imputed value scores and one imputed utility score were used). However, imputation of missing scores was necessary to provide the balanced data set required for fitting multi-linear models based on the fractional factorial design plan. To ensure that results from the fitted multiplicative and multi-linear models were based on the same data sets and, therefore directly comparable, imputed scores were included in the common data set for fitting both types of models.

General strategies for imputation are based, in part, on methods described in the literature (Little and Rubin 1987; and Rubin 1987). The general strategies are hot deck imputation, modelling, and cold deck imputation.

Hot deck imputation involves identifying the respondents with complete data who rated the same health state set or block, and who are similar to the respondent with missing scores.

There are different ways of defining "similar". One would be to plot some existing data of the respondent with missing data, against data from each other respondent, and selecting the most similar respondent. Another way to select similar respondents would be to match on the basis of one other data point. Once a similar respondent is identified, replace the missing score with those of the similar respondent if there is only one, or with those of a respondent randomly selected from a group of similar respondents. The modelling approach involves deriving replacement scores from models based on existing scores. Cold deck imputation is similar to hot deck imputation, but using a different data set.

For the HUI3-M survey, missing preference scores were imputed as follows.

1) For respondent 357, the missing SV6 value was imputed as a normalized value score using the hot-deck technique and randomly selecting a SV6 score from the set of HUI3-M respondents with the same lower anchor as respondent 357, who also rated vision, and whose SV5 score within 5 of the SV5 score for respondent 357.
2) The one missing value score for Death (ID 1771) was estimated from the respondent's own reported utility score for Death and from the respondent's own fitted utility/value power function. (The process is similar to that used for three Direct Survey respondents who rated Death at 100.)
3) For the seven missing value scores (out of 16) for B38, the distribution of the existing nine scores (normalized) was examined. Because the SD of the mean was small, sampling with replacement was used to replace each missing value from the existing data, in each case only from those respondents with the same lower anchor (Pits, Death, or Pits and Death equal).
4) For the missing value scores for all of one set B12 (ID 394), respondents with the same lower anchor and who also rated set B12 were identified. For each pair of respondents (394 and one other respondent), the difference between the values for each marker state (MA, MB, and MC) was calculated. The squares of the differences for each pair were summed, and the missing values were replaced with those of the respondent with the lowest sum of squares.
5) For respondent 2044 with missing MA utility, all respondents with the same lower anchor (on the Chance Board) were identified. For each pair of respondents (2044 and one other respondent) the difference between the utilities for each marker state (MB and MC) was calculated. The squares of the differences for each pair were summed, and replace the missing MA utility was replaced with that of the respondent with the lowest sum of squares.

Missing preference scores for HUI3-D respondents were imputed as follows.

6) For replacing missing utility scores of LE and MA from respondent 219 a hot-deck imputation technique was used. The respondents who rated set L and who selected the same lower anchor on the Chance Board were identified. For each pair of respondents (219 and one other respondent), the difference between the utilities for each of health states (LA, LB, LC, LD, MB, and MC) was calculated. The squares of the differences for each pair were summed, and the missing LE and MA utilities were replaced with those of the respondent with the lowest sum of squares.

7) Missing utility scores from respondent 598 for states JC, JD, JE, and MB were replaced using the hot-deck imputation technique described above. The respondents who rated set J and who selected the same lower anchor on the Chance Board were identified. For each pair of respondents (598 and one other respondent), the difference between the utilities for each of health states JA, JB, MA, and MC was calculated. The squares of the differences for each pair were summed, and the missing utilities were replaced with those of the respondent with the lowest sum of squares.

3.5.2 Modelling survey details

As described earlier, the MAUF was derived from a fitted multi-attribute disutility function. The general form of the multiplicative multi-attribute disutility function is presented in Section 4.2.12. The analytical strategy used to fit the HUI3 MAUF involved numerous steps. The content of each step and the combination of steps used to fit the HUI3 function is only one of many possible strategies we considered. A variety of methods for aggregating individual-level preference measurements into Person-Mean estimates and for fitting value-to-utility conversion models were investigated before finally selecting the analytical strategy described in detail below. For example, we investigated aggregating as early as possible using individual-level preference scores which had been standardized on the Pits = 0.00 to PH = 1.00 scale (eg., the aggregation technique used in fitting the HUI2 multiplicative multi-attribute utility function reported by Torrance et al. 1992 and 1996), and we fitted numerous value-to-utility conversion models using various linear and non-linear estimation methods (eg., spline functions). After selecting the basic analytical strategy the effect of EOSBA was assessed and determined to make an important contribution to the performance of the fitted function.

The analytical strategy finally selected for fitting the multiplicative multi-attribute utility function involves 15 major discrete steps, as illustrated in Figure 7:

1) normalization of value scores reported by each respondent (see following paragraph for details), to create commensurate scales for aggregating scores across respondents, such

that the lowest score is zero (ie., least desirable state = 0.00) and the highest score is 100 (ie., most desirable state = 100);

2) classification of respondents into groups, according to the state selected by each respondent as least desirable (ie., Group A respondents reported Pits as least desirable and Group B respondents reported Dead as least desirable);

3) aggregation of individual-level preference measures into Person-Mean(A) and Person-Mean(B) value and utility scores of the 3 marker states and 37 multi-attribute health states used to fit the multiplicative MAUF, using the trimmed mean measure of central tendency for each group (see section 4.2.4 for details);

4) apply EOSBA to Person-Mean(A) and Person-Mean(B) trimmed mean value scores (see section 4.2.5 for details);

5) fitting of Person-Mean(A) and Person-Mean(B) value/utility conversion models, using Person-Mean(A) and Person-Mean(B) value and utility scores for marker states MA and MB and MC (see section 4.2.6 for details and note that data from n=10 Person-Mean(B) respondents, who reported Dead less desirable than Pits on the Feeling Thermometer but Pits less desirable than Dead on the Chance Board, were not included in fitting the Person-Mean(B) value/utility conversion model);

6) calculate Person-Mean(A) and Person-Mean(B) utility scores for each of the 37 multi-attribute health states used to fit the multiplicative MAUF, using Person-Mean(A) and Person-Mean(B) value/utility conversion models respectively (see section 4.2.7 for details);

7) re-scale Person-Mean(B) calculated utility scores from Dead = 0.00 / PH = 1.00 scale to Pits = 0.00 / PH = 1.00 scale, using a positive linear transformation (see section 4.2.8 for details);

8) calculate overall Person-Mean utility scores, for each of the 37 multi-attribute health states used to fit the multiplicative MAUF on Pits = 0.00 / PH = 1.00 scale, as weighted means of Person-Mean(A) utilities from step #6 and Person-Mean(B) utilities from step #7 based on prevalence proportions in Person-Mean(A) (n=223) and Person-Mean(B) (n=33);

9) convert Person-Mean corner state utility scores to disutility scores (disutility = 1 - utility) to estimate c_j's, where j=1,2,...,8;

10) calculate C, by solving a set of simultaneous non-linear equations (see section 4.2.12 for details);

11) normalize 8 sets of Person-Mean single-attribute health state utility scores onto single-attribute utility scales using positive linear transformation, such that for each attribute the state with the lowest level has a utility score of 0.00 and the state with the highest level

(represented by the Perfect Health state) has a utility score of 1.00 (see section 4.2.11 for details);

12) convert normalized Person-Mean single-attribute utility scale scores (from step #11) to disutility (disutility = 1 - utility) scores to estimate u_j's, where j=1,2,...,8;

13) test whether the additive model fits, if not fit multiplicative form (see section 4.2.12 for details);

14) convert the fitted multi-attribute disutility function (MADUF) into a multi-attribute utility function (MAUF) - defined on Pits = 0.00 to Perfect Health = 1.00 scale (see Table 10 for details);

15) convert the MAUF defined on Pits = 0.00 to Perfect Health = 1.00 scale, to a simplified format that is defined on the conventional Dead = 0.00 to Perfect Health = 1.00 scale (see section 4.2.12 for details).

The steps involved in fitting the Person-Mean(A) and Person-Mean(B) value/utility conversion models are illustrated in Figure 8. Normalized value scores were calculated in step #1 using the following equation

$$Vn_{statei} = Vn_{PH} - (((Vr_{PH} - Vr_{statei}) \times Rn) / Rr)$$

where Vn_{statei} is the normalized score for state i,
Vn_{PH} is the normalized score for Perfect Health (defined as 100),
Vr_{PH} is score reported by respondent for Perfect Health,
Vr_{statei} is the score reported by respondent for state i,
Rn is the normalized scale range (defined as 100),
Rr is the range used by the respondent (the difference between the maximum and minimum scores reported by respondent), and
if $Vr_{statei} > Vr_{PH}$ then $Vn_{statei} = 100$.

A positive linear transformation (PLT) was used to re-scale Person-Mean(B) utility scores in step #7 above for three reasons. First, PLT had been used successfully in our previous work (Torrance et al. 1982, Torrance 1986 and Torrance et al. 1996a). Second, utilities are unique up to a positive linear transformation according to MAUT (von Neumann and Morgenstern 1944, 1947) and, therefore, the re-scaled Person-Mean(B) scores remain grounded in utility theory. Note that Step #7 did not involve measures of states rated worse than dead and, therefore, no consideration was given to use of the transformation proposed by Patrick et al. (1994).

3.5.3 Agreement between predicted and directly measured utility scores

If it was to be assumed that the directly measured standard gamble scores of comprehensive health states are "gold standard" preference measures representing truth, then it would be reasonable to consider the performance of the fitted HUI3 multiplicative MAUF in the context of the following model:

$$\text{Accuracy} = \text{Precision} + \text{Bias}.$$

However, this approach to assessing the performance of the HUI3 multiplicative MAUF has not been taken, for two reasons. First, because as argued by Fischer (1979), the decomposed modelling approach may provide more "accurate" preference scores than direct measurement because the cognitive burden for each of the decomposed tasks is much less than the cognitive burden of the comprehensive task. Second, lack of experience with utility scores defined on the Pits/Perfect Health scale limits the ability to judge the face-validity of the directly measured utility scores of health states used in assessing the performance of the function. Instead, a more agnostic approach has been adopted - assessing the level of agreement between scores calculated using the MAUF and directly measured standard gamble scores, without attempting to assess the overall accuracy of MAUF scores.

Agreement between utility scores predicted by the fitted MAUF in standard format (described in step 14 above) and directly measured utility scores from standard gambles involving Pits and Perfect Health states, was assessed for a variety of states directly measured in both the HUI3-M and HUI3-D surveys. Directly measured utility scores from the HUI3-M survey for the three methodological marker states were used to assess internal, or intra-survey agreement (ie., the extent of agreement between estimates obtained from the same set of respondents). Directly measured utility scores from the HUI3-D survey for the three methodological marker states and for 73 prevalent health states were used to assess external, or inter-survey agreement (ie., the extent of agreement between estimates obtained from a different set of respondents).

Agreement between predicted and directly measured utility scores on the Pits/Perfect Health scale was assessed using a number of statistics: mean difference, mean absolute difference, overall standard deviation, and intra-class correlation coefficient. Agreement assessment statistics were also calculated for utility scores weighted by the prevalence of the health states in the general population.

3.5.3.1 Intra-survey agreement: HUI3-M

Internal predictive validity (ie., the extent to which each model can predict utility scores for respondents *within* the HUI3-M preference survey) was assessed by comparing the utility

scores calculated using the MAUF for each of the three Marker health states to the mean of directly measured utility scores for these states, as reported by respondents to the HUI3-M preference survey.

3.5.3.2 Assessment of comparability of HUI3-M and HUI3-D survey respondent characteristics and "marker" state utility scores

Frequency distributions of respondent characteristics and the distributions of marker state scores were examined to determine the extent to which the Modelling Survey and Direct Survey respondent groups might be considered comparable, at least in terms of these variables. The expectation was that the frequency distributions for these variables would be quite similar between the two groups because respondents had been randomly assigned to either the Modelling Survey or the Direct Survey. Similar distributions would suggest that it may be reasonable to consider both groups of respondents as samples from the same underlying population and, therefore, these groups could be considered comparable for the purposes of assessing the generalizability of scores calculated using the HUI3 MAUF. Good agreement between scores calculated using the HUI3 MAUF (ie., from HUI3-M survey respondent data) and directly measured utility scores from HUI3-D survey respondent data) would constitute evidence of generalizability.

3.5.3.3 Inter-survey agreement: HUI3 MAUF versus HUI3-D utility scores

External agreement (ie., the extent to which each model can predict utility scores for a group of respondents *other than* the group whose preference scores were used to develop the model) was assessed by comparing the utility scores calculated using the MAUF for each of the 73 health states (three Marker states, five highly prevalent states, and 65 other prevalent states) to the mean of directly measured utility scores for these states, as reported by respondents to the HUI3-D preference survey.

The external agreement analyses were extended to weight the differences between MAUF calculated scores and directly measured utility scores for each of the 73 health states measured in the HUI3-D survey. The differences were weighted by the prevalence rates of the states in the 1991 General Social Survey. These calculations provide evidence about the sensitivity of results to the type of scoring method used in evaluating the health-related quality of life of a general population.

4.0 RESULTS

4.1 Field Work

4.1.1 Sampling results

A random sample of 3,000 households was drawn from the assessment rolls for the City of Hamilton Canada by staff of the Regional Municipality of Hamilton-Wentworth Planning Department. Information about each household included the name of the registered owner or tenant, street address and postal code. Letters, on McMaster University Faculty of Health Sciences letterhead, were sent to a total of 1,144 residents of the sample households. Interviewers attempted to make contact with household residents by phone or at the door, within one week of the letter being mailed. Figure 9 summarizes the outcomes of contacting the 1,144 households. Five hundred and four (504) interviews were completed, which represents 65% of eligible subjects who could be contacted after the households had been enumerated by a study interviewer.

4.1.2 Respondent characteristics

Summary statistics and frequency distributions for some demographic characteristics, including additional life expectancy, of survey respondents are presented in Table 2. The mean age of respondents was 43.3 years, 59% were female, 52% were married, 45% were employed full-time, and the mean expected time of remaining life was 37.0 years. Table 3 presents the frequency distributions of levels for each of the eight HUI3 attributes, based on HUI3-M survey respondents' reports about their own health status. Most levels of each attribute were reported by at least one HUI3-M respondent.

4.1.2.1 Generalizability of HUI3-M and HUI3-D respondent characteristics

The frequency distributions of HUI3-M and HUI3-D preference survey respondents' major demographic characteristics (eg., age, sex, income, education) have been compared with frequencies reported for Hamilton in the 1991 census. The respondent characteristic frequency distributions from HUI3-M and HUI3-D surveys are capable of representing the demographic, social and economic structure of the general population (Roberge 1996b).

4.2 Fitting MAUF

4.2.1 Modelling steps

The modelling steps were described in Section 3.5 and are illustrated in Figure 7. Recall that there are 15 major steps. Results for each step will be described briefly.

4.2.2 Normalization of respondent value scores

Each respondent's value scores for the single-attribute (including corner) states were normalized such that the least desirable health state was assigned a value score of 0.00 and the most desirable health state was assigned a value score of 1.00. The most desirable health state reported by all respondents was Perfect Health. Respondent preference measures (ie., value and utility scores) were then classified into one of two groups: Person-Mean(A), or Person-Mean(B).

4.2.3 Classification of respondents according to attitude about states worse than dead

Person-Mean(A) respondents (n=223) each reported Pits as being the least desirable state, and some reported Dead as being equally undesirable to Pits, in terms of value scores (from the Feeling Thermometer measurement task). All Person-Mean(A) respondents also reported Pits as being the least desirable state in terms of utility scores (from the Chance Board measurement task). Data from all Person-Mean(A) respondents were included in fitting the Person-Mean(A) value/utility conversion model, and for fitting the MAUF.

Person-Mean(B) respondents (n=33) each reported Dead as being the least desirable state in terms of value scores reported using the Feeling Thermometer, but a sub-group of Person-Mean(B) respondents (n=10) reported Pits as being the least desirable state in terms of utility scores (from the Chance Board measurement task). Data from this Person-Mean(B) sub-group of respondents were excluded from the fitting of the Person-Mean(B) value/utility conversion model, but their value score data were included in the calculation of Person-Mean(B) utility scores used for fitting the MAUF.

4.2.4 Person-Mean(A) and Person-Mean(B) preference scores

The 10% trimmed mean was selected as the measure of central tendency for health state value and utility scores reported by Person-Mean(A) and Person-Mean(B) respondents. (The 10% trimmed mean is the mean calculated after excluding 5% of observations from each end of the distribution of scores for the state.)

The 10% trimmed means of state i (an 8-element vector health state) for Person-Mean(A) respondents' value and utility scores are referred to as the state i Person-Mean(A) value and utility score, respectively. Similarly, the 10% trimmed mean of state i for Person-Mean(B) respondent scores is referred to as the Person-Mean(B) value or utility score for state i. The value scale of each Person-Mean(A) respondent was defined such that Pits = 0.00, Perfect Health = 1.00 and the value scores of all other states including dead are included in this interval. The value scale for Person-Mean(A) is, therefore, defined in the interval Pits = 0.00 and Perfect Health = 1.00. Correspondingly, the value scale of each Person-Mean(B) respondent was defined such that Dead = 0.00, Perfect Health = 1.00 and the value scores of all other states including Pits are included in this interval. The value scale for Person-Mean(B) is, therefore, defined in the interval Dead = 0.00 and Perfect Health = 1.00.

Person-Mean(A) and Person-Mean(B) value scores were calculated for each of the following 40 states: 3 Marker states (MA, MB, MC); 8 corner states; and 29 other attribute-level states. In addition, the Person-Mean(A) value score was calculated for Dead and the Person-Mean(B) value score was calculated for Pits state. These scores are presented in the second columns, labelled "10% Trd Mean Val", of Tables 4 and 5.

Corner states are defined as states having one, and only one, attribute described at the lowest level defined by the classification system (eg., the corner state for vision is "unable to see at all" with all other attributes at level one). Other single-attribute states are defined as states having one, and only one, attribute described at an intermediate level (ie., not the lowest level defined by the classification system for the attribute) with all other attributes at level one. The other single-attribute states are distributed across attributes as follows: 4 for vision, 4 for hearing, 3 for speech, 4 for ambulation, 4 for dexterity, 3 for emotion, 4 for cognition, 3 for pain.

Person-Mean(A) utility scores for each of the 3 Marker states were calculated from standard gamble measurements on Group A respondents on the Pits = 0.00 to Perfect Health = 1.00 scale. Person-Mean(B) utility scores for each of the 3 Marker states were calculated from standard gamble measurements on Group B respondents on the Dead = 0.00 to Perfect Health = 1.00 scale. These scores are presented in the columns labelled "Mean Utility" in sections 7.1 and 7.2 of Table 7, for Person-Mean(A) and Person-Mean(B) respectively.

4.2.5 EOSBA of Person-Mean(A) and Person-Mean(B) value scores

The Person-Mean(A) and Person-Mean(B) 10% trimmed mean value scores were adjusted for end-of-scale bias (EOSB) using an EOSB adjustment (EOSBA) factor of 1.78. The

EOSBA was applied to the value score of the health state ranked second (ie., ranked immediately below Perfect Health - the first ranked state) among a set of states, if and only if the value score of the state was greater than 75 (on a 0 to 100 VAS). Sets of states were defined based on the group of states rated by respondents on the feeling thermometer as a set. Nine sets of states were defined: 1) Scale anchor and Marker states; 2) Vision attribute-level and corner states; 3) Hearing attribute-level and corner states; 4) Speech attribute-level and corner states; 5) Ambulation attribute-level and corner states; 6) Dexterity attribute-level and corner states; 7) Emotion attribute-level and corner states; 8) Cognition attribute-level and corner states; 9) Pain attribute-level and corner states. The 10% trimmed mean value scores adjusted for EOSB are presented in the fourth column (labelled "EOSBA MA Val") of Tables 4 and 5, for Person-Mean(A) and Person-Mean(B).

EOSBA MA Val scores for attribute-level states were re-scaled, using a positive linear transformation, within the interval defined by the rank-2 state and corner state scores. Re-scaling within the rank-2 state to corner state interval was used to retain the relative-distance information provided in the original measures of single-attribute level scores. These scores are presented in the fifth columns, labelled "Re-Scald MA Val", of Tables 4 and 5. EOSBA value scores for Anchor and Marker states were not re-scaled after EOSBA.

4.2.6 Person-Mean(A) and Person-Mean(B) value to utility conversion models

We explored a variety of functional forms for the conversion of value (visual analogue scale feeling thermometer scores) scores to utility scores (scores derived from the standard gamble using the chance board). In previous studies we had used a simple power function to capture the non-linear relationship between dis-value and dis-utility scores. In this study we examined a variety of transformations (including natural logarithm), a variety of formulations (including regressing utility scores on value scores or alternatively, disutility scores on disvalue scores), nonlinear estimation techniques, and spline functions. After considerable exploration we concluded that the traditional power function did as good a job (or better) in predicting utility scores from value scores and was more convenient to use than alternative methods. It should be noted that the conversion model being reported here was a power function fitted to value and utility scores, rather than a power function conversion model fitted to dis-value and dis-utility scores as in our previous work, because in this application the agreement between predicted and measured utility scores was better using a conversion model fitted to value and utility scores. It is interesting to note that Stiggelbout et al. (1996) provides a concise review of literature reporting use of power functions to convert VAS scores to TTO scores and presents other data that led them to conclude that the power function transformation reported by Torrance (1976) is replica-

ble. (TTO scores, like standard gamble, are elicited using forced-choice type questions.) Thus the MAUF relies on a power function for the conversion of value to utility scores.

Data used to fit Person-Mean(A) and Person-Mean(B) value/utility models are presented in Table 7. Person-Mean(A) value/utility model was fitted using Person-Mean(A) value and Person-Mean(A) utility scores for states MA, MB and MC. The Person-Mean(A) value and utility scores for these states were each based on 223 observations. The Person-Mean(A) fitted value/utility relationship was

$$u = v^{0.559}$$

The fitting process used straight-line regression through the origin, on the natural log transformations of Person-Mean(A) value and utility scores (ie., $\ln(u) = \alpha\ln(v)$) The fit yielded an R^2 of 0.760, not corrected for the mean.

Person-Mean(B) value/utility model was also fitted using Person-Mean(B) value and Person-Mean(B) utility scores for states MA, MB and MC. Person-Mean(B) value and utility scores for these states were each based on 23 observations. Data from a sub-group (n=10) of Person-Mean(B) respondents were excluded from the fitting of the Person-Mean(B) value/utility conversion model because respondents in the sub-group reported inconsistent rankings of Dead and Pits and, therefore, these data were not appropriate for estimating the Person-Mean(B) value/utility conversion model for use in converting values on the Dead/PH scale to utilities on the Dead/PH scale. (These 10 respondents each reported Dead was less desirable than Pits on the Feeling Thermometer, but Dead was more desirable than Pits on the Chance Board question.) The Person-Mean(B) fitted value/utility relationship was

$$u = v^{0.474}$$

($R^2 = 0.974$, not corrected for the mean).

4.2.7 Conversion of Person-Mean(A) and Person-Mean(B) value scores to utility scores

Person-Mean(A) value scores for each attribute-level and corner state (column 5 of Table 4) were converted into Person-Mean(A) utility scores (column 6 of Table 4) using the Person-Mean(A) value/utility model. Similarly, the Person-Mean(B) value/utility model was used to calculate Person-Mean(B) utility scores (column 6 of Table 5) from Person-Mean(B) value scores for each attribute-level and corner health state (column 5 of Table 5).

4.2.8 Re-scaling of Person-Mean(B) utility scores

Person-Mean(B) utility scores for each health state were re-scaled using linear transformation, to facilitate fitting the MADUF. The MADUF is fitted on a scale defined such that Perfect Health has a disutility score of 0.00 (ie., a utility score of 1.00) and Pits has a disutility score of 1.00 (ie., a utility of 0.00). Person-Mean(A) utility scores are defined on the scale anchored by Pits = 0.00 and Perfect Health = 1.00 and, therefore, no re-scaling was required (column 6 Table 4 and column 3 Table 6). In contrast, Person-Mean(B) utility scores are defined on the scale anchored by Dead = 0.00 and Perfect Health = 1.00 and were re-scaled using linear transformation such that Person-Mean(B′) utility scores are defined on the scale Pits = 0.00 and Perfect Health = 1.00. Person-Mean(B′) utility scores for each of the attribute-level and corner states are presented in column 7 of Table 5.

4.2.9 Calculation of overall Person-Mean utility scores

The overall Person-Mean utility score was calculated as the weighted mean of Person-Mean(A) and Person-Mean(B′) utility scores, for each of the attribute-level and corner states. The Person-Mean(A) weighting factor of 0.9102 (233 Group A respondents / 256 total HUI3-M survey respondents) and the Person-Mean(B) weighting factor of 0.0898 (33 Group B respondents / 256 HUI3-M survey respondents) were applied to Person-Mean(A) (column 3 of Table 6) and Person-Mean(B′) (column 5 of Table 6), respectively. The overall Person-Mean, referred to as "Person-Mean", scores for each attribute-level and corner state are presented in column 6 of Table 6.

4.2.10 Person-Mean utility scores of corner states

To facilitate the collection of preference measurements, the study design specified that the multiplicative multi-attribute function would be fitted in disutility terms. Person-Mean utility scores (u) were converted into disutility scores ($\bar{u}$) using the formula $\bar{u} = 1 - u$. Person-Mean utility and disutility scores for each state are presented in Table 8. The lowest levels of each attribute represent the corner state for that attribute and, therefore, the underlined scores in Table 8 represent the disutility scores for each of the 8 corner states.

4.2.11 Person-Mean single-attribute utility functions

Table 9 presents the single-attribute utility functions on the Pits = 0.00 / PH = 1.00 scale (column 2), the single-attribute utility functions on the lowest level = 0.00 / highest level = 1.00 scale (column 3), and the single-attribute disutility functions on the lowest level = 1.00 / highest level = 0.00 scale (column 4). Note that "level 1" is the highest level for each of the attributes and is a scale anchor. The scores for the highest levels (1) are not presented, and are defined as equal to 1.00 in terms of utility and 0.00 in terms of disutility.

4.2.12 Person-Mean multi-attribute utility functions (MAUF) in standard and simplified formats

Table 10 presents the multi-attribute disutility function in standard format. The Person-Mean single-attribute disutility scores from Table 9 provide the $\bar{u}_j$'s. The c_j's are the disutility scores for each of the lowest attribute-level states (ie., corner states) on the Pits/PH scale and presented in Table 8, and c was calculated by iteratively solving the equation

$$1+c=\prod_{j=1}^{8}(1+c\,c_j)$$

where $\prod_{j=1}^{8}$ is the product of all $(1+c\,c_j)$ from c_1 to c_8.

The sum of the eight c_j constants indicates the search interval for the iterative solution. Specifically, if

$$\sum_{j=1}^{8} c_j > 1, \text{ then } -1 < c < 0;$$

$$\sum_{j=1}^{8} c_j = 1, \text{ then } c = 0, \text{ and the additive model holds; and}$$

$$\sum_{j=1}^{8} c_j < 1, \text{ then } c > 0.$$

The additive model was soundly rejected based on the sum of c_j being 3.55. The parameter c was calculated to be -0.991.

Table 10 includes the formulae for the multi-attribute disutility function, on the Perfect Health = 0.00 to Pits = 1.00 scale ($\bar{u}$) and on the Perfect Health = 0.00 to Dead = 1.00 scale ($\bar{u}^*$), in standard format.

Many applications require utility scores be based on the conventional scale where Dead = 0.00 and Perfect Health = 1.00. To accommodate this need we have provided the conversion formulae at the bottom of Table 10. To facilitate the use of the HUI3 multiplicative multi-

attribute function by most analysts we have also converted the standard format for this scale into the simpler to use format presented in Table 11.

It is acknowledged that the approach presented here is but one of a number of potential alternative approaches. The approach used here was selected to be conceptually consistent with our previous work (Torrance et al. 1996a) and to appropriately weight the preferences for dead, relative to living in various other health states, of the two major groups described earlier: Person-Mean(A) and Person-Mean(B).

The weighted mean utility score for Dead on the Pits/PH scale was used to convert the MADUF in standard format on the Pits/PH scale (Table 10) into two alternative formats: i) the MADUF on the Dead/PH scale (bottom of Table 10); and ii) the MAUF in simplified format on the Dead/PH scale (Table 11).

The weighted mean utility score for Dead was calculated using the 10% trimmed mean of directly measured standard gamble scores of Dead for those respondents (n=233) who provided standard gamble scores for Dead on the Pits/PH scale, and the utility score of Dead derived using linear transformation of the 10% trimmed mean of directly measured standard gamble scores of Pits for those respondents (n=23) who provided standard gamble scores for Pits on the Dead/PH scale. (The weighted mean utility score for dead could alternatively have been estimated based on VAS scores converted to utility scores using the power functions. The weighted mean utility scores for dead have been estimated using directly measured standard gamble scores because these scores provide the best available estimate of the utility of dead.)

The 10% trimmed mean of the directly measured standard gamble scores for Dead was 0.346 (n=233). The 10% trimmed mean of directly measured standard gamble scores for Pits was 0.362 (n=23), and using positive linear transformation the derived utility score for Dead for this group of respondents was calculated to be -0.567. The weighted mean score for Dead, based on the 10% trimmed mean of directly measured standard gamble scores and the derived utility of Dead, was 0.264 [ie., 0.264 = (233 / 256 * 0.346) + (23 / 256 * (-0.567))].

The constant terms of the MAUF in simplified format formula in Table 11, were calculated as follows.

1st constant = (-1.0 / c) / (1.0 - utility of Dead on Pits/PH scale)
= (-1.0 / -0.991) / (1.0 - 0.264)
= 1.371

2nd constant = ((1 + (1 / -0.991) - utility of Dead on Pits/PH scale)) /
(1.0 - utility of Dead on Pits/PH scale)
= ((1 + (1 / -0.991) - 0.264)) / (1.0 - 0.264)
= -0.27308 / 0.736
= -0.371

Scores calculated using the functions presented in Tables 10 and 11 should be interpreted as representing the point estimates of Person-Mean utility scores for a specific health state. Multi-attribute utility theory specifies that a deterministic, rather than statistical, approach be used to estimate the parameters of the appropriate functional form. This approach to model fitting does not involve analysis of variance and this distinction was an important design factor in selecting the appropriate methods for assessing the performance of the MAUF. A number of agreement assessments of HUI3 multiplicative MAUF scores and directly measured utility scores are described in the next section.

4.3 Assessment of the Performance of the MAUF

When assessing HUI3 MAUF scores it is necessary to consider three important factors: 1) that MAUF scores represent Person-Mean utility measures; 2) that the MAUF provides a broad range of scores for a large number of unique comprehensive health states; and 3) that there is no accepted "gold-standard" utility scores for comprehensive health states defined using the HUI3 classification system, or any other system for that matter. These three factors establish that "rater" should be defined by Person-Mean, that the performance of the MAUF should be assessed using numerous health states that span the health status and preference scoring space, and that performance should be assessed as agreement between two scoring methods that are considered equally valid.

Conventional intra- and inter-rater agreement terms consider the level of agreement between two or more types of individual-level measurements. The analogous approach presented here, given that the Person-Mean is the appropriate analytical level for assessing performance of the MAUF, might be appropriately labelled intra-survey and inter-survey agreement. In this case intra-survey agreement would be assessed according to the level of agreement between two different scoring methods each based on the same Person-Mean: Person-Mean HUI3-M. In this case, the data set consists of pairs of scores for various health states. A pair of scores for a specified health state consists of the 10% trimmed mean of the standard gamble score for the state and the calculated score for the same state using the MAUF. Intra-survey agreement is, therefore, somewhat analogous to conventional intra-rater agreement because it is assessed as the level of agreement between pairs of these scores for a collection of states such that both scores in the pair are based on measures from the same rater.

Similarly, inter-survey agreement can be considered somewhat analogous to conventional inter-rater agreement assessments. In this case, inter-survey agreement is assessed between pairs of scores for specified health states. However, the pairs of scores used for inter-rater agreement analyses differ by two factors: Person-Mean HUI3-M versus Person-Mean HUI3-D; and scoring method (directly measured and calculated using the MAUF). Specifically, a pair of scores consists of the calculated score for a specified state using the MAUF based on measures collected from HUI3-M survey respondents and the 10% trimmed mean of the standard gamble score for the same state provided by HUI3-D survey respondents.

In summary, for the purposes of the agreement analyses reported here to assess the performance of the MAUF, the underlying concept is that the variability in observed utility scores should be explained in terms of 3 major factors:

1) variability due to differences in the desirability of health states;
2) variability due to differences in scoring methods;
3) variability due to differences between the two sets of raters (HUI3-M and HUI3-D); and
4) error.

Intra-survey agreement involves analysis of variance due to factors 1, 2 and 4, while inter-survey agreement includes all four sources of variability but the effects of factors 2 and 3 are fully confounded. Inter-rater agreement would be expected to be lower than intra-rater agreement because there are two potential sources of observed inter-rater disagreement, rather than the one source available to explain intra-rater agreement. The overall design of the HUI3 preference measurement study provided for intra-rater agreement to be assessed using utility scores

from a set of three health states: MA, MB and MC. Inter-survey agreement was to be assessed using pairs of scores from two sets of health states: the set of MA and MB and MC; and 70 other states which were reported as being prevalent in the general population.

The following agreement statistics were calculated.

Mean difference

$= [\Sigma$ (predicted - 10% trimmed mean)/n]

Mean absolute difference

$= [\Sigma$ (|predicted - 10% trimmed mean|)/n]

Overall standard deviation

$= [(\Sigma$ (predicted - 10% trimmed mean)^2)/(n-1)]

Intra-class correlation coefficient.

$$= \frac{MS1 - MS3}{MS1 + MS3 + (2/n) \times (MS2 - MS3)}$$

Where MS1 is mean sums of squares for health states; MS2 is the mean sums of squares for scoring methods; and MS3 is the mean sums of squares for residual.

The mean difference (the sum of the differences divided by the number of states) provides one summary of the extent of agreement between the directly measured scores and scores for the same health states generated by the MAUF. If the sign of the mean difference (MD) is positive, then the MAUF over predicts scores (relative to directly measured); if the sign is negative, the MAUF under predicts. A small MD is an indication of agreement.

The mean difference statistic could, however, be misleading if large over predictions were offset by large under predictions. The mean absolute difference (MAD) addresses this potential limitation by taking the absolute value of the difference between predicted and directly measured scores. Again a small MAD is an indication of agreement.

The overall standard deviation (OSD) provides information of the variability of the difference between predicted and directly measured. Again, ideally one would like the OSD to be small.

The intra-class correlation coefficient is the proportion of explained variability attributable to differences in health state descriptions. High intra-class correlations for agreement between methods used to determine utility scores for health states would be associated with a large proportion of the explained variability being attributed to differences among health states.

Intra-survey agreement analyses are presented in detail to illustrate the process and interpretation of the results. Only the summary statistics are presented for inter-survey agreement results.

4.3.1 Intra-survey agreement

The Table 12 presents the MAUF scores and the mean directly measured standard gamble utility scores based on HUI3-M survey respondent data for states MA and MB and MC, and the descriptive statistics for the difference scores. To provide a benchmark for comparative purposes, analogous results (intra-survey) for the multiplicative HUI2 MAUF (data from Table 10 of Torrance et al. 1996a) and for the HUI1 MAUF (data from Torrance et al. 1982 and Drummond et al. 1987) are also provided. The HUI3 results indicate that the MAUF utility score was lower for two of the three states, that the MAUF score was 0.011 units higher than the directly measured SG scores, that the mean absolute difference between the two sets of scores was 0.067 units, and that the overall standard deviation was 0.084 units.

A more statistically rigorous method of assessing agreement is based on the analysis of variance framework and summarized using an intra-class correlation coefficient (Burdick and Graybill 1992). This technique involves partitioning the total variance into components due to 3 factors: health state; scoring method; and residual. Table 13 presents the ANOVA table of the intra-rater agreement analysis.

The intra-class correlation coefficient (ICC) is calculated using the following equation (Burdick and Graybill 1992):

$$ICC = \frac{MS1 - MS3}{MS1 + MS3 + (2/n) \times (MS2 - MS3)}$$

$$= \frac{0.04974 - 0.00349}{0.04974 + 0.00349 + (2/3) \times (0.00017 - 0.00349)}$$

$$= 0.907.$$

The ICC represents the proportion of utility score variability that is either unexplained (MS3) or attributed to differences among health state descriptions. In this case, the point estimate of the proportion total variability attributed to these types of differences exceeds 0.9. The remaining variability in utility scores, which is attributed to the effects of scoring method, is small in comparison and is interpreted as evidence of good agreement between scoring methods. Good intra-survey agreement between scoring methods implies that the HUI3 MAUF scores are robust to the method used to determine utility scores. As a comparison, the ICC for intra-survey agreement for HUI2 is 0.95.

4.3.2 Inter-survey agreement

The evidence suggests that intra-survey or within Person-Mean agreement is good. Inter-survey agreement evidence will provide information about the extent to which the MAUF might generalize and represent preferences of respondents not used to fit the MAUF, if we extrapolate the intra-survey agreement analyses results to assume that the differences due to method of scoring are small.

As described previously, two preference measurement surveys were conducted for HUI3 health states: HUI3-M; and HUI3-D. The respondents to these two surveys are mutually exclusive, were selected from the same sample frame, were randomly assigned to one of the two surveys, and the sample sizes are approximately equal. The expectation, therefore, was that there would be no important differences between the characteristics of the HUI3-M (n=256) and HUI3-D (n=248) respondent samples. It was expected that the two surveys could be treated as replicate samples of the underlying population. Good agreement between scores from two different samples would be evidence of stability in HUI3 scores between groups: HUI3-M respondent scores used to fit the HUI3 MAUF scores; and HUI3-D Person-Mean SG scores (Feeny et al. 1995b).

The frequency distributions of respondents' socio-economic factors and mean preference scores for a small sample of health states were compared across the two groups of survey respondents to determine whether there was any evidence to suggest that the independent characteristics of two groups were different. Differences in these variables might suggest a problem with the random allocation process that formed the two survey groups.

There was no significant difference between the HUI3-M and HUI3-D survey respondent group mean age ($p > 0.20$) and no significant differences in frequency distributions of categorical socio-economic variables: sex ($p > 0.10$); marital status ($p > 0.10$); education ($p > 0.50$);

religion ($p > 0.10$); employment status ($p > 0.10$); family income ($p > 0.10$); and perceived global health of self ($p > 0.50$). The mean utility scores, on the Pits = 0.00 to Perfect Health = 1.00 scale, for the marker states were not significantly different between the two surveys ($p > 0.10$). The mean utility scores for Marker State A in the HUI3-M and HUI3-D surveys were 0.794 and 0.771, for Marker State B 0.663 and 0.662, and for Marker State C 0.565 and 0.559.

Based on these results and for the purposes of interpreting inter-survey agreement statistics, the HUI3-M and HUI3-D preference-measurement data sets are considered as two samples of the same underlying population for three reasons. First, the sample frame for the two groups was the same and respondents were randomly allocated to either the HUI3-M or HUI3-D surveys. Second, there are no significant or important size differences between the mean utility scores of the two groups for the states measured in common (ie., the 3 Marker states). Third, there are no significant differences between the distributions of socio-economic characteristics between the two sets of survey respondents.

Inter-survey agreement statistics are presented in Table 14 for 73 health states reported to be prevalent in the general population. The first row of statistics present results based on the premise that each health state is equally important to the assessment of agreement (ie., the scores were not weighted to account for differences in the prevalence rates of the health states in a general population). This set of statistics is the most appropriate for the purposes of assessing the performance of the MAUF across the full utility scale. The mean difference is -0.008. The mean absolute difference and the overall standard deviation are each approximately 0.10. The intra-class correlation coefficient point estimate is 0.88. The 95% confidence bounds for the ICC point estimate is 0.49 and 0.92. In summary, agreement between scoring methods is high.

Inter-survey agreement statistics have also been calculated for use by analysts interested in the use of HUI3 in general population surveys. The prevalence rates of health states varies greatly in the general population and for the purposes of calculating indices of general population health it is more important that scores for highly prevalent states be estimated precisely and without bias than it is for low-prevalence rate states. Rows 2 and 3 of Table 14A present the summary agreement statistics for prevalence-weighted utility scores, with and without including scores for Perfect Health. The prevalence rates for the unique health states are for a non-institutionalized general population (personal communication from Edward Praught of Statistics Canada to William Furlong dated May 19, 1992) based on results from the 1991 Canadian General Social Survey (Statistics Canada 1994).

For agreement analyses including the prevalence rate of Perfect Health, the prevalence rates were normalized to account for the 86.18 percent coverage by the 73 health states. For agreement analyses excluding the prevalence rate of Perfect Health, the prevalence rates were normalized to account for the 56.46 percent coverage by the 73 health states. For the purposes of most analysts of general population data, the statistics including Perfect Health are the most appropriate because the prevalence of Perfect Health is high. Although it should be noted that by definition Perfect Health has a utility score of 1.00 and therefore there is no scope for a difference between scoring methods, the descriptive agreement statistics for both sets of analyses are similar which indicates that the performance of the HUI3 MAUF in the context of general population indices is robust to the inclusion/exclusion of unmeasured Perfect Health scores.

4.3.3 Summary of agreement evidence

The intra- and inter-survey agreement statistics are very similar. In brief, most of the variability in utility scores is explained by differences in health states. The combined effects of differences between raters (ie., Person-Mean HUI3-M versus Person-Mean HUI3-D) and differences due to scoring method (ie., MAUF or direct SG measurement) are small in comparison. However, given that the ICCs for intra- and inter-rater agreement are approximately equal, it is hypothesized that differences due to scoring method are more important than differences due to respondent sampling. In sum, the multiplicative multi-attribute utility function generates utility scores that are in close agreement with those obtained from direct measurement using the standard gamble technique.

5.0 DISCUSSION

The HUI3 is the latest member of the Health Utilities Index family of health status and health-related quality of life assessment systems. HUI1 has, for most purposes, been superseded by HUI2 and HUI3.

In many, but not all ways, the HUI3 represents an improvement over the HUI2 system. For example, the HUI3 health status classification system includes more attributes and levels, describes more health states, has greater structural independence, and includes vision and hearing and speech as distinct attributes. In addition, HUI3 provides a larger set of scores for states considered to be worse than dead. Further, the model fitting process included a number of innovative features such as more appropriate weightings of preferences of respondents who did not consider dead as the least preferable state and the inclusion of end-of-scale bias adjustment of value scores. The survey included many more states for assessing intra- and inter-survey agreement between utility scores calculated using the MAUF and mean directly measured standard gamble scores. However, it should also be recognized that the HUI2 system has withstood the test of time, and includes different constructs for attributes sharing the same name (eg., emotion and pain) or different attributes (ie., self-care and fertility) than HUI3. For these reasons, we recommend that many applications consider the HUI2 and HUI3 systems as complements, and collect sufficient data to classify the health status of each study subject according to both systems. Typically, the marginal costs of collecting the extra information are quite small.

This report presents the HUI3 utility function on the conventional Dead = 0.00 to Perfect Health = 1.00 scale, and the HUI3-specific Pits = 0.00 to Perfect Health = 1.00 scale. The conventional scale is most appropriate for comparisons involving preference scores from other measures that use this same scale, and for calculating aggregated indices of morbidity and mortality such as quality-adjusted life years (QALYs).

The HUI3 MAUF is based on well-established expected-utility theory (von Neuman and Morgenstern 1944 and 1947) and multi-attribute utility theory (Keeney and Raiffa 1976 and 1993). On the basis of expected-utility theory, the standard gamble technique used to collect the preference scores for fitting the HUI3 MAUF is the "gold standard" method for measuring preferences under conditions of uncertainty. The multiplicative multi-attribute utility function is one of the three basic forms defined by multi-attribute utility theory.

Scores for health states calculated using HUI3 health status classification data and the multiplicative scoring function described in this report represent mean utility scores for health states as assessed by a random sample of the general population. Interviewer assessments indicate that the vast majority of respondents understood the questions and tasks they were asked, and that the respondents were able to quantify their expected additional life expectancy without consulting life expectancy tables. The strategy for fitting the scoring function was developed to provide appropriate weightings for preference scores representing two major groups of respondents: those respondents who reported that living in some states would be worse than dead; and those respondents who reported that dead was less preferable than any of the health states evaluated.

In general, the sample of respondents to the HUI3 preference measurement surveys is reasonably representative of the adult general populations of Hamilton, the province of Ontario and Canada (ie., the univariate frequency distributions of common demographic variables for these four groups are similar). In addition, there is considerable evidence that little, if any, of the variability in utility scores for hypothetical states can be explained by socio-demographic factors (Sackett and Torrance 1978; Froberg and Kane 1989; Furlong 1996). For these reasons, the utility scores calculated using the HUI3 MAUF may be expected to be generally representative of many general populations. General population, or community, preference scores are appropriate for use in population health status indexes and in programme evaluation. It should be noted, however, that there is great heterogeneity of preferences for health states among individual respondents and therefore scores calculated using the HUI3 MAUF should not be considered as substitutes for utilities representing the preferences of specific individuals that might be used to inform clinical decisions for individual patients.

HUI3 MAUF seems to perform very well. The level of intra- and inter-survey agreement between utility scores calculated using the HUI3 multiplicative utility function and mean utility scores based directly on standard gamble measurements of numerous health states might be interpreted as evidence of both validity and reliability, depending upon theoretical perspective. From the point of view of utility theory, the level of agreement provides evidence of criterion validity because there is good agreement between scores calculated using the MAUF and the criterion method of measurement: directly measured standard gamble utility scores. In other words, the fitted MAUF models the structure of preferences among attributes sufficiently well to match mean directly measured standard gamble scores across a variety of health states. From an epidemiologic or psychometric viewpoint, agreement between two measurement methods (ie., utility scores calculated using the fitted MAUF and mean directly measured standard gamble

utility scores) may be considered evidence of reliability. *A Dictionary of Epidemiology* explains that "Lack of reliability may arise from divergences between observers or instruments of measurement or instability of the attribute being measured" (Last 1983). Streiner and Norman (1993) consider reliability in the context of psychometric generalizability theory and report that the "first step in providing evidence of the value of an instrument is to demonstrate that measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results". The multi-disciplinary differences in interpretation emphasizes that the results of the intra- and inter-survey agreement analyses provides important evidence about the performance characteristics of the HUI3 MAUF. This information should be considered as fundamental evidence toward establishing the usefulness of HUI3 MAUF scores.

Readers interested in using HUI3 in their own studies and readers interested in estimating multi-attribute utility functions on their own may be interested in the brief advice found in Appendix A (see pages 93-94).

6.0 CONCLUSION

The HUI3 utility functions provide estimates of preferences for health states from a community perspective. The community perspective is important for applications to meet methodologic guidelines regarding appropriate methods for economic evaluation in both Canada (Canadian Coordinating Office for Health Technology Assessment 1994 and 1997; Torrance et al. 1996b; Menon et al. 1996) and the U.S. (Gold et al. 1996; Siegel et al. 1997).

Gold et al. (1996, page 38) explain that a complete analysis of the costs and benefits of an intervention should include all costs, including lost productivity costs, and that whether the costs of lost productivity should be included in the numerator or the denominator of a cost-effectiveness analysis "depends, at least in part, on the framing of the question used to elicit utility weights for health states...If we choose the opposite convention, and explicitly exclude monetary costs from consideration in the utility assessment procedure... then these costs must be in the numerator." Respondents to the HUI3 preference measurement study were given explicit instructions to imagine that their income would remain the same, no matter what health state was being assessed. Therefore, the utility scores of HUI3 health states should be interpreted as not including the effects on income, and productivity costs should be counted separately in the numerator of cost-utility ratios calculated using HUI3 utility scores.

Utility scores for health states are interpreted in the context of accumulated knowledge about the distribution of scores associated with known groups and changes in scores of specific types of individuals over time. The meaning associated with HUI3 utility scores as measures of morbidity or HRQL is, therefore, informed by the interpretation of data collected from heterogeneous groups of subjects. Some results using utility scores from the HUI3 single- and multi-attribute functions have already been reported by us (Torrance et al. 1998) and, with the release of this document to other research groups, we expect many more reports in the near future. We encourage other researchers to contribute to the development of the HUI3 by including it in their own research studies.

The HUI3 classification system is applicable to a wide variety of subjects and studies (eg., general population and clinical settings). We suggest that others may wish to consider using the HUI3 multi-attribute classification system (Table 1) and the HUI3 utility scoring functions for their projects. The single-attribute utility functions presented in Table 9 can be used to convert categorical data to cardinal measures of limitations within attributes. The multi-attribute utility function presented in Table 11 provides single summary scores of HRQL for comprehensive

health states on the conventional Dead = 0.00 to Perfect Health = 1.00 scale. Inquiries about support services for applications (eg., sets of standardized questionnaires and procedures manuals) should be directed to the first author: William Furlong.

In summary, the HUI3 is a comprehensive, compact and efficient system for describing the comprehensive health status and for determining the health-related quality of life utility scores for a broad range of subjects. Evidence to date indicates that the HUI3 measurement system is responsive (Torrance et al. 1998), acceptable, reliable, valid, interpretable and useful (Feeny et al. 1998).

7.0 REFERENCES

Barr RD, Furlong W, Dawson S, Whitton AC, Strautmanis I, Pai M, Feeny D, Torrance GW. An assessment of global health status in survivors of acute lymphoblastic leukemia in childhood. *American Journal of Pediatric Hematology/Oncology* 1993;15(4):284-290.

Barr RD, Pai MKR, Weitzman S, Feeny D, Furlong W, Rosenbaum P, Torrance GW. A multi-attribute approach to health status measurement and clinical management - illustrated by an application to brain tumors in childhood. *International Journal of Oncology* 1994;4:639-648.

Barr RD, Feeny D, Furlong W, Weitzman S, Torrance GW. A preference-based approach to health-related quality of life in children with cancer. *International Journal of Pediatric Hematology/Oncology* 1995;2:305-315.

Barr RD, Petrie C, Furlong W, Rothney M, Feeny D. Health-related quality of life during post-induction chemotherapy in children with acute lymphoblastic leukemia in remission: an influence of corticosteroid therapy. *International Journal of Oncology* 1997;11:333-339.

Barr R. Development of a multiattribute health status measure. *Proceedings of Workshop on Quality of Life in Pediatric Oncology, Heriot-Watt University, Edinburgh*. Hertfordshire UK: Serono Laboratories 1998, pp 9-17.

Berthelot J-M, Roberge R, Wolfson M. The calculation of health-adjusted life expectancy for a Canadian province using a multi-attribute utility function: a first attempt. In JM Robine, CD Mathers, MR Bone, and I Romieu (eds). *Calculation of Health Expectancies: Harmonization, Consensus Achieved and Future Perspectives*. Montrouge, France: Colloque INSERM/John Libbey Eurotext Ltd 1993; 226: 161-172.

Billson AL, Walker DA. Assessment of health status in survivors of cancer. *Archives of Disease in Childhood* 1994;70:200-204.

Bosch JL, van Wijck EEE, Baum PL, Donaldson MC, van den Dungen JJAM, Hunink MGM. The McMaster Health Utilities Index (II) and the EuroQol 5D assessed in patients with peripheral arterial disease in the United States and the Netherlands. *Medical Decision Making* 1996;16(4):450(abstract).

Boyle MH, Furlong W, Feeny D, Torrance G, Hatcher J. Reliability of the Health Utilities Index Mark III used in the 1991 Cycle 6 General Social Survey health questionnaire. *Quality of Life Research* 1995;4(3):249-257.

Burdick RK, Graybill FA. *Confidence Intervals on Variance Components*. New York, NY: Marcel Dekker Inc 1992.

Bush JW, Chen MM, Patrick DL. Social indicators of health based on function status and prognosis. *Proceedings of the American Statistical Association, Social Statistics Section*; 1972 Aug: Montreal. Washington DC: American Statistical Association, 1972;71-80.

Cadman D, Goldsmith C, Torrance G, Boyle M, Furlong W. Development of a health status index for Ontario children. Final report to Ontario Ministry of Health for grant #DM648 (00633). McMaster University, December 1986.

Canadian Coordinating Office for Health Technology Assessment. *Guidelines for Economic Evaluation of Pharmaceuticals: Canada, First Edition.* Ottawa: Canadian Coordinating Office for Health Technology Assessment, November, 1994. Second Edition November, 1997.

Costet N, Le Gales C, Buron C, Kinkor F, Mesbah M, Chwalow J, Clinical and Economic Working Groups, Slama G. French cross-cultural adaptation of the Health Utilities Indexes Mark 2 (HUI2) and 3 (HUI3) classification systems. *Quality of Life Research* 1998;7:245-256.

Drummond MF, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford: Oxford University Press 1987.

Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes: Second Edition.* Oxford: Oxford University Press 1997.

Essink-Bot M-L, Stouthard MEA, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQol questionnaire. *Health Economics* 1993;2(3):237-246.

Fanshel S, Bush JW. A health status index and its application to health services outcomes. *Operations Research* 1970;18:1021-1066.

Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multi-attribute system for classifying the health status of survivors of childhood cancer. *Journal of Clinical Oncology* 1992;10(6):923-928.

Feeny DH, Leiper A, Barr RD, Furlong W, Torrance GW, Rosenbaum P, Weitzman S. The comprehensive assessment of health status in survivors of childhood cancer: application to high-risk acute lymphoblastic leukaemia. *British Journal of Cancer* 1993;67:1047-1052.

Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics* 1995a; 7(6):490-502.

Feeny D, Torrance GW, Furlong W, Goldsmith C, DePauw S, Boyle M, Denton M. Health Utilities Index Mark 3: utility scores for 74 prevalent health states. *Quality of Life Research* 1995b; 4(5): 424 (abstract).

Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In B Spilker (ed). *Quality of Life and Pharmacoeconomics in Clinical Trials: Second Edition.* Philadelphia: Lippincott-Raven Press 1996:85-95.

Feeny D, Furlong W, Barr RD. Multiattribute approach to assessment of health-related quality of life: Health Utilities Index. *Medical and Pediatric Oncology* 1998;Supplement 1:54-59.

Fischer GW. Utility models for multiple objective decisions: do they accurately represent human preferences? *Decision Sciences* 1979;10:451-479.

Froberg DG, Kane RL. Methodology for measuring health-state preferences - III: population and context effects. *Journal of Clinical Epidemiology* 1989;42(6):585-592.

Furlong WJ. Variability of utility scores for health states among general population groups. Hamilton, Canada: McMaster University, 1996. 140 pages. MSc Thesis.

Furlong W, Feeny D, Torrance G, Boyle M, Horsman J. Design and pilot testing of comprehensive health-status measurement system for the Ontario Health Survey. Final report to Ontario Ministry of Health for contract #12-907. McMaster University, June 1989.

Furlong W, Feeny D, Torrance GW, Barr RD, Horsman J. Guide to design and development of health-state utility instrumentation. Hamilton, Canada: McMaster University Centre for Health Economics and Policy Analysis, CHEPA Working Paper 90-9, 1990.

Furlong W, Torrance GW, Feeny D. Properties of Health Utilities Index: preliminary evidence. *Quality of Life Newsletter* 1995/6;13/14(June - January):3,4,10.

Gemke RJBJ, Bonsel GJ, van Vught AJ. Long term survival and state of health after paediatric intensive care. *Archives of Disease in Childhood* 1995;73(September):196-201.

Gemke RJBJ, Bonsel GJ. Reliability and validity of a comprehensive health status measure in a heterogenous population of children admitted to intensive care. *Journal of Clinical Epidemiology* 1996;49(3,March):327-333.

Glaser AW, Davies D, Walker D, Brazier D. Influence of proxy respondents and mode of administration on health status assessment following central nervous system tumours in childhood. *Quality of Life Research* 1997;6(1):43-53.

Gold, Martha R., Joanne E. Siegel, Louise B. Russell, and Milton C. Weinstein, eds. *Cost-Effectiveness in Health and Medicine*. New York: Oxford University Press 1996.

Grossman R, Mukherjee J, Vaughan D, Eastwood C, Cook R, LaForge J, Lampron N. A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis. *CHEST* 1998;113:131-141.

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Annals of Internal Medicine* 1993;118(8):622-629.

Hennessy CH, Moriarty DG, Zack MM, Scherr PA, Brackbill R. Measuring health-related quality of life for public health surveillance. *Public Health Reports* 1994;109(5):665-672.

Hood SC, Beaudet MP, Catlin G. A Healthy Outlook. *Health Reports* (Statistics Canada, Cat. No. 82-003) 1996;7(4):25-32.

Kanabar DJ, Attard-Montalto S, Saha V, Kingston JE, Malpas JE, Eden OB. Quality of life in survivors of childhood cancer after megatherapy with autologous bone marrow rescue. *Pediatric Hematology and Oncology* 1995;12:29-36.

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

Kaplan RM, Bush JW. Health related quality of life measurement for evaluation research and policy analysis. *Health Psychology* 1982;1:61-80.

Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs: Second Edition*. New York: Cambridge University Press 1993. First Edition: John Wiley & Sons 1976.

Kiltie AE, Gattamaneni HR. Survival and quality of life of paediatric intracranial germ cell tumour patients treated at the Christie Hospital, 1972-1993. *Medical and Pediatric Oncology* 1995;25:450-456.

Last JM (ed). *A Dictionary of Epidemiology*. An International Epidemiological Association Handbook. New York: Oxford University Press 1983.

Le Gales C, Buron C, Costet N, Kinkor F, Feeny D, Furlong W, Chwalow J, Slama PG. Assessment of the multi-attribute preference function for Health Utilities Index 3 in France: preliminary results. *Quality of Life Research* 1997;6(7/8):678(abstract).

Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: J Wiley and Sons, 1987.

Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL. Use of the Health Utilities Index with stroke patients and their caregivers. *Stroke* 1997;28(10):1888-1894.

Menon D, Schubert F, Torrance GW. Canada's new guidelines for the economic evaluation of pharmaceuticals. *Medical Care* 1996;34(12):DS77-DS86.

Ontario Ministry of Health. *Ontario Health Survey 1990: User's Guide Volume 1: Documentation*. Toronto: Ontario Ministry of Health and Premiers Council on Health, Well-Being and Social Justice; 1993.

Ontario Ministry of Health. *Ontario Guidelines for Economic Analysis of Pharmaceutical Products*. Toronto: Ontario Ministry of Health; September 1994.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Services Research* 1973a;8:228-248.

Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *Journal of Health and Social Behavior* 1973b;14:6-23.

Patrick DL, Erickson P. *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. New York: Oxford University Press 1993.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Medical Decision Making* 1994;14:9-18.

Roberge, Roger. Personal communication by facsimile message transmission addressed to Bill Furlong on June 27, 1996 in support of OECD questionnaire submission. Health Analysis and Modeling Section, Analytical Studies Branch, Statistics Canada, 1996a.

Roberge R. Health Utilities Index Mark 3: demographic characteristics of respondents to the preference measurement surveys. Draft technical report for HUI3 Utility Scoring Function Study dated February 23. Health Analysis and Modelling Group, Analytical Studies Branch, Statistics Canada, 1996b.

Roberge R, Berthelot J-M, Wolfson M. Health and socio-economic inequalities. *Canadian Social Trends* (Statistics Canada, Cat. No. 11-008E) 1995a;Summer:5-19.

Roberge R, Berthelot J-M, Wolfson M. The Health Utility Index: measuring health differences in Ontario by socioeconomic status. *Health Reports* (Statistics Canada, Cat. No. 82-003) 1995b;7(2, November):25-32.

Roberge R, Berthelot JM, Wolfson MC. Adjusting life expectancy to account for morbidity in a national population. *Quality of Life Newsletter* 1996;17(March-August):12-13.

Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: J Wiley and Sons, 1987.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *Journal of Chronic Disease* 1978;31:697-704.

Saigal S, Rosenbaum P, Stoskopf B, Hoult L, Furlong W, Feeny D, Burrows E, Torrance G. Comprehensive assessment of the health status of extremely low birthweight children at eight years of age: comparison with a reference group. *Journal of Pediatrics* 1994a;125(3):411-417.

Saigal S, Feeny D, Furlong W, Rosenbaum P, Burrows E, Torrance G. Comparison of the health-related quality of life of extremely low birthweight children and a reference group of children at age eight years. *Journal of Pediatrics* 1994b;125(3):418-425.

Saigal S, Feeny D, Rosenbaum P, Furlong W, Burrows E, Stoskopf B, Hoult L. Self-perceived health status and health-related quality of life of extremely low birthweight infants at adolescence. *Journal of the American Medical Association* 1996;276(6):453-459.

Statistical Engineering Laboratory. *Fractional Factorial Experiment Designs at Two Levels.* United States Department of Commerce, April 15, 1957.

Siegel JE, Torrance GW, Russell LB, Luce BR, Weinstein MC, Gold MR. Guidelines for pharmacoeconomic studies: recommendations from the Panel on Cost Effectiveness in Health and Medicine. *PharmacoEconomics* 1997;11(2):159-168.

Statistics Canada. *Health Status of Canadians: Report of the 1991 General Social Survey.* Ottawa: Statistics Canada, Cat. No. 11-612E #8, March 1994.

Statistics Canada. National Population Health Survey, 1996-97, Household Component, User's Guide for the Public Use Microdata Files. Ottawa: Statistics Canada, Cat. No. 82-M0009GPE, 1998a.

Statistics Canada. National Population Health Survey, 1996-97, Public Use Microdata Files. Ottawa: Statistics Canada, Cat. No. 82-M0009XCB, 1998b.

Statistics Canada and Human Resources Development Canada. 1994-95 National Longitudinal Survey of Children and Youth, User's Handbook and Microdata Guide. Ottawa: Statistics Canada Cat. No. 89M0015GPE (documentation) and Cat. No. 89M0015XDB (file), 1996a.

Statistics Canada and Human Resources Development Canada. *Growing up in Canada: National Longitudinal Survey of Children.* Ottawa: Statistics Canada Cat. No. 89-550-MPE, no. 1, 1996b.

Stiggelbout AM, Eijkemans MJC, Kiebert GM, Kievit J, Leer JH, De Haes HJCJM. The 'utility' of the visual analogue scale in medical decision making and technology assessment: is it an alternative to the time trade-off? *International Journal of Technology Assessment in Health Care* 1996;12(2 - Spring):291-298.

Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to their Development and Use.* Oxford: Oxford University Press 1993.

Torrance GW: Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. *Socio-Economic Planning Sciences* 1976;10(3):129-136.

Torrance W. Measurement of health state utilities for economic appraisal - a review. *Journal of Health Economics* 1986;5(1):1-30.

Torrance GW. Instrument and model precision estimates from previous studies. Technical notes for HUI3 Preference Modelling Study dated July 20, 1992. McMaster University, 1992.
Torrance GW. End-of-scale bias in feeling thermometer. Technical notes for HUI3 Preference Modelling Study dated October 24, 1996. McMaster University, 1996.

Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research* 1982;30(6):1042-1069.

Torrance GW, Furlong W, Feeny D, Boyle M. Provisional health status index for the Ontario Health Survey. Final report to Statistics Canada for project #44400900187. McMaster University, February 1992a.

Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R. Multi-attribute preference functions for a comprehensive health status classification system. McMaster University, Centre for Health Economics and Policy Analysis, CHEPA Working Paper 92-18, 1992b.

Torrance G, Feeny D, Boyle M, Goldsmith C, Denton M, Streiner D, Keeney R, Furlong W, DePauw S. Determining health status measures for Ontario Health Survey: general population preference scoring for the Health Utilities Index Mark III. Final report to the Ontario Ministry of Health for Grant #04020. McMaster University, November 1994.

Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health Utilities Index. *PharmacoEconomics* 1995a;7(6):503-520.

Torrance GW, Feeny D, Furlong W, Boyle M, Denton M, DePauw S, Goldsmith C. Health Utilities Index Mark 3: utility scores for prevalent states. *Medical Decision Making* 1995b; 15(4, Oct/Dec):435 (abstract).

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multi-attribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical Care* 1996a;34(7):702-722.

Torrance GW, Blaker D, Detsky A, Kennedy W, Schubert F, Menon D, Tugwell P, Konchak R, Hubbard E, Firestone T. Canadian guidelines for economic evaluation of pharmaceuticals. *PharmacoEconomics* 1996b;9(6):535-559.

Torrance GW, Feeny D, Furlong W, Barr RD, DePauw S. Health Utilities Index: further evidence of responsiveness. *Value in Health* 1998;1(1):41 (abstract).

von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior: Second Edition.* Princeton: Princeton University Press 1947. First Edition 1944.

Whitton AC, Barr RD, Rhydderch H, Case T, Feeny D, Furlong W, Torrance GW. Assessing the global health status of patients with brain tumours using a multi-attribute system. *Journal of Neuro-Oncology* 1994;18:191 (abstract).

Whitton AC, Rhydderch H, Furlong W, Feeny D, Barr RD. Self-reported comprehensive health status of adult brain tumor patients using the Health Utilities Index. *Cancer* 1997;80(2):258-265.

Wolfson MC. Health-Adjusted Life Expectancy. *Health Reports* (Statistics Canada Cat. No. 82-003) 1996;6(1): 41-46.

8.0 TABLES

Table 1

Multi-Attribute Health Status Classification System:
Health Utilities Index Mark 3 (HUI3)

Attribute		Level Description
VISION	1.	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses.
	2.	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses.
	3.	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses.
	4.	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses.
	5.	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses.
	6.	Unable to see at all.
HEARING	1.	Able to hear what is said in a group conversation with at least three other people, without a hearing aid.
	2.	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people.
	3.	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid.
	4.	Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hearing what is said in a group conversation with at least three other people even with a hearing aid.
	5.	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
	6.	Unable to hear at all.
SPEECH	1.	Able to be understood completely when speaking with strangers or friends.
	2.	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.
	3.	Able to be understood partially when speaking with strangers or people who know me well.
	4.	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.
	5.	Unable to be understood when speaking to other people (or unable to speak at all).

Table 1 Continued

AMBULATION	1.	Able to walk around the neighbourhood without difficulty, and without walking equipment.
	2.	Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person.
	3.	Able to walk around the neighbourhood with walking equipment, but without the help of another person.
	4.	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.
	5.	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.
	6.	Cannot walk at all.
DEXTERITY	1.	Full use of two hands and ten fingers.
	2.	Limitations in the use of hands or fingers, but does not require special tools or help of another person.
	3.	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person).
	4.	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools).
	5.	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools).
	6.	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools).
EMOTION	1.	Happy and interested in life.
	2.	Somewhat happy.
	3.	Somewhat unhappy.
	4.	Very unhappy.
	5.	So unhappy that life is not worthwhile.

Table 1 Continued

COGNITION	1.	Able to remember most things, think clearly and solve day to day problems.
	2.	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.
	3.	Somewhat forgetful, but able to think clearly and solve day to day problems.
	4.	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.
	5.	Very forgetful, and have great difficulty when trying to think or solve day to day problems.
	6.	Unable to remember anything at all, and unable to think or solve day to day problems.
PAIN	1.	Free of pain and discomfort.
	2.	Mild to moderate pain that prevents no activities.
	3.	Moderate pain that prevents a few activities.
	4.	Moderate to severe pain that prevents some activities.
	5.	Severe pain that prevents most activities.

Source: Feeny D, Furlong W, Boyle M, Torrance GW (1995a).

Note: Level descriptions are worded here exactly as presented to respondents in the HUI3 preference measurement surveys.

Table 2

Characteristics of Direct and Modelling Survey Respondents: Demographic Characteristics and Additional Life Expectancy

Characteristic	Direct	Modelling	Total
Total Number of Survey Respondents	248	256	504
Age (Years)			
n	248	256	504
Mean	43.9	42.7	43.3
SD	17.1	17.8	17.5
Min	16.0	16.1	16.0
Max	85.2	85.6	85.6
(Note: frequency distributions of 10 year age intervals are presented on next page)			
Expected Years of Remaining Life			
n	248	256	504
Mean	35.9	38.2	37.0
SD	16.9	16.7	16.9
Min	3.0	4.0	3.0
Max	78.0	78.0	78.0
Number people supported by family income			
n	246	255	501
Mean	2.6	2.5	2.6
SD	1.3	1.4	1.4
Min	1.0	1.0	1.0
Max	6.0	9.0	9.0
Interview duration (minutes)			
n	247	256	503
Mean	79.0	76.3	77.6
SD	15.6	18.3	17.0
Min	43.0	43.0	43.0
Max	128.0	165.0	165.0
Median	76.0	73.5	75.0

Table 2 Continued

Characteristic	Direct		Modelling		Total	
	n	(col%)	n	(col%)	n	(col%)
Total Number of Respondents	248		256		504	
Sex						
-female	140	(56.5)	157	(61.3)	297	(58.9)
-male	108	(43.5)	99	(38.7)	207	(41.1)
Age (10 year intervals)						
- 16 to 24 years	35	(14.1)	39	(15.2)	74	(14.7)
- 25 to 34 years	59	(23.8)	67	(26.2)	126	(25.0)
- 35 to 44 years	48	(19.3)	50	(19.5)	98	(19.4)
- 45 to 54 years	37	(14.9)	32	(12.5)	69	(13.7)
- 55 to 64 years	32	(12.9)	29	(11.3)	61	(12.1)
- 65 to 74 years	25	(10.1)	29	(11.3)	54	(10.7)
- 75+ years	12	(4.9)	10	(3.9)	22	(4.4)
Present Health						
-excellent	67	(27.0)	67	(26.2)	134	(26.6)
-very good	102	(41.1)	109	(42.6)	211	(41.9)
-good	54	(21.8)	55	(21.5)	109	(21.6)
-fair	18	(7.3)	22	(8.6)	40	(7.9)
-poor	7	(2.8)	3	(1.2)	10	(2.0)
-refused	0	(0.0)	0	(0.0)	0	(0.0)
-don't know	0	(0.0)	0	(0.0)	0	(0.0)
Marital Status						
-single	62	(25.0)	83	(32.4)	145	(28.8)
-married or common law	142	(57.3)	119	(46.5)	261	(51.8)
-divorced or separated	27	(10.9)	34	(13.3)	61	(12.1)
-widowed	16	(6.5)	20	(7.8)	36	(7.1)
-other	1	(0.4)	0	(0.0)	0	(0.0)
-refused	0	(0.0)	0	(0.0)	0	(0.0)
Know Disabled Person						
-yes	172	(69.4)	215	(84.0)	387	(76.8)
-no	76	(30.6)	41	(16.0)	117	(23.2)
-don't know	0	(0.0)	0	(0.0)	0	(0.0)
-refused	0	(0.0)	0	(0.0)	0	(0.0)

Table 2 Continued

Characteristic	Direct n	Direct (col%)	Modelling n	Modelling (col%)	Total n	Total (col%)
Total Number of Respondents	248		256		504	
Education achieved						
-no formal schooling	0	(0.0)	0	(0.0)	0	(0.0)
-some elementary schooling	3	(1.2)	4	(1.6)	7	(1.4)
-completed elementary schooling	13	(5.2)	12	(4.7)	25	(5.0)
-some secondary schooling	51	(20.6)	51	(19.9)	102	(20.2)
-secondary school graduation certificate	63	(25.4)	54	(21.1)	117	(23.2)
-apprenticeship or journeyman	4	(1.6)	3	(1.2)	7	(1.4)
-non-university certificate or diploma (some or completed)	51	(20.6)	61	(23.8)	112	(22.2)
-some university experience	22	(8.9)	26	(10.2)	48	(9.5)
-bachelor's degree	24	(9.7)	36	(14.1)	60	(11.9)
-degree in medicine, dentistry or veterinary medicine (MD, DDS, DMD, DVM)	2	(0.8)	0	(0.0)	2	(0.4)
-master's degree	9	(3.6)	5	(2.0)	14	(2.8)
-earned doctorate (example, PhD)	4	(1.6)	3	(1.2)	7	(1.4)
-other	1	(0.4)	1	(0.4)	2	(0.4)
-refused	1	(0.4)	0	(0.0)	1	(0.2)
-don't know	0	(0.0)	0	(0.0)	0	(0.0)
Religion						
-Roman Catholic	84	(33.9)	75	(29.3)	159	(31.5)
-United Church	28	(11.3)	36	(14.1)	64	(12.7)
-Anglican	32	(12.9)	37	(14.5)	69	(13.7)
-Presbyterian	14	(5.6)	11	(4.3)	25	(5.0)
-Other Protestant	17	(6.9)	15	(5.9)	32	(6.3)
-Eastern Orthodox	5	(2.0)	1	(0.4)	6	(1.2)
-Jewish	2	(0.8)	1	(0.4)	3	(0.6)
-No religious preference	30	(12.1)	40	(15.6)	70	(13.9)
-other	36	(14.5)	40	(15.6)	76	(15.1)
-refused	0	(0.0)	0	(0.0)	0	(0.0)
-don't know	0	(0.0)	0	(0.0)	0	(0.0)

Table 2 Continued

Characteristic	Direct n	Direct (col%)	Modelling n	Modelling (col%)	Total n	Total (col%)
Total Number of Respondents	248		256		504	
Employment Status						
-employed full-time (including self-employed)	114	(46.0)	111	(43.4)	225	(44.6)
-employed part-time (including self-employed)	29	(11.7)	29	(11.3)	58	(11.5)
-unemployed-looking for work	8	(3.2)	10	(3.9)	18	(3.6)
-unemployed-not looking for work	1	(0.4)	1	(0.4)	2	(0.4)
-unable to work because of health	13	(5.2)	15	(5.9)	28	(5.5)
-retired	30	(12.1)	37	(14.5)	67	(13.3)
-student	21	(8.5)	32	(12.5)	53	(10.5)
-keeping house	24	(9.7)	14	(5.5)	38	(7.5)
-other	8	(3.2)	7	(2.7)	15	(3.0)
Annual Family Income						
- under $5,000	7	(2.8)	4	(1.6)	11	(2.2)
- $5,000 to $9,999	15	(6.0)	14	(5.5)	29	(5.7)
- $10,000 to $14,999	19	(7.7)	25	(9.8)	44	(8.7)
- $15,000 to $19,999	17	(6.9)	17	(6.6)	34	(6.7)
- $20,000 to $29,999	43	(17.3)	32	(12.5)	75	(14.9)
- $30,000 to $39,999	27	(10.9)	36	(14.1)	63	(12.5)
- $40,000 to $49,999	21	(8.5)	35	(13.7)	56	(11.1)
- $50,000 to $59,999	31	(12.5)	18	(7.0)	49	(9.7)
- $60,000 to $69,999	26	(10.5)	22	(8.6)	48	(9.5)
- $70,000 to $79,999	12	(4.8)	13	(5.1)	25	(5.0)
- $80,000 to $89,999	11	(4.4)	16	(6.3)	27	(5.3)
- $90,000 to $99,999	4	(1.6)	2	(0.8)	6	(1.2)
- $100,000 and over	9	(3.6)	10	(3.9)	19	(3.8)
- refused	5	(2.0)	9	(3.5)	14	(2.8)
- don't know	1	(0.4)	2	(0.8)	3	(0.6)
- missing	0	(0.0)	1	(0.4)	1	(0.2)

Table 3

Health Status of Modelling Survey Respondents:
Frequency Distributions (%) of Attribute Levels

	ATTRIBUTE							
Levels	**Vision**	**Hearing**	**Speech**	**Ambulation**	**Dexterity**	**Emotion**	**Cognition**	**Pain**
1	39.8	91.4	92.2	86.3	92.2	71.9	68.4	44.5
2	52.7	6.3	6.6	11.3	7.4	23.0	12.9	38.3
3	3.9	0.0	1.2	2.3	0.0	4.3	17.6	9.0
4	2.3	1.6	0.0	0.0	0.4	0.4	1.2	5.9
5	1.2	0.4	0.0	0.0	0.0	0.0	0.0	1.2
6	0.0	0.4	n/a	0.0	0.0	n/a	0.0	n/a
Missing	0.0	0.0	0.0	0.0	0.0	0.4	0.0	1.2
Total	99.9	100.1	100.0	99.9	100.0	100.0	100.1	100.1

na - not applicable

Note: Totals not equal to 100.0 are due to rounding errors.

Table 4
Person-Mean(A) Preference Scores (Measured and Calculated)
for Attribute Level States, Corner States, Pits and Dead

		Multi-Attribute Pits=0.00 / PH=1.00 Scale			
State	n	10% Trd Mean Val*	EOSBA MA Val	Re-Scaled MA Val	Calculated Utility
Vision					
2	56	95.1	97.3	97.3	0.985
3	56	80.2	80.2	81.7	0.893
4	56	72.2	72.2	73.4	0.841
5	56	59.2	59.2	60.0	0.751
6	56	39.9	39.9	39.9	0.598
Hearing					
2	52	84.1	91.0	91.0	0.949
3	52	75.6	75.6	81.1	0.889
4	52	63.7	63.7	67.1	0.800
5	52	56.7	56.7	58.9	0.744
6	52	44.2	44.2	44.2	0.633
Speech					
2	61	80.1	88.8	88.8	0.936
3	61	72.9	72.9	79.8	0.882
4	61	61.3	61.3	65.2	0.787
5	61	45.7	45.7	45.7	0.645
Ambulation					
2	57	76.7	86.9	86.9	0.924
3	57	67.3	67.3	75.2	0.852
4	57	53.2	53.2	57.4	0.733
5	57	43.8	43.8	45.5	0.644
6	57	36.8	36.8	36.8	0.572
Dexterity					
2	57	83.0	90.5	90.5	0.946
3	57	73.3	73.3	79.2	0.878
4	57	56.3	56.3	59.5	0.748
5	57	44.4	44.4	45.8	0.646
6	57	35.6	35.6	35.6	0.562
Emotion					
2	58	85.1	91.6	91.6	0.952
3	58	70.0	70.0	74.9	0.851
4	58	41.5	41.5	43.2	0.625
5	58	26.0	26.0	26.0	0.471
Cognition					
2	51	79.6	79.6	85.3	0.915
3	51	85.6	91.9	91.9	0.954
4	51	66.1	66.1	70.5	0.822
5	51	38.2	38.2	39.9	0.598
6	51	21.6	21.6	21.6	0.424
Pain					
2	54	87.4	92.9	92.9	0.960
3	53	78.6	78.6	83.2	0.902
4	54	59.5	59.5	62.1	0.766
5	54	34.6	34.6	34.6	0.553
Dead	223	13.3	13.3	13.3	0.323
Pits	223	0.0	0.0	0.0	0.000

Legend: Underlined numbers are anchor value scores of the within-attribute interval subject to EOSBA and subsequent re-scaling and, therefore, these scores do not vary across the rows.

- See Glossary for short form definitions.

Table 5
Person-Mean(B) Preference Scores (Measured and Calculated)
for Attribute Level States, Corner States, Pits and Dead

		Multi-Attribute Dead=0.00/PH=1.00 Scale				Utility on
State	n	10% Trd Mean Val*	EOSBA MA Val	Re-Scaled MA Val	Calculated Utility	Pits=0.00/PH=1.00 Scale
Vision						
2	8	87.3	92.9	92.9	0.966	0.949
3	8	79.6	79.6	83.8	0.920	0.882
4	8	71.4	71.4	74.3	0.869	0.806
5	8	66.5	66.5	68.6	0.836	0.758
6	8	53.7	53.7	53.7	0.745	0.623
Hearing						
2	12	82.2	90.0	90.0	0.951	0.928
3	12	77.0	77.0	83.9	0.920	0.882
4	12	66.9	66.9	72.0	0.856	0.787
5	12	54.4	54.4	57.3	0.768	0.657
6	12	37.5	37.5	37.5	0.628	0.451
Speech						
2	3	96.0	97.8	97.8	0.989	0.984
3	3	93.3	93.3	94.7	0.975	0.963
4	3	92.3	92.3	93.6	0.969	0.954
5	3	83.0	83.0	83.0	0.916	0.875
Ambulation						
2	7	85.9	92.1	92.1	0.962	0.943
3	7	82.0	82.0	87.5	0.939	0.909
4	7	61.1	61.1	62.7	0.802	0.707
5	7	55.9	55.9	56.5	0.763	0.650
6	7	52.6	52.6	52.6	0.738	0.612
Dexterity						
2	7	90.9	94.9	94.9	0.975	0.964
3	7	82.1	82.1	85.4	0.928	0.894
4	7	71.7	71.7	74.1	0.867	0.804
5	7	55.4	55.4	56.4	0.762	0.649
6	7	44.1	44.1	44.1	0.679	0.525
Emotion						
2	6	84.5	91.3	91.3	0.958	0.938
3	6	75.2	75.2	80.8	0.904	0.858
4	6	61.2	61.2	65.0	0.816	0.728
5	6	30.8	30.8	30.8	0.572	0.368
Cognition						
2	13	85.5	85.5	91.7	0.960	0.941
3	13	85.8	92.0	92.0	0.961	0.943
4	13	75.2	75.2	80.2	0.901	0.854
5	13	51.6	51.6	53.9	0.746	0.625
6	13	31.2	31.2	31.2	0.576	0.374
Pain						
2	10	94.4	96.8	96.8	0.985	0.978
3	10	81.0	81.0	82.8	0.914	0.874
4	10	66.6	66.6	67.7	0.831	0.751
5	10	44.4	44.4	44.4	0.681	0.528
Dead	33	0.0	0.0	0.0	0.000	-0.477
Pits	33	9.2	9.2	9.2	0.323	0.000

Legend: Underlined numbers are anchor value scores of the within-attribute interval subject to EOSBA and subsequent re-scaling, and therefore, these scores do not vary across the rows.
- See glossary for short form definitions.

Table 6
Person-Mean Utility Scores on Pits/PH Scale
for Attribute Level States, Corner States, Pits and Dead

Health State	Person-Mean(A) n	Person-Mean(A) Utility	Person-Mean(B′) n	Person-Mean(B′) Utility	Person-Mean Utility
Vision					
2	56	0.985	8	0.949	0.980
3	56	0.893	8	0.882	0.892
4	56	0.841	8	0.806	0.837
5	56	0.751	8	0.758	0.752
6	56	0.598	8	0.623	0.602
Hearing					
2	52	0.949	12	0.928	0.946
3	52	0.889	12	0.882	0.888
4	52	0.800	12	0.787	0.799
5	52	0.744	12	0.658	0.733
6	52	0.633	12	0.451	0.610
Speech					
2	61	0.936	3	0.984	0.942
3	61	0.882	3	0.963	0.892
4	61	0.787	3	0.954	0.809
5	61	0.645	3	0.875	0.675
Ambulation					
2	57	0.924	7	0.943	0.927
3	57	0.852	7	0.909	0.860
4	57	0.733	7	0.707	0.730
5	57	0.644	7	0.650	0.645
6	57	0.572	7	0.612	0.577
Dexterity					
2	57	0.946	7	0.964	0.948
3	57	0.878	7	0.894	0.880
4	57	0.748	7	0.804	0.755
5	57	0.646	7	0.649	0.646
6	57	0.562	7	0.525	0.557
Emotion					
2	58	0.952	6	0.938	0.950
3	58	0.851	6	0.858	0.852
4	58	0.625	6	0.728	0.638
5	58	0.471	6	0.368	0.458
Cognition					
2	51	0.915	13	0.941	0.918
3	51	0.954	13	0.943	0.953
4	51	0.822	13	0.854	0.826
5	51	0.598	13	0.625	0.601
6	51	0.424	13	0.374	0.418
Pain					
2	54	0.960	10	0.978	0.962
3	53	0.902	10	0.874	0.898
4	54	0.766	10	0.751	0.764
5	54	0.553	10	0.528	0.549
Dead	223	0.323	33	-0.477	0.220
Pits	223	0.000	33	0.000	0.000

Legend: n - number of value measurements per state.
Note: the number of value measurements were not used as weights to calculate Person-Mean utility.

Table 7
Measured Value and Utility Scores for Marker and Anchor States, and Fitted Value to Utility Conversion Models: Person-Mean(A) and Person-Mean(B)

Table 7.1 - Group A (n= 223) on Pits=0.00 to Perfect Health=1.00 scale.

Health State	10% Trd Mean Value*	EOSBA MA Value	10% Trd Mean Utility
MA	0.883	0.934	0.806
MB	0.703	0.703	0.674
MC	0.441	0.441	0.564
Dead	0.133	0.133	0.353

Power function alpha = 0.559 (r^2 = 0.760)

Table 7.2 - Group B (n= 23) on Dead=0.00 to Perfect Health=1.00 scale.

Health State	10% Trd Mean Value	EOSBA MA Value	10% Trd Mean Utility
MA	0.874	0.929	0.879
MB	0.707	0.707	0.796
MC	0.504	0.504	0.748
Pits	0.106	0.106	0.346

Power function alpha = 0.474 (r^2 = 0.974)

Legend:

* See glossary for short form definitions.

Note: The 10% trimmed mean value scores reported in this table have been transformed from the 0.00 to 100.00 scale of the feeling thermometer measurement scale to the 0.00 to 1.00 scale (by dividing the value scores by 100), to be commensurate with the scaling of the utility scores and facilitate fitting the value/utility conversion models.

Table 8
Person-Mean Disutility Scores on PH/Pits Scale for Attribute Level States, Corner States, Pits and Dead

State	Weighted Trimmed Mean Utility: Person Mean Utility (Pits/PH Scale)	Person Mean Disutility on Pits=1.00/PH=0.00
Vision		
2	0.980	0.020
3	0.892	0.108
4	0.837	0.163
5	0.752	0.248
6	0.602	0.398
Hearing		
2	0.946	0.054
3	0.888	0.112
4	0.799	0.201
5	0.733	0.267
6	0.610	0.390
Speech		
2	0.942	0.058
3	0.892	0.108
4	0.809	0.191
5	0.675	0.325
Ambulation		
2	0.927	0.073
3	0.860	0.140
4	0.730	0.270
5	0.645	0.355
6	0.577	0.423
Dexterity		
2	0.948	0.052
3	0.880	0.120
4	0.755	0.245
5	0.646	0.354
6	0.557	0.443
Emotion		
2	0.950	0.050
3	0.852	0.148
4	0.638	0.362
5	0.458	0.542
Cognition		
2	0.918	0.082
3	0.953	0.047
4	0.826	0.174
5	0.601	0.399
6	0.418	0.582
Pain		
2	0.962	0.038
3	0.898	0.102
4	0.764	0.236
5	0.549	0.451
Dead	0.220	0.780
Pits	0.000	1.000

Note: The underlined scores represent the disutility scores for each at the 8 corner states (ie. c_j).

Table 9
Person-Mean Single-Attribute Utility and Disutility Functions

State	Person-Mean Utility (Pits/PH scale)	Person-Mean Utility (low/high scale)	Person-Mean Disutility (low/high scale)
Vision			
2	0.980	0.950	0.050
3	0.892	0.728	0.272
4	0.837	0.590	0.410
5	0.752	0.378	0.622
6	0.602	0.000	1.000
Hearing			
2	0.946	0.862	0.138
3	0.888	0.714	0.286
4	0.799	0.484	0.516
5	0.733	0.315	0.685
6	0.610	0.000	1.000
Speech			
2	0.942	0.821	0.179
3	0.892	0.668	0.332
4	0.809	0.412	0.588
5	0.675	0.000	1.000
Ambulation			
2	0.927	0.827	0.173
3	0.860	0.668	0.332
4	0.730	0.360	0.640
5	0.645	0.160	0.840
6	0.577	0.000	1.000
Dexterity			
2	0.948	0.882	0.118
3	0.880	0.729	0.271
4	0.755	0.448	0.552
5	0.646	0.202	0.798
6	0.557	0.000	1.000
Emotion			
2	0.950	0.908	0.092
3	0.852	0.726	0.274
4	0.638	0.333	0.667
5	0.458	0.000	1.000
Cognition			
2	0.918	0.860	0.140
3	0.953	0.919	0.081
4	0.826	0.702	0.298
5	0.601	0.315	0.685
6	0.418	0.000	1.000
Pain			
2	0.962	0.916	0.084
3	0.898	0.774	0.226
4	0.764	0.476	0.524
5	0.549	0.000	1.000

Note: Underlined scores are scale anchor scores, fixed by definition.

Table 10
Multi-Attribute Disutility Function: Standard Format on Pits/PH Scale

Vision		Hearing		Speech		Ambulation		Dexterity		Emotion		Cognition		Pain	
x_1	$\tilde{u}_1$	x_2	$\tilde{u}_2$	x_3	$\tilde{u}_3$	x_4	$\tilde{u}_4$	x_5	$\tilde{u}_5$	x_6	$\tilde{u}_6$	x_7	$\tilde{u}_7$	x_8	$\tilde{u}_8$
1	0.00	1	0.00	1	0.00	1	0.00	1	0.00	1	0.00	1	0.00	1	0.00
2	0.05	2	0.14	2	0.18	2	0.17	2	0.12	2	0.09	2	0.14	2	0.08
3	0.27	3	0.29	3	0.33	3	0.33	3	0.27	3	0.27	3	0.08	3	0.23
4	0.41	4	0.52	4	0.59	4	0.64	4	0.55	4	0.67	4	0.30	4	0.52
5	0.62	5	0.69	5	1.00	5	0.84	5	0.80	5	1.00	5	0.69	5	1.00
6	1.00	6	1.00	6	n/a	6	1.00	6	1.00	6	n/a	6	1.00	6	n/a

Parameter estimates for multi-attribute disutility function

c =	-0.991						
c1 =	0.40	c2 =	0.39	c3 =	0.33	c4 =	0.42
c5 =	0.44	c6 =	0.54	c7 =	0.58	c8 =	0.45

Formula (Pits/ PH Scale)

MADUF*:

$$\tilde{u} = [1/c]\left[\prod_{j=1}^{8}(1 + c * c_j * \tilde{u}_j) - 1\right]$$

$$\tilde{u} = [1/(-0.991)] *[[1+(-0.991) * 0.40 * \tilde{u}_1] * [1+(-0.991) * 0.39 * \tilde{u}_2] * [1+(-0.991) * 0.33 * \tilde{u}_3] * [1+(-0.991) * 0.42 * \tilde{u}_4] * [1+(-0.991) * 0.44 * \tilde{u}_5] * [1+(-0.991) * 0.54 * \tilde{u}_6] * [1+(-0.991) * 0.58 * \tilde{u}_7] * [1+(-0.991) * 0.45 * \tilde{u}_8] - 1]$$

MAUF:

$u = 1 - \tilde{u}$

Conversion to Dead/PH Scale

$$\tilde{u}^* = \tilde{u} / \tilde{u}_{Dead} = \tilde{u} / (1 - 0.264) = \tilde{u} / 0.736$$

$$u^* = 1 - \tilde{u}^* = 1 - (1-u) / 0.736$$

* See glossary for short form definitions.

Table 11
Multi-Attribute Utility Function:
Simplified Format on Dead/PH Scale

Vision		**Hearing**		**Speech**		**Ambulation**		**Dexterity**		**Emotion**		**Cognition**		**Pain**	
x_1	b_1	x_2	b_2	x_3	b_3	x_4	b_4	x_5	b_5	x_6	b_6	x_7	b_7	x_8	b_8
1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
2	0.98	2	0.95	2	0.94	2	0.93	2	0.95	2	0.95	2	0.92	2	0.96
3	0.89	3	0.89	3	0.89	3	0.86	3	0.88	3	0.85	3	0.95	3	0.90
4	0.84	4	0.80	4	0.81	4	0.73	4	0.76	4	0.64	4	0.83	4	0.77
5	0.75	5	0.74	5	0.68	5	0.65	5	0.65	5	0.46	5	0.60	5	0.55
6	0.61	6	0.61	6	n/a	6	0.58	6	0.56	6	n/a	6	0.42	6	n/a

Formula (Dead/PH Scale)

$$u^* = 1.371 (b_1 * b_2 * b_3 * b_4 * b_5 * b_6 * b_7 * b_8) - 0.371$$

where u^* is the utility of a chronic health state[1] on the utility scale where dead[2] has a utility of 0.00, and healthy[1] has a utility of 1.00.

Notes:

1. chronic states, and healthy state, are here defined as lasting for a lifetime.
2. dead is defined as immediate.

Table 12
Intra-Survey Agreement: HUI3, HUI2 and HUI1.

Health State	MAUF* Score	SG Score	HUI3 MAUF-SG Difference	HUI2 MAUF-SG Difference	HUI1 MAUF-MEASURED Difference
HUI3					
MA	0.88	0.81	+0.07		
MB	0.72	0.68	+0.04	n/a	n/a
MC	0.49	0.57	-0.08		
HUI2					
M3/5	0.84	0.78		+0.06	
F3/3	0.89	0.88	n/a	+0.01	n/a
I1	0.68	0.76		-0.08	
I3	0.51	0.51		0.00	
HUI1					
P2R2H4	0.69	0.67	n/a	n/a	+0.02
P5R2H5	0.28	0.31			-0.03
# of states			3	4	2
sum			+0.03	-0.01	-0.01
mean difference			+0.010	-0.003	-0.005
mean absolute difference			0.067	0.038	0.025
overall standard deviation			0.084	0.058	0.036

Sources:

HUI2 predicted (MAUF) and directly measured utility scores from Table 10 of Torrance et al. 1996a.

HUI1 predicted (MAUF) scores from formula on page 121 of Drummond et al. 1987 and directly measured utility scores from page 1060 of Torrance et al. 1982.

- See glossary for short form definitions.

Table 13
Intra-Survey Agreement ANOVA Table

Source	df	Sums of Squares(SS)	Mean SS	F	P-value
States	2	0.09949	0.04974	14.27	0.065
Method	1	0.00017	0.00017	0.05	0.844
Residual	2	0.00697	0.00349		
Total	5	0.10664			

Table 14
Agreement Between Calculated and Directly Measured Utility Scores: External and Internal Assessments

Table 14A **HUI3* MAUF Scores versus HUI3 Validation Survey Scores (or External Agreement)**

Survey	States	MD	MAD	Overall SD	ICC (95% CI)
HUI3	73 (not weighted by GP prevalence)	- 0.008	0.087	0.1032	0.88 (0.49, 0.92)
HUI3	73 (weighted by GP prevalence, excluding PH)	+0.001	0.002	0.0061	
HUI3	74 (weighted by GP prevalence, including PH)	+0.001	0.001	0.0040	

Table 14B **Multiplicative MAUF Scores versus Modelling Survey Scores (or Internal Agreement)**

Survey	States	MD	MAD	Overall SD	ICC (95% CI)
HUI3	3 (markers: MA,MB,MC)	+0.010	0.067	0.084	0.91 (0.00, 1.00)
HUI2	4 (M3/5, F3/3,I1,I3)	- 0.003	0.038	0.058	0.95 (0.74, 1.00)

Legend:
CI = confidence interval
GP = general population (from 1991 General Social Survey)
ICC = intra-class correlation coefficient
MAD = mean absolute difference
$= [\sum(|\text{predicted} - 10\% \text{ trimmed mean}|)/n]$
MD = mean difference
$= [\sum(\text{predicted} - 10\% \text{ trimmed mean})/n]$
Overall SD = overall standard deviation
$= \sqrt{[(\sum(\text{predicted} - 10\% \text{ trimmed mean})^2)/(n-1)]}$

- See glossary for short form definitions.

9.0 FIGURES

Figure 1
Two-Sided Feeling Thermometer

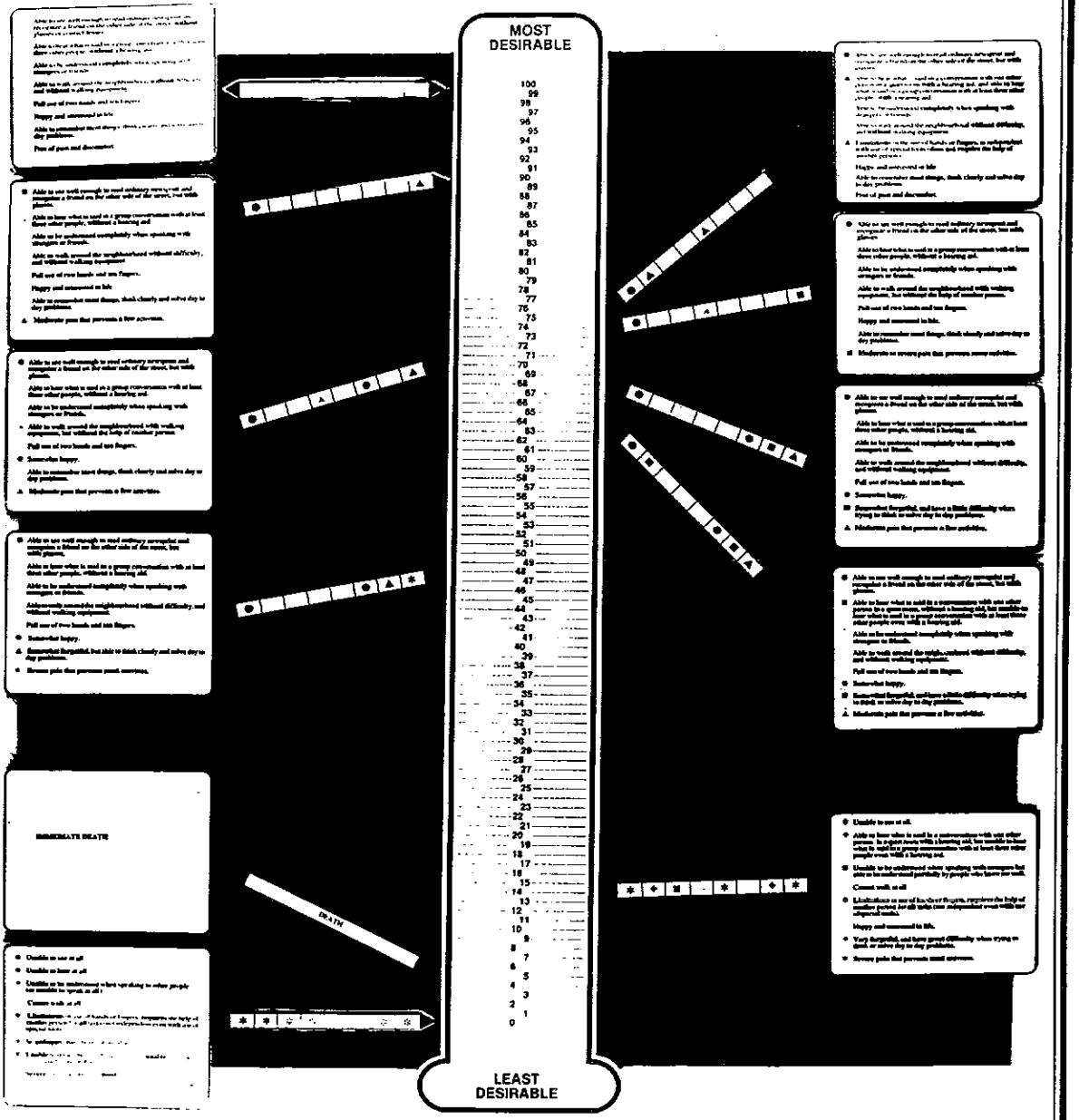


Figure 2
Flip-Card Chance Board

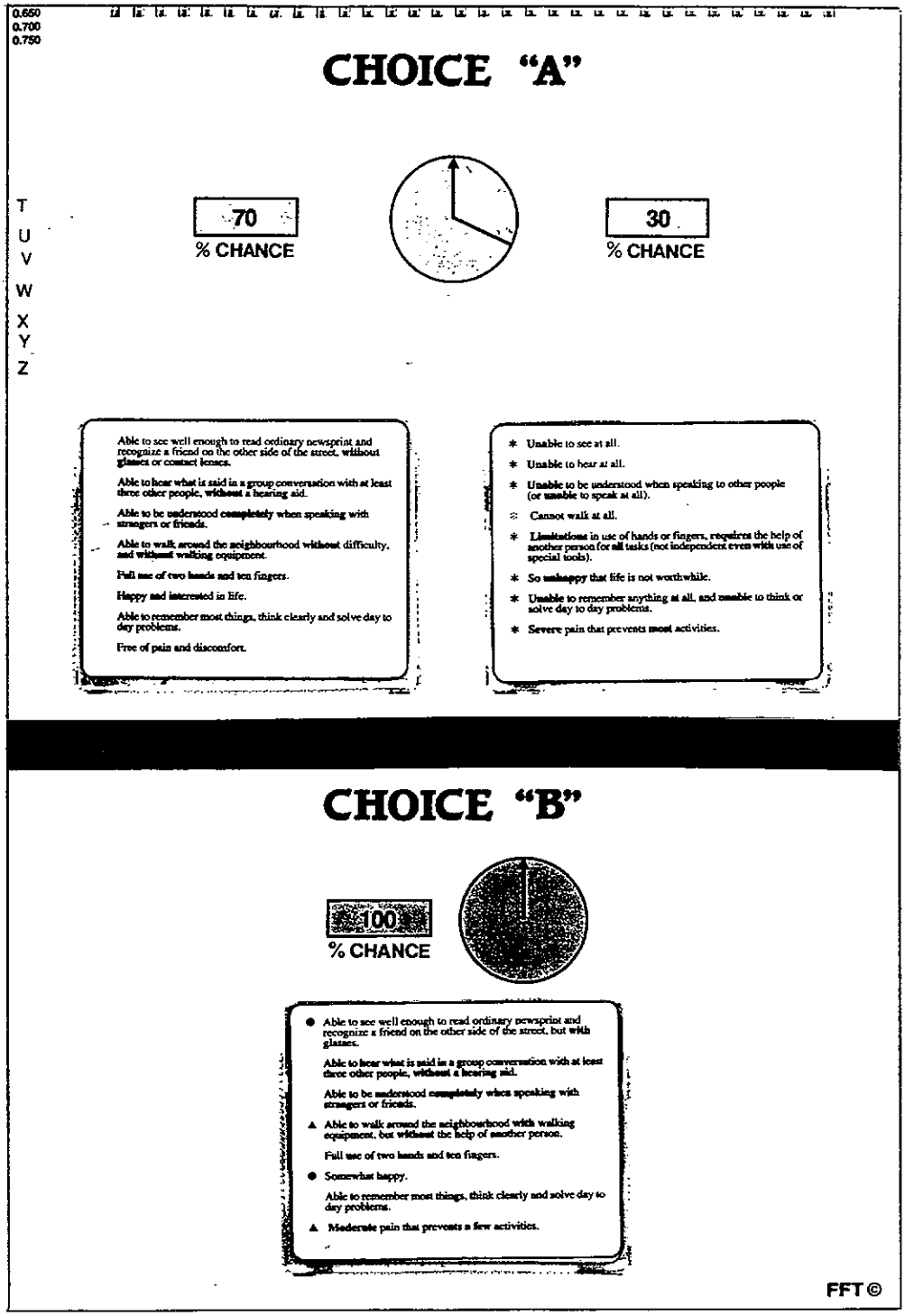
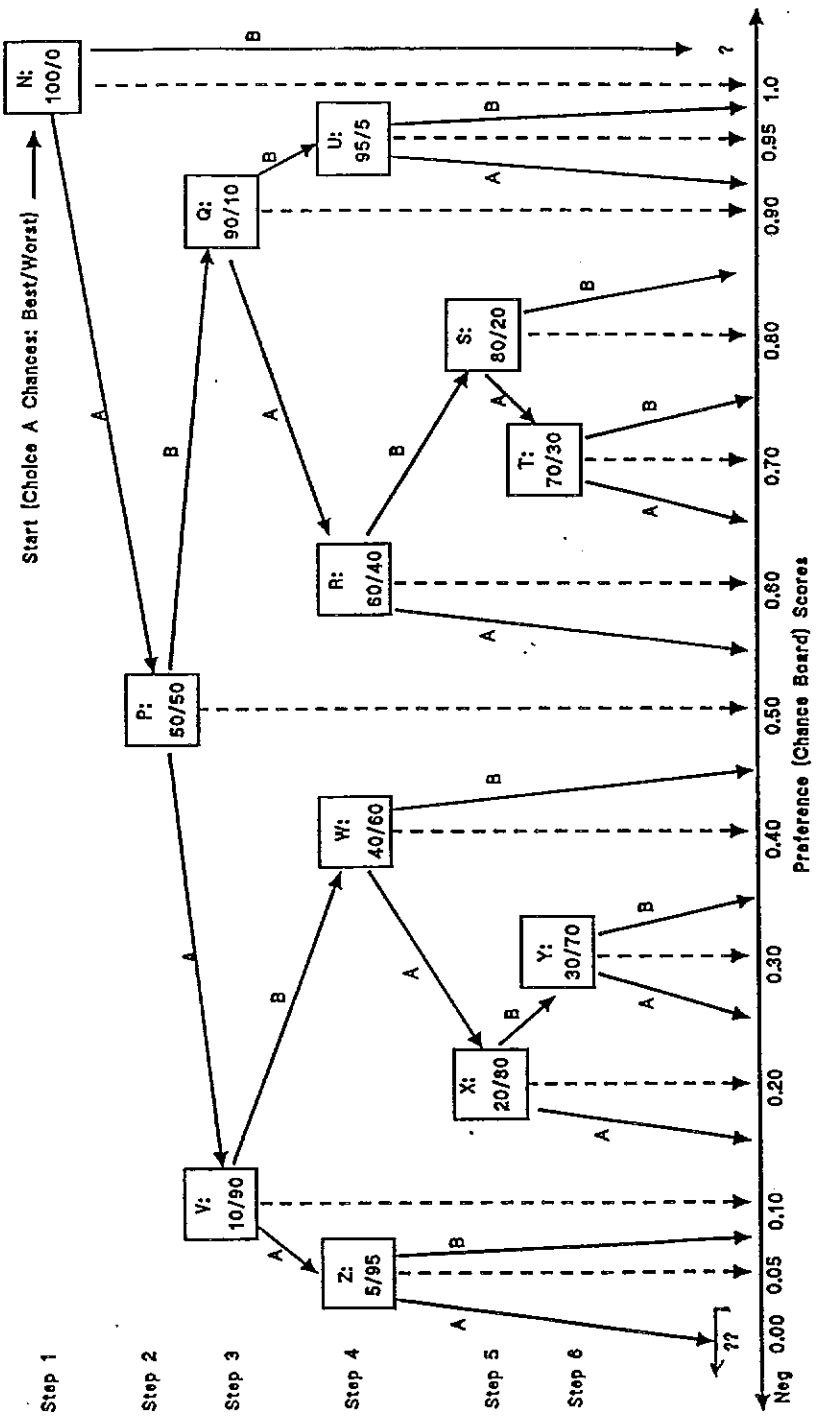


Figure 3

Schematic of Steps in Determining Preference Scores from Chance Board



Notes: Dotted lines indicate scores for indifference to Choices A and B for each decision node.
Letters in boxes represent Chance Board Flip Cards.

Figure 4
HUI3-M Survey Interview Strategy

Interview Section		Number of Measurements For Section	Number of Measurements Per State
I.	Introduction		
II.	Respondent HUI3 health status & estimated additional life expectancy	256	
III.	Feeling Thermometer Value Measures		
	On left-hand side (remain on board):		
	1) 3 anchor states (Pits, Dead, PH).	256	256
	2) 3 marker states.	256	256
	On right-hand side (removed after each step):		
	3) first set of randomly selected single-attribute states (worst level first, intermediate levels in random order).	256	64
	4) second set of randomly selected single-attribute states (worst level first, intermediate levels in random order).	256	64
	5) first 4 block states selected at random from block "x" (block "x" randomly selected 8-state block from 16 block fractional factorial plan 2.8.8).	256	16
	6) second 4 block states selected at random from block "x".	256	16
IV.	Standard Gamble:		
	7) 3 marker states (in random order) each measured on Pits/PH scale.	256	256
	8) Dead measured on Pits/PH scale or Pits measured on Dead/PH scale.	256	256
V.	Respondent Demographics (disabilities in family or friends, global rating of current health, marital status, date of birth, educational level, religion, employment status, income).	256	
VI.	Respondent Evaluation of Interview.	256	
VII.	Interviewer Evaluation of Interview.	256	

Uses of HUI3-M preference measurement data:

A. Fitting of value to utility conversion model.
B. Fitting of MAUT multiplicative multi-attribute utility function.
C. Fitting of statistical multi-linear utility functions (topic of a future report).
D. Assessment of intra-survey validity of multi-attribute utility functions.

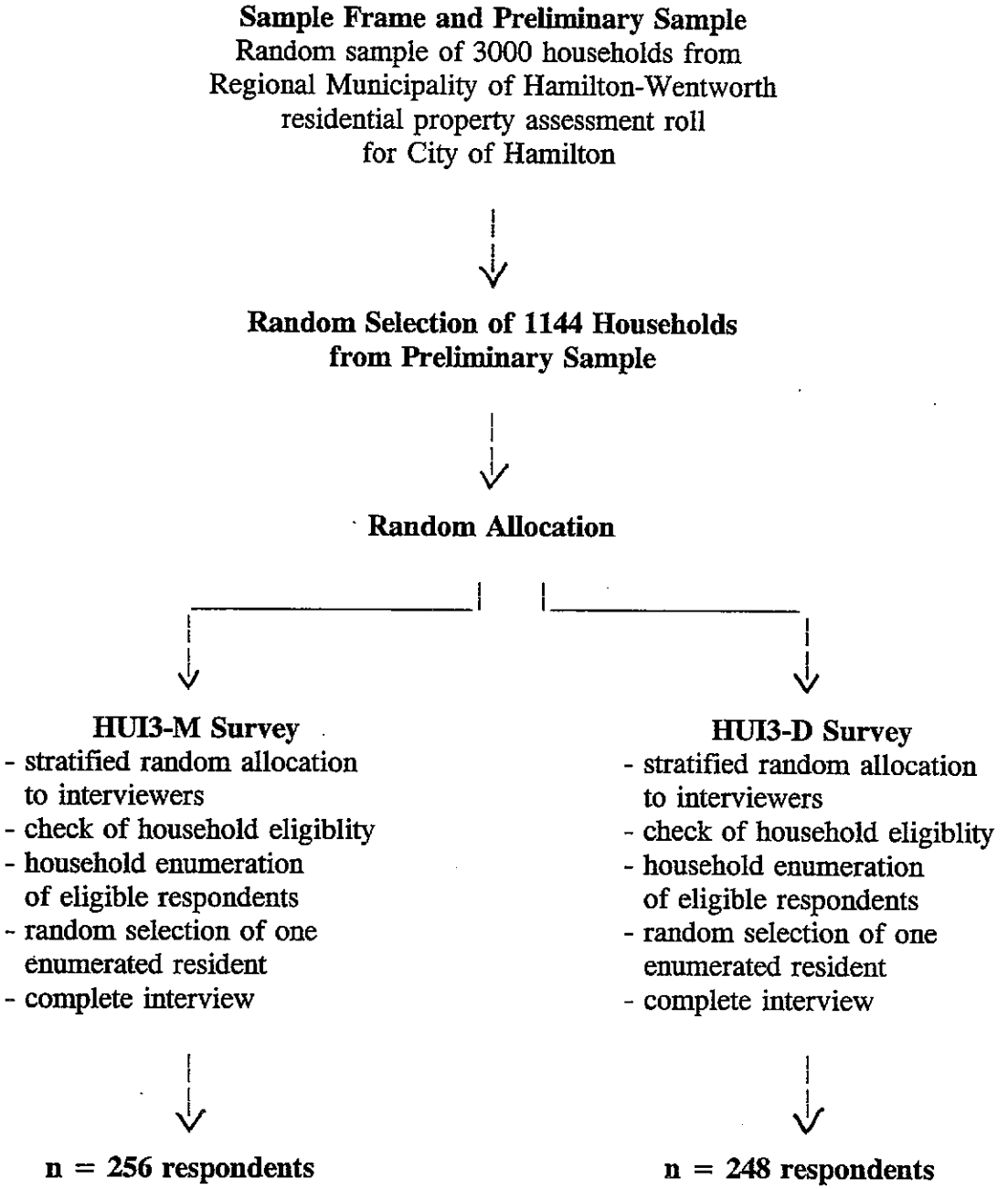
Figure 5
HUI3-D Survey Interview Strategy

Interview Section		Number of Measurements For Section	Number of Measurements Per State
I.	Introduction		
II.	Respondent HUI3 health status & estimated additional life expectancy	248	
III.	Feeling Thermometer Value Measures		
	On left-hand side (remain on board):		
	1) 3 anchor states (Pits, Dead, PH).	248	248
	2) 3 marker states.	248	248
	On right-hand side (steps #3 and #4 in random order, cards removed after each step):		
	3) set of 5 most prevalent ("z") health states (presented in random order).	248	248
	4) set of 5 other prevalent health states (randomly selected from 13 sets of states and presented in random order).	248	19
IV.	Standard Gamble:		
	5) 3 marker states (in random order) each measured on Pits/PH scale.	248	248
	6) set of 5 most prevalent ("z") health states or set of 5 "other" prevalent health states used in step #4 above (presented in random order).	75 172	75 13
	7) Dead measured on Pits/PH scale or Pits measured on Dead/PH scale.	248	248
V.	Respondent Demographics (disabilities in family or friends, global rating of current health, marital status, date of birth, educational level, religion, employment status, income).	248	
VI.	Respondent Evaluation of Interview.	248	
VII.	Interviewer Evaluation of Interview.	248	

Use of HUI3-D preference measurement data:

A. Mean standard gamble measures of states were used to assess the validity of scores calculated from multi-attribute utility functions fitted from data collected in the HUI3-M survey, using agreement statistics of data weighted and unweighted by the prevalence of HUI3-D health states in the general population.

Figure 6
HUI3 Preference Study Sampling Schematic



Note: Details of eligiblity and contact rates are presented in section 3.4.11 of the text and in Figure 9.

Figure 7
Schematic of Analytical Steps for Fitting HUI3
Multiplicative Multi-Attribute Utility Function

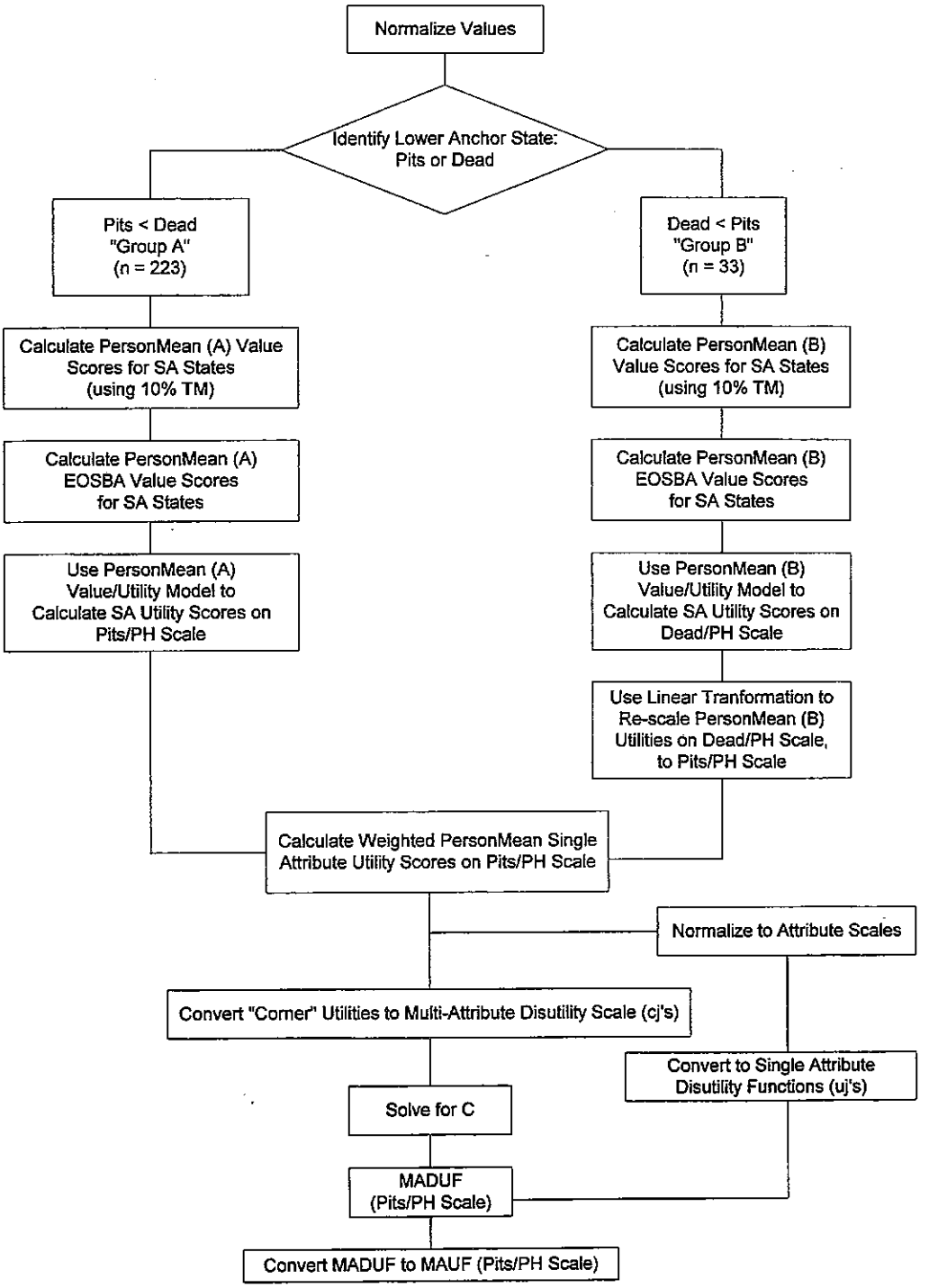


Figure 8
Schematic of Analytical Steps for Fitting Person-Mean(A) and Person-Mean(B) Value to Utility Conversion Models

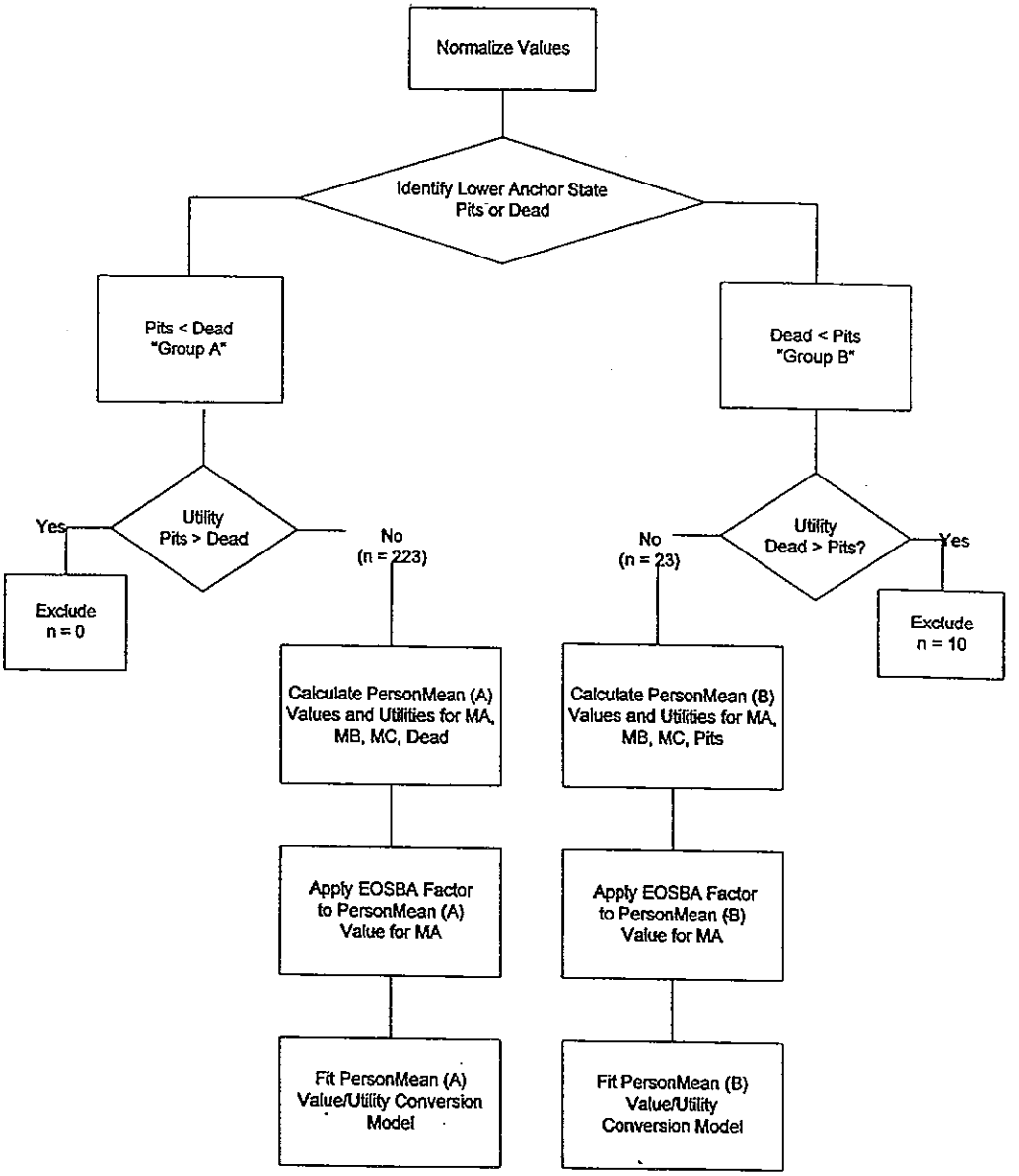
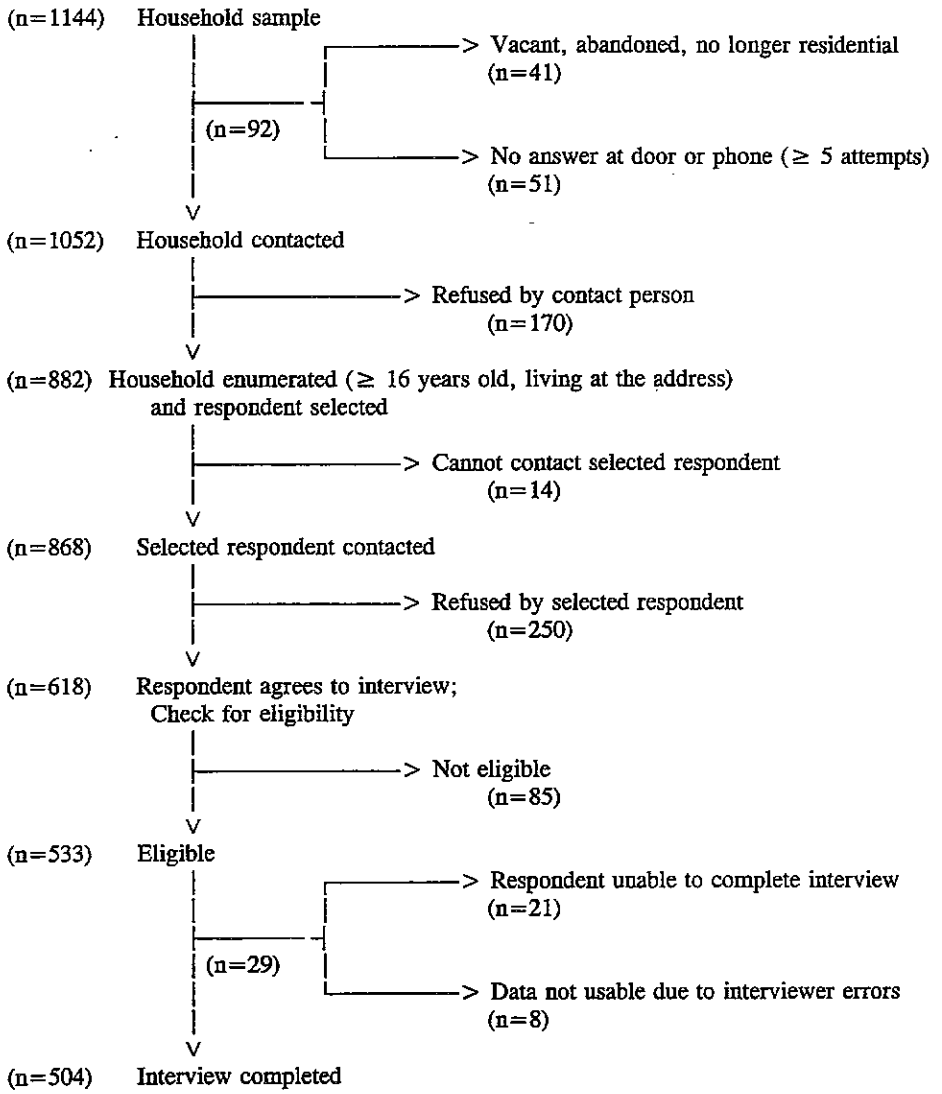


Figure 9

Final Interviewing Status Report Overview Diagram



interviewed / households approached, 504/1144, or 44.1%
interviewed / contacted and presumed eligible, 504/1052, or 47.9%
interviewed / eligible, contacted subjects in enumerated households, 504/(882-14-92) or 64.9%

10.0 APPENDICES

APPENDIX A

Advice for Applications of HUI3 Utility Functions and for Estimating Multi-Attribute Utility Functions

Analysts wishing to use the HUI3 in their own studies should note that it is a four step process. First, relevant health status information must be collected for each subject. Second, the health status information is used to identify the most appropriate level on each of the eight attributes for each subject. Third, the eight attribute levels are used in conjunction with the single- and multi-attribute utility functions to calculate morbidity and health-related quality of life scores for each subject. Finally, statistical techniques are used to describe the health status and health-related quality of life of groups of subjects (ie., summary statistics) or to assess differences among groups. The contents of the last step is generally very dependent on the objectives and design of each specific study. The first three steps are facilitated by use of standard questionnaires and procedures manuals (see third and fourth paragraphs of Conclusion).

A worked example, for three subjects in a clinical study, is provided below to help the uninitiated reader better understand how to apply the HUI3. Let us assume that we have collected information that has enabled us to determine the appropriate levels of the 8 HUI3 attributes for each patient as follows:

Patient	**Attribute Levels**							
ID	**Vision**	**Hearing**	**Speech**	**Ambulation**	**Dexterity**	**Emotion**	**Cognition**	**Pain**
1	2	1	1	1	1	2	3	2
2	1	2	1	2	1	1	1	3
3	6	4	3	4	5	4	5	4

The single-attribute utility scores of morbidity from Table 9 and the multi-attribute utility score of comprehensive health-related quality of life calculated from the HUI3 MAUF in Table 11 for each patient are as follows:

Patient ID	Single-Attribute Utility Scores								Multi-Attribute Score
	Vision	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain	
1	0.95	1.00	1.00	1.00	1.00	0.91	0.92	0.92	0.79
2	1.00	0.86	1.00	0.83	1.00	1.00	1.00	0.77	0.72
3	0.00	0.48	0.67	0.36	0.20	0.33	0.32	0.48	-0.29

Many standard questionnaires and procedures manuals are available from the HUI Service Centre. The questionnaires were designed to collect sufficient information to describe the health status of subjects according to both HUI2 and HUI3 classification systems. Questionnaires are available in formats for self-complete or interviewer-administration, for self-reporting or proxy-reporting respondents, and for various recall assessment periods (eg., past one week or past two weeks or past 4 weeks or "usual health status"). The HUI Service Centre operates on a fee-for-service basis. It is recommended that users ensure that project budgets include the cost of HUI Service Centre support. Additional information is available upon request (contact William Furlong, email furlongb@mcmaster.ca). It should be noted, that standardized and well-documented algorithms for coding attribute levels from questionnaire response combinations are not readily available for many questionnaires (eg., 1990 Ontario Health Survey, 1991 Canadian General Social Survey, various versions of the 1994 and ongoing Canadian National Population Health Survey questionnaires).

Other researchers may be interested in replicating the design or methods used to develop the HUI3 MAUF. We would encourage this type of activity. It should be noted, however, that to undertake such studies in a rigorous manner requires a great deal of time and effort. Most key members of our team have been involved in developing the HUI family or closely related activities for more than 15 years (including some of the interviewers), more than two dozen individuals contributed to the data management or analyses of the study, and development of HUI3 was initiated almost a decade ago. Replication of the design and implementation of this study would also require detailed knowledge of the concepts and survey procedures, and a group of very highly trained professional interviewers. (We caution others it is very easy to under-budget the time and resources required to achieve credible results.)

APPENDIX B

Health Utilities Index Mark 3 (HUI3) Multiplicative Multi-Attribute and Single-Attribute Utility Functions for Scoring Applications

Furlong W, Feeny D, Torrance GW, Goldsmith C, DePauw S, Zhu Z, Denton M, Boyle M.
Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3
(HUI3) System: A Technical Report
McMaster University Centre for Health Economics and Policy Analysis Working Paper 98-11, 1998.

Introduction

Three McMaster Health Utilities Index systems have been developed: Health Utilities Index Mark 1 (HUI1); Health Utilities Index Mark 2 (HUI2); and Health Utilities Index Mark 3 (HUI3). This document describes the multiplicative multi-attribute and single-attribute utility functions for HUI3.

Attribute levels and comprehensive health states, defined by the HUI3 health status classification system, are categorical variables and are useful for describing the health status of individuals. The comprehensive health state of an individual is defined as the combination of one level from each of the 8 attributes in the HUI3 system.

It is important to note that attribute level codes represent functional classes within each attribute and do not have interval scale properties. Utility scores have interval scale properties. Scores having interval scale properties allow for the use of powerful statistical methods (eg., parametric procedures) for making comparisons of HRQL and functional capacities between groups of subjects, or to assess changes in HRQL and functional capacities within individuals and groups.

The multiplicative multi-attribute utility function (see next page) facilitates the calculation of health-related quality of life (HRQL) scores on the conventional Dead=0.00 to Perfect Health=1.00 scale, for comprehensive HUI3 health states described by 8-element vectors. The single-attribute functions (see page following next) present utility scores of functional capacity on 8 scales, one scale for each of the 8 attributes. Each single-attribute scale is defined on a scale from 0.00 to 1.00, such that lack of functional capacity in an attribute (lowest level for that attribute) has a single-attribute utility score of 0.00 and full function (level 1) for an attribute has a single-attribute score of 1.00.

This appendix of the full technical report presents the HUI3 utility scoring systems in a concise section for use by data managers and analysts interested only in applying the HUI3 preference scoring functions. The full technical report was prepared as a reference document for details about the development of the multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) health status classification system. The details of the study design and results will be important for readers interested in the development process, assessing the validity of the scoring functions or replicating the study.

APPENDIX B (cont'd)

of Furlong W, Feeny D, Torrance GW, Goldsmith C, DePauw S, Zhu Z, Denton M, Boyle M. Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report. McMaster University Centre for Health Economics and Policy Analysis Working Paper 98-11, 1998.

1. Multiplicative multi-attribute utility scores on the Dead/Perfect Health scale

The HUI3 multi-attribute utility score for a health state is calculated according to the following, for the Dead/Perfect Health scale:

$$u^* = 1.371\ (b_1 \times b_2 \times b_3 \times b_4 \times b_5 \times b_6 \times b_7 \times b_8) - 0.371$$

where u^* is the utility of a chronic health state on the utility scale where dead has a utility of 0.00, and Perfect Health has a utility of 1.00. The b_j's are substituted from Table 1 for the appropriate attribute and level (x_i).

Table 1

Vision		Hearing		Speech		Ambulation		Dexterity		Emotion		Cognition		Pain	
x_1	b_1	x_2	b_2	x_3	b_3	x_4	b_4	x_5	b_5	x_6	b_6	x_7	b_7	x_8	b_8
1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
2	0.98	2	0.95	2	0.94	2	0.93	2	0.95	2	0.95	2	0.92	2	0.96
3	0.89	3	0.89	3	0.89	3	0.86	3	0.88	3	0.85	3	0.95	3	0.90
4	0.84	4	0.80	4	0.81	4	0.73	4	0.76	4	0.64	4	0.83	4	0.77
5	0.75	5	0.74	5	0.68	5	0.65	5	0.65	5	0.46	5	0.60	5	0.55
6	0.61	6	0.61	-	-	6	0.58	6	0.56	-	-	6	0.42	-	-

Example calculation:

A patient reports their health status as follows:

	Vision	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain
Level	2	1	1	2	1	2	1	3

Referring to the table above, substitute the appropriate scores for b_j for each attribute as follows:

$$u^* = 1.371\ (0.98 \times 1.00 \times 1.00 \times 0.93 \times 1.00 \times 0.95 \times 1.00 \times 0.90) - 0.371$$

$$= 0.70$$

The utility score for this individual's health state is 0.70 on the Dead/Perfect Health scale.

APPENDIX B (cont'd)

of Furlong W, Feeny D, Torrance GW, Goldsmith C, DePauw S, Zhu Z, Denton M, Boyle M. Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report. McMaster University Centre for Health Economics and Policy Analysis Working Paper 98-11, 1998.

2. Single-attribute utility functions for HUI3

Single-attribute utility functions consist of eight sets of utility scores defining the relative desirability for levels of function within each attribute of the HUI3. Table 2 shows the HUI3 single-attribute utility scores (y_i) for all levels of the eight attributes (x_i).

Table 2: HUI3 single attribute utility scores

Vision		Hearing		Speech		Ambulation		Dexterity		Emotion		Cognition		Pain	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8
1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
2	0.95	2	0.86	2	0.82	2	0.83	2	0.88	2	0.91	2	0.86	2	0.92
3	0.73	3	0.71	3	0.67	3	0.67	3	0.73	3	0.73	3	0.92	3	0.77
4	0.59	4	0.48	4	0.41	4	0.36	4	0.45	4	0.33	4	0.70	4	0.48
5	0.38	5	0.32	5	0.00	5	0.16	5	0.20	5	0.00	5	0.32	5	0.00
6	0.00	6	0.00	-	-	6	0.00	6	0.00	-	-	6	0.00	-	-

For example, the single-attribute utility score for speech, level 4, is 0.41.