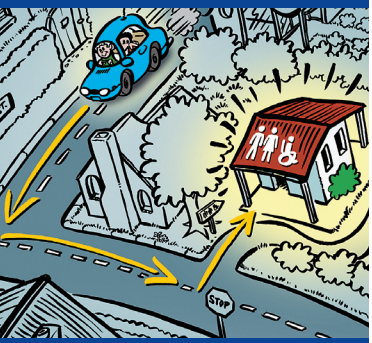
An Australian Government Initiative

Measuring Incontinence in Australia

2006



www.toiletmap.gov.au

A PROJECT OF THE NATIONAL CONTINENCE MANAGEMENT STRATEGY:
AN AUSTRALIAN GOVERNMENT INITIATIVE

Measuring Incontinence in Australia

2006

By
A/Professor Graeme Hawthorne

The Department of Psychiatry
The University of Melbourne

ISBN: 0 642 82980 2

Online ISBN: 0 642 82981 0

Publications Approval Number: 3864

Contents

List of Tables

List of Figures

Reading this Report

This report covers three aspects of incontinence in Australia. It provides population prevalences (section 3), estimates of the impact of incontinence of peoples' lives (section 4), and an estimate of the effect of incontinence on peoples' health using the SF-36 Version 2 (section 5).

Because there are epistemiological issues behind each of these aspects, each part of the report contains detailed material that examines the basis on which the findings rest. Although much of this material is technical in nature, to appreciate the uncertainties of the findings it is important for readers to have a basic awareness of the limitations implicit in the procedures used throughout the report. Wherever possible readers should try to read the full report.

However, it is recognised that many readers will have neither the time nor technical expertise to assess all the issues raised in the report. The following suggestions are offered as a guide to the different sections in the report and which sections may be of interest to different readers.

- Readers who want a quick overview should read the executive summary.
- Readers who are interested in incontinence prevalence only should read section 3.
- For those with an interest in the impact of incontinence on peoples' lives, they should read section 4. For readers who have no background or understanding of utility theory, it is strongly suggested that they read Appendix A before reading section 4 as this provides a more detailed introduction to utility theory and how quality of life within a utility paradigm is measured.
- Readers with an interest in the impact of incontinence on health status should read section 5. For those who have an interest in the SF-36 Version 2 and how it may be scored in an Australian context, this section presents some evidence suggesting that Australians interpret health differently to Americans; therefore both the published US weights and equivalent Australian weights for the SF-36 Version 2 are presented.
- The summary conclusion draws together the key study findings and recommendations.

Executive Summary

Incontinence is a common health problem affecting over 2 million Australians. As a major public health initiative, the Commonwealth Government resourced the National Continence Management Strategy (NCMS) to improve continence treatment and management so that more Australians can live and participate in their communities with dignity and confidence.

This study is part of this project. It provides prevalence estimates, examines the psychometric properties of the instruments used to assess incontinence and quality of life (including the impact of incontinence on quality of life), and it presents Australian population norms for the SF36 Version 2.

Method

As part of the NCMS the Commonwealth Department of Health and Ageing funded a special version of the South Australian Health Omnibus Survey (SAHOS) in 2004. The SAHOS is a population-based user-pays health survey (2). In brief, the 2004 survey involved interview with sampled households throughout South Australia. The total number of participants interviewed was 3015, giving a within scope response rate of 72%. The obtained data were weighted by Australian Bureau of Statistics population estimates to achieve representativeness.

Urinary incontinence was measured by the Incontinence Severity Index (ISI) and the Urogenital Distress Inventory – Short Form (UDI-6). Faecal incontinence was assessed by the Wexner Continence Grading Scale (Wexner). Soiling was measured by two additional questions.

Quality of life was assessed by utility, which is the value of quality of life to a person. Utility scales use 1.00 to represent the best possible quality of life, and 0.00 represents death-equivalent states. The utility scales used in this study were the Assessment of Quality of Life (AQoL), the EQ5D, the Health Utilities Index – 3 (HUI3), the 15D and the SF6D (derived from the SF36). The psychometric properties of each of these instruments was assessed, along with their sensitivity to incontinence.

Health status was assessed with the SF36V2 (SF-36 Version 2). Australian norms are provided, as are estimates of the association between incontinence and health status.

Results

Incontinence Prevalence

For urinary and faecal incontinence, the best estimate based on the ISI and Wexner measures was that the prevalence of any incontinence is 27% (95%CI: 26% – 29%). For females it is 40% (38% – 43%) and for males 14% (12% – 15%).

Based on self-report of any symptoms of urinary leakage, the ISI estimated prevalence of urinary incontinence was 24% (95%CI: 23% – 26%) overall. When broken down by gender, it was 38% (95%CI: 36% – 41%) for females and 10% (95%CI: 9% – 12%) for males. When measured by the UDI-6, which measures being bothered by symptoms, the overall prevalence of urinary incontinence was 47% (95%CI: 45% – 48%); for females it was 60% (95%CI: 58% – 63%) and for males 33% (30% – 35%). These estimates for the UDI-6 are confounded due to its poor psychometric properties; thus the ISI estimates are preferred.

For faecal incontinence, the standard Wexner Scale data suggested that the prevalence was 35% (95%CI: 33% – 36%). For females this was 38% (95%CI: 35% – 40%), and for males it was 32% (95%CI: 29% – 34%). However, the Wexner includes flatus, which is excluded from the current International Continence Society faecal incontinence definition. If the flatus question is excluded from the Wexner, the data show that the prevalence would be 8% (95%CI: 7% – 9%). For females this would be 10% (95%CI: 8% – 11%) and 6% (5% – 7%) for males. In the interests of consistency with international definitions, these modified prevalence estimates are preferred.

The Utility of Incontinence

The psychometric properties of the five utility instruments were examined using a combination of classic, modern and econometric test theory. The results suggested that there were particular measurement difficulties with the 15D, because it is not weighted with a preference-based technique, it uses an additive scale which prevents loss of utility for severe health states, and the data from respondents was found to provide a poor fit to the 15D utility model. There were also measurement difficulties with the SF6D due to the restricted scoring range. The lower boundary for the SF6D is 0.30, which implies that while scores are well reported for those with 'healthy' conditions, for those with severe health conditions there is an ever-increasing gap between the theoretical utility model (score range 0.00 to 1.00) and obtained scores. For the EQ5D two measurement problems were observed. Examination of its internal structure suggested that the 5 items were measuring two different constructs which led to difficulties with the underlying measurement model. A second issue concerned the obtained data distribution: the scores were 'lumpy' and clustered around certain values. This lumpiness is caused by the presence of an additional weight that comes into effect whenever a person endorses the worst health state level on any EQ5D item. The effect of this additional weight is to cause an increase/decrease of utility between 0.1 and 0.3. The impact of this additional EQ5D weight is to confer increased sensitivity on the EQ5D whenever a respondent moves from a level-3 endorsement to a level-2 endorsement. It also, however, has the effect of undermining the necessary interval property needed for use during cost-utility analysis.

The two better performing instruments were the AQoL and HUI3. Both possessed good psychometric properties, with the AQoL performing slightly better (e.g. it was the more reliable of the two and had the better data to model fit indices). No particular problems were identified for either of these two measures.

Population norms for all five measures were computed. For the AQoL the mean utility was 0.81 (SD = 0.20), for the EQ5D it was 0.82 (0.22), for the HUI3 it was 0.82 (0.21), for the 15D it was 0.93 (0.08) and for the SF6D it was 0.81 (0.14).

When the utility measures were examined by incontinence status, the data showed that incontinence has a small to mild effect upon quality of life. The range in disutility (i.e. loss of quality of life) for those with moderate urinary incontinence on the ISI was between 0.08 (15D) and 0.14 (AQoL). For those with weekly faecal incontinence the range was from 0.07 (15D) to 0.15 (EQ5D). When the utility instruments were assessed by responsiveness to incontinence, it was observed that the most sensitive instrument for urinary incontinence was the 15D, then the HUI3 and AQoL. The EQ5D and SF6D were less sensitive. For faecal incontinence the most sensitive instruments were the 15D, AQoL and EQ5D. The HUI3 and SF6D were less sensitive. Overall, urinary incontinence as measured by the ISI explained between 2-7% of the variance in utility scores, and faecal incontinence as measured by the Wexner between 5-13%.

In short, there were substantial differences in scores between the MAU-instruments such that utilities obtained from one measure cannot be assumed to be compatible with those from the other measures. These differences reflect different descriptive systems, assigned weights, and scoring mechanisms. That these deliver utilities that are statistically significantly different across a wide range of values, suggests the results for the different instruments cannot all be right, and that study results may depend upon the instrument chosen as much as actual treatment benefits.

Incontinence and Health Status

Examination of the psychometric properties of the SF36V2 suggested that there were important differences between the Australian and US versions, both in the descriptive systems and in the obtained scale scores. In addition, when Australian factor weights for the two summary scales (PCS (physical health) and MCS (mental health)) were computed using the identical methods used by the SF36V2 developers differences were also observed. Given the limitations of these methods, it is quite likely a better model could be constructed using more sophisticated methods. These findings, however, suggest that there are differences between the US samples used for the SF36V2 weights and the Australian sample reported in this study. Consequently, Australian weights were used in reporting the study findings. A feature of the SF36V2, when compared with the SF36V1, is that all scale scores are reported as T-scores. Based on the Australian weights, therefore, all of the eight sub-scales and the two summary scales have population norms of 50 and standard deviations of ±10 points.

When examined by age and gender, for physical function (PF), role physical (RP), bodily pain (BP) and general health (GH), although there are differences between males and females, in general there are progressive declines over the lifetime. For the other scales (vitality (VI), social function (SF), role emotion (RE) and mental health (ME)) there were small variations over the lifetime. On the physical summary scale (PCS) for both genders there was a progressive decline over the lifetime, but this was not evidenced for the mental summary scale (MCS).

In addition to these population norms, the proportion of cases within scale score deciles were examined. This revealed that for the role emotion (RE) scale 79% of all cases fell within the top decile, as did 64% for the role physical (RP) scale, 61% for the social function (SF) scale and 54% for the physical function (PF) scale. These findings are suggestive of extreme skew on these scales, and it is recommended that researchers should either transform their data prior to analysis or report medians rather than means.

When the association between incontinence status and health status as measured by the SF36V2 scales was examined, the results showed that as incontinence severity increased health status deteriorated. This was the case for all four measures of incontinence and for both males and females, although there were different patterns of decline in health status by gender. Generally, for those with severe urinary or faecal incontinence their health status was 1 standard deviation or more below the health status of those with no urinary incontinence symptoms. This finding was consistent with that of the utility instruments suggesting that severe urinary incontinence has a similar effect as severe faecal incontinence.

The SF36V2 was shown to be suitable for measuring health status in incontinence studies.

Recommendations

The results of this study suggest that the preferred urinary incontinence measure is the ISI. It was found to possess superior measurement properties than the UDI-6. Because the UDI-6 measures the impact of urinary incontinence on peoples' lives rather than incontinence per se, it may overstate incontinence prevalence and the impact of this on peoples' lives (defined as their health status and their quality of life). Given its poor psychometric properties, there is a prima facie case for major revision of the UDI-6. Although the ISI is the preferred measure, because it violates the assumptions of classic psychometric theory relating to scale stability, further research into its properties is also recommended.

For faecal incontinence the current definition by the International Continence Society excludes flatus, yet this is included in the Wexner. In addition to this definitional inconsistency, the evidence from this study suggested that the inclusion of flatus led to overestimates of faecal incontinence prevalence. It is recommended that further work on the Wexner is undertaken to remove flatus and to improve its measurement properties.

A key finding of the study was that different utility measures provided such different estimates of disutility that the outcomes from cost-utility studies were just as likely to be a function of instrument choice as intervention effect. It is therefore recommended that two utility measures should be included in any particular study and that both sets of results should be reported with appropriate sensitivity analyses. The preferred instrument would be the Australian AQoL since it performed at least as well if not slightly better than any of the other MAU-instruments and because it is weighted with Australian time trade-off (TTO) values. The instrument of second choice would be the HUI3. Where direct comparison between Australian and international data is required, the EQ5D could be used. Because of its measurement shortcomings the EQ5D should not be used alone.

Given that all five utility instruments are contained within the SAHOS dataset, further research into similarities and differences between the utility measures could be undertaken with the objective of providing standardized algorithms for the development of a common scoring metric enabling imputation of scores from each instrument to each other instrument.

When the SF36V2 was closely examined, the data showed that the structure of Australian responses from the SAHOS participants was significantly different to that of the published US samples. The implication is that Australians conceptualize health differently to their US counterparts. This observation suggests that Australian researchers should use Australian weights when scoring the SF36V2, and suitable weights are provided in this report. These weights were derived using the identical methods to those used by the SF36V2 developers. The shortcomings of these

methods are acknowledged, and it is recommended that further work on scoring the SF36V2 be undertaken.

Additionally, it was observed that SF36V2 data are extremely skewed, and it is recommended that researchers should either transform their data prior to analysis or report medians rather than means.

Based on Australian weights derived from SAHOS participants, the SF36V2 scales proved sensitive to incontinence status. The SF36V2 is a suitable measure for assessing the impact of incontinence on health status.

Acknowledgements

This project was supported by a grant from the Community Care Branch, Australian Commonwealth Department of Health and Ageing, as part of the National Continence Management Strategy. The collection of data was carried out by Harrison Health Research under the auspice of the Population Research and Outcome Studies Unit, South Australian Department of Health.

I would like to thank Ms Jan Sansoni from the Australian Health Outcomes Collaboration, University of Wollongong who put forward the idea of conducting this study. Ms Anne Taylor and Ms Eleonora Dal Grande from the Population Research and Outcome Studies Unit, South Australian Department of Health, I thank for their assistance and support throughout the project. I would also like to thank Dr Richard Osborne from The University of Melbourne for his advice on early drafts of the manuscript, and Ms Konstancja Densley for her statistical assistance. I would like to thank the interviewers who conducted interviews involving questions which probed intimate aspects of responders lives. Finally, my thanks go to those South Australians who gave their time to complete a very long interview.

Where instruments are not in the public domain for research use, permissions to use instruments were obtained from the instrument developers.

Abbreviations and Terms

15D	Utility instrument developed in Finland (3, 4).
ADF	Asymptotic distribution free structural equation model. Structural equation model suitable for the analysis of non-normally distributed data.
AGFI	Adjusted goodness of fit index for assessing the proportion of variance explained by a structural equation model. Good fitting models will have AGFIs >0.90.
ANOVA	Analysis of variance.
AQoL	Assessment of Quality of Life instrument (5, 6). Utility instrument developed in Australia.
BP	Bodily Pain scale of the SF36.
Cohen's d	Estimate of effect size.
Cohen's q	Statistical test for determining whether two correlation coefficients are significantly different.
Cronbach α	Estimate of the internal consistency of items within a scale. A high level of consistency indicates that people endorse similar responses to items in the scale. Where α is between 0.70 -0.90 this is accepted as indicating good internal consistency or reliability for group assessments.
Disutility	The value of a person's quality of life state to them is measured in utilities, which are scored on a scale where 0.00 represents a quality of life state equivalent to death, and 1.00 represents the best possible quality of life state. Most people report a utility of about 0.80 (the 'norm'). Disutility is the difference between the norm and the quality of life state of interest. For example, if the norm is 0.75 and for those with urinary incontinence it is 0.68, then the disutility associated with urinary incontinence would be 0.75-0.68 = 0.07. This disutility provides an estimate of the loss of quality of life due to incontinence.
EQ5D	Utility instrument developed in Europe (7, 8). Formerly known as the EuroQol.
GH	General Health scale of the SF36.
HRQoL	Health-related quality of life.
HUI3	Health Utilities Index-3 (9, 10). Version 3 of the Health Utilities Index, a Canadian utility instrument.
ICC	Intra-class correlation.
ICS	International Continence Society.
IQOLA	International Quality of Life Assessment project carried out in the early 1990s to internationalise the SF36.
ISI	Incontinence Severity Index (11). Norwegian instrument for measuring urinary incontinence. The four-level version is used in this study (12).
Kappa (κ)	Describes the extent to which two observations on a categorical measure (e.g. Yes/No) are in agreement. Good agreement is where κ is greater than 0.60.
MAU	Multi-attribute utility instrument. This describes utility instruments.
MCS	Mental component summary score. Index score for mental health measured by the SF36.
MH	Mental Health scale of the SF36.
ML	Maximum likelihood.
NCMS	National Continence Management Strategy

OR Odds ratio. An OR is the ratio of odds of an event occurring in the group of interest (for example reporting symptoms of urinary incontinence) when compared with the odds of the event occurring in the comparator group. This is represented by:

	Event	
	Observed	**Not observed**
Group of interest	a	b
Comparator group	c	d

The formula for calculating the odds ratio is:

$$OR = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c}$$

An OR > 1.00 (for example OR = 1.5) indicates that the event is more common among the interest group when compared with the comparator group, whereas an OR < 1.00 (for example OR = 0.7) indicates the event is less common among the interest group. Where OR = 1.00 the event is observed the same relative number of times among both groups.

The critical question about ORs relates to how precisely the OR measures the difference between groups. The calculation of the confidence interval (CI) studies this. CIs represent the upper and lower bounds of the event that would occur in a specified proportion of repeated studies. The CIs were set at the conventional proportion of 95%; this means that the researcher can be confident the true OR lies between the calculated 95%CI upper and lower boundaries. The width of these boundaries determines the precision of the estimate; the narrower the width the more precise the estimate. If the 95%CIs are both greater than 1.00, then the reseacher is confident the event is truly more common among the interest group. If the 95%CIs are both less than 1.00, then the event is less common among the interest group. If the 95%CIs cross 1.00 then it can be concluded there is no significant difference between the two groups.[1]

PCS Physical component summary score. Index score for physical health measured by the SF36.

PF Physical Functioning scale of the SF36.

QALY Quality adjusted life year. Refers to the value gained as the result of an intervention expressed in utilities over time. For example, if physiotherapy for incontinence led to a 0.10 utility gain in quality of life and this was maintained over time, say for 10 years, then the gain would be 0.10 x 10 = 1.00 Quality adjusted life year (QALY).

r Pearson correlation

r_s Spearman correlation

RE Relative efficiency.

RE Role Emotion scale of the SF36.

RMANOVA Repeated measures ANOVA.

RMSEA Root mean square error of approximation. Statistical test used in structural equation models to assess the goodness of model fit to the data. For a good model fit the RMSEA should be <0.06.

RP Role Physical scale of the SF36.

SAHOS South Australian Health Omnibus Survey.

SEM Structural equation modelling

[1] If the lower or upper boundary is 1.00 then the result is statistically significant; expressed as a p-value a boundary of 1.00 is the equivalent of p = 0.05.

SF	Social Functioning scale of the SF36.
SF36	Short Form-36 (13-15). Health status instrument developed in the US.
SF6D	Short Form 6 Dimension (16, 17). Utility instrument derived from 12 items of the SF-36. The SF6D was developed in the UK.
T-score	Describes statistically transformed scores where the mean is always represented by a score of 50 and the standard deviations are always represented by scores of 10.
UDI	Urogenital Distress Inventory (18). The original form had 19 items, but a 6 item version was quickly constructed (19). This study used the short form, which is referred to as the UDI-6.
UK	United Kingdom.
US	United States of America.
Utility	Refers to the value that a person places on some object, or the preferences someone has for an object. In quality of life research utility refers to a person's preferences for a given quality of life state.
Utility instrument	Is an instrument designed to measure utilities and disutilities for use in evaluation studies, particularly cost-utility studies. Utility instruments, also known as multi-attribute utility (MAU) instruments or preference instruments, consist of two parts. The first part is the descriptive system, which consists of the items that a respondent completes. The second part is where the responses are weighted by preference weights, once weighted responses are then combined into a single score on a scale where 0.00 represents a quality of life state equivalent to death, and 1.00 represents the best possible quality of life state.
VAS	Visual analog scale.
VI (VT)	Vitality scale of the SF36.
WHOQOL	World Health Organization's quality of life instruments.
Wexner	Wexner Continence Grading Scale (20). The Wexner was designed to assist clinicians with assessing faecal incontinence. It is also known as the Cleveland Clinic Florida Fecal Incontinence Score.

1. Introduction

Incontinence is a common health problem that is thought to affect over 2 million Australians of all ages and backgrounds. The overall prevalence was estimated by Avery et al (21) to be 26% in 1998; urinary incontinence was reported to be 20% in 1998 and 21% in 2001. Five percent of respondents reported symptoms of both urinary and faecal incontinence.

For males, Avery et al (21) reported any self-reported symptoms prevalence at 11%. For urinary incontinence of all types it was 4% and 2% for faecal incontinence. Flatus was reported at 7%. In another study among Sydney men aged 40+ years, 15% were reported with either stress or urge incontinence in the last month (22). In a study of Italian-born men aged 40-80 years in Sydney, urge incontinence was reported to be 3% (23).

For females Avery et al (21) reported that the prevalence for any symptoms was 40%. For urinary incontinence the self-reported prevalence was 35% and for faecal incontinence it was 4%. For flatus it was 11%. The Women's Health Australia project, based on a community sample representative of the Australian population, reported the previous year prevalence of any urinary incontinence was 13% for those aged 18-23 years, 36% 45-50 years and 35% 70-75 years (24). Among Sydney females 40+ years the previous month prevalence was 46% of females with either stress or urge incontinence (22). These recent estimates may be compared with a study published in the 1980s which reported that the prevalence of urinary incontinence among women over the age of 10 years was 34% (25).

As the figures above make clear, incontinence is more likely to be reported by women and that as people age the incidence of incontinence rises. Indeed, an early study showed that the prevalence of incontinence among older adults living in nursing homes was 50% (26).

To address this important issue, the Commonwealth Government resourced the National Continence Management Strategy (NCMS). Through the NCMS the Government aims to improve continence treatment and management so that more Australians can live and participate in their communities with dignity and confidence.

As part of the NCMS, a report on the possibility of a national suite of outcome measures to be used by Australian clinicians and researchers working in the continence field was commissioned, the Continence Outcomes Measurement Suite Project (27). Although this report contained a review of both incontinence and utility instruments, no definitive conclusion was reached and it was recommended that further studies be undertaken.

One of these studies was to examine the Thomas et al recommended instruments using Australian data. To enable this, the Commonwealth Department of Health and Ageing funded a special version of the South Australian Health Omnibus Survey in 2004. This report details the outcomes from that survey. The study was designed to report on four important incontinence issues: (a) it provides current prevalence estimates of incontinence in the Australian general community, (b) it provides psychometric insights into those incontinence assessment instruments recommended in the Thomas et al report, (c) it reports Australian population norms for the leading five utility instruments and the impact of incontinence on respondents' lives, and (d) it provides Australian population norms and Australian-derived weights for the SF-36 Version 2.

2. Methods

2.1 Participants

The current study uses data collected from 3015 South Australians who participated in the 2004 South Australian Health Omnibus Survey (SAHOS) (28). The SAHOS is a population-based user-pays health survey which has been carried out every year since 1991. A full description of the methodology can be found in Wilson et al (2). In brief, the 2004 survey involved interview with sampled households throughout South Australia, including all country towns with a population of 1000 or greater. For the metropolitan sample, the Australian Bureau of Statistics collectors districts for the 2001 Census were sampled based on probability of selection proportional to size (n=363 districts). Within districts, using a 'skip' pattern of every 4th household, 10 dwellings were chosen and one person (aged 15 or more years) from each dwelling interviewed, based on closest last birthday to interview day. Similar procedures were also used to select the country sample, based on 107 districts.

Four thousand seven hundred dwellings were selected, 127 were vacant, 366 dwellings were non-contactable after six visits, 39 dwellings could not be accessed, in 82 dwellings the respondent was unable to speak English, 58 cases were absent during the data collection phase, 62 were incapacitated due to illness, and 945 refused to participate. The total number of participants interviewed was 3015, giving a within scope response rate of 72% (3015/(4700-366)).

2.2 Defining Incontinence

Urinary Incontinence

Abrams et al in their report from the Standardisation Sub-committee of the International Continence Society (ICS) defined urinary incontinence as the complaint of any involuntary leakage of urine. They further defined three key types of urinary incontinence as stress urinary incontinence (the complaint of involuntary urinary leakage on effort or exertion, including sneezing or coughing), urge urinary incontinence (the complaint of involuntary leakage preceded or accompanied by urgency) and mixed urinary incontinence (where both urgency and exertion leakage occurred) (29). In general, this definition is widely accepted (e.g. it is the definition used in Abrams (30, 31), and also by Avery et al (21) in their report on incontinence prevalence in South Australia). Earlier definitions required that the report of involuntary loss or urine was a social or hygienic problem, and in yet earlier definitions that this was to be objectively demonstrable (32). Thomas et al (27), however, in their report on the measurement of incontinence, noted that the most recent definition cuts across these earlier definitions. They averred that the current definition (that given above) was appropriate for epidemiologic studies, such as this study, whereas the ICS's previous definitions were more appropriate for clinicians. The ICS, however, did not draw this distinction in its most recent definition. Rather it suggested that there were different levels of measurement (29). These included:

- Symptoms, which were qualitative reports volunteered by the patient. These were defined as the patient's subjective perception of their condition. Importantly, the ICS noted that symptoms cannot be used for diagnosis of incontinence. Evidence for this position is that self-report of incontinence is not highly correlated with interference in quality of life, clinical assessments or urine pad tests (e.g. see (12)). For example, a British study reported that 69% of women who responded to a population-based survey reported some loss, 30% reported that their state interfered with their social life, and 29% reported this was a hygienic problem based on the use of pads or the soiling of underwear. The researchers concluded that it may be regarded as 'normal' for women to experience some urinary leakage (33).
- Signs, which referred to clinically observable verification of urinary leakage. Methods of assessing signs of incontinence included frequency volume charts, pad tests and quality of life questionnaires (29). The inclusion of the latter seems misplaced because these are generally self-reports from the patient perspective and are therefore not objective tests.
- Urodynamic observations, which referred to the outcomes of urodynamic tests carried out under clinical supervision. It was argued by the ICS that urodynamic observations did not provide diagnosis evidence due to the special circumstances of testing.

Faecal Incontinence

As with urinary incontinence, different definitions have been advanced at different times. The most recent ICS definition is that faecal incontinence is the involuntary loss of liquid or solid stool that is a social or hygienic problem (34). This definition explicitly excludes involuntary flatus. The exclusion of flatus marks a major departure from some previous definitions (35), yet it is consistent with that of other researchers (36). For example, the Royal College of Physicians (37) defined faecal incontinence as the involuntary or inappropriate passage of faeces, thereby excluding flatus.

This definition is also inconsistent with that used in previous research in South Australia. Avery et al (21) used the definition of unwanted release of faeces or gas, which was taken from Nelson et al's 1995 study (35). They also included in their definition the occasional staining of underwear, loss of loose stool or inadvertent loss of formed stool.

This study reports incontinence prevalence rates based on self-reported symptoms as assessed during interview by the respondent completing the standard Incontinence Severity Index (ISI), the Urogenital Distress Inventory – Short Form (UDI-6) and the Wexner Continence Grading Scale (Wexner) measures. Consequently the reported rates are a function of the descriptive systems of the measures, the self-assessments of respondents, and the interview situation. It is possible that different instruments completed in different settings may provide different estimates.

2.3 Description of the Instruments included in this Study

The instruments included in the SAHOS and used in this report fall into three groups: those measuring incontinence, health status and utility. Each is described.

2.3.1 Incontinence Instruments

The Incontinence Severity Index (ISI)

The Incontinence Severity Index (ISI) originally consisted of two items, one with 4 response levels and the other with two response levels (11). In 2000 the instrument developers altered the second item's response scales from 2 to 3 levels, known as the four-level severity index (12). The four-level severity index is reported here. The ISI comprises two items: *How often do you experience urine leakage (*response scale: *Less than once a month (1) /1-several times a month (2)/1-several times a week (3)/Every day/night (4));* and *How much urine do you lose* (response scale: *None (0)/ A few drops (1)/A little(2)/More(3))* (11). Scoring is through multiplication of endorsed response levels giving a score range from 0 to 12. Higher scores denote more severe urinary incontinence. Validity of the ISI was assessed against a 48-hour pad test; the correlation was r =0.59 (11). In a second study, also against a 48-hour pad test, the correlation was r =0.54 (12). Test-retest kappa at 3-days was 0.69 and 0.83 for the two items (38). Sandvik et al (12) recommended that when using the four-level severity index the interpretations were scores 1-2 a slight problem, 3-6 moderate, 7-9 severe and 10-12 very severe. The ISI has been used in population surveys (11, 39). The standard scoring system described above does not discriminate between those with no incontinence symptoms and those with slight symptoms. The ISI was therefore modified through inclusion of a 'never' category; thus *Never/Less than once a month/1-several times a month/1-several times a week/Every day/night*, which added an extra category, '0', describing those with no symptoms. This procedure is that recommended by Sandvik et al (12). Where classification was needed, ISI scores were recoded into (score range) *None(0) /Slight (1to 2)/ Moderate (3 to 6)/ Severe (7 to 9)/ Very Severe (10 to 12)* levels.

The Urogenital Distress Inventory (short form; UDI-6)

The Urogenital Distress Inventory (UDI) was developed to assess the impact of urinary incontinence symptoms upon quality of life for women (18). The original form had 19 items, but a 6 item version of the UDI-6 was quickly constructed (19). The UDI-6 consists of six items measuring urination frequency, leakage due to urgency, leakage due to physical activity, small leakages, emptying bladder difficulties, and pain or discomfort. A typical item is: *Do you experience, and if so, how much are you bothered by urine leakage related to physical activity, coughing, or sneezing?*. Responses are on a 4-point Guttman scale *(Not at all/ Slightly/ Moderately/ Greatly)*. Scores from the items are summed. In the present study, where needed, UDI-6 scores were classified into *No incontinence symptoms (score range: 0)/Slight problem (1 to 3)/Moderate problem*

(4 to 6)/Problem (7 to 9)/Major problem (10+). UDI-6 items have been previously shown to predict incontinence symptoms (40), and it has been used in previous population surveys (41).

Wexner Continence Grading Scale (Wexner)[2]

The Jorge and Wexner faecal continence grading scale was developed to provide clinicians with a means of assessing faecal incontinence severity (20). The Wexner requires assessment on leakage/accidental faeces for solid, liquid, and gas, the need to wear a pad and alterations to lifestyle. There is no mention in the Wexner of urge incontinence. Each item is assessed on a Guttman scale *(Never/Rarely/Sometimes/Usually/Always)*. Scores are determined by a simple summation of endorsements. The range is from 0 to 20 and the higher the score the worse the faecal incontinence. Scores can also be classified into categorical levels (faecal incontinence *Never/Rarely (1 episode in past month)/Sometimes (2-4 episodes)/Weekly (>1 week – <1 day episodes)/ Daily (1 or more daily episodes))*. The obvious difficulty with the Wexner is that it is unconventional to sum symptoms and symptom effects; a procedure that gives rise to double counting. In the Wexner, stool leakage and its consequence (wearing a pad) are both counted. Vaizey et al (42) also criticized the Wexner for the pad wearing question, arguing this was a measure of patient fastidiousness or urinary comorbidity. In test-retest at 2-week, the reliability of the Wexner was r =0.75 (42).

Other Faecal Continence Questions

Faecal Behaviours

Three questions were included measuring bowel pattern, the number of weekly bowel movements and bowel movement urge.

Soiling

Because of concerns with the Wexner, it was supplemented by two additional items, *Do you leak if you don't get to a toilet on time?* and *Does stool leak so that you have to change your underwear?*. For each the response scale was *Never/Rarely/Sometimes/Often/Always*, defined as for the Wexner.

Reliability analysis of these two questions revealed that the correlation between the two items was $r_s = 0.61$. If interpreted as a scale measuring soiling, the Cronbach $\alpha = 0.78$, and the explained variance was 83%. For convenience, the responses to the two soiling items were added to form a single score.

2.3.2 Health Status

SF36V2

To assess participants' health status, the SF-36 Version 2 (Australian version; SF36V2) was administered. Following growing awareness of the limitations of the SF-36 V1 (SF36V1), between 1996-2000 Ware et al developed the "international version" of the SF-36 – the SF-36 Version 2 (hereafter SF36V2) (hereafter SF36V2 15). The changes were designed to make it easier to understand, to reduce missing data, improve the sensitivity of the two role function scales, and to simplify the response categories for the health and vitality scales. The SF36V2 consists of 36 items probing functional health status. The items are organised into 8 scales measuring Physical Function, Physical Role, Bodily Pain, General Health, Vitality, Social Function, Role Emotion, and Mental Health. For the SF36V2, scores on these scales are presented as T-scores (43), whereas in SF36V1 these were presented as percentile scores (14). As with the SF36V1, scores can be aggregated up into Physical Component Summary (PCS) and Mental Component Summary (MCS) scores. These scores are also presented as T-scores. Following release of the SF36V2, Sansoni and Costi (44) released the Australian SF36V2. The differences between Australian and US versions of the SF36V2 are discussed in section 5 of this report.

[2] Wexner has started calling this the Cleveland Clinic Florida Fecal Incontinence Score. Since this measure is not widely known by this name it is not used here.

2.3.3 Utility Instruments

Brief descriptions of the utility instruments included in the study are provided here; more detailed descriptions can be found in Appendix A.

Assessment of Quality of Life (AQoL)

The AQoL describes utility from a 'handicap' perspective, and the descriptive system has 15 items of which 12 are used in computing the index (45, 46). Each item has 4 levels. There are five dimensions: Illness (not used in utility computation), Independent Living, Social Relationships, Physical Senses and Psychological Well-being (5). A multiplicative model is used to compute the utility index (45). The upper boundary is 1.00, and the lower boundary is –0.04: it permits health state values worse than death.

EQ5D (formerly the EuroQol)

The EQ5D (formerly the EuroQoL) has 5 items, each with 3 response levels, measuring Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression (7). The utility weights used in this study are from a British population random sample (n = 3395 respondents, response rate 56%) based on the TTO (time trade-off) for 42 marker health states using a 10 year timeframe (47). The index is computed using an econometric regression model. The upper boundary is 1.00, and the lower boundary is –0.59: it permits health state values worse than death.

Health Utilities Index, Mark 3 (HUI3)

The Health Utilities Index (HUI3) measures utility from a 'within the skin' functional perspective (10), adopted to enhance its use in clinical studies (48). Social aspects of HRQoL are not measured. Items have 4–6 response levels. Twelve of the 15 items form 8 attributes (Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition and Pain). A multiplicative function combines the attributes into the utility score (49, 50). The upper boundary is 1.00, and the lower boundary is –0.36, permitting health states worse than death.

15D

The Finnish 15D is concerned with impairment and disability of 'within the skin' functions (3, 51). There are 15 items, each with 5 levels, measuring Mobility, Vision, Hearing, Breathing, Sleeping, Eating, Speech, Elimination, Usual Activities, Mental Function, Discomfort & Symptoms, Depression, Distress, Vitality and Sexual Function (3). Responses are combined using an additive model (4, 52). The upper boundary is 1.00, and the lower boundary is +0.11: death-equivalent and worse than death health states are not allowed.

SF6D

Two different algorithms have been published for deriving preference-based values from the SF-36 (16, 17). They are referred to as the SF6D-1 and SF6D-2; the SF6D-2 is used in this study. It uses 10 items from the SF-36: three from the physical functioning scale, one from physical role limitation, one from emotional role limitation, one from social functioning, two bodily pain items, two mental health items and one vitality item. These form 6 dimensions: Physical Functioning (PF: 6 levels), Role Limitation (RL: 4 levels), Social Functioning (SF: 5 levels), Pain (PA: 6 levels), Mental Health (MH: 5 levels) and Vitality (VI: 5 levels). An additive econometric model is used to compute the utility index. The endpoints for the SF6D are 1.00, and 0.30 for the worst possible health state.

2.4 Data Analysis

For each part of the study, the detailed data analysis procedures are given in the relevant section. The comments here are more general.

Weighting the Data

The data were weighted to achieve population representativeness. Although researchers strive to achieve population representativeness through either taking random or stratified samples of the population, usually certain groups are over- or under-represented. For example, it is well known that in random samples more females than males will choose to participate. For example, in the

SAHOS 58% of actual participants were female. To correct for any bias this may cause in study findings, the data can be weighted so that, for example, the results reflect what would have been the case if there were equal numbers of males and females.

In the SAHOS study the data were weighted by the inverse of the probability of selection, and then re-weighted to population benchmarks from the 2002 Estimated Resident Population for South Australia. The effect is that the number of cases reported reported in any table is a probability estimate which will vary depending upon the analysis and the proportion of missing data. For example, in the age group 15-29 years there were 83 actual male cases. In Table 1 (page 9) the weighted number of probable male cases is reported as 129, in Table 2 (page 10) the weighted number of probable males cases is 130. Because the reported numbers are probabilities, the results are usually reported as percentages which have been rounded up to whole integers.

Missing Data

The data were collected through interview, consequently there was very little missing data. Where data were missing, every attempt was made to collect these data through followup telephone contact at the end of regular data collection. Where data were still missing, one of two procedures was followed:

(a) Missing data within scales were imputed using horizontal mean substitution (53); and

(b) Missing datum on individual items was left as a missing datum. Thus the default position adopted in this study was listwise deletion. This means that where a datum on an individual item was missing that case was removed from the analysis. This is the reason there are slight discrepancies in the numbers in some tables. This decision was made because the level of missing data was <1%.

Statistical Procedures

To compare between scales with different ranges, the scales were converted to McCall's T-scores prior to analysis (43).

Agreement between two variables was tested with Kappa (κ). Categorical data were examined with χ^2. To examine scale relationships, Spearman's ρ was used where the data were significantly skewed. Elsewhere Pearson's ρ was used and Cohen's θ calculated. The internal reliability of scales was assessed using Cronbach α and the internal structure examined using exploratory factor analysis and structural equation modelling.

Odds ratios with 95% confidence intervals were used to examine the relative frequency of events. Cohen's d effect size was used to compare across classifications, and for comparisons between instruments the relative efficiency (RE) statistic was computed.

Means and standard deviations are reported. Analysis of variance (ANOVA) was used to examine differences between cohorts on the various measures. Where severe data skew was observed the Kruskall-Wallis χ^2 or the Wilcoxon Signed Ranks test was used as indicated. For multi-variate comparisons, repeated measures analysis of variance (RMANOVA) was used and for post hoc comparisons the Tukey-Kramer multiple comparisons test value computed.

All percentages have been rounded up to the nearest integer; therefore summaries in tables may not always add up to 100%.

3. Incontinence Prevalence

3.1 Summary

The prevalence of urinary incontinence varies according to how the data are interpreted.

The best estimate for urinary incontinence of all types based on the ICS definition for incontinence symptoms, namely the self-report of any symptoms of urinary leakage, would be the ISI estimated prevalence of urinary incontinence at 24% (95%CI: 23% – 26%) overall. When broken down by gender, it would be 38% (95%CI: 36% – 41%) for females and 10% (95%CI: 9% – 12%) for males. On the other hand, if measured by the UDI-6, which measures being bothered by symptoms, the overall prevalence of urinary incontinence would be 47% (95%CI: 45% – 48%); for females it would be 60% (95%CI: 58% – 63%) and for males 33% (30% – 35%). For the reasons outlined in the discussion to this section, the ISI estimates are preferred.

For faecal incontinence, the standard Wexner data suggested that the prevalence was 35% (95%CI: 33% – 36%). For females this was 38% (95%CI: 35% – 40%), and for males it was 32% (95%CI: 29% – 34%). However, it should be noted that the Wexner includes flatus, which is excluded from the current ICS faecal incontinence definition (34). If the flatus question is excluded from the Wexner, the data show that the prevalence would be 8% (95%CI: 7% – 9%). For females this would have been 10% (95%CI: 8% – 11%) and for males 6% (5% – 7%). For the reasons outlined in the discussion to this section, the modified Wexner scale excluding flatus prevalence estimates are preferred.

Based on the preferred methods of measuring urinary and faecal incontinence, it is estimated that the prevalence of any incontinence is 27% (95%CI: 26% – 29%). For females it is 40% (38% – 43%) and for males 14% (12% – 15%).

It should, however, be noted that these findings do not provide evidence of diagnosed urinary or faecal incontinence prevalence. To establish this level of evidence, clinical assessments would be needed (29).

3.2 Introduction

Incontinence is a common health problem that affects many Australians. A recent report stated that the overall prevalence was 26%. When examined by incontinence type, Avery et al (21) reported prevalence for those aged over 15 years as 20% for any symptoms of urinary incontinence and 11% for any symptoms of faecal incontinence. Kalantar (54) et al also reported a 12-month prevalence of 11% for any faecal incontinence. Five percent of Australians are reported as having symptoms of both urinary and faecal incontinence (21, 55).

When examined by gender, Avery et al (21) reported that females were eleven times more likely to report symptoms of urinary incontinence when compared with males (odds ratio (OR): 11.8, 95%CI: 8.9 – 15.5), and were twice as likely to report any symptoms of faecal incontinence (OR: 1.7, 95%CI: 1.3 – 2.2). They reported that, based on the 1998 and 2001 South Australian Health Omnibus Surveys (SAHOS), 35% of females reported symptoms for any urinary incontinence and 4% for faecal incontinence (excluding flatus). This estimate of urinary incontinence was almost identical to that reported 20 years earlier when the prevalence of urinary incontinence among Australian women over the age of 10 years was estimated to be 34% (25). Regarding faecal incontinence (excluding flatus), the prevalence was much lower than an earlier estimate which had reported that the 12-month prevalence among women was 12% (54).

Additionally, there is Australian evidence that prevalence for women increases across the lifespan. Avery et al (21) reported increases in stress (urge) urinary incontinence from 18% (for urge incontinence it was 5%) for women aged 15-39 years, to 45% (17%) for those aged 40-59 years and 43% (28%) for those aged over 60 years. For faecal incontinence the increases were 2%, 4% and 6% respectively. The urinary prevalence estimate for middle-aged women (aged 40+ years) participating in the NSW Health Survey was 46% with either stress or urge incontinence in the last month (22). The Women's Health Australia project reported that urinary incontinence prevalence increased from 13% for women aged 18-23 years, to 36% for those aged 45-50 years and to 35% for those aged 70-75 years (24).

A possible reason for the slightly lower prevalence among older women is that incontinence is a major reason for admission to residential care amongst older adults (56). Approximately 70% to 80% of nursing home residents are female and the rate of incontinence (both urinary and faecal) is between 44% to 50% (26, 57). This might partly explain, for example, the lower 12-month prevalence estimates reported by Liu & Andrews (58) for a sample of those aged 70+ years (males and females). The overall urinary incontinence prevalence was 21% (17% for urge and 4% for stress incontinence).

Prevalence rates for males are lower. Avery et al (21) reported a rate of 11% for any symptoms. For urinary incontinence the prevalence was 4% and 2% for faecal incontinence (excluding flatus). The 1997 NSW Health Survey reported that for Sydney men aged 40+ years, 15% of males were classified with either stress or urge urinary incontinence in the last month (22). Urge incontinence prevalence was reported at 3% among Italian-born men aged 40-80 years in Sydney (23). Among middle-aged to older men (40-80 years), although more than half of all respondents reported at least one urinary symptom the prevalence of urge incontinence was reported at 4% (23, 59). For faecal incontinence (excluding flatus), Kalantar et al (54) reported a 12-month prevalence of 11%.

Similarly to females, incontinence symptoms increase over the lifespan. Avery et al (21) reported that for males aged 15-39 years the prevalence of stress (urge) incontinence was 1% (1%), for those aged 40-59 years it was 1% (1%), and for those aged 60+ years it was 7% (11%). For faecal incontinence (excluding flatus) the figures were 2%, 4% and 6% respectively.

Against this background, this study reports incontinence prevalence rates from the 2004 South Australian Health Omnibus Survey.

3.3 Missing Data

For reasons of privacy, all respondents were given the opportunity to refuse to answer questions on incontinence. In the event, few respondents refused; the number of refusals was between 4 to 7 cases, i.e. <1% of all respondents. These missing data were not imputed, which accounts for some of the discrepancies in the tables.

3.4 The Reliability of the Scales

The reliability of the three incontinence measures was assessed using internal consistency (Cronbach α). The estimates for the ISI and UDI-6 fell within the conventional range (>0.70), but that for the Wexner was unacceptable because it fell well short of conventional practice (60-62).

The reliability of the ISI was Cronbach α = 0.89. The relationship between the two items of the ISI was examined using kappa (κ) and found to be 0.74 indicating good agreement between the two items (63). There was perfect agreement between the two items for 90% of all respondents. Thus the ISI gains its high level reliability through replication; the r_s = 0.96 and the proportion of explained variance was 92%. Although this analysis suggests the ISI has excellent measurement properties, that it consists of just two items suggests that it violates classical test theory which postulates that at least 3 items are needed for stable measurement interpretation (64).

The reliability of the UDI-6 was Cronbach α = 0.78. Two items, the last two questions *How much are you bothered by pain or discomfort in the lower abdominal or genital area?* and *Do you have difficulty emptying your bladder?* did not fit well with the other items (the item-total-correlation was r =0.31 and r =0.37 respectively). Deletion of these items would have improved the α to 0.81 and the explained variance from 49% to 66%.

The reliability of the Wexner was Cronbach α = 0.53, which was well below that normally considered acceptable for good measurement (65). The difficulty with the Wexner was in relation to the item measuring flatus: *Do you leak, have accidents or lose control with gas (flatus or wind)?*. Removal of this item would have improved the α to 0.76 and the explained variance from 49% to 58%.

In the interests of comparability with other studies, the changes suggested by the analyses above were not made to the instruments for this study.

3.5 Urinary Incontinence Prevalence

Table 1 presents urinary incontinence severity as measured by the ISI. This shows that for those aged 15-19 years 94% reported no symptoms, declining to 63% for those aged 80+ years. However, the decline across the age groups was not monotonic (e.g. 64% of those aged 50-59 years reported no symptoms compared with 69% for those aged 60-69 years). For females, the age group with the highest proportion was the 50-59 age group (55%) and for males it was those aged 80+ years (30%).

Table 1: Urinary Incontinence assessed by the ISI, by Gender and Age Group, Percentages

Age group (years)	*Gender*	*Number*	*ISI*						
			Any symptoms			*Classification if symptomatic*			
			None	*Any*	*95%CI*	*Slight*	*Moderate*	*Severe*	*Very severe*
15-19	Female	123	89%	11%	(5% – 16%)	11%	0%	0%	0%
	Male	129	98%	2%	(0% – 5%)	2%	0%	0%	0%
	All	253	94%	6%	(3% – 9%)	6%	0%	0%	0%
20-29	Female	230	82%	18%	(13% – 23%)	14%	3%	1%	0%
	Male	244	95%	5%	(2% – 8%)	5%	0%	0%	0%
	All	474	89%	11%	(9% – 14%)	10%	1%	0%	0%
30-39	Female	263	60%	40%	(34% – 46%)	32%	6%	1%	0%
	Male	267	96%	4%	(2% – 7%)	3%	0%	0%	0%
	All	530	78%	22%	(18% – 25%)	18%	3%	1%	0%
40-49	Female	278	56%	44%	(38% – 49%)	34%	9%	0%	1%
	Male	273	94%	6%	(3% – 9%)	5%	1%	0%	0%
	All	550	75%	25%	(21% – 28%)	20%	5%	0%	0%
50-59	Female	243	45%	55%	(49% – 61%)	40%	11%	3%	1%
	Male	236	83%	17%	(12% – 22%)	12%	4%	1%	0%
	All	479	64%	36%	(32% – 41%)	27%	8%	2%	1%
60-69	Female	162	52%	48%	(40% – 56%)	30%	14%	3%	1%
	Male	157	87%	13%	(8% – 19%)	12%	2%	0%	0%
	All	318	69%	31%	(26% – 36%)	21%	8%	1%	1%
70-79	Female	158	60%	40%	(32% – 48%)	24%	11%	4%	1%
	Male	125	74%	26%	(19% – 34%)	21%	5%	1%	0%
	All	283	66%	34%	(28% – 39%)	23%	9%	3%	0%
80+	Female	75	59%	41%	(30% – 52%)	23%	8%	8%	3%
	Male	44	70%	30%	(16% – 43%)	18%	7%	5%	0%
	All	119	63%	37%	(28% – 46%)	21%	7%	7%	3%
All	**Female**	**1531**	**62%**	**38%**	**(36% – 41%)**	**28%**	**8%**	**2%**	**1%**
	Male	**1475**	**90%**	**10%**	**(9% – 12%)**	**8%**	**2%**	**0%**	**0%**
All		**3007**	**76%**	**24%**	**(23% – 26%)**	**18%**	**5%**	**1%**	**0%**

The slight discrepancy in table numbers is because of missing data.

For those classified as having slight incontinence, the highest proportion was for females aged 50-59 years (40%), and for males it was for those aged 70-79 years (21%). For those classified as being moderately incontinent, for females the highest proportion was for those aged 60-69 years, and for males if was for those aged 80+ years (7%). For those classified as having severe urinary incontinence, for both genders the highest proportion was those aged 80+ years (females: 8%, and males: 5%). Very few cases were classified as having very severe urinary incontinence (1% overall).

The prevalence of urinary incontinence identified by the UDI-6 is given in Table 2. For those aged 15-19 years, 81% reported no symptoms, compared with 32% for those aged 80+. Regarding the

Table 2: Urinary Incontinence assessed by the UDI-6, by Gender and Age Group, Percentages

Age group (years)	*Gender*	*Number*	*ISI*						
			Any symptoms			*Classification if symptomatic*			
			None	*Any*	*95%CI*	*Slight*	*Moderate*	*Severe*	*Very severe*
15-19	Female	124	74%	26%	(18% – 34%)	195	7%	0%	0%
	Male	130	88%	12%	(7% – 18%)	12%	0%	0%	0%
	All	253	81%	19%	(14% – 24%)	16%	3%	0%	0%
20-29	Female	231	51%	49%	(42% – 55%)	35%	10%	3%	1%
	Male	243	83%	17%	(12% – 22%)	16%	1%	0%	0%
	All	474	68%	32%	(28% – 37%)	25%	6%	1%	0%
30-39	Female	263	43%	57%	(51% – 63%)	42%	11%	1%	2%
	Male	267	77%	23%	(18% – 28%)	20%	2%	0%	0%
	All	530	60%	40%	(35% – 44%)	31%	7%	1%	1%
40-49	Female	278	36%	64%	(58% – 70%)	40%	17%	3%	5%
	Male	273	75%	25%	(20% – 31%)	24%	1%	0%	0%
	All	551	55%	45%	(41% – 49%)	32%	9%	2%	2%
50-59	Female	243	28%	72%	(66% – 78%)	41%	21%	6%	5%
	Male	237	56%	44%	(37% – 50%)	30%	10%	3%	0%
	All	480	42%	58%	(54% – 62%)	36%	15%	4%	3%
60-69	Female	161	26%	74%	(67% – 81%)	43%	24%	5%	3%
	Male	157	52%	48%	(41% – 56%)	38%	8%	3%	0%
	All	317	39%	61%	(56% – 67%)	40%	16%	4%	1%
70-79	Female	157	36%	64%	(57% – 72%)	31%	24%	6%	3%
	Male	126	31%	69%	(61% – 77%)	52%	10%	6%	2%
	All	282	34%	66%	(61% – 72%)	40%	18%	6%	3%
80+	Female	77	30%	70%	(60% – 80%)	40%	17%	9%	5%
	Male	44	39%	61%	(47% – 76%)	41%	11%	5%	5%
	All	121	32%	68%	(60% – 76%)	41%	15%	7%	5%
All	**Female**	**1534**	**40%**	**60%**	**(58% – 63%)**	**37%**	**16%**	**4%**	**3%**
	Male	**1476**	**67%**	**33%**	**(30% – 35%)**	**26%**	**4%**	**1%**	**0%**
All		**3005**	**54%**	**47%**	**(45% – 48%)**	**32%**	**10%**	**3%**	**2%**

The slight discrepancy in table numbers is because of missing data.

reporting of any symptoms, for females the highest proportion was the 50-59 age group (72%), and for males it was for those aged 70-79 years (69%).

For those with slight incontinence, the highest proportion was males aged 70-79 years (52%); for those with moderate incontinence it was females aged 60-69 years and 70-79 years (both 24%), and for those with incontinence problems the highest proportion was females aged 80+ years. As with the ISI, very few cases were classified as having severe incontinence.

Regarding the relationship between the two estimates of urinary incontinence, the ISI and UDI-6, this was examined after conversion to McCall's T-scores to compensate for the different scale ranges used by the two measures. The correlation was $r_s = 0.75$ (n=3005, $p<0.01$). This suggests that the ISI and UDI-6 are measuring similar and related aspects of incontinence. A scatterplot of this T-score relationship is presented in Figure 1. It suggests that the correlation between the two measures is primarily leveraged by those classified as having no incontinence symptoms. The figure also suggests that the UDI-6 was more sensitive for those with minor or moderate symptoms (in the T-score range 41 to 70), and that the ISI may have classified too many cases at the ceiling (T-score 115 for the ISI). The diagonal line in Figure 1 is the line where all cases would fall if the relationship between the ISI and UDI-6 was perfect. As shown, however, it appears that the ISI classified more cases at the floor (i.e. those with no symptoms) and also more cases towards the ceiling (i.e. the greatest incontinence severity) when compared with the UDI-6.

Table 3 presents these similarities and differences by the proportion of cases, where the criterion was the decile score range for each instrument. This analysis confirms the impression conveyed by Figure 1. The ISI classified 76% of cases compared with the UDI-6's 54% of cases as continent (with no symptoms). In the worst 50% of each scale (i.e. 51-100% of the scales ranges, indicating increasing levels of incontinence) were classified 3% of all cases for the ISI compared with 1% for the UDI-6.

Figure 1: Scatterplot of the UDI-6 and ISI, T-scores

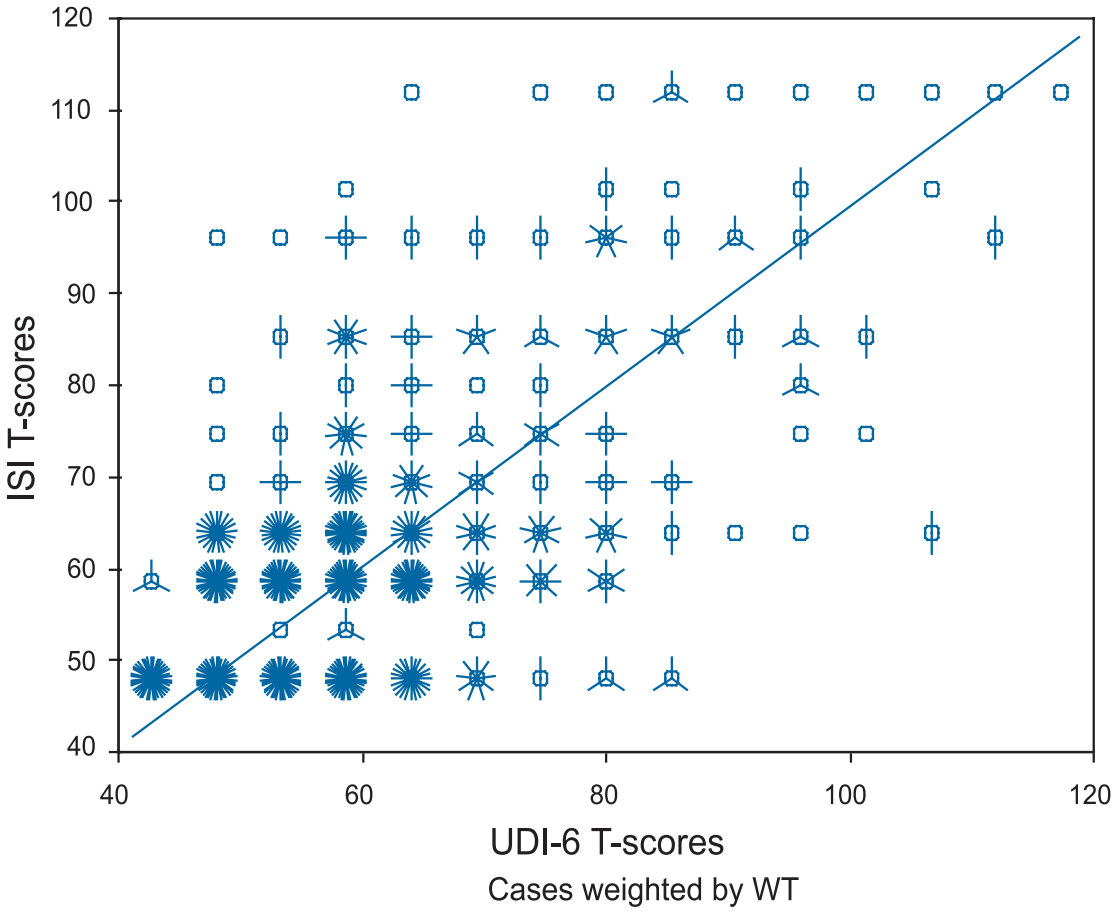


Table 3: Proportion of Cases with Urinary Incontinence Symptoms by ISI and UDI-6 Decile Scores

Score range deciles	*ISI*			*UDI-6*		
	Score range	*N. cases*	*%*	*Score range*	*N. cases*	*%*
No symptoms	1	2273	76%	0	1609	54%
1-10%	2-3	35	1%	1-2	749	25%
11-20%	4-5	374	12%	3-4	365	12%
21-30%	6-7	141	5%	5-6	158	5%
31-40%	8-9	83	3%	7-8	61	2%
41-50%	10-11	13	0%	9-10	33	1%
51-60%	12-13	42	1%	11-12	20	1%
61-70%	14-15	27	1%	13-14	4	0%
71-80%	16-17	7	0%	15-16	7	0%
81-100% (a)	18-20	12	0%	17-18	2	0%

Notes The slight discrepancy in table numbers is because of missing data.

a = Combined category due to small numbers.

Statistics: $\chi 2 = 805.41$, df = 8, $p < 0.01$ (Categories 71-80% and 81-100% collapsed to ensure adequate cell size).

3.6 Faecal Incontinence Prevalence

Faecal or anal incontinence was measured with the Wexner. Across the entire sample 66% of respondents did not report any faecal incontinence symptoms; thus any faecal incontinence symptoms were reported by 34%. Among those with symptoms, 20% reported these rarely, 9% sometimes, 3% weekly and 2% daily. When broken down by age group the data showed there was a continuous increase in the proportion with any symptoms until the age of 59 years, after which the proportion with symptoms declined. The highest proportion with symptoms was females aged 50-59 years (50%); the highest proportion for males was 43% for those aged 50 to 69 years. The details are given in Table 4.

As discussed above, the current ICS definition of faecal incontinence excludes flatus. The standard Wexner scale reported in Table 4, however, includes flatus. If the flatus question is removed from the Wexner, Wexner scores are compatible with the current ICS definition. This has been done in Table 5.

The results show that the prevalence of faecal incontinence would have been 8% (95%CI: 7% – 9%). For females this would have been 10% (95%CI: 8% – 11%) and for males 6% (5% – 7%). The highest prevalence rate would have been for females aged 70+ years (17%) and for males aged 70-79 years (15%).

Table 4: Faecal Incontinence assessed by the Wexner, by Gender and Age Group, Percentages

Age group (years)	Gender	Number	Wexner						
			Any symptoms			Classification if symptomatic			
			None	Any	95%CI	Slight	Moderate	Severe	Very severe
15-19	Female	123	87%	13%	(7% – 19%)	11%	2%	0%	0%
	Male	130	78%	22%	(14% – 29%)	14%	5%	2%	2%
	All	253	83%	17%	(13% – 22%)	12%	4%	1%	1%
20-29	Female	231	77%	23%	(17% – 28%)	16%	4%	2%	0%
	Male	243	77%	23%	(18% – 28%)	17%	3%	2%	2%
	All	474	77%	23%	(19% – 27%)	17%	4%	2%	1%
30-39	Female	263	60%	40%	(34% – 45%)	24%	10%	3%	2%
	Male	267	78%	22%	(17% – 27%)	11%	6%	3%	1%
	All	529	69%	31%	(27% – 35%)	18%	8%	3%	1%
40-49	Female	278	62%	39%	(33% – 44%)	23%	8%	3%	4%
	Male	275	66%	34%	(28% – 39%)	20%	8%	4%	2%
	All	553	64%	36%	(32% – 40%)	22%	8%	3%	3%
50-59	Female	244	50%	50%	(44% – 56%)	24%	17%	6%	3%
	Male	238	57%	43%	(37% – 49%)	23%	13%	6%	2%
	All	482	53%	47%	(42% – 51%)	23%	15%	6%	2%
60-69	Female	161	55%	45%	(38% – 53%)	19%	15%	8%	4%
	Male	156	57%	43%	(35% – 51%)	29%	10%	2%	3%
	All	318	56%	44%	(39% – 50%)	23%	12%	5%	4%
70-79	Female	158	55%	45%	(37% – 53%)	22%	15%	3%	5%
	Male	126	62%	38%	(30% – 47%)	23%	8%	4%	3%
	All	285	58%	42%	(36% – 48%)	22%	12%	4%	4%
80+	Female	78	61%	39%	(28% – 49%)	23%	8%	4%	4%
	Male	44	70%	30%	(16% – 43%)	16%	7%	0%	7%
	All	122	65%	35%	(27% – 44%)	21%	7%	3%	5%
All	**Female**	**1534**	**62%**	**38%**	**(35% – 40%)**	**21%**	**10%**	**4%**	**3%**
	Male	**1478**	**68%**	**32%**	**(29% – 34%)**	**19%**	**8%**	**3%**	**2%**
All		**3016**	**66%**	**35%**	**(33% – 36%)**	**20%**	**9%**	**3%**	**2%**

The slight discrepancy in table numbers is because of missing data.

3.7 Soiling

Soiling of clothes is not a clinical indicator of incontinence itself; rather it is a consequence of urge incontinence. As such it is difficult to interpret soiling reports with any degree of certainty. To assist with understanding the meaning of soiling, the relationship between faecal incontinence status and soiling was examined. When compared with those who reported no faecal incontinence on the Wexner, those with rare faecal incontinence were 11 times more likely to report soiling (OR: 10.64, 95%CI: 6.42 – 17.78), those classified as having faecal incontinence sometimes were 23 times more likely (OR: 22.94, 95%CI: 13.50 – 39.23), those with weekly faecal incontinence were 22 times more likely (OR: 21.65, 95%CI: 11.01 – 42.56) and those with daily faecal incontinence

Table 5: Faecal Incontinence assessed by the Modified Wexner (excluding Flatus), by Gender and Age Group, Percentages

Age group (years)	*Gender*	*Number*	*Wexner*						
			Any symptoms			*Classification if symptomatic*			
			None	*Any*	*95%CI*	*Slight*	*Moderate*	*Severe*	*Very severe*
15-19	Female	124	94%	6%	(2% – 10%)	6%	0%	0%	0%
	Male	130	98%	2%	(0% – 5%)	2%	0%	0%	0%
	All	253	96%	4%	(2% – 6%)	4%	0%	0%	0%
20-29	Female	231	96%	4%	(1% – 6%)	3%	1%	1%	0%
	Male	243	95%	5%	(2% – 5%)	5%	0%	0%	0%
	All	474	96%	4%	(3% – 6%)	4%	1%	0%	0%
30-39	Female	263	92%	8%	(4% – 11%)	5%	2%	0%	0%
	Male	267	94%	6%	(3% – 8%)	5%	0%	1%	0%
	All	529	93%	7%	(5% – 9%)	4%	0%	0%	0%
40-49	Female	278	92%	8%	(5% – 11%)	6%	0%	1%	1%
	Male	275	98%	2%	(1% – 4%)	2%	0%	0%	0%
	All	553	95%	5%	(3% – 7%)	8%	3%	0%	0%
50-59	Female	243	86%	14%	(10% – 19%)	10%	5%	0%	0%
	Male	239	93%	7%	(4% – 11%)	6%	1%	0%	1%
	All	482	89%	11%	(8% – 14%)	6%	1%	1%	2%
60-69	Female	162	89%	11%	(6% – 16%)	6%	2%	1%	3%
	Male	157	92%	8%	(4% – 12%)	6%	1%	0%	1%
	All	317	91%	9%	(6% – 12%)	10%	4%%	1%	1%%
70-79	Female	158	83%	17%	(11% – 23%)	10%	4%	1%	2%
	Male	126	85%	15%	(9% – 21%)	11%	5%	0%	0%
	All	284	84%	16%	(12% – 21%)	6%	3%	2%	4%
80+	Female	77	83%	17%	(9% – 25%)	5%	5%	3%	4%
	Male	44	91%	9%	(1% – 18%)	5%	1%	0%	0%
	All	122	85%	15%	(3% – 13%)	6%	2%	0%	1%
All	**Female**	**1536**	**90%**	**10%**	**(8% – 11%)**	**6%**	**2%**	**1%**	**1%**
	Male	**1481**	**94%**	**6%**	**(5% – 7%)**	**5%**	**1%**	**0%**	**0%**
All		**3014**	**92%**	**8%**	**(7% – 9%)**	**6%**	**2%**	**0%**	**1%**

The slight discrepancy in table numbers is because of missing data.

were 53 times more likely to report soiling (OR: 53.07, 95%CI: 26.94 – 104.06). These results suggest that soiling probably reflects a situation where a person has been unable to control their faecal incontinence through behaviour modification (e.g. where a person may be out on a social occasion and is unable to reach a toilet) or the use of aids

The relationship between soiling and urge was examined. This showed that, when compared with those who reported no faecal urge, those who reported at least monthly faecal urge were 12 times more likely to report soiling (N = 63 who reported both urge and soiling; OR: 12.65, 95%CI: 7.89 – 20.33), those who reported urge often were 30 times more likely to report soiling (N = 31; OR: 30.06, 95%CI: 16.41 – 55.16), and those who report daily urge were 70 times more likely to report soiling (N = 11; OR: 70.66, 95%CI: 24.25 – 209.46). When interpreting these findings,

Table 6: Faecal Incontinence assessed by Soiling, by Gender and Age Group, Percentages

Age group (years)	*Gender*	*Number*	*Soiling*						
			Any symptoms			*Classification if symptomatic*			
			None	*Any*	*95%CI*	*Slight*	*Moderate*	*Severe*	*Very severe*
15-19	Female	124	94%	6%	(2% – 10%)	6%	0%	0%	0%
	Male	129	99%	1%	(0% – 2%)	1%	0%	0%	0%
	All	253	97%	3%	(1% – 5%)	3%	0%	0%	0%
20-29	Female	230	97%	3%	1% – 5%)	3%	0%	0%	0%
	Male	244	100%	0%	N/A	0%	0%	0%	0%
	All	474	98%	2%	(0% – 3%)	1%	0%	0%	0%
30-39	Female	262	97%	3%	(1% – 6%)	3%	0%	0%	0%
	Male	267	96%	4%	(2% – 7%)	4%	0%	0%	0%
	All	529	96%	4%	(2% – 5%)	3%	0%	0%	0%
40-49	Female	277	95%	5%	(3% – 8%)	3%	1%	0%	0%
	Male	271	98%	2%	(1% – 3%)	2%	0%	0%	0%
	All	550	97%	4%	(2% – 5%)	3%	1%	0%	0%
50-59	Female	243	88%	12%	(8% – 16%)	11%	1%	0%	0%
	Male	237	93%	7%	(4% – 11%)	6%	1%	0%	0%
	All	479	91%	9%	(7% – 12%)	9%	1%	0%	0%
60-69	Female	161	90%	10%	(5% – 15%)	8%	1%	1%	0%
	Male	154	90%	10%	(5% – 14%)	10%	0%	0%	0%
	All	315	90%	10%	(7% – 13%)	9%	1%	1%	0%
70-79	Female	158	82%	18%	(12% – 24%)	14%	3%	2%	1%
	Male	127	84%	16%	(9% – 22%)	15%	1%	0%	0%
	All	285	83%	17%	(13% – 22%)	14%	2%	1%	1%
80+	Female	77	84%	16%	(8% – 24%)	8%	4%	4%	1%
	Male	43	93%	7%	(0% – 15%)	2%	0%	2%	0%
	All	121	87%	13%	(7% – 19%)	6%	3%	4%	1%
All	Female	1532	92%	8%	(7% – 9%)	6%	1%	1%	0%
	Male	1472	95%	5%	(4% – 6%)	4%	0%	0%	0%
All	3006	93%	7%(6% – 7%)		5%	1%	0%	0%	

The slight discrepancy in table numbers is because of missing data.

however, it should be borne in mind that the numbers were very small.

The key feature of the soiling data shown in Table 6 was that 93% of respondents indicated no soiling on either of the two items. The range was from 100% of males aged 20-29 years to 82% of females aged 70-79 years. Across these two questions soiling, where it was reported, was reported as occurring rarely. The exception was for older adults aged over 70 years. The highest proportion was for females aged 70-79 years (18%). Details can be found in Table 6.

3.8 Estimated Incontinence Prevalence

By combining the symptom data from the ISI and modified Wexner (see Tables 1 and 5) it was possible to estimate symptom incontinence prevalence for urinary, faecal and both conditions combined. Table 7 presents the results. This shows that the estimated prevalence of any incontinence was 27% (95%CI: 26% – 29%). For females it was 40% (38% – 43%) and for males 14% (12% – 15%). The prevalences for females were estimated to be 30% for symptoms of urinary continence only, 2% for symptoms of faecal continence only, and 8% for women suffering symptoms of both conditions. For males the estimated symptom prevalences were 8% for urinary continence only, 3% for faecal incontinence only, and 3% for both.

3.9 Discussion

This study reports data from the 2004 South Australian Health Omnibus Survey, a community survey weighted by Australian Bureau of Statistics data to achieve population representativeness. The data were collected during face-to-face interviews. As such the data are at the level of subjective symptom reports rather than confirmed diagnoses. This distinction is important because it profoundly affects the interpretation of the data. Previous literature has shown that the relationship between self-report and clinical assessment of incontinence is, at best, moderate (66).

The prevalence estimates provided here for any urinary incontinence range from 24% to 47%. For females it was from 38% to 60% and for males from 10% to 33%. For any faecal incontinence it was 35%; for females it was 38% and for males 32%. Regarding soiling, 7% of respondents reported this (8% of females and 5% of males). For any incontinence symptoms at all it was 27% overall, and 40% for females and 14% for males. Eight percent of females and 3% of males reported symptoms of both urinary and faecal incontinence.

These prevalence ranges suggest that the different measures used to assess urinary incontinence prevalence are not fully compatible. The ISI measures frequency and quantity, whereas the UDI-6 measures the bothersomeness of urinary incontinence symptoms upon quality of life. Following transformation to T-scores (43), the Spearman correlation between the two measures in this study was $r_s = 0.75$. More importantly, if the two scales are treated as different observations of the same phenomenon (urinary incontinence), and cases dichotomized (no symptoms/any symptoms) the level of agreement between the two measures was moderate ($\kappa = 0.53$) (63). There was agreement between the ISI and UDI-6 for just 77% of cases. Of the 23% of cases where there was disagreement in classification, 22% were where no symptoms were reported in the ISI but were classified as symptomatic on the UDI-6. Given that the ISI measures frequency and quantity and that the UDI-6 measures impact on wellbeing or quality of life (the UDI-6 stem is *"...how much are you bothered by..."*), it is difficult to avoid the conclusion that the UDI-6 is measuring conditions other than urinary incontinence.

The UDI-6 questions regarding the frequency of urination (Q1) and suffering pain or discomfort in the lower abdominal or genital area (Q6) may be gaining endorsement from non-urinary incontinence conditions. Evidence for this is that if these two questions are omitted from the UDI-6, agreement between the ISI and UDI-6 rises to 87% of all cases, and the kappa value to $\kappa = 0.71$. The prevalence rate would have been reported as 36% rather than the 47% for the full UDI-6. Given that the psychometric analysis of the UDI-6 at the beginning of this section suggested there were difficulties with two items (one of which was also identified in this comparison with the ISI), it would seem that there is a prima facie case for revision of the UDI-6.

There are also difficulties with the ISI. Table 3 reveals an anomaly in the scoring of the ISI because just 1% of cases fell within the 1st scale range decile (1-19%; those obtaining raw ISI scores of 2-3) compared with 25% of cases for the UDI-6. This is almost certainly an artefact of the multiplicative model of the ISI. To obtain a score in this range a respondent must have endorsed the best possible

Table 7: Estimated Incontinence Prevalence, by Gender and Age Group, Percentages

Age group (years)	Gender	Number	Incontinence status					
			Any symptoms			Incontinence type		
			None	Any	95%CI	Urinary only	Faecal only	Both
15-19	Female	123	89%	11%	(5% – 16%)	5%	0%	6%
	Male	129	95%	5%	(1% – 8%)	2%	2%	0%
	All	253	93%	7%	(4% – 11%)	4%	1%	3%
20-29	Female	230	82%	18%	(13% – 23%)	15%	1%	3%
	Male	243	90%	10%	(6% – 14%)	5%	5%	0%
	All	475	86%	14%	(11% – 17%)	10%	3%	2%
30-39	Female	262	58%	42%	(36% – 48%)	34%	2%	5%
	Male	266	91%	9%	(6% – 12%)	4%	5%	0%
	All	529	75%	25%	(22% – 29%)	19%	4%	3%
40-49	Female	278	55%	45%	(39% – 51%)	37%	1%	6%
	Male	274	93%	7%	(4% – 10%)	5%	1%	1%
	All	553	74%	26%	(22% – 30%)	21%	1%	4%
50-59	Female	243	43%	57%	(51% – 63%)	43%	2%	12%
	Male	238	80%	20%	(15% – 25%)	12%	3%	5%
	All	482	61%	39%	(34% – 43%)	28%	3%	9%
60-69	Female	161	48%	52%	(44% – 59%)	40%	4%	8%
	Male	157	82%	18%	(12% – 24%)	10%	5%	3%
	All	318	65%	35%	(30% – 40%)	26%	4%	5%
70-79	Female	158	56%	44%	(37% – 52%)	27%	4%	13%
	Male	127	70%	30%	(21% – 38%)	14%	3%	12%
	All	284	62%	38%	(32% – 43%)	22%	4%	12%
80+	Female	78	56%	44%	(33% – 55%)	27%	4%	14%
	Male	44	70%	30%	(16% – 43%)	21%	0%	9%
	All	121	61%	39%	(30% – 48%)	24%	3%	12%
All	Female	1533	60%	40%	(38% – 43%)	30%	2%	8%
	Male	1478	86%	14%	(12% – 15%)	8%	3%	3%
	All	3015	73%	27%	(26% – 29%)	19%	3%	5%

The slight discrepancy in table numbers is because of missing data.

category on one question and the slight incontinence option on the other question. For example, on the question *How often is urine leakage experienced?* a person could have selected *Never*, and on the question *How much urine was lost?* have endorsed *A few drops or Small splashes*. Alternately, they could have endorsed *<once a month* or *A few times a month* to the first question and have endorsed *None* to the second question. These combinations, however, are all inconsistent which may indicate a problem within the two question ISI. Although such problems may also exist within the UDI-6, because of the greater number of questions they are less obvious and will have had a smaller influence on the findings. This judgement reflects the classic psychometric axiom that item errors cancel each other out where there are multiple items and random samples. Generally, this implies that the minimum number of items needed to form a reliable scale is between three

to five (64, 67). With just two items, the ISI falls below this standard, although Moran et al have argued on an empirical basis that 2 items is the reductionist limit (68).

On balance, however, it seems likely that urinary incontinence prevalence rates from the ISI are to be preferred. These indicate that for any urinary incontinence the current prevalence was 24%, and that it was 38% for females and 10% for males.

For faecal incontinence, a very different scenario applied. The standard Wexner scale includes a question probing flatus, however this is excluded from the current ICS definition of faecal incontinence and this item was identified as problematic during the psychometric examination of the Wexner (see above). There were, therefore, good grounds for the removal of this item from the Wexner. Tables 4 and 5 present Wexner data with and without this item. The differences in prevalence estimates are striking: 35% versus 8%. Importantly, the estimate of 35% is higher than that of the ISI for urinary incontinence – a situation that is inconsistent with the incontinence literature.

It would seem, therefore, that the standard Wexner systematically overclassifies cases as having faecal incontinence due to the inclusion of flatus. Based on the data in Tables 4 and 5, the effect of this is to probably inflate faecal incontinence estimates by a factor of four. It is therefore recommended that the modified estimate excluding flatus is preferred. On the other hand, there is no mention in the Wexner of faecal urge incontinence, yet this study has shown that this is an important predictor of soiling. There is a prima facie case for revision of the Wexner scale so that it reflects the current definition of faecal incontinence and takes account of faecal urge incontinence.

The best estimate for urinary incontinence of all types based on the ICS definition for incontinence symptoms, namely the self-report of any symptoms of urinary leakage, would be the ISI estimated prevalence of urinary incontinence at 24% (95%CI: 23% – 26%) overall. When broken down by gender, it would be 38% (95%CI: 36% – 41%) for females and 10% (95%CI: 9% – 12%) for males. Based on the Wexner, but excluding flatus, for faecal incontinence the prevalence would be 8% (95%CI: 7% – 9%). For females this is 10% (95%CI: 8% – 11%) and for males 6% (5% – 7%).

In general, these findings are consistent with the other Australian prevalence estimates reported in the introduction. There are, however, two important considerations. The prevalence rates for older adults are lower than those for the middle-aged. Almost certainly this is to do with the sampling strategy: those in residential care would not have been recruited into the study. Thus the prevalences for older adults will systematically understate the true prevalence rate because incontinence is a known predictor of residential care (26, 56, 57). Second, it is possible that the findings may also be underreported due to embarrassment, since there is a general social reluctance to talk about incontinence (69).

Both these issues are important as potential predictors of the changes in reported symptoms of incontinence over time. If the rates of residential care for older adults increases, thereby removing from the population persons with a higher probability of being incontinent, then the estimated prevalence rates should decline. This, however, could be easily offset by increased health literacy making it more socially acceptable to report incontinence symptoms. The extent to which these factors affected the current study is unknown.

3.10 Recommendations

The incontinence prevalence estimates reported in the SAHOS are consistent with the literature in general and suggest that urinary incontinence is a common condition, particularly among females. To adequately quantify this for medical decision-making and policy direction, there is need for an excess burden of disease study. These data would also suggest the need for trials evaluating the relative impacts of preventive programs (e.g. pelvic floor exercises, health literacy) and acute interventions (e.g. surgery).

There is, however, considerable uncertainty over the measurement of incontinence. As this study has shown, none of the existing measures – whether for urinary or faecal incontinence – could be used with a great deal of confidence. Depending upon which instrument was used, or which items were included or excluded, there were very different prevalence estimates. This implies that all of the measures, to some degree, provided misleading estimates. It is recommended, therefore, that based on the SAHOS dataset a full psychometric evaluation of the measures is undertaken with the intent of developing better measures, and that these revised measures are then tested in future incontinence studies.

4. The utility of Incontinence

Utility was assessed by five different multi-attribute utility (MAU)-instruments, the AQoL, EQ5D, HUI3, 15D and SF6D. Basic descriptions of these instruments can be found in section 2.3; more detailed descriptions are contained in Appendix A. A literature review of studies reporting the use of utility measures in incontinence can be found in Appendix A. Table 8 provides a summary of the characteristics of each of the five MAU-instruments.

Table 8: Summary of Properties of MAU-instruments used in this Study, from the Published Literature

	AQoL	*EQ5D*	*HUI3*	*15D*	*SF6D*
Country of origin	Australia	UK	Canada	Finland	UK
Preference weights sample	Population	Population	Population	Population	Population
Coverage (a)	✓✓	✗	✓✓	✓✓	✓✓
Type of HRQoL emphasis (b)	Handicap	Impairment/ Disability	Impairment	Impairment/ Disability	Handicap
N. dimensions (c)	4	5	8	15	6
N. items (c)	12	5	12	15	11
N. health states (d)	>100,000	243	>100,000	>100,000	9,000
Utility preference method (e)	TTO	TTO	SG/VAS	VAS	SG
Combination rule	Multiplicative	Regression/ Additive	Multiplicative	Additive	Additive
Utility scale range (f)	-0.04 – 1.00	-0.59 – 1.00	-0.36 – 1.00	+0.11 – 1.00	+0.30 – 1.00
Other than English versions	✓	✓✓✓	✓✓	✓✓	✓✓✓
Completion time (minutes)	5-10	1-2	5-10	5-10	10-15
Construct validity evidence (g)	✓	✗	✓	✗	✓
Validation studies (h)	✓	✗	✓	✓	✗
Reliability evidence (i)	✓✓	✓	✓	✓✓	✓
Normative data (j)	✓	✓	✓	✗	✗
Sensitivity (k)	✓	✗	✓	✓	✓

Notes: Scoring system used in this table: ✗✗ = very poor, ✗ = poor/not available, ✓ = good or available, ✓✓ = very good, ✓✓✓ = excellent.

a = See Table 11.

b = Based on WHO classification of impairments and diseases (1)

c = Utility-contributing items only.

d = Where >100,000 deemed irrelevant for all practical purposes

e = SG = Standard gamble, TTO = Time trade-off, VAS = Visual analog scale

f = Lower and upper boundaries shown, where 0.00 = Death equivalent state and 1.00 = Best possible state.

g = Where the descriptive system was constructed following standard psychometric procedures for instrument construction.

h = Validation studies reported using psychometric procedures for instrument validation

i = Test-retest or internal consistency

j = Population norms published

k = Minimum important or clinically important differences published

Source: Adapted from Hawthorne et al. (44)

4.1 Introduction

In general, the literature comparing different utility instruments is disappointing because the emphasis has been firmly on comparing between two or three MAU-instruments assessing which instrument is the more/most sensitive to the health condition of interest (70-74) or the most practical to use (75-77). The implicit assumption is that greater sensitivity, responsiveness, or practicality indicates the 'better' instrument. For example, Stavem et al (78) compared the EQ5D, 15D and SF6D among HIV/AIDS patients and reported that there was no major difference between the measures, despite that fact that the 15D gave systematically higher scores. These assessments, however, ignore both psychometric and econometric requirements for utility measurement. Hawthorne et al (45) noted that the selection of instruments solely on the basis of sensitivity could lead to overstating the value of interventions during cost-utility analyses.

Several other research teams have examined the implication of differences between the MAU-instruments for cost-utility analysis (which is, after all the reason for the existence of these instruments) on grounds of scale range, instrument coverage, and the preference weights used. These studies show major differences between MAU-instruments (79-84). Consistent with Stavem et al (85), it would seem that choice of MAU-instrument has the potential to have a major impact on economic evaluations. For a discussion of how these issues may affect studies in incontinence, see Appendix A.

This brief summary of the comparative literature suggests the need for more consistent criteria to be used when comparing across MAU-instruments. Although rarely used, appropriate criteria for reliable and valid measurement have been outlined by Hawthorne and Richardson (45) and Brazier and Deverill (86). Although there were differences in emphasis by these two research teams, in general they both argued that the criteria were evidence of preference measurement, reliability, validity, responsiveness (sensitivity) and practicality.

A critical issue in comparing different MAU-instruments is that of scale range. The five different MAU-instruments compared here have five different life-death scoring ranges, as reported in Table 8 and further discussed in Appendix A. Although it has been argued that because of this different MAU-instruments cannot be directly compared without transformation onto a common scale (75), this argument is rejected here because the sine qua non of utility instruments is the calculation of QALYs (quality adjusted life years) for use in cost-utility analysis. This demands the assumption of both 'weak' and 'strong' interval properties (87); i.e. that the differences in obtained utilities from a condition or an intervention represent actual interval differences such that the difference between 0.60 and 0.80 -- a gain of 0.20 – implies that treating 5 people with this gain is the equivalent of treating 1 person in a death-equivalent HRQoL state (0.00) and returning him/her to best possible HRQoL (1.00) (88). This axiom of utility theory can be met only where MAU-instruments are scored on an inviolable life-death scale. Therefore the different ranges of the various MAU-instruments represent the extent to which each covers this underlying theoretical life-death scale: if there are disagreements between different MAU-instruments on the value gained as a result of an intervention, then one (or both) of the MAU-instruments is invalid. In this study multivariate outliers were identified and removed from the dataset when making MAU-instrument comparisons, but these outliers illustrate this issue. One of the removed cases obtained utility scores of 0.92 on the AQoL, 0.85 for the EQ5D, -0.01 for the HUI3, 0.70 for the 15D and 0.45 for the SF6D. Clearly these scores are inconsistent and some (or all) are obviously invalid. Transformation to ensure scale equivalent ranges for different MAU-instruments is a direct violation of this axiom because the life-death utility scale already possesses this property, i.e. standard MAU-instrument scores have already been transformed into a common scale – at least in theory.

4.1.1 Preference Measurement

The key issues for preference measurement relate to credible assumptions about the model used during elicitation of preferences, the method of valuation used to elicit preferences, whether elicited preferences are consist with the underlying preference model (for a discussion of this issue see Hawthorne et al (89)), the quality of the data obtained (including missing data, evidence

[3] Multi-attribute utility (MAU) instruments are generic instruments designed for use across all life conditions that may impact on health-related quality of life. Therefore they must be able to measure the effect of all types of HRQoL interventions, from, say, a quit smoking health promotion campaign to surgery for heart/lung transplant. This implies that the full range of the life-death scale must be measured, from death equivalent to best possible HRQoL states.

of response bias, variation in the data, inconsistent valuations) and the method of interpolating planned missing values for intermediate health states. Additionally, there are issues concerning the combination rule for deriving the utility index, double-counting, and evidence of both the weak and strong interval measurement. Finally, the utility algorithm must provide coverage of the full spectrum of HRQoL values, from full health states to values representing states worse than death.[3]

A review of MAU-instruments against these axioms was presented in the Thomas et al report (27). The relevant section of that report can be found in Appendix A of this report. In general, this review suggests important differences between the MAU-instruments, particularly with respect to the 15D which uses a visual analog scale for the utility weights.

Because this study is based on an examination of completed instruments scored using the standard scoring algorithms, the issues listed under this criterion are not considered further here. The interested reader is referred to Appendix A.

4.1.2 Reliability

This refers to the stability of the measurement either over time (where no important event has taken place between administrations) or between two different instruments measuring the same thing. The former, test-retest, is usually assessed using intra-class correlation (ICC) or correlation (r), and the latter, equivalent tests or internal consistency, by correlation between different measures of the same construct which are administered at the same time or by using a test of the within-instrument item relationships, most commonly Cronbach α (alpha) for scaled response sets.

For the AQoL, Hawthorne (90), using random population sampling and mail/telephone comparisons reported the test-retest ICC = 0.83. Based on community and hospital samples, the range of internal consistency estimates for the AQoL have been reported between Cronbach α = 0.73 – 0.84 (6, 91-96).

EQ5D test-retest reliability at 2-week interval among those with chronic obstructive pulmonary disease was ICC = 0.78 (85). In a study of stroke patients, Dorman et al (97) reported test-retest reliability estimates for the EQ5D of ICC = 0.83. In a Dutch population study of the EQ5D where test-retest was carried out at 10-month interval, the test-retest reliability correlation coefficient was reported as r = 0.90 (98). At 1-week test-retest the ICC for the EQ5D was reported to be 0.70 for those with osteoarthritis (99).

Table 9: Summary Table of Reported Utility Instrument Reliability

Instrument	*Cronbach* α	*Test-retest*	
		r	*ICC*
AQoL	0.73-0.84	0.80	0.83
EQ5D	0.69	0.73	0.67-0.90
HUI3	0.74-0.81	0.77	0.75-0.91
15D	0.84	0.90-0.94	–
SF6D	–	0.88	0.94

For the 15D, Stavem et al (85) reported that the 2-week test-retest for the 15D was r = 0.90 among those with chronic obstructive pulmonary disease. Sintonen (52) reported that the groups' means between samples provided a reliability estimate of Spearman r_s = 0.94.

For the HUI3, Furlong et al reported an ICC = 0.91 for the construction sample (49), which was identical to that reported by Ruiz et al (100). Test-retest reliability assessed using telephone administration at a one month interval was ICC = 0.77 (10, 101). In a study involving children (>9.5 years) the ICC was reported to be >0.50 (102). An ICC = 0.75 at 7-day interval was reported for those with rheumatic disease (103). Internal consistency for the HUI3 was reported for a French cross-cultural adaptation study at Cronbach α = 0.81 (104). Cronbach α = 0.79 was reported in the Spanish validation study (100). The reliability of case completion versus proxy completion has

been reported to be ICC = 0.60–0.70 (105), suggesting that cases and proxies complete the HUI3 somewhat differently. Regarding different modes of administration, a Dutch study indicated that self-completion is to be preferred (106).

For the SF6D, based on 2-week test-retest Stavem et al (85) reported the ICC = 0.94 for those with chronic obstructive pulmonary disease. Test-retest at 3-months for those whose health state was stable was reported by Conner-Spady and Suarez-Almazor to be 0.88 (79).

A summary of reliability estimates for the different MAU-instruments is presented in Table 9. This study presents further evidence on MAU-instrument reliability.

4.1.3 Validity

This is the extent to which scores on a measure represent the underlying model the measure is supposed to be assessing.[4] Generally, three different types of validity are described:

- Content validity refers to adequate coverage of the dimensions of HRQoL deemed to be important. These are usually defined as comprising at least physical, mental, social and somatic sensations (eg. pain).[5]
- Construct validity is where scores on an instrument are interpreted as representative of a latent construct, such as HRQoL. The ideal situation is where there is an isomorphic relationship between the manifest instrument scores and the degree to which respondents possess the latent construct.[6] This implies there is a latent construct that the researcher has defined, and that this definition is an adequate description of what is being measured.
- Criterion validity relates to the relationship between scale scores and either other independent measures (criteria) or other specific measures (predictors). *Concurrent validation* is where the criterion data are collected at the same time as the instrument data (e.g. in a cross-sectional survey). *Convergent/divergent validation* is where a measure is explicitly tested against similar measures of the same construct or against measures of other constructs that are not related to the construct of interest.

The review of utility instruments in Appendix A provides a detailed assessment of the validity of MAU-instruments against these criteria. Other reviews of their validity can be found in the three papers by Hawthorne et al (45, 107, 108), and in the review by Brazier et al (109).

Content validity is not discussed further here; instead the reader is referred to Appendix A where there is a substantive review of MAU-instrument content. Tests of construct and criterion validity are presented.

4.1.4 Responsiveness (Sensitivity) & Practicality

Instruments must be sensitive to the states of interest. This is usually tested by examining standardised differences in scores by different levels of the state of interest. Practicality describes how easy it is to use an instrument, including how long it takes respondents to complete a measure, the response rate and the missing data rate.

A general review of MAU-instruments against these criteria is given in Appendix A, based on instrument examination and the literature. Other reviews can be found in the three validation papers published by Hawthorne et al (45, 107, 108), and in the review by Brazier et al (109).

This report tests responsiveness through examination of utility scores by differing levels of incontinence.

---

[4] Because scores on an instrument are a function of both the instrument descriptive system and respondent endorsement of particular item response categories, validity can never be established. What can be established is that this descriptive system when used in this sample of people exhibits these characteristics. It is assumed that the identified characteristics are transferable to other settings and populations or samples. This implies that each time a measure is used in a different setting, population or sample basic validity tests should be performed to verify that the measure is appropriate.

[5] Because MAU-instruments are generic, they must be able to be used in all types of HRQoL interventions, as noted above (footnote #3).

[6] If no adequate construct is defined, the content of the instrument defines the construct that is being measured.

4.2 Methods

Regarding missing data, the highest proportion was for the HUI3 (22/3015 cases), the 15D (8 cases) and the SF6D (1 case). There were no missing data for the AQoL or EQ5D. These low levels of missing data are not regarded as important, although it should be noted that these missing data levels cannot be taken as representative of missing data on these instruments in general since all data were collected during interview. Missing data at the item-level were imputed using horizontal mean substitution, which has been recommended for within scale imputation (53).

Since all five MAU-instruments' scores were skewed – as is expected since most people have a good quality of life – multivariate outliers were detected based on Mahalanobis distance and deleted for the comparative analyses (64). Likewise, for comparative analysis all cases with missing data on one or more measures were deleted from the dataset, thus ensuring both equal sample sizes and equivalent characteristics across all five MAU-instruments. For the reasons outlined above in the introduction concerning transformations, the data were not transformed. Tabachnick and Fidell (64) report that provided there is sufficient sample size (defined as at least 20 cases in the smallest analysis cell) and that the sample sizes being compared are equal in numbers, F-tests are robust. Both these conditions applied in this study.

Instrument reliability was assessed by Cronbach α, and the precision of α assessed.

The relationship between the five MAU-instruments was examined using principal component analysis. Structural equation modelling (SEM) was used to examine the extent to which each MAU-instrument measurement model matched with the SAHOS data. After Byrne (110), because of the extreme skew in individual items, bootstrapping was used to supplement the dataset. Depending on the resulting matrix, either maximum likelihood (ML) or asymptotic distribution free (ADF) models were used. Because these analyses were testing instrument models, modification indices to improve model fit were not used. Two fit statistics were used to assess the models. The adjusted goodness-of-fit (AGFI) estimate was used to assess the proportion of variance in the dataset explained by the model, and the root mean square error of approximation (RMSEA) was used to assess the model fit against a perfect (saturated) model. For assessing AGFI model fit the conventional criteria of >0.90 was accepted, while for the RMSEA values <0.05 indicated good fit and that models <0.08 were acceptable (64, 110, 111). Correlations between utility instruments were assessed through the Pearson ρ and the significance of correlation was assessed by Cohen's q. To compare the sensitivity of MAU-instruments, Cohen's d (112) for related samples, viz.,

$$d = \frac{m_A - m_B}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}}$$

where m_A and m_B represent the mean score of the two samples and σ_A and σ_B represent the standard deviations on the assumption that the sample standard deviations equal the population standard deviations. Cohen provided the following classification for interpreting d: 0.20 = a small effect, 0.50 = a moderate effect, and 0.80 = a large effect.

The relative efficiency (RE) statistic was used to quantify differences between measures (113).

Data analyses were carried out in SPSS V13 (114), AMOS V4 (111), PRISM V4 (115) and InStat V3 (116).

4.3 Results

4.3.1 Reliability

Table 10: MAU-instrument Reliability

	N. items	*Inter-item correlation*		*Standardized*
		Mean	*Range*	*Cronbach* α
AQoL	15	0.23	0.05 – 0.80	0.82
EQ5D	5	0.33	0.12 – 0.59	0.71
HUI3	15	0.16	0.01 – 0.82	0.74
15D	15	0.24	0.02 – 0.61	0.83
SF6D	11	0.35	0.05 – 0.73	0.85

The reliability of each of the five MAU-instruments administered in the 2004 SAHOS is given in Table 10. This shows that the items forming the SF6D and EQ5D had the highest average inter-item correlations, whereas the lowest was reported for the HUI3. The standardized Cronbach α was just within the accepted range for the EQ5D and HUI3, and within the accepted range for group data for the other instruments. None of the instruments met the standards that are accepted for individual assessment (117).

A feature of MAU-instruments is the extent to which they are multi-dimensional. Generally, where multi-dimensionality is present, Cronbach α will perform poorly because it reflects inter-item correlations. This may explain the modest reliability results for the EQ5D and HUI3. If all inter-item correlations are identical, then a scale would have just one vector and would be, by definition, unidimensional. Under this axiom it follows that α is a function of the principal component in a scale. If an instrument measures several vectors, then there will be a corresponding reduction in estimated reliability. Calculation of the precision of α enables an estimate of whether there is a single general factor (or a single latent construct) that a scale is measuring (61). Precision of 0.00 reflects the situation where all items are identically correlated. The further the departure from 0.00, the greater the presence of multidimensionality based on greater variability in inter-item correlation.

The precision estimates for each of the MAU-instruments were computed. For the AQoL, precision of α was 0.01, for the EQ5D it was 0.05, for the HUI3 it was 0.01, for the 15D 0.02, and for the SF6D 0.02. These findings suggest that that AQoL, HUI3, 15D and SF6D are unidimensional in the sense that they are each measuring a single underlying construct. From this it follows that the lower

Cronbach α for the HUI3 is not a function of instrument multi-dimensionality, but may reflect an inconsistency within the instrument itself. The findings for the EQ5D, however, suggest that it is not measuring a single underlying construct, despite the fact that it has the highest average inter-item correlation. Rather it is multi-dimensional: items 1, 3 and 4 (mobility, usual activities and pain) correlated r >0.50 whereas items 2 and 5 (self-care and anxiety) correlated with the other items r<0.35. It is also the least reliable of the five MAU-instruments.

The conclusion is that the AQoL, 15D and SF6D – which have similar numbers of items to the HUI3 – provide more reliable measurement than either the HUI3 or EQ5D.

4.3.2 Validity

Construct Validity

Table 11, adapted from Hawthorne & Richardson (45), provides an overview of those aspects of life that have been deemed to contribute to the construct HRQoL. As shown in the table, the extent to which the different MAU-instruments reported in this study cover this latent construct varies considerably. The issue for construct validity is to assess the extent to which the five MAU-instruments measure an underlying common construct, which would be presumed to be 'health-related quality of life'.

Table 11: Content of MAU-instruments (a)

HRQoL dimensions (b)	*AQoL*	*EQ5D*	*HUI3*	*15D*	*SF6D*
Relative to the body					
Anxiety/depression/distress	*	*		**	**
Bodily care	*	*			*
Cognitive ability			*	*	
General health					
Memory			*		
Mobility	*	*	*	*	**
Pain	*	*	*	*	*
Physical ability/vitality/disability			*	*	*
Rest and fatigue	*			*	**
Sensory functions	**		****	*****	
Social expression					
Activities of daily living	*	*		*	*
Communication	*		**	*	
Emotional fullfilment			*		
Environment					
Family role	*				
Intimacy/Isolation	*				
Medical aids use					
Medical treatment					
Sexual relationships				*	
Social function	*				*
Work function					*

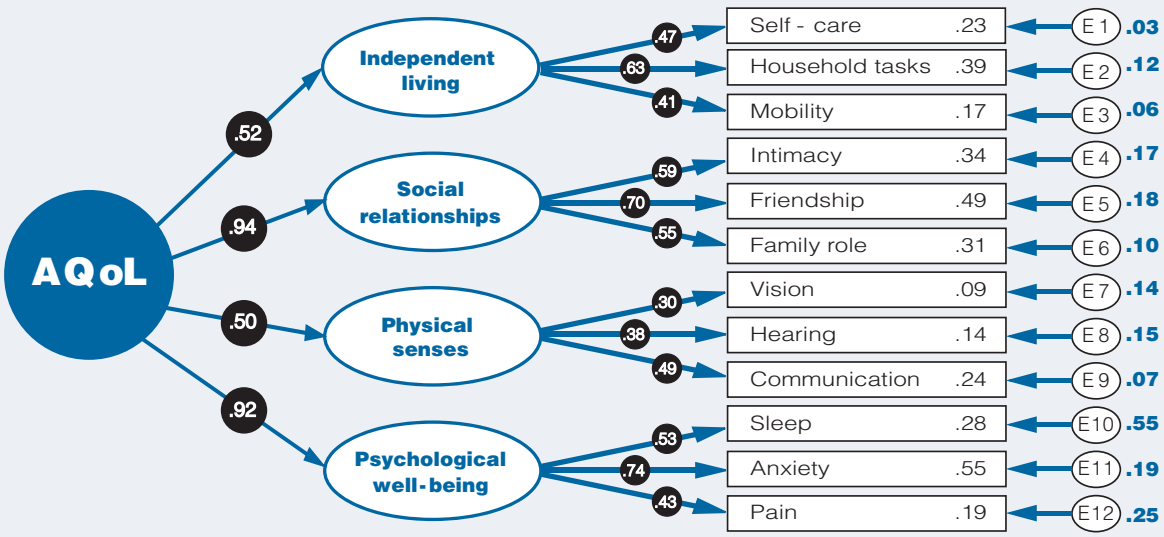
Note: a = Table shows only those items used in calculation of utility scores.
Each asterisk represents an item. Based on item content examination.
b = Dimensions of HRQoL defined by a review of 14 HRQoL instruments, 1971–1993.

Source: Adapted from Hawthorne et al. (44)

To test for this, the five instrument utility scores were entered into a principal component analysis. The results showed a single factor explaining 76% of the variance (eigenvalue = 3.81). The loadings on the principal component were: 0.90 (AQoL), 0.88 (HUI3), 0.88 (15D), 0.87 (EQ5D) and 0.84 (SF6D). The five MAU-instruments are thus measuring a single underlying construct.

Each instrument was then separately examined using structural equation modelling (SEM) as described in the methods section. The results are presented in Figures 2, 3, 4, 5, and 6. The best fitting model was the AQoL (RMSEA: 0.04), then the SF6D and the HUI3 (0.07 respectively), the EQ5D (0.08). The only instrument with an unacceptable fit was the 15D (0.12). Regarding the explanatory power of the instruments, the instrument with the highest AGFI was the EQ5D (AGFI: 0.97), the HUI3 (0.93), the SF6D (0.90) and the AQoL (0.89). Unsurprisingly, given the poor RMSEA fit, the AGFI for the 15D was also poor (AGFI = 0.78).

Figure 2: SEM of the AQoL (Utility-Contribution Items only)

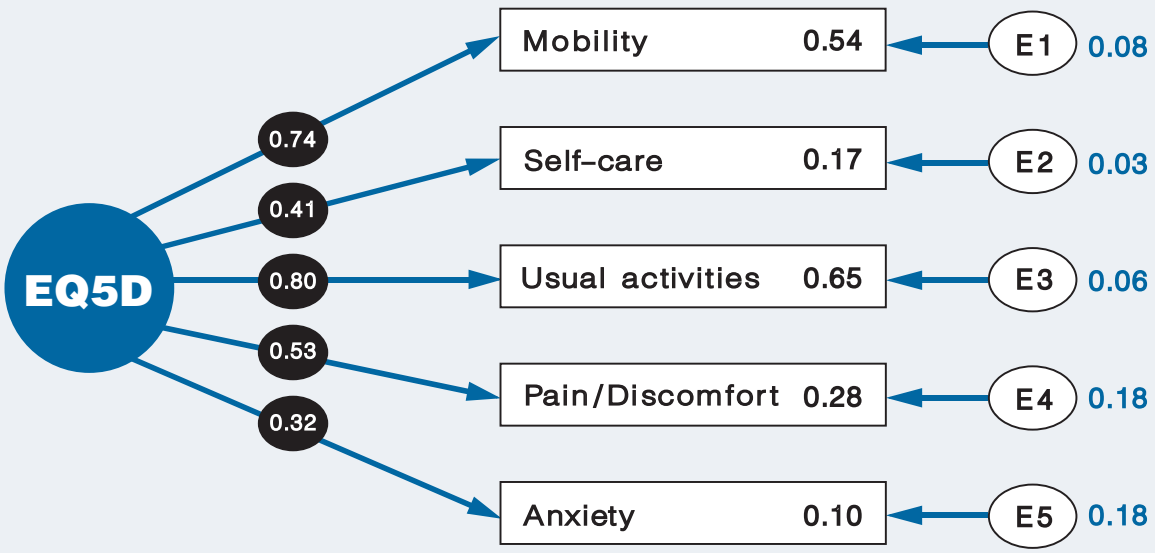


Model shows standardized regression weights and squared multiple correlations.

Statistics: N = 3015, model = ADF, χ^2 = 320.53, df = 50, p < 0.01
AGFI = 0.892, RMSEA = 0.042 (95%CI: 0.038 – 0.047)
Hoelter, p = 0.05, N = 636

In terms of understanding the meaning of these analyses the results suggest that, in the SAHOS sample, the AQoL utility was primarily influenced by social relationships and psychological wellbeing, the EQ5D by usual activities and mobility, the HUI3 by cognition and pain, the 15D by vitality and usual activities, and the SF6D by limitations in functional role and physical capacity.

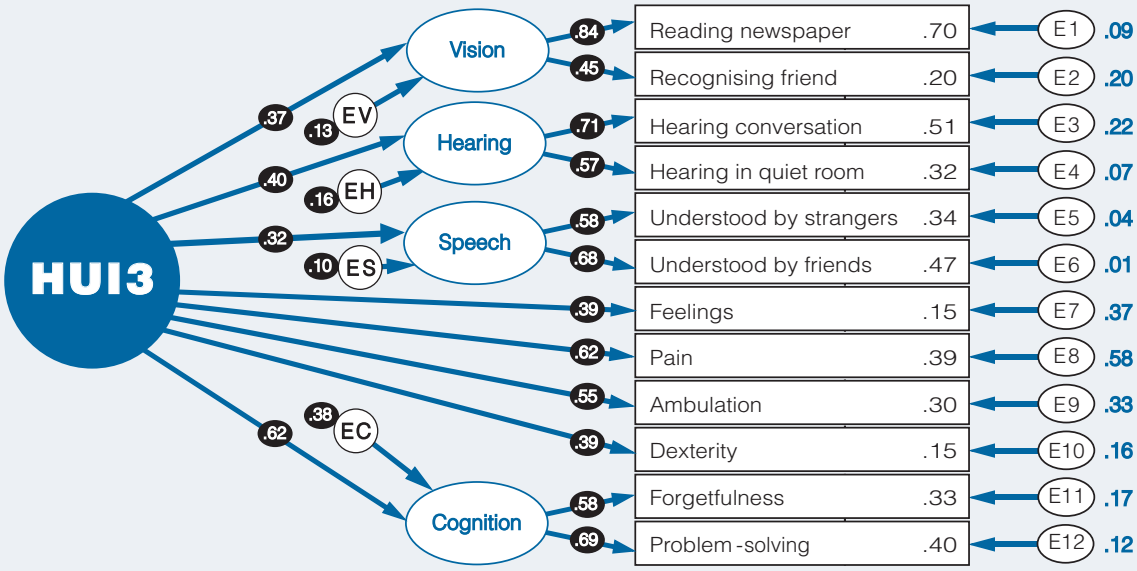
Figure 3: SEM of the EQ5D



Model shows standardized regression weights and squared multiple correlations.

Statistics: N = 3015, model = ML, χ^2 = 90.90, df = 5, p < 0.01
AGFI = 0.966, RMSEA = 0.075 (95%CI: 0.062 – 0.089)
Hoelter, p = 0.05, N = 368

Figure 4: SEM of the HUI3 (Utility-Contributing Items Only)



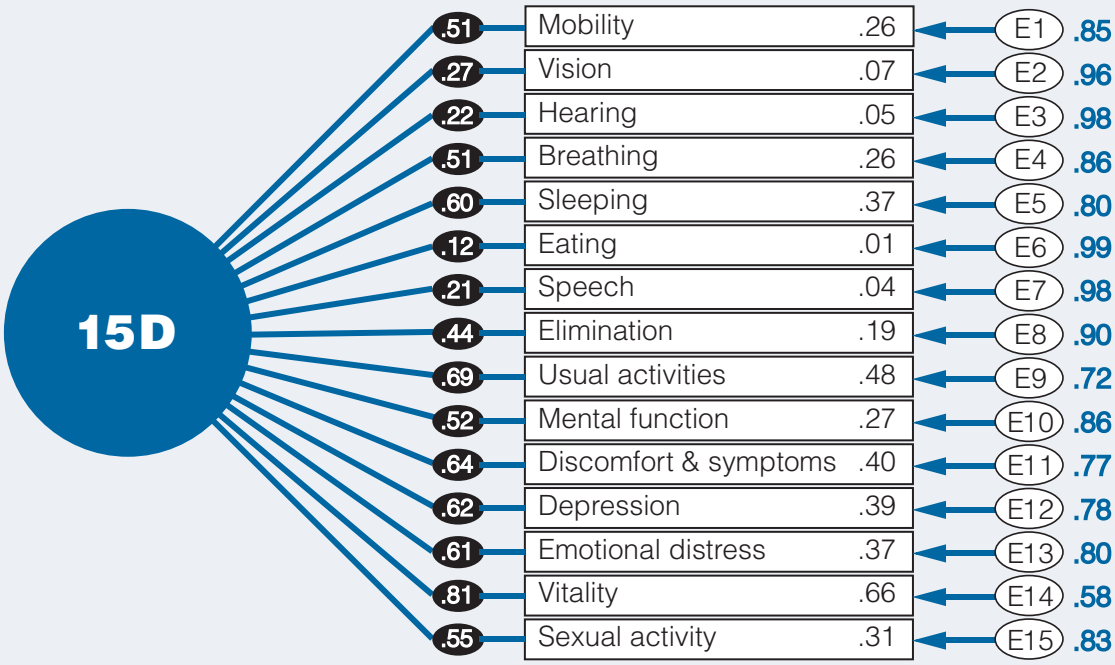
Model shows standardized regression weights and squared multiple correlations.

Statistics: N = 3015, model = ML, χ^2 = 854.97, df = 50, p < 0.01

AGFI = 0.926, RMSEA = 0.073 (95%CI: 0.069 – 0.077)

Hoelter, p = 0.05, N = 238

Figure 5: SEM of the 15D



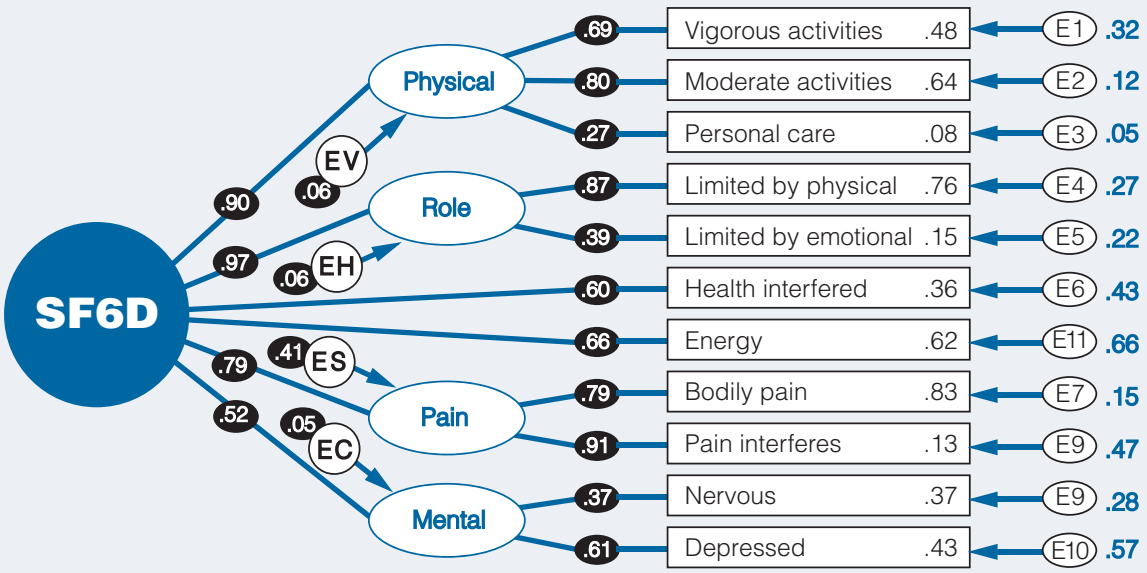
Model shows standardized regression weights and squared multiple correlations.

Statistics: N = 3015, model = ML, χ^2 = 3711.04, df = 90, p<0.01

AGFI = 0.782, RMSEA = 0.12 (95%CI: 0.11 – 0.12)

Hoelter, p = 0.05, N = 92

Figure 6: SEM of the SF6D (Utility-Contributing Items Only)



Model shows standardized regression weights and squared multiple correlations.

Statistics: N = 3015, model = ADF, χ^2 = 560.66, df = 40, p < 0.01
AGFI = 0.896, RMSEA = 0.066 (95%CI: 0.061 – 0.071)
Hoelter, p = 0.05, N = 300

Criterion Validity: the Relationship between the Utility Measures

Table 12 shows the correlation between the five MAU-instruments, while Table 13 shows statistically significant differences in the correlation-pairs, based on Cohen's q. This analysis suggests that there are significant differences in the slopes of the utility scores for the different instruments across the utility score range. The AQoL was significantly more highly correlated with the 15D than with the EQ5D or the SF6D; the HUI3 was significantly more highly correlated with the AQoL than with the SF6D, it was more highly correlated with the 15D than with the EQ5D and the SF6D.

This analysis is, of course, based on the assumption of linearity. Furthermore, it does not reveal which utility instruments report similar utility values. Subject to these two caveats, it would suggest that the divergent instruments (in terms of greatest variation in utility slope across the utility score range) would be the SF6D and the 15D. The scatterplot in Figure 7 illustrates this phenomenon. It shows the relationship between the HUI3 and the 15D. The dotted line represents the 'perfect' relationship if the utility scores from the two instruments were isomorphic. The solid line shows the line of best fit (r = 0.76). Despite this high correlation between the two measures (r = 0.75), as the utility scores deteriorate Figure 7: Scatterplot of the 15D and the HUI3 utility scores

Table 12: Correlations between the Utility Instruments

	AQoL	*EQ5D*	*HUI3*	*15D*
EQ5D	0.73			
HUI3	0.75	0.68		
15D	0.73	0.69	0.75	
SF6D	0.69	0.69	0.65	0.65

Notes: Statistics: Pearson r.
All correlations significant, p <0.01

Figure 7: Scatterplot of the 15D and the HU13 Utility Scores

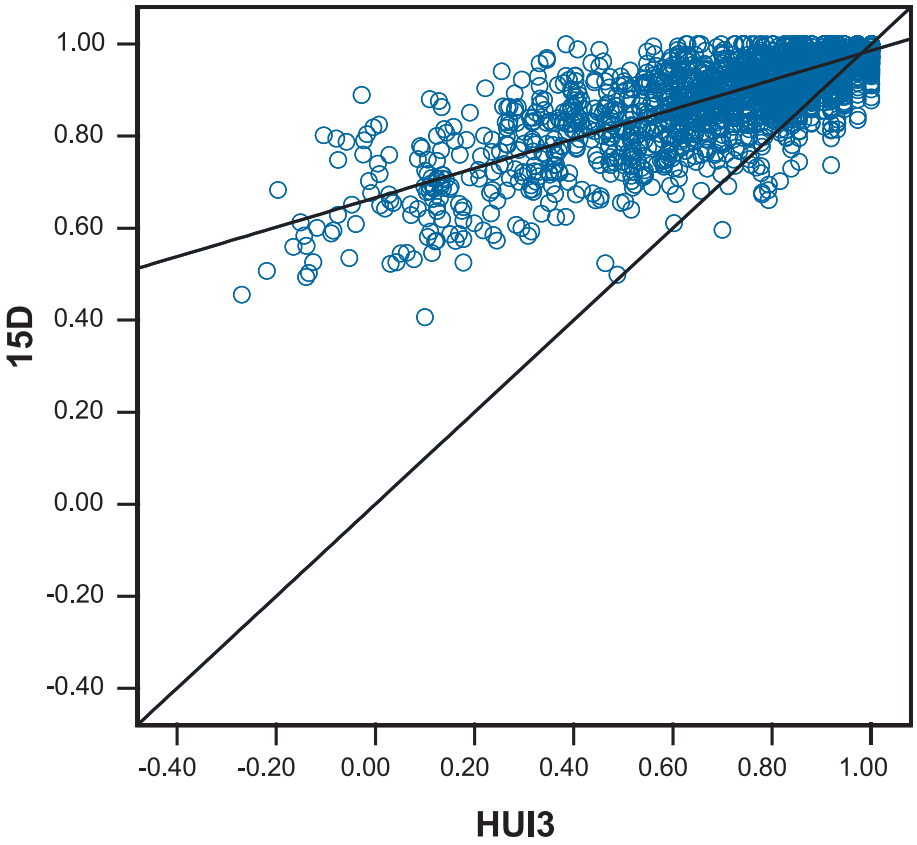


Table 13: Cohen's θ Analysis of the Correlations between Utility Instruments, Table 12

	AQoL/ EQ5D	*AQoL/ HUI3*	*AQoL/ 15D*	*AQoL/ SF6D*	*EQ5D HUI3*	*EQ5D 15D*	*EQ5D SF6D*	*HUI3/ 15D*	*HUI3/ SF6D*	*15D/ SF6D*
AQoL	–	–	–	–	0.02	0.09*	0.08	0.07	0.10	0.16*
EQ5D	–	0.06	0.02	0.04	–	–	–	0.08	0.02	0.06
HUI3	0.08	–	0.07	0.17*	–	0.14*	0.09*	–	–	0.23*
15D	0.08	0.00	–	0.23*	0.07	–	0.06	–	0.13*	–
SF6D	0.04	0.07	0.04	–	0.11*	0.00	–	0.11*	–	–

Notes: * = Cohen's θ; for statistically significant differences in correlations between pairs, $\theta \geq 0.09$.

(i.e. indicate worse HRQoL states), there is ever less agreement between the two measures (indicated by the increasing divergence between the line of best fit and the 'perfect' line). The effect of this is easily seen: for the 15D there are no cases with scores in the range <0.40 utilities, whereas for the HUI3 6.4% of all cases were classified with utility scores <0.40.

Table 14 quantifies the effect of the different utility ranges by reporting the number of cases within utility deciles for each of the MAU-instruments. If it is assumed that there should be a monotonic decline in the number of cases from best HRQoL (0.90-1.00) to worst HRQoL (<0.00) on the grounds that progressively fewer people suffer greater levels of illness, then the MAU-instruments with reasonable data distributions would be the AQoL and the HUI3. The difficulty with the EQ5D is the large inconsistent number of cases within the intervals 0.81-0.90, 0.51-0.60, 0.41-0.50 and 0.31-0.40. For example, for the EQ5D 0.3% of cases fell within the range 0.41-0.50. Although the data distribution for the 15D was monotonic, it classified 74% of cases within the top decile (0.91-1.00), which was greatly in excess of that of any other instrument (the instrument with the next highest proportion in the top decile was the HUI3 with 51% of cases). The effect of the truncated ranges for the 15D and the SF6D is apparent: very few cases were classified below 0.50 on either of these (0.0% for the 15D and 3.2% for the SF6D)

Table 14: Distribution of MAU-instrument Utility Scores, by Utility Decile

Utility deciles	MAU-instrument				
	AQoL	EQ5D	HUI3	15D	SF6D
<0.00	2	30	15		
0.01-0.10	24	37	15		
0.11-0.20	33	49	47		
0.21-0.30	41	39	37		2
0.31-0.40	66	20	74		14
0.41-0.50	89	8	79		61
0.51-0.60	141	32	134	19	208
0.61-0.70	232	322	222	64	424
0.71-0.80	413	915	264	184	517
0.81-0.90	689	257	573	493	887
0.91-1.00	1250	1272	1522	2219	883

The slight discrepancy in table numbers is because of missing data.

4.4 Population Utility

Population norms for each of the MAU-instruments are presented in Table 15, broken down by age group and gender. For examining differences between the instruments, the data for each age group are summarized in Figure 8. For each MAU-instrument there were statistically significant differences by age groups by gender (ANOVA, F_{range} = 128.89 to 261.49, $p < 0.01$ for all comparisons).

Effect sizes for each instrument over time were computed (youngest cohort, 15-19, versus oldest cohort, 80+). The results showed that the HUI3 was most sensitive to age group (d = 1.05), followed by the 15D (0.93), the AQoL (0.85), EQ5D (0.81) and the SF6D (0.62).

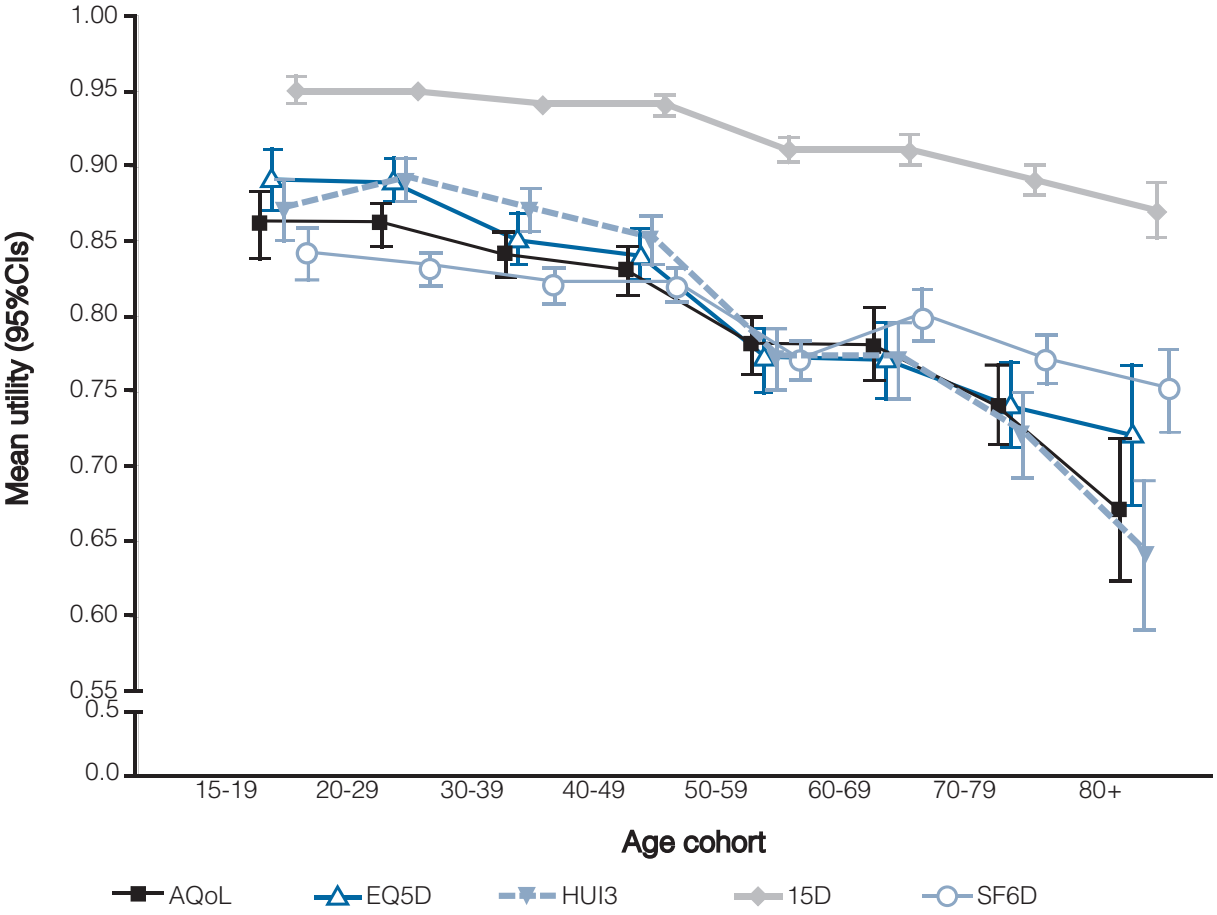
Regarding differences within age groups, however, the data showed that the 15D consistently assigned scores that were significantly higher than those of any other instrument. The other instrument with a different trajectory over the age cohorts was the SF6D. For the younger age cohorts it assigned significantly lower utility scores when compared with the EQ5D (15-19 cohort), AQoL and HUI3 (20-29 age cohort), the HUI3 and 15D (30-39 age cohort), but statistically higher scores when compared with the AQoL, EQ5D and HUI3 (60-69 age cohort), the HUI3 (70-79 age cohort) and the AQoL and HUI3 (80+ age cohort). Almost certainly the reason for this compression of scores on the SF6D across the age groups relates to the limited range of scores available; hence when compared with other MAU-instruments, as mean utility scores decline it will assign comparatively higher scores.

Table 15: Population norms for Selected MAU-instruments, by Gender and Age Group

Age group (years)	Gender	N	MAU-instrument													
			AQoL			EQ5D			HUI3			15D			SF6D		
			M	sd	95%CI	M	sd	95%CI	M	sd	95%CI	M	sd	95%CI	M	sd	95%CI
15-19	Male	129	0.90	0.13	0.88 – 0.92	0.95	0.09	0.94 – 0.97	0.93	0.10	0.91 – 0.95	0.97	0.04	0.97 – 0.99	0.90	0.11	0.88 – 0.92
	Female	123	0.82	0.21	0.78 – 0.86	0.84	0.19	0.81 – 0.87	0.82	0.21	0.78 – 0.86	0.93	0.09	0.92 – 0.95	0.78	0.14	0.76 – 0.81
	All	251	0.86	0.18	0.84 – 0.88	0.89	0.16	0.87 – 0.91	0.87	0.17	0.85 – 0.89	0.95	0.07	0.94 – 0.96	0.84	0.14	0.82 – 0.86
20-29	Male	244	0.89	0.14	0.87 – 0.91	0.91	0.13	0.89 – 0.93	0.91	0.12	0.90 – 0.93	0.97	0.05	0.97 – 0.98	0.85	0.12	0.84 – 0.87
	Female	230	0.84	0.18	0.82 – 0.86	0.87	0.17	0.85 – 0.89	0.87	0.17	0.85 – 0.89	0.94	0.07	0.93 – 0.95	0.80	0.14	0.78 – 0.82
	All	474	0.86	0.16	0.85 – 0.88	0.89	0.15	0.88 – 0.90	0.89	0.15	0.88 – 0.90	0.95	0.06	0.95 – 0.96	0.83	0.13	0.82 – 0.84
30-39	Male	266	0.84	0.18	0.82 – 0.86	0.85	0.22	0.82 – 0.88	0.87	0.17	0.85 – 0.89	0.95	0.07	0.94 – 0.96	0.84	0.13	0.83 – 0.86
	Female	260	0.85	0.16	0.83 – 0.87	0.86	0.17	0.84 – 0.88	0.88	0.16	0.86 – 0.90	0.94	0.07	0.93 – 0.95	0.81	0.13	0.80 – 0.83
	All	526	0.84	0.17	0.83 – 0.86	0.86	0.20	0.84 – 0.88	0.87	0.17	0.86 – 0.89	0.94	0.07	0.94 – 0.95	0.83	0.13	0.82 – 0.84
40-49	Male	269	0.85	0.16	0.83 – 0.87	0.87	0.18	0.85 – 0.89	0.87	0.17	0.85 – 0.89	0.95	0.07	0.94 – 0.96	0.84	0.12	0.83 – 0.86
	Female	275	0.83	0.21	0.81 – 0.86	0.82	0.24	0.80 – 0.85	0.83	0.21	0.81 – 0.86	0.93	0.09	0.92 – 0.94	0.80	0.15	0.78 – 0.82
	All	544	0.84	0.19	0.82 – 0.85	0.84	0.21	0.82 – 0.86	0.85	0.19	0.83 – 0.87	0.94	0.08	0.93 – 0.95	0.82	0.14	0.81 – 0.83
50-59	Male	235	0.78	0.21	0.75 – 0.81	0.79	0.24	0.76 – 0.82	0.78	0.22	0.75 – 0.81	0.92	0.09	0.91 – 0.93	0.78	0.15	0.76 – 0.80
	Female	243	0.78	0.20	0.76 – 0.81	0.75	0.24	0.72 – 0.78	0.77	0.24	0.74 – 0.80	0.90	0.10	0.89 – 0.91	0.76	0.15	0.74 – 0.78
	All	478	0.78	0.21	0.76 – 0.80	0.77	0.24	0.75 – 0.79	0.77	0.23	0.75 – 0.79	0.91	0.09	0.90 – 0.92	0.77	0.15	0.76 – 0.78
60-69	Male	152	0.79	0.21	0.76 – 0.82	0.78	0.23	0.74 – 0.82	0.75	0.23	0.71 – 0.79	0.92	0.09	0.91 – 0.94	0.81	0.15	0.79 – 0.83
	Female	160	0.77	0.23	0.74 – 0.81	0.77	0.22	0.74 – 0.80	0.78	0.23	0.75 – 0.82	0.91	0.09	0.90 – 0.93	0.79	0.15	0.77 – 0.81
	All	312	0.78	0.22	0.76 – 0.81	0.77	0.23	0.75 – 0.80	0.77	0.23	0.75 – 0.80	0.91	0.09	0.90 – 0.92	0.80	0.15	0.78 – 0.82
70-79	Male	125	0.79	0.20	0.76 – 0.83	0.79	0.19	0.76 – 0.82	0.75	0.23	0.71 – 0.79	0.91	0.08	0.90 – 0.93	0.79	0.13	0.77 – 0.81
	Female	154	0.69	0.24	0.65 – 0.73	0.70	0.27	0.66 – 0.74	0.69	0.26	0.65 – 0.73	0.88	0.10	0.87 – 0.90	0.75	0.15	0.73 – 0.77
	All	279	0.74	0.23	0.71 – 0.77	0.74	0.24	0.71 – 0.77	0.72	0.25	0.69 – 0.75	0.89	0.10	0.88 – 0.90	0.77	0.14	0.75 – 0.79
80+	Male	41	0.75	0.22	0.68 – 0.82	0.78	0.24	0.71 – 0.85	0.71	0.25	0.63 – 0.79	0.89	0.09	0.86 – 0.92	0.79	0.16	0.74 – 0.84
	Female	74	0.63	0.27	0.57 – 0.69	0.69	0.25	0.63 – 0.75	0.60	0.26	0.54 – 0.66	0.86	0.09	0.84 – 0.88	0.73	0.14	0.67– 0.76
	All	115	0.67	0.26	0.62 – 0.72	0.72	0.25	0.68 – 0.77	0.64	0.26	0.59 – 0.69	0.87	0.10	0.86 – 0.88	0.75	0.15	0.72 – 0.78
Total	Male	1459	0.83	0.18	0.82 – 0.84	0.85	0.20	0.84 – 0.86	0.84	0.19	0.83 – 0.85	0.94	0.07	0.94 – 0.95	0.83	0.14	0.82 – 0.84
	Female	1519	0.79	0.21	0.78 – 0.80	0.80	0.23	0.79 – 0.81	0.80	0.22	0.79 – 0.81	0.92	0.09	0.92 – 0.93	0.79	0.15	0.78 – 0.80
	All	2978	0.81	0.20	0.80 – 0.82	0.82	0.22	0.81 – 0.83	0.82	0.21	0.81 – 0.83	0.93	0.08	0.93 – 0.93	0.81	0.14	0.81 – 0.82

The slight discrepancy in table numbers is because of missing data.

Figure 8 : Summary of Population norms for five MAU-instruments, by Age Cohort



Statistics: RMANOVA: Huynh-Feldt adjustment. Between instruments F = 679.94, p < 0.01. Age-cohort F= 42.42, p < 0.01. For within age groups analysis adjusting for the number of comparisons: Tukey-Kramer Multiple Comparisons Test, for p ≤ 0.05, q >3.86:

15-19: AQoL ≠ EQ5D*; AQoL ≠ 15D***; EQ5D ≠15D***; EQ5D ≠ SF6D***; HUI3 ≠ 15D***; 15D ≠ SF6D***.

20-29: AQoL ≠ EQ5D*; AQoL ≠ HUI3*; AQoL ≠ 15D***; AQoL ≠ SF6D***; EQ5D ≠ 15D***; EQ5D ≠ SF6D***; HUI3 ≠ 15D***; HUI3 ≠ SF6D***; 15D ≠ SF6D***.

30-39: AQoL ≠ HUI3*; AQoL ≠ 15D***; EQ5D ≠ 15D***; HUI3 ≠ 15D***; HUI3 ≠ SF6D***; 15D ≠ SF6D***.

40-49: AQoL ≠ 15D***; EQ5D ≠ 15D***; HUI3 ≠ 15D***; 15D ≠ SF6D***.

50-59: AQoL ≠ EQ5D*; AQoL ≠ HUI3*; AQoL ≠ 15D***; AQoL ≠ SF6D***; EQ5D ≠ 15D***; HUI3 ≠ 15D***; 15D ≠ SF6D***.

60-69: AQoL ≠ 15D***; AQoL ≠ SF6D***; EQ5D ≠ 15D***; EQ5D ≠ SF6D***; HUI3 ≠ 15D***; HUI3 ≠ SF6D***; 15D ≠ SF6D***.

70-79: AQoL ≠ 15D***; EQ5D ≠ 15D***; HUI3 ≠ 15D***; HUI3 ≠ SF6D*; 15D ≠ SF6D***.

80+: AQoL ≠ 15D***; AQoL ≠ SF6D*; EQ5D ≠ 15D***; EQ5D ≠ HUI3*; HUI3 ≠ 15D***; HUI3 ≠ SF6D**; 15D ≠ SF6D***.

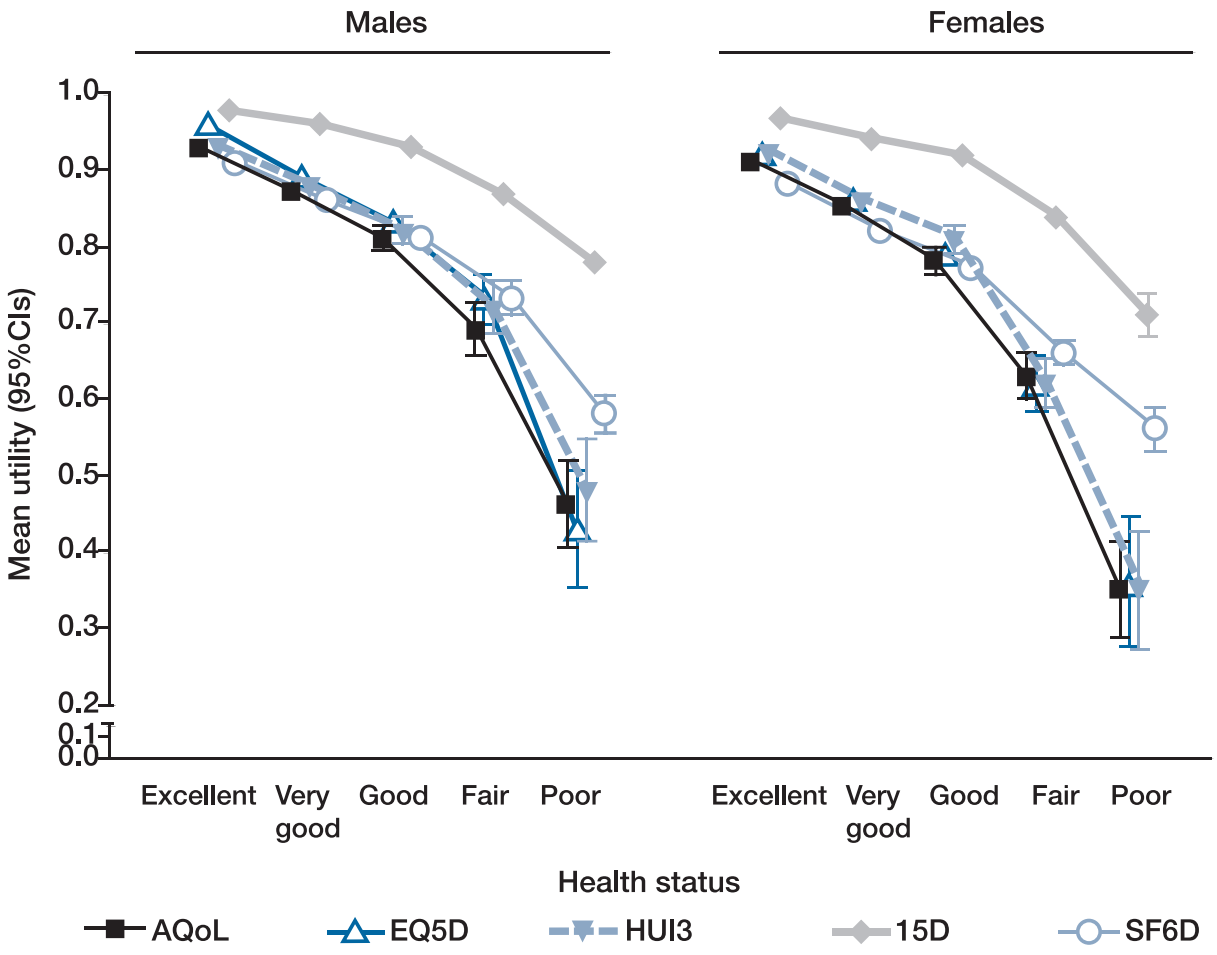
* = p< 0.05, ** = p< 0.01, *** = p < 0.001.

Apart from the statistically significant differences shown in the notes to Figure 8, there were also differences in monotonicity. Table 15 and Figure 8 show that at the age cohort level the AQoL, EQ5D and 15D assigned monotonically declining scores, whereas for the HUI3 those aged 20-29 were assigned utility scores higher than those aged 15-19, and for the SF6D those aged 60-69 were assigned scores higher than those aged 50-59.

Whether these differences matter is important because the results suggest two important issues. First, that none of the MAU-instruments possesses the 'weak' interval property which is a pre-requisite for calculating quality adjusted life years (QALYs) (see Appendix A for a discussion of this requirement). This conclusion is not based on the variable change scores per se, but on the differential relationships between the five measures. If the instruments possessed interval properties, then it would be expected that there would be parallel changes in scores by age group. This did not occur. Second, these findings suggest that under some circumstances, the choice of instrument would play a critical part in determining the cost-utility ratio from an intervention. Obviously the best way of resolving this would be to test each MAU-instrument in a range of conditions. Since this, however, is not generally feasible, the health status measure from the SF-36V2 was used as the criterion, and the data were examined by gender. The results are given in Figure 9.

This analysis shows that all five MAU-instruments were responsive to health status, showing monotonically increasing declines in utility scores as reported health status also declined (ANOVA, F_{range} = 128.80 – 274.99). When responsiveness to health status (by gender) was examined, the average relative efficiency (RE) of each of the five instruments was 1.07 for the HUI3, 1.10 for the SF6D, 1.14 for the EQ5D, 1.27 for the AQoL and 1.53 for the 15D. For males the REs were 1.00 for the HUI3, 1.20 for the EQ5D, 1.21 for the SF6D, 1.23 for the AQoL and 1.46 for the 15D; for females they were 1.00 for the SF6D, 1.08 for the EQ5D, 1.13 for the HUI3, 1.31 for the AQoL and 1.60 for the 15D. Thus, on average, the 15D and AQoL were the most responsive instruments to differences in self-reported health status.

Figure 9: Utility Value from Five MAU-instruments by Health Status, by Gender



4.5 Sensitivity of MAU-instruments to Incontinence Status

Although the population norms presented in section 4.4 are essential information for those using the MAU-instruments in that they provide benchmark data against which study findings can be interpreted, they provide limited information regarding which utility instruments would be preferred in studies of incontinence.

Three incontinence measures were included in the SAHOS, as described in part 3 of this report: the UDI-6, ISI and Wexner. To assess the impact of incontinence on people's lives, the MAU-instruments described above were examined by incontinence status as determined by these three measures.

4.5.1 The impact of Urinary Incontinence on HRQoL

The score range on the UDI-6 was from 0 through 18, where the higher scores indicate increasing urinary incontinence severity. Because of the small numbers with severe urinary incontinence (there were just 31 cases with scores >10), UDI-6 scores were recoded into the standard UDI-6 classifications: 0 = no symptoms, 1 = slight, 2 = moderate, 3 = problem, and 4 = major problem. Similar data distribution problems were evident for the ISI, which was also recoded, viz.: 0 = no symptoms, 1 = slight, 2 = moderate, 3 = severe/very severe.

Table 16: The impact of Urinary Incontinence as assessed by the UDI-6 on HRQoL, by Gender

UDI-6 status	*Gender*	*N.*	*MAU-instrument*									
			AQoL		*EQ5D*		*HUI3*		*15D*		*SF6D*	
			M	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
None	Male	986	0.87	0.16	0.89	0.17	0.88	0.15	0.96	0.06	0.85	0.13
	Female	610	0.86	0.18	0.86	0.19	0.86	0.19	0.95	0.07	0.82	0.14
	All	1596	0.86	0.17	0.88	0.18	0.87	0.17	0.95	0.06	0.84	0.13
Slight	Male	384	0.78	0.20	0.79	0.22	0.78	0.21	0.92	0.08	0.79	0.13
	Female	570	0.80	0.18	0.81	0.20	0.81	0.20	0.92	0.08	0.79	0.14
	All	955	0.79	0.19	0.80	0.20	0.80	0.20	0.92	0.08	0.79	0.14
Moderate	Male	62	0.72	0.21	0.78	0.20	0.69	0.27	0.88	0.10	0.75	0.15
	Female	241	0.71	0.24	0.73	0.24	0.73	0.25	0.88	0.10	0.74	0.14
	All	303	0.71	0.23	0.74	0.23	0.72	0.25	0.88	0.10	0.74	0.14
Problem	Male (a)	24	0.60	0.25	0.56	0.32	0.47	0.30	0.78	0.11	0.64	0.16
	Female	55	0.61	0.27	0.59	0.32	0.61	0.26	0.82	0.12	0.67	0.15
	All	73	0.60	0.26	0.58	0.32	0.58	0.27	0.81	0.12	0.66	0.15
Major problem	Male		N/A		N/A		N/A		N/A		N/A	
	Female	40	0.55	0.31	0.56	0.39	0.54	0.34	0.77	0.13	0.67	0.17
	All	46	0.56	0.31	0.56	0.38	0.52	0.34	0.77	0.13	0.66	0.14

The slight discrepancy in table numbers is because of missing data.

Notes: a = For males includes 'Major problem' since N = 6 cases.

Statistics:	ANOVA, F-values. All $p < 0.001$					
	Male	47.87	47.16	83.33	111.38	51.99
	Female	52.67	43.49	46.73	85.96	34.57
	All	96.22	87.34	107.83	183.15	85.31
Relative efficiency						
	Male	1.02	1.00	1.77	2.36	1.10
	Female	1.52	1.26	1.35	2.49	1.00
	All	1.13	1.02	1.26	2.15	1.00

The impact of urinary incontinence on HRQoL as measured by the UDI-6 is shown in Table 16. Whilst this shows that all five MAU-instruments delivered monotonically declining utility scores as incontinence status deteriorated, the relative efficiency statistic shows that there were differences in instrument sensitivity by incontinence status by gender. For females, the least sensitive instrument was the SF6D, followed by the EQ5D, HUI3, AQoL and I5D. For males, the least to most sensitive instruments were the EQ5D, AQoL, SF6D, HUI3 and 15D. For all cases, the least to most sensitive instruments were the SF6D, the EQ5D, AQoL, HUI3 and 15D. Particularly important was the failure of the SF6D to discriminate between females and all cases for those classified as having urinary incontinence problems and major problems (the mean SF6D score was 0.67 for both these groups). This may have implications for its use in clinical trials.

The second urinary incontinence measure in the study was the ISI. Table 17 shows utility scores by ISI urinary incontinence status. As with the UDI-6, there was a monotonic decline in utility for all MAU-instruments. The RE statistic suggested that for males the least sensitive measure was the EQ5D, followed by the SF6D, AQoL, HUI3 and 15D. For females the least sensitive measure was the SF6D, then the EQ5D, HUI3, AQoL and 15D. For all cases, the least sensitive measure was the SF6D, then the EQ5D, AQoL, HUI3 and 15D.

Table 17: The impact of Urinary Incontinence as assessed by the ISI on HRQoL, by Gender

ISI status	*Gender*	*N.*	*MAU-instrument*									
			AQoL		*EQ5D*		*HUI3*		*15D*		*SF6D*	
			M	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
None	Male	1310	0.84	0.17	0.86	0.19	0.85	0.18	0.95	0.07	0.84	0.13
	Female	943	0.83	0.19	0.83	0.20	0.84	0.20	0.93	0.08	0.81	0.14
	All	2253	0.84	0.18	0.85	0.20	0.85	0.19	0.94	0.07	0.82	0.14
Slight	Male	118	0.74	0.22	0.76	0.26	0.72	0.25	0.89	0.10	0.77	0.16
	Female	422	0.77	0.21	0.78	0.22	0.78	0.23	0.91	0.09	0.77	0.14
	All	541	0.77	0.21	0.77	0.23	0.77	0.23	0.90	0.09	0.77	0.15
Moderate	Male (a)	28	0.66	0.21	0.71	0.22	0.62	0.26	0.82	0.10	0.69	0.12
	Female	114	0.70	0.26	0.72	0.28	0.73	0.26	0.86	0.11	0.74	0.15
	All	138	0.70	0.25	0.72	0.26	0.72	0.25	0.86	0.11	0.73	0.14
Severe/ Very severe	Male		N/A		N/A		N/A		N/A		N/A	
	Female	37	0.52	0.28	0.54	0.35	0.49	0.28	0.79	0.12	0.66	0.15
	All	41	0.52	0.27	0.53	0.34	0.47	0.28	0.78	0.12	0.66	0.15

The slight discrepancy in table numbers is because of missing data.

Notes: a = For males includes 'Severe/Very severe' (N = 4).

Statistics:	ANOVA, F-values. All p < 0.001					
	Male	30.33	22.99	47.02	71.00	29.99
	Female	41.57	31.95	40.70	61.30	21.90
	All	72.43	61.62	79.45	129.16	55.87
Relative efficiency						
	Male	1.32	1.00	2.05	3.09	1.31
	Female	1.90	1.46	1.86	2.80	1.00
	All	1.30	1.10	1.42	2.31	1.00

4.5.2 The Impact of Faecal Incontinence on HRQoL

The score range on the Wexner was from 0 through 16, where the higher scores indicate increasing faecal incontinence severity. Because of the small numbers with severe faecal incontinence (there were just 33 cases with scores >6), Wexner scores were recoded: 0 = never, 1 = rarely, 2 = sometimes, 3 = weekly and 4 = daily faecal incontinence.

Table 18: The Impact of Faecal Incontinence as assessed by the Wexner on HRQoL, by Gender

Wexner status	*Gender*	*N.*	*MAU-instrument*									
			AQoL		*EQ5D*		*HUI3*		*15D*		*SF6D*	
			M	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
Never	Male	1003	0.85	0.17	0.87	0.19	0.86	0.17	0.95	0.06	0.85	0.13
	Female	953	0.83	0.19	0.83	0.20	0.83	0.21	0.93	0.08	0.80	0.14
	All	1956	0.84	0.18	0.85	0.20	0.85	0.19	0.94	0.07	0.83	0.14
Rarely	Male	274	0.81	0.20	0.81	0.22	0.81	0.22	0.93	0.08	0.81	0.14
	Female	317	0.78	0.20	0.79	0.22	0.79	0.22	0.91	0.09	0.77	0.14
	All	591	0.79	0.20	0.80	0.22	0.80	0.22	0.92	0.08	0.79	0.14
Sometimes	Male	109	0.79	0.18	0.83	0.14	0.80	0.22	0.91	0.08	0.79	0.14
	Female	154	0.71	0.23	0.73	0.25	0.72	0.26	0.87	0.11	0.74	0.15
	All	263	0.74	0.22	0.77	0.22	0.75	0.25	0.89	0.10	0.76	0.15
Weekly	Male	45	0.72	0.24	0.73	0.24	0.75	0.23	0.89	0.09	0.76	0.16
	Female	57	0.68	0.29	0.68	0.28	0.69	0.27	0.86	0.12	0.74	0.17
	All	102	0.70	0.27	0.70	0.26	0.72	0.25	0.87	0.11	0.75	0.16
Daily	Male	27	0.66	0.21	0.67	0.27	0.64	0.29	0.85	0.11	0.70	0.15
	Female	39	0.57	0.26	0.60	0.33	0.65	0.26	0.82	0.12	0.68	0.16
	All	66	0.61	0.24	0.63	0.31	0.64	0.27	0.84	0.12	0.69	0.15

The slight discrepancy in table numbers is because of missing data.

Notes:

Statistics: ANOVA, F-values. All $p < 0.001$						
	Male	17.38	16.43	17.44	29.06	17.73
	Female	28.41	20.96	18.94	38.20	13.13
	All	47.36	37.80	36.38	69.66	32.17
Relative efficiency						
	Male	1.06	1.00	1.06	1.77	1.08
	Female	2.16	1.60	1.44	2.91	1.00
	All	1.47	1.18	1.13	2.17	1.00

Table 18 shows the impact of faecal incontinence on HRQoL by each of the five MAU-instruments by gender. There was a monotonic decline in utility value by increasing faecal incontinence for all instruments, with the exception of the EQ5D for males where those classified as being sometimes incontinent obtained higher EQ5D scores than did those classified as being rarely incontinent. When the instruments were compared using the RE statistic, for males the least sensitive measure was the EQ5D, then the AQoL, HUI3, SF6D and 15D. For females the least sensitive measure was the SF6D, then the HUI3, EQ5D, AQoL and 15D. Across all cases, the least to most sensitive measures were the SF6D, and then the HUI3, EQ5D, AQoL and the most sensitive was the 15D.

Table 19: The Impact of Soiling on HRQoL, by Gender

Wexner status	Gender	N.	MAU-instrument									
			AQoL		EQ5D		HUI3		15D		SF6D	
			M	sd	M	sd	M	sd	M	sd	M	sd
Never	Male	1386	0.84	0.18	0.86	0.20	0.85	0.18	0.95	0.07	0.83	0.13
	Female	1399	0.81	0.20	0.81	0.21	0.82	0.21	0.92	0.08	0.79	0.14
	All	2785	0.83	0.19	0.84	0.21	0.83	0.20	0.93	0.08	0.81	0.14
Symptoms	Male	69	0.69	0.22	0.70	0.22	0.64	0.26	0.85	0.10	0.71	0.14
	Female	119	0.62	0.27	0.65	0.29	0.65	0.27	0.84	0.12	0.71	0.15
	All	188	0.64	0.25	0.67	0.27	0.65	0.27	0.84	0.11	0.71	0.15

The slight discrepancy in table numbers is because of missing data.

Notes:

Statistics: ANOVA, F-values. All $p < 0.001$						
	Male	46.89	44.06	79.02	120.42	54.78
	Female	93.01	62.29	61.22	101.28	39.89
	All	150.71	115.38	141.29	226.59	98.91
Relative efficiency						
	Male	1.06	1.00	1.79	2.73	1.24
	Female	2.39	1.60	1.57	2.60	1.00
	All	1.52	1.17	1.43	2.29	1.00

The effect of soiling of clothes on HRQoL was examined through the two questions on soiling. The data distributions were poor: 34 cases (5 of whom were males) reported soiling sometimes, often or always. Soiling was therefore dichotomized into 0 = never, 1 = symptoms (any soiling). The results are presented in Table 19. This shows that females were significantly more likely to report soiling when compared with males ($\chi^2 = 12.03$, df = 1, $p < 0.01$). For males, the least sensitive measure was the EQ5D, then the AQoL, SF6D, HUI3 and 15D. For females the SF6D was the least sensitive measure, followed by the HUI3, EQ5D, AQoL and 15D. Overall, the least sensitive measure was the SF6D, EQ5D, HUI3, AQoL and 15D.

The results presented in Tables 16 to 19 are suggestive of differences in sensitivity among the MAU-instruments by gender and incontinence status. The effect of gender was further investigated through relative efficiency (RE) analysis of the F-values across the four incontinence measures. The RE statistics were computed for females, where males were the denominator in all tests. The results showed that there were statistically significant differences by sensitivity to gender for the AQoL versus the HUI3, the 15D and the SF6D, and the EQ5D versus the SF6D. These findings suggest that there may be a gender effect by MAU-instrument. Based on this analysis, for incontinence the AQoL may be more sensitive to females when compared to the HUI3 and SF6D (both of which may be more sensitive to males) and the 15D (which appears to have no gender bias). Likewise, the EQ5D may also favour females when compared with the SF6D. The details are given in Table 20.

4.5.3 Predicting Utility from Incontinence Status

Each of the utility measures was regressed on each of the incontinence measures. The results suggested that although the MAU-instruments were sensitive to incontinence state, as shown in the previous section, incontinence state predicted small to moderate changes in HRQoL.

With the exception of the 15D, urinary incontinence as measured by the UDI-6 explained between 8% to 15% of the variance in utility scores. This can be compared with the ISI, which explained

Table 20: Relative Efficiency Analysis of five MAU-instruments, by Gender

	Utility instrument					
	UDI	*ISI*	*Wexner*	*Soiling*	*Mean*	*SD*
AQoL	1.10	1.37	1.63	1.98	1.52	0.33
EQ5D	0.92	1.39	1.28	1.41	1.25	0.20
HUI3	0.56	0.87	1.09	0.77	0.82	0.19
15D	0.77	0.86	1.31	0.84	0.95	0.21
SF6D	0.66	0.73	0.74	0.73	0.72	0.03

Notes: One way ANOVA, F = 9.41, df = 4,15, p = <0.01

Tukey-Kramer Multiple Comparisons Test, q > 4.37, p < 0.05.

AQoL vs. HUI3, q = 6.53, p < 0.01

AQoL vs. 15D, q = 5.32, p < 0.05

AQoL vs. SF6D, p < 0.01

EQ5D vs. SF6D, q = 4.92, p < 0.05

between 2% to 7%. The reason for these differences is almost certainly to do with the inclusion in the UDI-6 of questions probing the consequences of urinary incontinence, whereas the ISI is more a measure of incontinence per se. The UDI-6 and ISI explained between 10% to 21% of the variance in 15D scores; almost certainly a function of the elimination question in the 15D.

For faecal incontinence a similar pattern was observed. The Wexner explained between 5% and 10% of HRQoL, except for the 15D where 11% to 13% of utility variance was explained. For soiling, however, the proportion of explained variance was between 3% to 7% for all utility measures. The details are given in Table 21.

Table 21: Predicting the Impact of Incontinence on HRQoL, by Gender

Type	*Gender*	*MAU-instrument*									
		AQoL		*EQ5D*		*HUI3*		*15D*		*SF6D*	
		r^2	β *(a)*	r^2	β *(a)*	r^2	β *(a)*	r^2	β *(a)*	r^2	β *(a)*
UDI-6	Male	0.08	-0.29	0.08	-0.28	0.15	-0.38	0.19	-0.43	0.09	-0.30
	Female	0.13	-0.37	0.11	-0.33	0.13	-0.36	0.20	-0.45	0.09	-0.30
	All	0.12	-0.35	0.10	-0.33	0.14	-0.37	0.21	-0.46	0.10	-0.32
ISI	Male	0.03	-0.18	0.02	-0.15	0.05	-0.22	0.08	-0.28	0.03	-0.16
	Female	0.07	-0.26	0.05	-0.22	0.07	-0.26	0.10	-0.32	0.03	-0.18
	All	0.05	-0.24	0.04	-0.21	0.06	-0.25	0.10	-0.31	0.03	-0.18
Wexner	Male	0.05	-0.23	0.05	-0.22	0.07	-0.26	0.11	-0.33	0.06	-0.24
	Female	0.10	-0.31	0.07	-0.27	0.07	-0.27	0.13	-0.36	0.05	-0.21
	All	0.08	-0.28	0.07	-0.26	0.07	-0.27	0.13	-0.36	0.05	-0.23
Soiling	Male	0.03	-0.18	0.03	-0.18	0.03	-0.18	0.03	-0.18	0.03	-0.18
	Female	0.07	-0.27	0.06	-0.23	0.06	-0.23	0.06	-0.23	0.06	-0.23
	All	0.06	-0.24	0.05	-0.22	0.05	-0.22	0.05	-0.22	0.05	-0.22

Notes: a = Standardized coefficient

4.6 Discussion

This study compared five leading MAU-instruments in a general population sample with particular analysis by incontinence condition.

The principal components analysis, between instruments correlations (Table 12) and SEM analyses (Figures 2, 3, 4, 5 and 6) suggest that although the five MAU-instruments measure the same underlying construct, they are measuring different vectors within that construct. This finding was fully consistent with that of Hawthorne et al (107) in their comparison of MAU-instruments.

Regarding the different emphases by instruments, the findings from this study are consistent with those from the earlier Hawthorne et al study (107). Like that earlier analysis, the AQoL was found to have a strong emphasis on social relationships and psychological wellbeing, whereas the EQ5D is heavily influenced by usual activities and mobility. In this study, the HUI3 was found to emphasize cognition and pain, compared with Hawthorne et al's (107) finding that it emphasized cognition, ambulation and pain. Both studies found that the 15D was heavily influenced by vitality and usual activities. The biggest difference between the findings from this study and Hawthorne et al's study was for the SF6D. This study found that the SF6D was mainly influenced by functional role and physical capacity; whereas Hawthorne et al reported it was mainly influenced by social and physical functioning.[7]

Other than the earlier Hawthorne et al study (107), there have been almost no studies of construct validation of MAU-instruments. Generally, the literature reflects instrument developers' claims about the instrument in question. For example, the developers of the HUI3 referred to it as measuring a 'within the skin' perspective to capture functional capacity, thus rendering it particularly suitable for use in clinical trials (10, 48) – a perspective that Richardson and Zumbo reported was simple physical impairment rather than HRQoL (118).

A second key issue for construct validity is preference measurement and the available scale range. For the reasons outlined in the introduction to this section, and in Appendix A, MAU-instruments must be weighted using preference-based techniques. As noted in Appendix A, the 15D has not been so weighted; it's validity as an instrument suitable for use in cost-utility analysis must therefore be questioned. There are additional issues around negative utilities which are of concern for the EQ5D and HUI3, as discussed in Appendix A. Although these matters were not explicitly explored in this study, the instrument developers' decisions on these matters played an important role in the study findings. For example, the restriction in the scale range for the SF6D particularly affected its scores by age group and health status (see Table 14 and Figures 8 and 9). This restriction may also have been responsible for the finding that the SF6D was the least sensitive instrument to incontinence status (Tables 16 through 19).

Regarding the test of criterion validity, viz., examination of the distribution of utility scores against the utility life-death scale, difficulties were observed in the distribution of scores for the 15D, the SF6D and the EQ5D.

A key finding, which may help to explain why the 15D scores were so different to those of the other utility instruments, was that there was an unacceptable fit between the 15D's measurement model and the obtained data (Figure 5). An important reason for this lack of fit relates to the poor standardized regression coefficients for eating, speech, hearing and vision in the 15D; the inclusion of these in the descriptive model confounds measurement, particularly in a population sample where difficulties with these aspects of life are rare (e.g. most people who cannot feed themselves will be in institutional care). The lack of fit of these items to the 15D model is compounded by its scoring algorithm: this is a simple additive model which predetermines the allowable disutility for any item to a maximum of -0.067. The implication is that unless a health state affects all or most items of the 15D, the utility score will always be high. As shown in Table 15 and Figure 8 the consequence is that the 15D will consistently understate disutility. The general conclusion from the construct validation tests used in this study would suggest that there are particular difficulties with the 15D.

For the SF6D, the observed difficulties are a function of the restricted scoring range. The lower boundary is 0.30, imposed because the absolute lower boundary was limited in the underlying theoretical model to 0.00. The effect is to spread scores out within the available bandwidth with

[7] This difference may be attributed to different versions of the SF6D. The Hawthorne et al study used the original SF6D algorithm, whereas this study used the revised algorithm.

the consequence that as worse HRQoL states are reported, there will be an ever increasing 'gap' between the theoretical utility (based on the full life-death utility scale, 1.00 to 0.00) and the SF6D manifest scores. This restriction is clearly seen in Figures 8 and 9, where there is a systematic increase in the gap between the utility scores of the AQoL, EQ5D and HUI3 and those of the SF6D as age increases or health state decreases.

For the EQ5D, there is a different criterion validity problem, exemplified by the data in Table 14. This shows that there are certain utility score ranges within which it is difficult to obtain EQ5D scores, thus the EQ5D data distribution is 'lumpy'. For example, in the score range 0.41-0.50, the number of assigned cases on the EQ5D was 1/10 that of those for the AQoL or HUI3 and 1/7 of those for the SF6D. This distributional difficulty is a function of two issues. First, the EQ5D's descriptive system is the simplest of any instrument at just 5 items with 3 levels each. Second, the scoring algorithm for the EQ5D incorporates an additional weight that comes into effect whenever a person endorses the lowest possible level on any item. The effect of this additional weight is to cause an increase/decrease of utility between 0.1 and 0.3. For a fuller discussion of the impact of this term on EQ5D scores, see Brazier et al (72). The impact of this additional weight within the EQ5D algorithm is to confer increased sensitivity on the EQ5D whenever a respondent moves from a level-3 endorsement to a level-2 endorsement[8] – hence the poor distribution of EQ5D scores in the utility region 0.30 to 0.60 as shown in Table 14. The implication of this is that the EQ5D utility scores may not possess the necessary interval properties needed to meet the axioms of utility theory.[9] Because none of the mean scores for incontinence severity fell within this region, there was no obvious effect of this problem on the incontinence sensitivity analyses presented in Tables 16 through 19.

Based on sensitivity criteria alone, the instrument of choice for incontinence studies would be the 15D. It was the most sensitive instrument on all four incontinence measures, being between twice to thrice more sensitive than the least sensitive instrument (see Tables 16 to 19). The SF6D was the least sensitive instrument across the four incontinence measures. Generally, there was little difference in sensitivity between the other three measures. Perhaps the HUI3 was slightly more sensitive than the AQoL, and both were perhaps more sensitive than the EQ5D. Elsewhere, the AQoL has been reported as being less sensitive than the EQ5D in older adults (76), but this was not confirmed in this study.

The sensitivity of the 15D, however, must be balanced by the fact that 15D scores were consistently inconsistent with those of the other four MAU-instruments. As discussed above, they were substantially higher on all four incontinence measures. This was also the situation for age group and general health analyses (see Figures 8 and 9). Although the most sensitive instrument was the 15D, the mean scores on the 15D for all types of incontinence are implausible. For example, according to the 15D those who were classified as having daily faecal incontinence experienced a quality of life that was the same as those with no faecal incontinence symptoms on all the other four MAU-instruments (Table 18). The same phenomenon is evident, but not to the same extent, for those with urinary incontinence (Tables 16 and 17). When interpreting the sensitivity of the 15D, this should be kept in mind.

Obviously, the differences in utility classification shown in Table 15 would have a major impact in a cost-utility analysis. For example, based on the data shown in Figure 7, if a treatment for faecal incontinence caused an improvement on the HUI3 from 0.40 to 0.60 in HRQoL utilities and this was maintained for 10 years the QALY benefit would be 2.00 QALYs (0.20*10 = 2.00). For the 15D, using the mean 15D scores for HUI3 utility scores of 0.40 and 0.60, the utility gain would be from 0.83 to 0.88 resulting in a 0.5 QALY gain (0.05*10 = 0.5). These estimates are obviously not equivalent and imply that the results of any particular study would be more influenced by the choice of utility instrument than the intervention if sensitivity were the only criteria for instrument selection.

---

[8] For an example where this may have affected EQ5D scores, the reader should see Holland et al (75). They reported that the EQ5D was more sensitive than the AQoL in detecting differences between groups and over time. However, at baseline the mean score on the EQ5D was 0.61, implying that many cases at baseline would have been endorsing level-3 on at least 1 EQ5D item. Their data also showed a similar lumpy distribution to that reported in this study.

[9] Hawthorne et al (88) examined this problem more generally noting that preferences were collected on theoretical models which incorporated interval measurement (with standard gamble or time trade-off the underlying scale forms a continuum). Consequently if scores are non-interval that is a property of the sample rather than the scale per se. The problem for the EQ5D is different. The weights are based on the TTO, so the underlying scale possesses the necessary interval properties, but the scoring algorithm prevents interval measurement due to the presence of the additional weight for those who endorse level-3 response categories.

In terms of practicality, there was almost no evidence to support the use of one measure against another. Because of its brevity, obviously the EQ5D is the most practical measure at just 5 items. Regarding missing data, because the data were collected in face-to-face interviews no meaningful differences in missing data rates were reported.

As shown in Table 20, there was some evidence that the AQoL and EQ5D may be more sensitive to the impact of incontinence on females than on males, and that the HUI3 and SF6D may be more sensitive to its impact on males. Gender differences like this have not been previously reported in the literature, but they may reflect the different descriptive systems of the utility measures as analysed in Figures 2 through 6. For the AQoL, the SEM analysis showed that social relationships and psychological wellbeing exerted the greatest influence on utility, for the EQ5D usual activities exerted the greatest influence. For the HUI3 the greatest influences were cognition and pain, while for the SF6D these were physical activities and role.

4.7 Conclusion

This evaluation of five MAU-instruments has presented tests of reliability, construct and criterion validity. It has also presented population norms that may be used by other researchers as benchmarks against which to interpret their work. In general, the data showed that incontinence has a small to mild effect upon HRQoL (Table 21).

The analyses suggested that although the five instruments were all measuring the same latent construct, they were measuring different aspects of it. The data also suggested that there were important differences in mean utility scores between the measures that were inconsistent by selected criteria, such as age group, health status or incontinence status.

The AQoL generally performed well in tests of reliability, construct and criterion validity, and in tests of incontinence sensitivity. It possessed good reliability, was well correlated with the other MAU-instruments, the population norms were almost identical with those previously reported (6) and it was among the more sensitive measures to incontinence status. The obvious shortcoming of the AQoL was in relation to the observed gender bias. It appeared to be more sensitive to incontinence states in females than males when compared with the HUI3, 15D or SF6D (Table 20). This finding has not been previously reported and warrants further investigation.

The attraction of the EQ5D is its simplicity: with just 5 items each of 3 levels, it is the most practical of the measures. As shown in Table 15 and Figure 8, general population norm values for the EQ5D are similar to those of HUI3 – an instrument which is over twice as long. However, analysis of its internal structure suggested that it possessed the lowest reliability of any of the MAU-instruments. Additionally, there was substantial evidence that scores on the EQ5D were 'lumpy' and that there were obvious gaps in utility value that were inconsistent with utility scores from the other instruments. Thus, despite its attractions, it is difficult to recommend the EQ5D by itself.

The HUI3 generally performed well overall. In terms of reliability, the slightly low Cronbach suggested there may be an internal inconsistency; a finding that was consistent with the SEM. For population norms, the HUI3 reported that those aged 20-29 years enjoyed a higher HRQoL when compared with those aged 15-19 years. Generally, the HUI3 was sensitive to those with differing levels of incontinence, although it is possible that it was more sensitive to differences in status for males than females which may reflect the lack of social relationships measurement.

Despite its superior sensitivity to incontinence status, particular difficulties were encountered for the 15D. These included a poor internal structure, inconsistent score ranges when compared with those from the other measures, and implausible values. In addition to these empirical findings, there are also theoretical difficulties with the 15D in relation to whether the weights used, which were derived from a visual analog scale, reflect preferences. For a fuller discussion see Appendix A. Consequent upon these shortcomings, it cannot be recommended as an instrument of choice.

The SF6D primarily measured physical capacity, such as being able to do vigorous or moderate activities, and limitations in one's life role, such as being limited by physical or emotional conditions. The restriction in utility range had the effect of limiting scores to a smaller range-width than was the case for any other MAU-instrument. The consequent restriction in utilities for those experiencing poorer HRQoL was particularly noticeable, including among those suffering severe incontinence. The SF6D also appeared to be more sensitive to the impact of incontinence on males when compared with females, which may reflect the emphasis on physical function.

Given this the SF6D was the least sensitive instrument in discriminating between incontinence status. For these reasons, the SF6D cannot be recommended as an instrument of choice.

4.8 Recommendations

This study has shown that there are substantial differences in manifest scores between five leading generic MAU-instruments. The differences are such that utilities obtained from one measure cannot be assumed to be compatible with those from the other measures. (The two measures which provided the most compatible scores are the EQ5D and HUI3.) This key finding provides empirical evidence supporting Thomas et al's (27) review, which came to the same conclusion based on examination of the published literature (this review is reproduced in Appendix A).

The inconsistencies reported in this study reflect different descriptive systems, assigned weights, and scoring mechanisms. That these deliver utilities that are statistically significantly different across a wide range of values, suggests the results for the different instruments cannot all be right. When taken in conjunction with the differences in implied QALYs, effect sizes and relative efficiencies, they are suggestive that study results may depend upon the instrument chosen rather than actual treatment benefits.

Regarding recommendations, the results of this study support those reached by Thomas et al (27), viz., that two utility measures should be included in any particular study and that both sets of results should be reported with appropriate sensitivity analyses. The preferred instrument would be the Australian AQoL since it performed at least as well as any of the other MAU-instruments and because it is weighted with Australian TTO-values. The instrument of second choice would be the HUI3. Where direct comparison between Australian and international data is required, the EQ5D could be used. Because of its measurement shortcomings it should not be used alone.

Given that all five utility instruments are contained within the SAHOS dataset, further research into similarities and differences between the utility measures could be undertaken with the objective of providing standardized algorithms for the development of a common scoring metric enabling imputation of scores from each instrument to each other instrument.

5. Australian SF-36 V2 Norms and the Impact of Incontinence on Health Status

The SF36 Version 2 (SF36V2) replaces the SF36 Version 1 (SF36V1), and will become the world's ubiquitous health status measure over the next few years. It is important that guidelines for interpreting scores are available to researchers.

Based on a random sample of Australians (n=3014), this study reports differences between the US norms published by Ware et al (15) and Australians. Significant differences were observed on 7 of the 8 scales and on the mental health summary scale. Although the cause of these differences is unknown, cross-cultural emic effects cannot be ruled out. Australian weights were therefore derived and have been used. Estimates from using the standard US-weights are reported for those making international comparisons.

Population norms by age cohort, gender and health status are reported by T-score as recommended by the instrument developers. Additionally, the proportions of cases within SF36V2 T-score deciles are presented. The findings suggest there are statistical artefacts associated with the use of T-scores that have implications for how the data from the SF36V2 are interpreted and analysed.

The procedures reported in this study may be used by other researchers where emic effects are suspected and who wish to develop local weights for the SF36V2. The population norms presented may also be of interest.

5.1 Introduction

The SF-36V1 (14), released in 1988, is the world's ubiquitous health status measure; a simple search of PubMed (May 2005) identified 4,029 references. Of these, 115 were Australian studies, far more than for any other health status measure used in Australia. The implication is that the SF36V1 is also the ubiquitous health status measure used by Australian researchers. Further evidence regarding its popularity is that there is an Australian version of the SF36V1 (119), and there have been several Australian validation studies (120-123), including the publication of Australian population norms for the SF36V1 (124).

Despite this popularity, the instrument developers acknowledged some of the criticisms levelled at the SF36V1 and between 1996-2000 developed the "international version" of the SF36 – the SF36 Version 2 (hereafter SF36V2, 15). These shortcomings included cross-cultural adaptation issues, difficulties with some word meanings, possible double negatives, scale floor and ceiling effects, problems in the two role function scales, and confusion caused by the standard layout, particularly among older adults, leading to unnecessarily high missing data and response errors (15, 125-130).

The changes, then, were designed to make it easier to understand, to reduce missing data, improve the sensitivity of the two role function scales, and to simplify the response categories for the health and vitality scales. An important reason for these changes was the finding from the International Quality of Life Assessment (IQOLA) project that the SF36V1 encountered cultural differences during translation and between language groups (131-133). A key aim was to ensure that the SF36V2 was more cross-culturally valid than had been the SF36V1 (15). A further change involved presenting the eight health attribute scale scores as T-scores (43), whereas in SF36V1 these were presented as percentile scores (14).

Since publication of the SF36V2, in addition to the US population norms provided by the SF36V2 developers, there have been two population validation studies. Regarding data from the US, Ware et al (15) administered the SF36V1 and SF36V2 to a population sample (n=6742), with random instrument allocation. The increased response choices in the Role Physical (RP) and Role Emotion (RE) scales increased their sensitivity and mean scores (for RP the mean score on the SF36V1 was 75.1 and on SF36V2 it was 80.8, for the RE scale these were 83.7 and 86.3 respectively), and reduced floor and ceiling effects (RP from 62% to 47% and 14% to 2%, and RE from 74% to 60% and 9% to 1% respectively). The reliability of these two scales also increased (RP from Cronbach 0.88 to 0.95, and RE 0.82 to 0.93). There were no significant effects on any of the other scales. In general, the effects of the revisions were to improve the measurement properties of the SF36 without any loss of internal structure.

Taft et al (134) administered the SF36V2 to a 2185 18-75 year old randomly chosen national sample of Swedes and reported similar better measurement results to those reported by Ware et al. Jenkinson et al (135) collected responses from 8889 Britons of working age, again reporting the better measurement properties of the SF36V2. Jenkinson et al reported basic norms by gender and social class. Unlike the Taft et al study, however, Jenkinson et al also published factor score coefficients which could be used as British weights during scoring of the two summary scales, rather than the US weights published by Ware et al. Neither of these two studies, however, included population norms for the two summary scales.

Following release of the SF36V2, the Australian SF36V2 was provided by QualityMetric (44). Table 22 describes the descriptive system differences between the Australian SF36V1, the US SF36V2 and the Australian SF36V2. This reveals that, when compared with the SF36V1, the SF36V2 differs with respect to the number of response categories for questions 4, 5 and 9. For questions 3G, 3H, and 3I there are also differences in the distances asked: the US version refers to distances in miles and yards compared with the Australian version asking about kilometres and metres. For 3G, for example, the response categories might suggest that respondents using the US version would have to be able to walk 60% further than their Australian counterparts to indicate equivalent health status (1 mile, which is 1.6 kilometres, versus 1 kilometre). Whether these differences matter is unknown, although in the case of a person with chronic obstructive pulmonary disease this difference may mean the difference between being able to do their own shopping and needing a carer to do the shopping. Other situations could easily be found.

As shown in the table, however, the international and Australian versions of the SF36V2 differ only on questions 3G, 3H, and 3I. There are, however, differences in the instructions, including that the Australian version includes a practice question.

Although Sansoni and Costi (44) acknowledged the superiority of the SF36V2 over the SF36V1, they also noted that the absence of Australian normative data limited its use by Australian researchers. This study rectifies this situation by providing normative data based on Australian weights derived from similar analysis procedures used to derive the US weights reported by Ware et al (15). The norms reported in this paper may be used by Australian researchers as benchmarks for the interpretation of their SF36V2 data, and they may be of interest to other researchers using the SF36V2.

Table 22: Differences between the Australian SF36V1, International SF36V2 and Australian SF36V2 in Item Wording and Response Categories

Question	*Question part*	*Australian SF36V1*	*International SF36V2*	*Australian SF36V2*
3G	Stem	Walking more than one kilometre	Walking more than a mile	Walking more than a kilometre
3H	Stem	Walking half a kilometre	Walking several hundred yards	Walking several hundred metres
3I	Stem	Walking 100 metres	Walking one hundred yards	Walking 100 metres
4A, B, C, D, 5A, B, C	Response	Yes/No	All of the time/ Most of the time/ Some of the time/ A little of the time/ None of the time	All of the time/ Most of the time/ Some of the time/ A little of the time/ None of the time
5C	Stem	Didn't do work or other activities as carefully as usual	Did work or other activities less carefully than usual	Did work or other activities less carefully than usual
7	Response	No bodily pain	None	None
9A, B, C, D, E, F, G, H, I	Response	All of the time/ Most of the time/A good bit of the time/Some of the time/A little of the time/None of the time	All of the time/ Most of the time/ Some of the time/ A little of the time/ None of the time	All of the time/ Most of the time/ Some of the time/ A little of the time/ None of the time

5.2 Methods

5.2.1 Participants

Regarding sample size for the population norms for the SF36, the International Quality of Life Assessment (IQOLA) project determined that these should be set at a minimum of 2500-3000 respondents to enable comparisons by gender and 10-year age groups. It was argued that samples should be representative of the general population, and the use of sampling weights to achieve population representativeness was encouraged (136).

The current study uses data collected from 3015 South Australians who participated in the 2004 South Australian Health Omnibus Survey (SAHOS), weighted by population characteristics to achieve representativeness (28). Full details are given in section 2.1.

5.2.2 Materials

Thirteen different research groups participated in the SAHOS sponsoring the use of 32 different measures. The SF36V2 was the first measure in the questionnaire. This study reports the use of the SF36V2, questions on demographics and question 16 from the HUI3.

Demographics

The demographic items used in this study were gender, age, birth country, partnership status, education attainment, and workforce participation.

The SF36V2

The SF36V2 (15) is a health status (function) instrument, the descriptive system of which comprises 36 items which are organised into 8 scales (Physical Functioning (PF), Role Physical (RP), Bodily Pain (BP), General Health (GH), Vitality (VI), Social Functioning (SF), Role Emotion (RE) and Mental Health (MH). There is also a health transition item (*Compared to one year ago, how would you rate your health in general now?*). The 8 scales can be combined into 2 summary measures, providing overall estimates of physical health (Physical component score, PCS) and mental health (Mental component score, MCS).

The most important change between the SF36V1 and the SF36V2 relates to how the scale scores are presented. In the SF36V1, scale scores are presented on percentage scales (0-100), whereas for SF36V2 although percentage scores are computed, Ware et al recommend that scale scores are presented as T-scores (15). This extension of the T-score presentation from the PCS and MCS summary scales to the 8 scales means that each of the eight scales forming the SF36V2 can be reported at three levels. Each is described.

SF36V2 data can be presented as unweighted raw scale scores. When scored like this, after item reversing (n = 9 items), the items contributing to a scale are simply summed and the raw score presented. Essentially, this scoring method treats the weight of each item as '1.00'. Because the response categories are not equi-interval this scoring is not recommended and this scoring is not used in this study.

It can also be presented as weighted percentage scores. This involves a two-step procedure. Items are weighted to achieve equal interval values and reversed where needed (15, 131), and then percentage scale scores are computed based on:

$$S_S = \frac{(S - S_m)}{S_r} \cdot 100$$

where S_S is the weighted scale score percentage, S is the weighted raw score, S_m is the minimum possible score for the scale of interest, and S_r is the possible raw score range.

Finally, Ware et al recommend that SF36V2 data are presented as T-scores (43), where the mean scale score is 50 and the standard deviation 10-points. SF36V2 T-scores are computed by first computing z-scores and then converting these to T-scores:

$$T_S = 50 + \left(\left(\frac{S_s - \overline{S_s}}{sd_{S_S}} \right) \cdot 10 \right)$$

where T_S is the T-score and sd_{SS} the standard deviation for the scale of interest. The expression

$$\left(\left(\frac{S_s - \overline{S_s}}{sd_{S_S}}\right)\right)$$

computes the z-score. These T-scores are described in the SF36V2 manual as "norm-based" scores (15). Effectively, the mean scores ($\overline{S_s}$) and sd_{SS} provide differential weights for scale scores. For the SF36V2, these weights were derived from the US population survey carried out in 1998 (137).

The two summary scores, PCS and MCS, use the sum of the eight dimension z-scores weighted by factor score coefficients. The factor score coefficients are derived from US 1990 general population estimates (15, 138).

In the present study, for reasons described below in the results section, the US weights described above have been replaced with Australian weights derived from the SAHOS dataset, except for the item equal interval weights. For those wishing to use the US weighted versions of the SF36V2, the normed data are presented in the Supplementary material at the end of this section.

HUI3

The HUI3 is a multi-attribute utility measure comprising 15 items (139). A 16th item asks respondents *Overall, how would you rate your usual health in the past four weeks? Excellent/Very good/Good/Fair/Poor*. This is the only item from the HUI3 used in this analysis.

5.2.3 Data Analysis

The data were weighted by inverse of the probability of selection, and then re-weighted to benchmarks from the 2001 Census to achieve representativeness. All data were double-entered and verified prior to analysis. Missing data were collected by follow up telephone interview. There was one case with missing SF36V2 scores; no attempt was made to impute the values for this case.

Exploratory factor analysis with orthogonal rotation was used to extract factor coefficients, as recommended by the SF36 developers (15, 138). For comparison between US and Australian data, given that only summaries were available for the US data, 95% confidence intervals were computed and significance was assumed where these did not overlap. Due to data non-normality, Kruskall-Wallis χ^2 was used to compare between groups.

The data were analysed in SPSS Version 13.1 (114).

5.3 Results

Participants

Females comprised 50.9% of the sample, and the mean age was 45.29 years (SD = 18.69 years). For country of birth, 74.4% were Australian-born and 12.2% UK/Ireland born. For partnership status, 61.9% were in a relationship, 24.1% had never married, 8.5% were separated or divorced, and 5.6% were widowed. Primary school education only was reported by 18.4%, 32.5% had completed high school, 12.6% held a trade qualification, 22.6% a certificate or diploma and 13.9% a university degree. Full time employment was reported by 38.8%, part-time employment by 16.9%, being unemployed by 2.1%, home duties by 11.0%, being retired by 18.7%, 9.5% were studying and 3.1% were not employment status classified.

SF36V2 weights

Table 23 presents the percentage score scale means, standard deviations and 95% confidence intervals from both the US 1998 general survey and the Australian SAHOS survey, for each of the 8 scales. The key features are that on every scale the SAHOS mean percentage scores were statistically significantly higher than the US scores and that the Australian standard deviations are smaller, with the exception of the GH and VI scales. That there are so few overlapping 95%CIs (they occur only for the PF and GH scales) suggests that there are significant differences between the US and SAHOS samples.

These data are shown graphically in Figure 10.

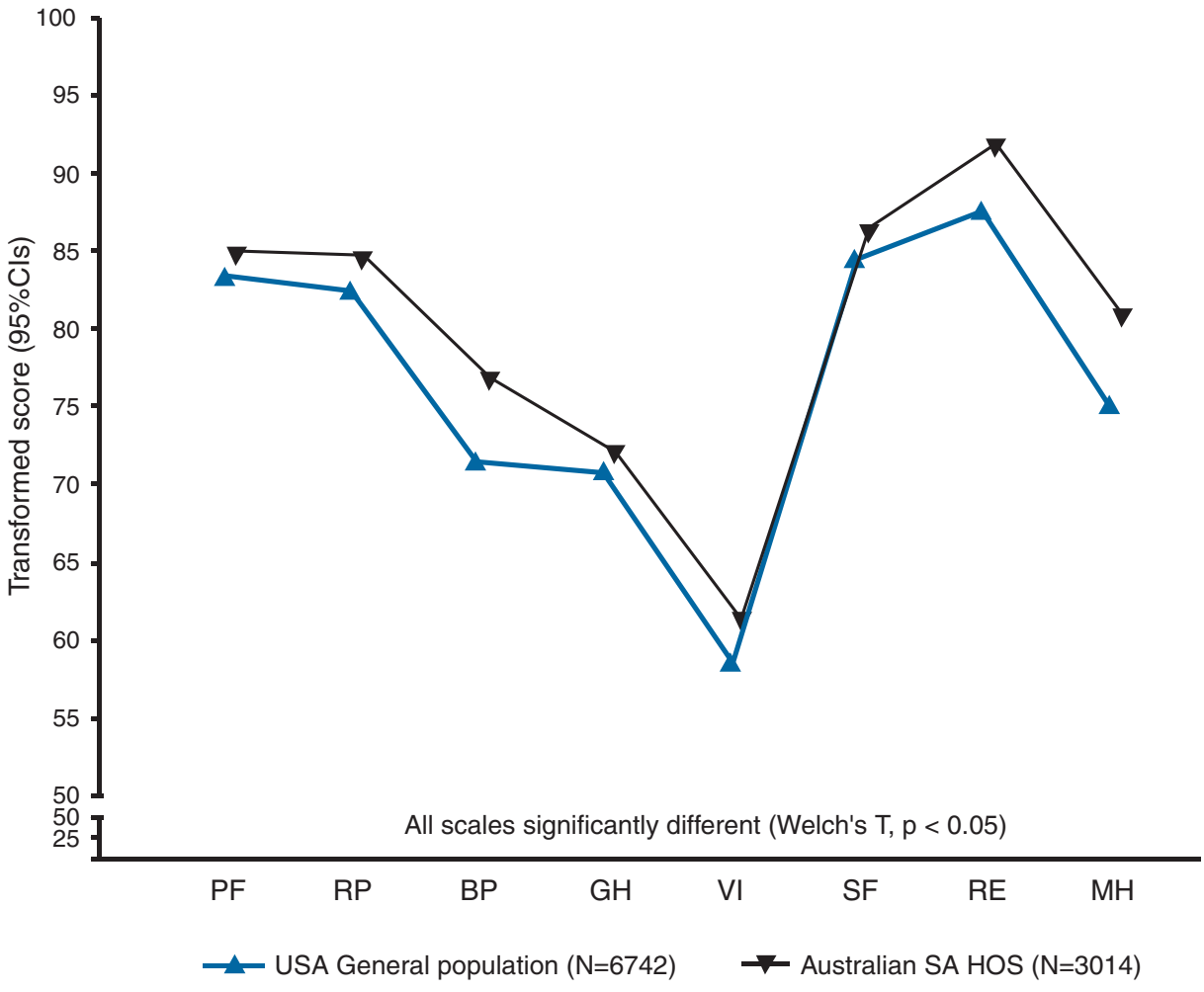
Table 23: SF36V2 Mean Scale Percentile Scores, Standard Deviations and 95% Confidence Intervals, from the US and SAHOS Surveys

SF36 scales	*Percentage scores*							
	US data (a)				*SAHOS data*			
	N	*Mean*	*sd*	*95%CI*	*N*	*Mean*	*sd*	*95%CI*
Physical function	6742	83.29	23.76	82.72–83.86	3015	84.64	21.86	83.85–85.41
Role physical	6742	82.51	25.52	81.90–83.12	3014	84.41	25.13	83.51–85.31
Bodily pain	6742	71.33	23.66	70.77–71.90	3014	76.45	21.24	75.69–77.21
General health	6742	70.85	20.98	70.35–71.35	3015	71.90	21.88	71.12–72.68
Vitality	6742	58.31	20.02	57.83–58.79	3014	61.12	20.80	60.38–61.86
Social functioning	6742	84.30	22.92	83.75–84.85	3014	86.19	22.33	85.39–86.99
Role emotional	6742	87.40	21.44	86.89–87.91	3014	91.59	17.50	90.97–92.22
Mental health	6742	74.99	17.76	74.57–75.41	3014	80.63	16.99	79.62–80.84

The slight discrepancy in table numbers is because of missing data.

Notes: a = computed from Ware et al (14)

Figure 10: SF36V2 Mean Scale Scores, by Country



For the two summary scales, the PCS and MCS, Table 24 shows the factor score coefficients for both the US 1990 data as reported by Ware (15) and the coefficients derived from the SAHOS data. As shown, there are differences in the coefficients, implying that the relative weighting of the 8 scales within the PCS and MCS scoring systems are different between the two countries. Particularly important is that the direction of the loadings is different for GH on the MCS and SF on the PCS. There are also differences in the coefficients for RP and VI on the MCS, and for BP and VI on the PCS.

The effect of these differences is shown in Table 25 which compares the Australian SAHOS with the US data, based on using the US weights reported in Tables 23 & 24. That there is no overlap in the 95%CIs for 8 of the 10 scales suggests there are differences between Australian and US data on the SF36V2. The two scales where there is overlap are the GH and the PCS summary scales.

Population Norms

The findings reported in Tables 23, 24 and 25 suggest there are potential cross cultural emic differences between the US and Australia. Therefore the Australian population normed T-scores for the 8 scales were derived from the Australian percentage score data in Table 23, and the two summary scales were weighted with the coefficients reported in Table 24. This has been done for the reporting of Australian population norms in Tables 26, 27, 28, 29, 30 and 31. There is an elaboration of this in the discussion section.

Based on Australian weights, Table 26 provides Australian normed mean T-scores for the 8 SF36V2 scales, by 10-year age groups and gender. A key feature of the table is that on 7 of the 8 scales, males obtained significantly higher scores, indicating better health, when compared with females (Kruskall-Wallis χ^2 range = 26.72 to 98.57, all $p < 0.001$). The exception was for general health (GH) where there was no significant gender difference (Kruskall-Wallis $\chi^2 = 1.80$, $p = 0.18$). Likewise, for age group there were significant differences on 7 scales (Kruskall-Wallis χ^2 range = 21.18 to 836.05, all $p < 0.001$). The exception was for the RE scale, where there were no significant differences by age (Kruskall-Wallis $\chi^2 = 13.81$, $p = 0.06$).

Table 27 shows the two summary scales by age group and gender, based on Australian weights. There were significant differences by both gender (Kruskall-Wallis $\chi^2 = 37.83$ and 20.89 for PCS and MCS, both $p < 0.001$) and age (Kruskall-Wallis $\chi^2 = 486.92$ and 182.36 for PCS and MCS, both $p < 0.001$). Tables 28 and 29 depict SF36V2 T-scores by deciles and show the proportion of cases which fell within each decile. As shown, there were very substantial ceiling effects. For example, for RE 79% of all cases fell within the top decile; indeed, on 4 of the 8 scales more than 50% of cases fell within the top decile. Conversely, there is almost no evidence of a floor effect. No scale had more than 3% of cases in the bottom decile. Finally, no cases fell within the 4th and 9th deciles for BP and within the 6th decile for SF. The reason is that these two scales have very limited scoring ranges (the ranges are 10 and 8 points respectively (Table 6.11, 15)).

Table 24: Factor Score Coefficient Weights for the SF36V2 PCS and MCS Summary Scales, from the US and SAHOS Population Surveys

SF36 dimension	*Factor score coefficients*			
	US (a)		*Australia SAHOS*	
	PCS	*MCS*	*PCS*	*MCS*
Physical functioning	0.42402	–0.22999	0.40931	-0.22383
Role physical	0.35119	–0.12329	0.32517	-0.09553
Bodily pain	0.31754	–0.09731	0.28912	-0.10501
General health	0.24954	–0.01571	0.23124	0.00074
Vitality	0.02877	0.23534	0.10596	0.15709
Social functioning	–0.00753	0.26876	0.01428	0.24907
Role emotional	–0.19206	0.43407	-0.18286	0.44909
Mental health	–0.22069	0.48581	-0.20470	0.47558

Notes: a = From Ware et al (14)

Table 25: Australian SF36V2 Percentage Scale Mean Scores and Summary Scale Scores, based on US Weights

SF36 dimension	*N*	*USA (a)*					*N*	*Australian SAHOS*				
		Percentage scores		*T-scores (normed scores)*				*Percentage scores*		*T-scores (normed scores)*		
		Mean	*sd*	*Mean*	*sd*	*95%CI*		*Mean*	*sd*	*Mean*	*sd*	*95%CI*
Physical functioning	6742	83.29	23.76	50.00	10.00	49.76-50.24	3014	84.64	21.86	50.57	9.20	50.24-50.90
Role physical	6742	82.51	25.52	50.00	10.00	49.76-50.24	3014	84.41	25.13	50.75	9.85	50.40-51.10
Bodily pain	6742	71.33	23.66	50.00	10.00	49.76-50.24	3014	76.45	21.24	52.16	8.98	51.84-52.48
General health	6742	70.85	20.98	50.00	10.00	49.76-50.24	3015	71.90	21.88	50.50	10.43	50.13-50.87
Vitality	6742	58.31	20.02	50.00	10.00	49.76-50.24	3014	61.12	20.80	51.40	10.39	51.03-51.77
Social functioning	6742	84.30	22.92	50.00	10.00	49.76-50.24	3014	86.19	22.33	50.82	9.74	50.47-51.17
Role emotional	6742	87.40	21.44	50.00	10.00	49.76-50.24	3014	91.59	17.50	51.96	8.16	51.67-52.25
Mental health	6742	74.99	17.76	50.00	10.00	49.76-50.24	3014	80.63	16.99	53.18	9.57	52.84-53.52
PCS	6742	N/A	N/A	50.00	10.00	49.76-50.24	3013	N/A	N/A	50.27	9.70	49.92-50.62
MCS	6742			50.00	10.00	49.76-50.24	3013			52.92	10.17	52.56-53.28

The slight discrepancy in table numbers is because of missing data.

Note: a = From Ware et al (14)

Table 26: Australian normed T-scores for the SF36V2 Scales, Australian Weights, by Age and Gender

Age	Gender	N	PF		RP		BP		GH		VI		SF		RE		MH	
			Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
15-19	M	129	55.77	4.42	54.27	4.28	54.56	8.00	54.21	6.51	55.71	8.33	53.48	5.89	52.99	5.62	53.94	6.28
	F	124	53.22	4.45	52.32	6.42	50.50	9.26	49.17	9.10	49.23	8.59	48.61	8.42	47.51	9.80	46.50	11.83
	All	253	54.53	4.61	53.31	5.51	52.58	8.86	51.75	8.26	52.56	9.04	51.10	7.62	50.31	8.39	50.31	10.10
20-29	M	244	55.55	2.73	53.47	5.20	53.61	8.23	53.23	7.45	53.36	7.20	52.24	6.31	51.91	6.06	51.29	7.68
	F	230	54.44	4.61	51.50	7.41	50.48	9.88	50.47	10.08	48.50	9.79	48.33	10.31	48.11	11.50	47.69	10.72
	All	474	55.01	3.80	52.51	6.44	52.09	9.19	51.89	8.93	51.00	8.89	50.34	8.71	50.06	9.30	49.54	9.45
30-39	M	266	53.75	6.03	51.83	8.91	50.64	9.88	52.05	8.99	52.17	8.21	51.04	9.06	50.81	8.36	51.30	8.96
	F	263	52.81	6.61	52.17	7.20	51.47	9.63	53.03	8.36	49.30	9.34	50.46	8.92	48.65	11.36	48.65	9.89
	All	529	53.19	6.33	52.00	8.10	51.05	9.75	52.54	8.69	50.75	8.90	50.75	8.99	49.74	10.02	49.98	9.52
40-49	M	275	53.22	5.68	52.43	8.21	51.81	8.92	51.04	8.46	51.28	9.90	51.52	9.62	51.45	8.28	50.07	9.75
	F	278	49.94	10.06	50.33	10.01	48.94	10.35	50.21	10.62	48.65	9.96	49.23	11.12	48.73	12.31	48.56	10.87
	All	553	51.57	8.34	51.37	9.21	50.36	9.76	50.62	9.61	49.95	10.01	50.37	10.46	50.08	10.58	49.31	10.35
50-59	M	239	49.69	9.32	48.58	10.96	48.53	10.56	47.57	10.94	49.50	10.51	49.11	10.06	50.12	10.30	49.69	10.23
	F	244	46.89	10.26	48.05	11.53	47.16	10.42	48.51	11.36	47.08	10.27	48.40	11.28	49.04	11.11	48.57	10.75
	All	482	48.27	9.89	48.31	11.24	47.84	10.05	48.05	11.15	48.28	10.45	48.75	10.69	49.58	10.72	49.12	10.50
60-69	M	156	47.42	10.44	48.22	11.77	48.58	9.95	47.57	10.69	51.52	10.47	50.90	10.02	50.85	9.84	52.02	9.03
	F	161	45.11	12.00	47.77	11.17	48.19	10.17	48.58	11.02	49.79	10.96	50.22	10.66	50.12	10.40	50.55	10.64
	All	318	46.25	11.30	47.99	11.45	48.38	10.05	48.08	10.85	50.64	10.74	50.55	10.34	50.48	10.12	51.27	9.89
70-79	M	127	45.24	10.54	46.24	11.63	49.65	9.68	46.90	9.77	50.70	10.28	50.49	10.14	49.80	9.10	52.92	9.68
	F	158	39.18	13.77	44.19	12.84	46.56	10.72	46.09	11.22	46.56	11.25	47.89	11.59	49.63	10.27	49.47	10.47
	All	285	41.88	12.78	45.10	12.34	47.93	10.37	46.45	10.48	48.40	11.01	49.05	11.03	49.70	10.13	51.00	10.25
80+	M	44	41.13	13.31	44.20	12.67	49.89	10.05	46.88	9.77	49.69	11.15	50.35	12.21	50.67	11.43	53.81	9.68
	F	78	32.58	12.87	40.40	13.13	46.74	11.08	45.81	10.71	44.86	11.93	45.40	13.86	51.16	8.24	51.28	9.08
	All	121	35.67	13.61	41.77	13.04	47.88	10.79	46.20	10.35	46.61	11.84	47.19	13.45	50.98	9.47	52.19	9.35
Total	M	1480	51.67	8.36	50.81	9.50	51.02	9.61	50.46	9.43	51.81	9.46	51.15	9.10	51.10	8.62	51.33	9.11
	F (a)	1534	48.39	11.13	49.22	10.40	49.01	10.27	49.56	10.50	48.25	10.20	48.89	10.69	48.94	11.07	48.72	10.64
	All (b)	3014	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00	50.00	10.00

The slight discrepancy in table numbers is because of missing data.

a = For PF and GH there were 1535 females; b = For PF and GH the total was 3015.

Table 27: Australian normed T-scores for the SF36V2 Summary Scales, Australian Weights, by Age and Gender

Age	Gender	N	PCS		MCS	
			Mean	sd	Mean	sd
15-19	M	129	55.44	4.59	52.78	6.30
	F	124	53.22	6.06	45.89	10.19
	All	253	54.35	5.46	49.42	9.09
20-29	M	244	55.06	5.18	50.63	6.48
	F	230	53.32	7.00	46.33	11.57
	All	474	54.21	6.19	48.54	9.55
30-39	M	266	52.61	8.34	50.60	8.93
	F	263	53.53	7.42	47.85	10.60
	All	529	53.07	7.90	49.23	9.88
40-49	M	275	52.78	6.99	50.16	9.01
	F	278	50.22	9.61	48.51	11.29
	All	553	51.49	8.50	49.33	10.25
50-59	M	239	48.37	10.67	49.98	9.60
	F	244	47.00	11.28	49.24	10.90
	All	482	47.68	10.99	49.61	10.27
60-69	M	156	46.89	11.35	52.62	8.96
	F	161	46.15	11.65	51.75	10.11
	All	318	46.52	11.49	52.18	9.56
70-79	M	127	45.41	10.10	52.78	9.42
	F	158	41.36	12.91	51.66	9.52
	All	285	43.16	11.90	52.16	9.47
80+	M	44	42.60	12.03	54.36	9.84
	F	78	36.47	12.92	53.92	7.99
	All	121	38.68	12.90	54.08	8.67
Total	M	1480	51.09	9.25	51.13	8.65
	F	1534	48.95	10.93	48.91	10.83
	All	3014	50.00	10.20	50.00	9.88

The slight discrepancy in table numbers is because of missing data.

Table 29 shows deciles for the two summary scales. Again, there was no evidence of a floor effect, and there was little evidence of ceiling effects. The data, however, were skewed with high proportions of cases obtaining scores in the top 20% of the scoring range (27% for the PCS and 42% for the MCS obtained scores in the range 81-100).

The effect of health status is reported in Table 30, broken down by gender. This reveals a monotonic decline for both males and females across all 8 SF36V2 scales. It should be noted that the magnitude of decline increases with poorer health levels.

Table 28: SF36V2 Scales, T-score Percentile Deciles with Proportions in Deciles, Australian Weights, by Gender

Decile	Gender	PF		RP		BP		GH		VI		SF		RE		MH	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
0-10	M	14	1.0%	24	1.6%	0	0.0%	14	0.9%	13	0.9%	9	0.6%	2	0.2%	1	0.1%
	F	35	2.2%	38	2.5%	4	0.3%	27	1.8%	37	2.4%	16	1.0%	10	0.7%	6	0.4%
	All	49	1.6%	63	2.1%	4	0.1%	41	1.4%	50	1.7%	25	0.8%	13	0.4%	8	0.2%
11-20	M	2	0.1%	29	1.9%	27	1.8%	24	1.6%	36	2.5%	15	1.0%	1	0.1%	4	0.3%
	F	20	1.3%	21	1.4%	35	2.3%	40	2.6%	74	4.8%	24	1.5%	6	0.4%	16	1.0%
	All	22	0.7%	50	1.7%	62	2.0%	64	2.1%	111	3.7%	39	1.3%	7	0.2%	20	0.7%
21-30	M	19	1.3%	28	1.9%	106	7.2%	40	2.7%	28	1.9%	22	1.5%	13	0.9%	13	0.9%
	F	36	2.3%	46	3.0%	182	11.8%	66	4.3%	57	3.7%	46	3.0%	17	1.1%	26	1.7%
	All	54	1.8%	74	2.5%	288	9.5%	106	3.5%	84	2.8%	68	2.3%	30	1.0%	40	1.3%
31-40	M	9	0.6%	37	2.5%	140	9.5%	60	4.0%	100	6.8%	35	2.3%	7	0.5%	23	1.6%
	F	20	1.3%	60	3.9%	174	11.3%	66	4.3%	146	9.5%	38	2.5%	16	1.0%	27	1.8%
	All	29	1.0%	97	3.2%	314	10.4%	126	4.2%	246	8.2%	73	2.4%	23	0.7%	51	1.7%
41-50	M	36	2.5%	64	4.3%	0	0.0%	75	5.0%	186	12.6%	55	3.7%	44	3.0%	51	3.4%
	F	56	3.6%	74	4.8%	0	0.0%	90	5.9%	247	16.1%	91	5.9%	77	5.0%	78	5.1%
	All	92	3.1%	138	4.6%	0	0.0%	165	5.5%	432	14.3%	147	4.9%	122	4.0%	129	4.3%
51-60	M	55	3.7%	20	1.3%	147	9.9%	109	7.4%	119	8.0%	0	0.0%	27	1.8%	64	4.4%
	F	108	7.0%	37	2.4%	194	12.6%	117	7.6%	182	11.9%	0	0.0%	42	2.7%	114	7.5%
	All	163	5.4%	57	1.9%	340	11.3%	227	7.5%	301	10.0%	0	0.0%	68	2.3%	179	5.9%
61-70	M	71	4.8%	53	3.6%	246	16.6%	205	13.8%	409	27.6%	61	4.1%	38	2.5%	129	8.7%
	F	107	6.9%	94	6.1%	212	13.8%	180	11.7%	366	23.8%	115	7.5%	53	3.5%	166	10.8%
	All	178	5.9%	147	4.9%	458	15.2%	385	12.8%	774	25.7%	176	5.8%	91	3.0%	295	9.8%
71-80	M	97	6.5%	72	4.9%	277	18.7%	335	22.6%	225	15.2%	141	9.6%	52	3.5%	256	17.3%
	F	140	9.1%	90	5.8%	269	17.6%	307	20.0%	198	12.9%	187	12.2%	91	5.9%	305	19.9%
	All	237	7.9%	162	5.4%	547	18.1%	642	21.3%	423	14.0%	329	10.9%	143	4.8%	561	18.6%
81-90	M	292	19.8%	152	10.3%	0	0.0%	310	20.9%	251	17.0%	167	11.3%	64	4.3%	550	37.2%
	F	283	18.4%	158	10.3%	0	0.0%	320	20.9%	176	11.5%	152	9.9%	78	5.0%	478	31.2%
	All	576	10.1%	310	10.3%	0	0.0%	630	20.9%	428	14.2%	319	10.6%	142	4.7%	1028	34.1%
91-100	M	884	59.7%	999	67.5%	537	36.3%	309	20.9%	113	7.6%	975	65.9%	1232	83.2%	387	26.1%
	F	730	47.6%	917	59.7%	465	10.3%	321	20.9%	52	3.4%	865	56.4%	1144	74.5%	317	20.7%
	All	1614	53.5%	1916	63.6%	1002	33.2%	630	20.9%	165	5.5%	1840	61.0%	2376	78.8%	704	23.4%

The slight discrepancy in table numbers is because of missing data.

Table 29: SF36V2 Summary Scales, T-score Percentile Deciles with Proportions in Deciles, Australian Weights, by Gender

Decile	*Gender*	*PCS*		*MCS*	
		N	*%*	*N*	*%*
0-10	M	4	0.3%	0	0.0%
	F	5	0.3%	1	0.1%
	All	9	0.3%	1	0.0%
11-20	M	17	1.2%	1	0.1%
	F	38	2.5%	8	0.5%
	All	55	1.8%	9	0.3%
21-30	M	17	1.2%	8	0.5%
	F	49	3.2%	18	1.2%
	All	66	2.2%	26	0.9%
31-40	M	48	3.2%	17	1.2%
	F	86	5.6%	34	2.2%
	All	133	4.4%	52	1.7%
41-50	M	87	5.9%	43	2.9%
	F	73	4.8%	56	3.7%
	All	161	5.3%	100	3.3%
51-60	M	112	7.6%	56	3.8%
	F	158	10.3%	106	6.9%
	All	270	9.0%	163	5.4%
61-70	M	234	15.8%	141	9.5%
	F	275	17.9%	186	12.1%
	All	510	16.9%	327	10.9%
71-80	M	528	35.7%	533	36.0%
	F	459	29.9%	546	35.6%
	All	987	32.7%	1080	35.8%
81-90	M	414	28.0%	657	44.4%
	F	368	24.0%	536	35.0%
	All	783	26.0%	1193	39.6%
91-100	M	17	1.1%	23	1.5%
	F	24	1.5%	40	2.6%
	All	40	1.3%	63	2.1%

The slight discrepancy in table numbers is because of missing data.

Table 30: Australian normed SF36V2 Scale T-scores, Australian Weights, by Self-reported Health Status

Gender	Health (a)	N	PF		RP		BP		GH		VI		SF		RE		MH	
			Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Male	Excellent	353	55.93	2.44	55.21	3.33	56.03	6.83	57.63	5.54	57.88	6.98	54.54	5.46	53.85	3.94	55.47	5.72
	Very good	582	53.16	6.42	53.28	6.26	52.93	8.13	52.45	6.29	53.74	6.97	53.01	6.18	52.61	5.45	52.89	6.91
	Good	350	50.78	7.30	49.59	8.88	47.71	9.33	47.74	7.35	49.21	7.91	50.88	8.35	51.02	8.10	49.56	8.94
	Fair	149	42.49	11.23	39.55	13.15	43.32	10.66	37.43	9.39	41.46	9.88	42.30	12.33	41.91	14.61	42.84	12.74
	Poor	44	35.92	14.02	30.65	11.76	38.14	8.67	32.07	7.79	33.30	6.80	31.10	13.13	40.99	16.23	40.26	12.55
Female	Excellent	326	54.35	4.81	54.73	4.37	56.11	7.08	58.22	4.88	55.37	7.07	54.36	4.86	53.45	4.45	54.32	6.08
	Very good	559	51.69	7.47	52.56	6.80	50.94	8.92	53.09	6.54	51.49	7.55	51.71	7.19	51.20	7.48	51.04	7.78
	Good	384	46.23	11.19	48.46	9.76	47.06	9.87	46.68	8.18	45.35	8.52	48.04	10.32	48.56	10.96	46.46	10.18
	Fair	198	39.75	12.75	39.21	11.44	39.56	8.08	37.30	9.53	38.73	8.90	40.45	12.25	41.05	15.38	41.95	13.67
	Poor	64	29.53	14.47	28.07	9.22	36.99	7.33	30.32	9.37	30.96	8.70	27.67	12.97	33.10	17.38	34.39	14.66
All	Excellent	679	55.17	3.85	54.98	3.87	56.07	6.95	57.91	5.24	56.67	7.13	54.45	5.18	53.66	4.19	54.92	5.92
	Very good	1142	52.44	6.99	52.93	6.54	51.95	8.58	52.76	6.42	52.64	7.34	52.37	6.72	51.92	6.56	51.98	7.41
	Good	734	48.40	9.80	49.00	9.36	47.37	9.61	47.19	7.81	47.19	8.45	49.40	9.53	49.73	9.77	47.94	9.73
	Fair	347	40.93	12.18	39.35	12.19	41.18	9.45	37.36	9.46	39.91	9.42	41.24	12.30	41.42	15.04	42.33	13.27
	Poor	108	32.12	14.56	29.12	10.36	37.46	7.89	31.03	8.77	31.91	8.03	29.06	13.09	36.31	17.29	36.77	14.08

The slight discrepancy in table numbers is because of missing data.
a = General health question from the HUI-3.

Table 31 presents health status by gender for the two summary scales. The same pattern as for Table 9 is also evident here: monotonic and accelerating declines for deteriorating health status, for both genders.

Table 31: Australian normed T-scores for the SF36V2 Summary Scales, Australian Weights, by Self-reported Health Status

Gender	*Health (a)*	*N*	*PCS*		*MCS*	
			Mean	*sd*	*Mean*	*sd*
Male	Excellent	353	56.83	3.95	54.17	4.85
	Very good	582	53.16	6.68	52.53	6.11
	Good	350	48.75	7.69	50.53	8.41
	Fair	149	40.54	11.98	43.04	14.01
	Poor	44	31.92	10.17	40.11	12.79
Female	Excellent	326	56.20	4.55	53.38	5.13
	Very good	559	52.28	7.10	51.02	7.69
	Good	384	46.72	10.37	47.81	10.88
	Fair	198	38.15	11.84	42.49	15.16
	Poor	64	30.15	11.10	34.31	15.85
All	Excellent	679	56.53	4.26	53.79	5.00
	Very good	1142	52.73	6.90	51.79	6.97
	Good	734	47.69	9.24	49.11	9.87
	Fair	347	39.18	11.94	42.73	14.66
	Poor	108	30.87	10.72	36.67	14.90

The slight discrepancy in table numbers is because of missing data.

5.4 The impact of Incontinence on Health Status

Finally, the effect of incontinence on health status was examined, where health status was defined by the eight SF36V2 scales, Australian weighted.

Urinary Incontinence

For urinary incontinence the details are presented in Tables 32 and 33. These show that on all eight of the SF36V2 scales there were significant declines in health status for both males and females when assessed by both the UDI-6 and the ISI. For the UDI-6 estimates the effect sizes (Cohen's d) ranged from 0.56 to 1.58, whereas for the ISI the range was from 0.67 to 1.41. For all incontinence conditions assessed, there was a monotonic decline in health status, with the exception of males on the role emotion (RE) scale for those with ISI classifications of slight and moderate incontinence.

The largest F-values were observed for the PF scale, for both the UDI-6 and the ISI, with the exception of males on the UDI-6 for GH. The smallest F-values for males were observed on the RE scale for the ISI and for females on the MH scale for the ISI.

Table 32: The Impact of Urinary Incontinence as assessed by the UDI-6 on Heatlh Status, by Gender

UDI-6 status	Gender	N.	SF36V2 scale T-scores, Australian weights															
			PF		RF		BP		GH		VI		SF		RE		MH	
			M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd
None	Male	995	53.32	6.61	52.38	8.09	52.44	9.24	52.39	8.28	53.49	8.83	52.03	8.15	52.11	6.99	52.65	7.93
	Female	614	51.12	9.42	51.35	9.12	51.67	9.76	51.87	9.71	50.43	9.88	50.13	9.87	50.58	9.21	50.42	10.00
	All	1609	52.48	7.87	51.98	8.51	52.14	9.45	52.19	8.85	52.32	9.36	51.30	8.89	51.53	7.94	51.80	8.84
Slight	Male	388	49.20	9.23	48.78	10.40	48.78	10.40	47.56	9.65	49.20	9.33	50.36	9.79	50.51	9.39	49.51	9.66
	Female	572	48.81	10.27	49.42	9.77	48.59	9.86	50.02	9.82	48.51	9.47	49.78	9.78	48.51	9.47	49.15	9.84
	All	960	48.97	9.86	49.16	10.03	48.59	9.67	49.03	9.82	48.79	9.42	50.01	9.79	49.92	9.83	49.30	9.76
Moderate	Male	65	47.24	12.05	45.86	11.95	45.86	11.95	44.36	11.96	46.61	10.04	46.77	11.29	44.16	14.65	46.91	12.10
	Female	246	45.28	11.47	47.43	10.67	46.32	10.11	46.74	10.13	45.84	10.01	46.55	11.45	45.84	10.01	45.98	10.82
	All	312	45.69	11.68	47.10	10.95	46.51	10.62	46.34	10.56	46.00	10.01	46.59	11.40	46.21	12.90	46.18	11.09
Problem	Male (a)	26	38.67	12.80	35.76	12.25	35.76	12.25	36.86	11.21	41.68	11.93	41.08	14.13	42.39	16.14	42.73	14.99
	Female	56	39.41	14.12	41.78	12.89	42.64	10.71	42.31	12.32	41.86	10.67	43.62	13.01	41.37	15.83	43.13	13.47
	All	76	39.45	13.45	40.06	12.97	42.86	10.62	41.02	12.13	41.62	11.01	43.12	12.98	41.11	16.06	42.80	13.97
Major problem	Male		N/A		N/A		N/A		N/A		N/A		N/A		N/A		N/A	
	Female	43	35.06	15.98	37.36	14.74	41.17	10.74	37.05	12.99	36.83	10.59	40.36	16.26	39.27	18.34	42.26	15.08
	All	50	35.17	15.90	37.41	14.44	41.29	10.32	36.85	12.82	37.76	10.90	40.21	16.30	40.47	17.89	42.68	14.81

The slight discrepancy in table numbers is because of missing data.

Notes: a = For males includes 'Major problem' since N = 6 cases.

Statistics: ANOVA, F-values. All $p < 0.001$

	Male	58.72	47.16	26.17	59.82	39.97	20.63	30.38	26.85	
	Female	43.48	32.20	28.42	37.54	32.12	17.04	22.53	16.88	
	All	103.76	68.82	55.89	76.61	76.99	38.73	50.16	45.11	

Table 33: The Impact of Urinary Incontinence as assessed by the ISI on Health Status, by Gender

ISI status	Gender	N.	SF36V2 scale T-scores, Australian weights															
			PF		RF		BP		GH		VI		SF		RE		MH	
			M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd
None	Male	1325	52.23	7.65	51.39	8.96	51.40	9.53	51.12	8.91	52.48	9.06	51.53	8.65	51.49	7.99	51.82	8.49
	Female	948	49.97	10.13	50.25	9.71	50.50	10.08	50.75	10.02	49.36	0.87	49.61	10.21	49.93	9.87	49.67	10.24
	All	2273	51.28	8.84	50.91	9.29	51.03	9.77	50.97	9.39	51.18	9.61	50.73	9.38	50.84	8,85	50.92	9.32
Slight	Male	121	48.66	10.08	47.32	11.56	48.05	9.48	45.43	11.47	46.48	11.00	48.70	11.77	47.61	13.02	47.79	12.51
	Female	426	48.11	10.44	49.23	9.86	47.19	9.83	49.18	10.08	47.73	9.63	48.64	10.38	48.10	11.48	47.83	10.48
	All	547	48.23	10.35	48.81	10.28	47.38	9.75	48.35	10.51	47.45	9.95	48.65	10.69	47.99	11.83	47.82	10.95
Moderate	Male (a)	29	40.62	14.48	40.94	12.47	46.31	9.29	42.77	12.72	45.24	9.82	44.83	11.26	48.66	9.90	47.16	10.88
	Female	117	42.83	13.26	45.97	12.21	46.59	10.33	45.33	11.94	45.03	10.75	47.17	12.03	46.96	13.75	46.58	11.20
	All	141	43.17	12.80	45.56	11.93	46.82	11.00	45.32	11.76	45.43	10.34	47.23	11.46	47.13	13.22	46.79	11.07
Severe/Very	Male		N/A		N/A		N/A		N/A		N/A		N/A		N/A		N/A	
severe	Female	40	32.40	14.36	35.44	13.70	41.58	11.09	38.64	12.13	38.01	10.01	40.81	15.56	41.50	17.41	43.29	14.73
	All	45	31.08	14.78	34.65	13.82	41.24	10.69	37.76	12.46	37.65	10.15	39.81	15.62	42.65	16.85	43.37	14.43

The slight discrepancy in table numbers is because of missing data.

Notes: a = For males includes 'Severe/Very severe' since N = 5 cases.

Statistics: ANOVA, F-values. All $p < 0.001$

	Male	38.83	27.53	10.49	31.39	30.89	12.78	12.64	14.59	
	Female	48.13	32.40	21.12	26.25	22.40	10.41	10.95	8.75	
	All	105.00	57.03	38.31	46.93	58.35	27.34	25.40	27.53	

Generally, with the exception of males on the UDI-6, it wasn't until respondents were classified at the worst incontinence level that SF36V2 scale scores were below 1-standard deviation from those classified with no urinary incontinence symptoms. The exception was for the UDI-6 for males where on the PF, RF, BP and GH scales those who were classified with urinary problems obtained SF36V2 scores <40.00. The reason for this was related to the amalgamation of "problem" with "major problem" for these males brought about by the very small number of cases classified with "major problems".

The apparent greater sensitivity of the SF36V2 for UDI-6 status when compared with ISI status may be explained by the fact that the UDI-6 measures both incontinence and the effects of incontinence, whereas the ISI is a purer measure of incontinence per se. It is likely, therefore, that the UDI-6 overstates the effect of urinary incontinence on health.

Faecal Incontinence

The impact of faecal incontinence on health status was also examined (Table 34). Although, as the table shows, there were significant differences by faecal incontinence status on all eight SF36V2 scales, for males there was a lack of monotonicity on the BP and GH scales. It was also apparent that there was very little impact on males' PF for those classified as suffering weekly faecal incontinence when compared with those with no symptoms. The decline in PF scores was almost entirely due to males with daily faecal incontinence. The effect sizes, when compared with those with no symptoms, were d = 0.29 for those with weekly faecal incontinence and 0.77 for those with daily incontinence. For females the declines in SF36V2 scale scores were more even across faecal incontinence levels.

In general, however, the impact of faecal incontinence on health status failed to reach more than 1 standard deviation from the T-score norm, which suggests that faecal incontinence has about the same impact, or even perhaps a slightly smaller impact, than does urinary incontinence as measured by the ISI. This observation is consistent with the findings from the utility instruments reported in section 4 (see Tables 17 and 18).

Soiling

The impact of soiling was examined and showed that, in general, soiling was associated with a loss of about 0.5 of a T-score standard deviation from the norm. This was consistent for both males and females across all SF36V2 scales, as shown in Table 35.

Table 34: The Impact of Faecal Incontinence as assessed by the Wexner on Health Status, by Gender

Faecal status	*Gender*	*N.*	*SF36V2 scale T-scores, Australian weights*													
			PF		*RF*		*BP*		*GH*		*VI*		*SF*		*RE*		*MH*	
			M	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*
Never	Male	1014	52.35	8.03	51.73	8.89	52.06	9.26	51.63	8.85	52.74	9.24	52.04	8.12	51.89	7.44	52.23	8.53
	Female	960	49.96	9.91	50.41	9.42	50.50	9.82	50.71	9.98	49.46	9.84	49.56	10.28	49.97	9.86	49.53	10.04
	All	1974	51.19	9.07	51.09	9.17	51.30	9.57	51.18	9.43	51.15	9.68	50.83	9.32	50.96	8.75	50.91	9.39
Rarely	Male	280	50.63	8.97	49.90	9.75	49.17	9.85	48.79	9.32	50.81	9.22	49.93	10.37	50.82	8.74	50.57	8.83
	Female	320	47.97	11.32	48.89	10.60	47.83	10.57	49.49	10.28	47.53	10.16	48.63	10.45	48.44	11.30	48.15	10.81
	All	600	49.21	10.37	49.35	10.21	48.45	10.25	49.16	9.85	49.06	9.86	49.24	10.42	49.55	10.25	49.28	10.00
Sometimes	Male	112	50.53	7.77	49.12	9.95	49.74	9.91	47.19	11.00	49.75	8.46	49.67	9.12	49.25	10.96	48.13	10.58
	Female	156	44.65	12.28	46.62	11.78	45.17	10.26	46.09	11.52	45.31	10.07	48.10	11.04	46.77	13.08	46.94	11.59
	All	267	47.11	11.00	47.67	11.10	47.08	10.35	46.55	11.30	47.16	9.97	48.75	10.29	47.81	12.28	47.43	11.18
Weekly	Male	45	50.12	7.31	47.67	9.95	47.35	10.04	47.28	11.26	47.79	10.12	49.13	9.86	47.01	11.78	47.40	12.0
	Female	58	42.35	14.63	44.19	13.26	45.84	10.34	44.88	12.15	44.94	11.91	45.32	13.25	44.68	15.52	46.37	12.81
	All	103	45.75	12.55	45.71	12.01	46.50	10.19	45.93	11.77	46.18	11.20	46.98	42.25	45.69	14.00	46.82	12.45
Daily	Male	29	44.64	11.77	39.29	14.52	43.49	9.87	43.12	11.32	43.24	13.48	40.73	15.92	40.00	16.51	45.87	12.84
	Female	40	37.59	14.10	40.78	12.95	42.51	10.24	43.00	11.00	41.40	10.53	43.36	13.82	42.85	15.44	44.23	13.27
	All	70	40.56	13.54	40.15	13.55	42.92	10.03	43.05	11.05	42.18	11.81	42.25	14.68	41.65	15.84	44.92	13.02

The slight discrepancy in table numbers is because of missing data.

Notes:

Statistic: ANOVA, F-values. All $p < 0.001$

	Male	9.07	16.61	12.63	15.80	13.06	15.17	19.15	11.45	
	Female	24.46	16.35	17.72	14.44	13.53	5.62	9.14	5.28	
	All	35.22	33.16	30.73	29.70	28.64	18.59	25.83	16.77	

Table 35: The Impact of Soiling on Health Status, by Gender

Soiling status	Gender	N.	SF36V2 scale T-scores, Australian weights															
			PF		RF		BP		GH		VI		SF		RE		MH	
			M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd	M	sd
Never	Male	1390	52.07	8.02	51.19	9.15	51.36	9.45	50.86	9.21	52.21	9.27	51.42	8.79	51.33	8.38	51.73	8.72
	Female	1406	49.14	10.47	49.73	10.00	49.51	10.11	50.16	10.19	48.71	10.03	49.31	10.37	49.35	10.61	49.11	10.37
	All	2796	50.60	9.45	50.46	9.61	50.43	9.83	50.51	9.72	50.45	9.82	50.36	9.67	50.34	9.61	50.41	9.67
Symptoms	Male	83	45.44	9.82	44.97	12.11	45.62	10.30	44.24	10.76	45.82	10.18	46.74	12.21	47.43	11.43	45.82	11.88
	Female	127	40.32	14.45	43.65	12.85	43.61	10.42	43.04	11.69	43.28	10.69	44.55	12.87	44.50	14.35	44.62	12.29
	All	210	42.34	13.04	44.17	12.55	44.41	10.40	43.51	11.32	44.28	10.54	45.41	12.63	45.65	13.32	45.09	12.12

The slight discrepancy in table numbers is because of missing data.

Notes:

Statistics: ANOVA, F-values. All $p < 0.001$

Male	52.24	34.74	28.68	39.72	36.73	21.10	16.23	34.43
Female	77.01	41.01	39.52	55.47	33.84	23.49	22.89	21.18
All	140.50	79.69	72.92	98.80	76.29	48.60	43.56	56.91

Table 36: Australian normed T-scores, based on the Published US Weights, for the SF36V2 Scales, by Age and Gender

Age	Gender	N	PF		RP		BP		GH		VI		SF		RE		MH	
			Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
15-19	M	129	55.88	4.07	54.95	4.22	56.26	7.18	54.89	6.79	57.34	8.66	54.22	5.74	54.40	4.59	56.95	6.01
	F	124	53.53	4.09	53.03	6.33	51.61	8.32	49.63	9.49	50.63	8.92	49.47	8.20	49.92	8.00	49.83	11.32
	All	253	54.73	4.24	54.01	5.43	54.48	7.95	52.32	8.62	54.06	9.39	51.90	7.43	52.21	6.85	53.47	9.97
20-29	M	244	55.67	2.52	54.16	5.13	55.41	7.39	53.87	7.77	54.89	7.48	53.01	6.15	53.52	4.94	54.41	7.35
	F	230	54.65	4.25	52.22	7.29	52.60	8.32	50.99	10.52	49.84	10.17	49.20	10.05	50.42	9.39	50.97	10.26
	All	474	55.18	3.50	53.22	6.34	54.04	8.25	52.47	9.31	52.43	9.23	51.15	8.48	52.01	7.60	52.74	9.04
30-39	M	266	53.85	5.55	52.54	8.78	52.73	8.89	52.64	9.38	53.66	8.53	51.83	8.83	52.62	6.83	54.42	8.57
	F	263	53.15	6.08	52.88	7.09	53.49	8.64	53.66	8.72	50.68	9.71	51.27	8.69	50.85	9.28	51.88	9.47
	All	529	53.51	5.83	52.71	7.98	53.11	8.76	53.15	9.06	52.18	9.24	51.56	8.76	51.74	8.18	53.16	9.11
40-49	M	275	53.53	5.23	53.14	8.09	53.78	8.00	51.58	8.82	52.73	10.28	52.31	9.37	53.14	7.76	53.24	9.33
	F	278	50.52	9.26	51.07	9.86	51.21	9.29	50.72	11.08	49.99	10.35	50.07	10.83	50.92	10.05	51.80	10.40
	All	553	52.01	7.67	52.10	9.07	52.49	8.76	51.15	10.02	51.35	10.40	51.18	10.19	52.03	8.64	52.34	9.90
50-59	M	239	50.28	8.57	49.35	10.80	50.84	9.48	47.97	11.41	50.88	10.92	49.96	9.81	52.05	8.41	52.88	9.79
	F	244	47.70	9.44	48.83	11.35	49.62	9.35	48.95	11.84	48.37	10.67	49.26	10.99	51.18	9.07	51.81	10.40
	All	482	48.98	9.10	49.08	11.07	50.23	9.42	48.46	11.63	49.62	10.85	49.61	10.41	51.61	8.75	54.39	10.05
60-69	M	156	48.19	9.61	48.99	11.59	50.89	8.93	47.97	11.15	52.98	10.88	51.70	9.76	52.65	8.04	55.81	9.26
	F	161	46.07	11.04	48.55	11.00	50.54	9.13	49.02	11.49	51.18	11.39	51.04	10.39	52.05	8.49	53.70	10.18
	All	318	47.12	10.40	48.77	11.28	50.71	9.02	48.50	11.32	52.07	11.16	51.36	10.08	52.35	8.26	54.14	9.46
70-79	M	127	46.19	9.70	47.05	11.46	51.85	8.68	47.27	9.92	52.13	10.68	51.31	9.88	51.79	8.16	55.97	9.26
	F	158	40.61	12.67	45.03	12.64	49.07	9.63	46.42	11.70	47.82	11.69	48.77	11.29	51.65	8.38	52.67	10.02
	All	285	43.10	11.76	45.92	12.15	50.31	9.31	46.80	10.93	49.74	1.44	49.90	10.74	51.71	8.27	54.14	9.81
80+	M	44	42.41	12.25	45.03	12.47	52.07	9.02	47.25	10.19	51.08	11.58	51.17	11.90	52.50	9.33	56.83	9.26
	F	78	34.54	11.84	41.29	12.93	49.24	9.95	46.13	11.17	46.06	12.39	46.35	13.51	52.91	6.72	54.40	8.70
	All	121	37.38	12.53	42.64	12.84	50.26	9.68	46.53	10.80	47.87	12.30	48.09	13.11	52.76	7.73	55.28	8.94
Total	M	1480	52.10	7.69	51.55	9.36	53.08	8.62	50.98	9.84	53.28	9.83	51.94	8.86	52.86	7.04	54.45	8.72
	F (a)	1534	49.09	10.24	49.97	10.24	51.28	9.22	50.04	10.95	49.59	10.59	49.75	10.41	51.09	9.04	51.95	10.18
	All (b)	3014	50.57	9.20	50.75	9.85	52.16	8.98	50.50	10.43	51.40	10.39	50.82	9.74	51.96	8.16	53.18	9.57

The slight discrepancy in table numbers is because of missing data.

a = For PF and GH there were 1535 females; b = For PF and GH the total was 3015.

Table 37: Australian normed T-scores for the SF36V2 Summary Scales, based on US Weights, by Age and Gender

Age	Gender	N	PCS		MCS	
			Mean	sd	Mean	sd
15-19	M	129	55.24	4.34	55.50	6.48
	F	124	53.37	5.82	48.45	10.07
	All	253	54.33	5.19	52.06	9.12
20-29	M	244	55.02	8.08	53.22	6.45
	F	230	53.52	5.82	48.77	11.26
	All	474	54.29	5.87	51.06	9.37
30-39	M	266	52.67	8.08	53.13	8.74
	F	263	53.80	7.05	50.31	10.23
	All	529	53.23	7.60	51.73	9.60
40-49	M	275	52.94	6.65	52.61	8.95
	F	278	50.58	9.17	50.91	10.84
	All	553	51.75	8.10	51.76	9.97
50-59	M	239	48.65	10.09	52.45	9.28
	F	244	47.56	10.81	51.53	10.50
	All	482	48.10	10.46	51.99	9.92
60-69	M	156	47.09	10.81	55.28	8.68
	F	161	46.56	10.99	54.29	9.76
	All	318	46.82	10.89	54.78	9.24
70-79	M	127	45.64	9.53	55.63	8.99
	F	158	42.13	12.26	54.09	9.29
	All	285	43.69	11.25	54.77	9.17
80+	M	44	43.04	11.35	57.17	9.38
	F	78	37.56	12.13	56.26	7.85
	All	121	39.54	12.10	56.59	8.41
Total	M	1480	51.21	8.82	53.71	8.48
	F	1534	49.37	10.40	51.34	10.50
	All	3014	50.27	9.70	52.50	9.63

The slight discrepancy in table numbers is because of missing data.

5.5 Supplementary Material

Although this report has presented SF36V2 T-score Australian population norms based on using Australian-derived weights which reflect how respondents perceive health, there are occasions when it will be preferable to report scores based on the original US weights. Such circumstances would include Australian-US studies where a common metric is needed.

Readers, however, should be aware that for the 8 SF36V2 scales, the use of Australian weights affects only the T-scores for the 8 scales and for the 2 summary measures, the PCS and MCS. If percentage scores are reported, the issue of weights does not apply.

The data reported in Table 23 provides percentage scale norms for the 8 scales; these data are non-weighted and may be used for direct comparison with data from the US or any other country. Similarly, the data presented in Table 25 are based entirely upon the published US weights. Thus, these data are directly comparable with that computed using the published US weights. Tables 26 and 27, however, present T-score norms based on derived Australian weights. For those who would prefer to use T-score norms based on the US weights, the data in Tables 36 and 37 should be used.

Comparison of the Australian and US-weighted population norms is presented in Table 38. This shows that there were statistically significant differences on all 8 SF36V2 scales between

the Australian- and US-weighted scores. The magnitude of these differences was described by Cohen's d (112).

The results suggest that for scales measuring physical health there is little difference between scores weighted by Australian and US weights. The mean d across those scales positively contributing the PCS was 0.10 (sd = 0.09), and the largest d was 0.23 for BP. In contrast, for those scales contributing positively to mental health there appears to be small d sizes between the Australian and US weighted scores. The mean d across the scales contributing to the MCS was 0.19 (sd = 0.10), and the largest d was 0.32 for the MH scale, which is between a small to moderate effect.

These findings suggest that there are small but important differences in weighted health states between the Australian and US populations, as measured by the SF36V2. The implication is that Australians and Americans have somewhat different underlying health constructs, particularly regarding mental health.

5.6 Discussion

This part has presented Australian weighted population norms for the SF36V2, based on data collected in the 2004 SAHOS. For those wishing to use US-weighted norms, the tables in the Supplementary Material section should be consulted.

Since the publication of the SF36V2, there have been two other validation population-based studies. Neither, however, reported detailed population norms for all SF36V2 scales and summary scales or proportions within deciles, although some norms for the eight scales were presented in Jenkinson et al (135). Although considerable data are provided in the SF36V2 user manual, these data are based on US values only and, as discussed below, there may be good reasons for researchers to derive their own local weights. This study demonstrates how this can be done and reports some implications, using Australian data as an exemplar.

To make the Australian norms useful to researchers, the results have been presented in considerable detail, by age group, gender, T-score decile and respondent health status, and also by percentage scores as well as the recommended T-scores.

Table 38: Differences between Australian-weighted and US-weighted SF36V2 Scale Scores

SF36V2 scale	*Statistics (a)*		*Cohen's d*
	z	*p*	
PF	-48.49	<0.0001	0.05
RP	-49.33	<0.0001	0.08
BP	-48.44	<0.0001	0.23
GH	-41.23	<0.0001	0.05
VI	-48.24	<0.0001	0.14
SF	-49.62	<0.0001	0.08
RE	-50.70	<0.0001	0.21
MH	-48.28	<0.0001	0.32

Note: a = Wilcoxon Signed Ranks Test

As shown in Tables 23, 24 and 25, there are important differences between the percentage, factor coefficients and normed scores obtained in the SAHOS survey when compared with the US norms published in the SF36V2 manual (15). The critical issue is to correctly interpret these differences with respect to both the use of the SF26V2 and the computation of population norms. Although the SF36V2 developers have argued that the SF36V2 is an international version, the weights behind both the items and scales were not derived from international samples; they were derived exclusively from US samples. Of particular concern is the use of the 1990 factor weights for computing the PCS and MCS summary scales given that Ware et al report that there were significant differences between the 1990 and 1998 surveys, including sampling bias in 1990 (15). Whether these US weights should be accepted as the international standard is thus open to discussion.

The US weights – and those reported in this study – were derived from exploratory factor analysis (EFA) factor loadings. EFA is based on identifying common patterns within a dataset through grouping variables which share similar response patterns. The axiom behind this procedure is that the mental processes used by respondents when answering questions possess an orderly structure which is meaningfully reflected in the data, and which can be extracted through statistical treatment of the intercorrelations between items. The patterns thus identified reflect an underlying construct which is described by a series of 'vectors'. Each vector represents a dimension within the construct (e.g. within the construct health, there may be different vectors to describe physical, social and mental health). The extent to which any particular item is related to a given vector is represented by its factor loading: the higher the factor loading the greater the association between the item and the vector (e.g. an item measuring 'running' might be highly related to the physical health vector, but not to the mental health vector). Because the vector structure within a construct is concomitant with factor loadings (i.e. the two are mathematically related and occur at the same time) the implication is that the vector structure describes how people understand the construct of interest (140). If two different samples have the same vector construct and factor loadings (or weights) for individual items, then it is argued the two samples come from a single underlying population.

Where, however, two samples have different vector constructs and factor loadings, then the implication is that they have different understandings of what is being measured. In the current study, as shown in Table 24, there appear to be small but important differences in how health is conceptualised between the Australian SAHOS sample and the US samples from which the SF36V2 was derived.

Differences between samples like this are quite common in cross-cultural research, and it cannot be assumed that different cultural groups share common constructs. Generally, there are two approaches to this problem of cross-cultural equivalence. On the one hand it has been argued that a rigorous approach to translation, reliability and equivalence across cultures ensures cross-cultural equivalence. This is essentially the position adopted by the IQOLA group and the SF36 developers (126, 136, 141). This was the position adopted by Sanson-Fisher and Perkis for the Australian version of the SF36V1 (119). The difficulty is that the descriptive system itself may be culture-bound. For example, in the case of the Australian coordinated care trials the SF36V1 was deemed unsuitable for use with remote Aboriginal communities due to inappropriate items, such as a person's ability to climb stairs. The alternative position is that for cross-culture validity the descriptive system of a measure must be internationally developed, but perhaps scored with local variations. This is the position behind the World Health Organization's quality of life (WHOQOL) instruments (142-144).

The relevance of this discussion to the present study is in relation to the adoption of local weights based on factor loadings for scoring the SF36V2 and the production of Australian population norms. In the case of the item response weights, referred to as 'recalibration' in the SF36V2 manual, the procedures and values were replicated in the IQOLA international studies (131). Verification of these was not part of the SAHOS, and these weights have been accepted. For the coefficient weights, the data suggest there are differences between the US and Australian samples (Table 24). The source of these differences is unknown and it could be any of the following (or several in combination): differences in cultural perception of the descriptive system, differences in the population samples (quota sampling based on age, gender and income in the US versus list sampling based on geographic location in the SAHOS), demographic characteristics (e.g. gender, age, birth country, race, education); differences in data collection procedures (mail administration in the US versus interview in the SAHOS); differential item functioning (DIF); or to actual differences in health status. Collectively, these suggest there may be emic differences between the US and Australia, thus there is a prima facie case for Australian weights to be used in scoring the SF36V2 when used in Australian samples. This position has been adopted in this paper, and it is the position that Jenkinson et al implicitly supported in their UK validation study of the SF36V2 where British factor coefficients were presented (135).

The use of local means and standard deviations, however, does have ramifications for the calculation of T-scores. Although T-scores enable the comparison of different tests on a common metric, they do not remove skew from data. T-scores are based on estimates of data distribution; as such the range of T-values is a function of the variance expressed as the standard deviation. Whilst this is not a problem where data are normally distributed, it artificially inflates reported score ranges where there are ceiling or floor effects which limit the standard deviation. This is the case with the SF36V2 scales, as shown in Tables 28 and 29. The better the health of a sample, the

higher the proportion that will achieve a ceiling-score, thus restricting the standard deviation. For those in poor health, their T-scores, expressed in standard deviations from the mean (50), will be highly skewed. Because it is the convention to report mean SF36V2 scores (see Tables 26 and 27), these effects (ceiling restricted standard deviations and skew) are largely hidden whereas they are made explicitly clear in the decile tables (Tables 28 and 29). Incidentally, that they are also present in the US norm data can be inferred from Tables 8.2 to 8.4 in the SF36V2 user manual where T-scores are identical for several scales between the 50th and 75th percentiles (p63-72, 15). These observations about the effect of conversion of percentile to T-scores calls into question the decision by Ware et al to report SF36V2 scales as T-scores (15) because T-scores may, unwittingly, lead some to assume data are normally distributed and to use parametric tests where non-parametric tests would be more appropriate. Perhaps, based on these results, medians should be the standard for reporting the SF36V2 rather than means.

In the case of the Australian scores presented in this study, based on the means and standard deviations presented in Table 25, it was observed that the standard deviations were smaller than those reported by Ware et al (15). This pattern is almost certainly due to the higher mean scores obtained from the SAHOS sample, thus restricting the standard deviations because of ceiling effects. Conversion to T-scores has had the effect of spiralling downwards the scores of those in poor health (because these cases' scores are based on the number of standard deviations away from the mean). This statistical phenomenon partly explains the highly skewed data distributions reported in Tables 28 and 29 which are not obvious when means are reported, as in Tables 26 and 27. Researchers should therefore test their SF36V2 data for evidence of skewness prior to conversion to T-scores. Where significant skewness is present, the data should either be transformed to achieve normality prior to data analysis or non-parametric analysis methods used.

When these results are compared with the results for two other patient-outcome measures which have had Australian norms reported recently, the WHOQOL-Brèf (145) and the AQoL (6) instruments, skewed data and ceiling effects appear to be more of a problem for the SF36V2 (the range was 5% to 79%; half of the scales with >50% in the top decile): for the WHOQOL-Brèf the proportion in the top decile ranged from 14% to 17%, and it was 45% for the AQoL. Researchers looking for better data distributions might consider these other instruments or reporting the two SF36V2 summary scales in preference to the 8 scales.

Regarding Australian population norms, these are presented in Tables 26 and 27, based on a representative sample of the population. To enhance the usefulness of the data to researchers, they have been presented by age group and gender, in accordance with the IQOLA recommendations (136). Additionally, the data have also been presented by health status, broken down by gender (Tables 30 and 31).

Table 26 shows that for PF there is a small and consistent decline for males between 15 to 50 years followed by an accelerated decline, particularly after 70 years. For females the decline starts at about 40 years followed by a more rapid decline, again particularly after 70 years. The same pattern is evident among males on the RP scale, although without the acceleration in older age. For females, the decline starts at about 40 years, and then accelerates at about 70 years. For BP, for males there is an increase in pain (i.e. a decrease in scores) during the 30s but which seems to stabilize during the 50s. For females, there appears to be a slight increase in the 40s, followed by further deterioration in the 70s. For GH, for males there is a small and consistent decline between 15 to 50 years, followed by a drop in GH which is then stabilized. For females, GH improves until the age of about 40 when it starts a small and consistent decline. Regarding vitality, there are small variations across the lifespan for males, but there doesn't appear to be any particular trend. For females, VI scores are consistent with small variations until about the age of 70 years when decreases are noticeable. For the SF scale, for males there is minor variation across the lifespan, but there doesn't appear to be any particular pattern, other than a dip during their 50s. For females, there is small variation across the lifespan, with a decline in their 80s. For RE, for males there are small variations across the lifespan, but no particular pattern in this variation. For females, there seems to be a consistent slight increase over the lifespan. For MH, for both males and females there appears to be a slight U-shape to the data with the highest values being reported for young and old.

Table 27 provides similar norms for the two summary scales. For the PCS, for males there is a small but consistent decline until 50 years, when the decline accelerates. For females, the decline seems to start about 10 years earlier at about age 40. For the MCS, for males there seems to be a U-shaped pattern, with a progressive decline until about 50 years, followed by a progressive

increase. For females there is a small and consistent increase across the lifespan.

Tables 30 and 31 present mean (SDs) scores by self-reported health status, based on the health status question from the HUI-3 (139). In all cases there was a monotonic decline, which accelerated with poorer health status.

When the association between health status (measured by the SF36V2 scales) and incontinence classification was examined, in general, monotonic and statistically significant declines in health status were observed on all eight SF36V2 scales. This was the case across all four incontinence measures and it was the case for both males and females (Tables 32 to 35). An important finding was that the SF36V2 scales were more sensitive to the UDI6 classification than to the ISI classification, and it is thought that this reflects that the UDI-6 measures the impact of urinary incontinence rather than incontinence alone. A second important finding was that faecal incontinence, in general, appeared to have less of an association with health status than did urinary incontinence.

5.7 Conclusion

Although many researchers will still be using the SF36V1, particularly in longitudinal studies, the use of the SF36V2 will rise rapidly given its superior psychometric properties. It will replace the SF36V1 as the world's ubiquitous health status measure. Given that this is claimed to be an international instrument, it is important that the descriptive system and weights used can be demonstrated to be culture-free. If there are emic effects, then local weights should be published and used.

Although this study was not a validation of the SF36V2 per se, the findings suggest that there are important differences between the US samples used for the SF36V2 weights and the Australian sample reported here. Consequently, local weights were also derived using the identical methods of the instrument developers and have been used to report local population norms. Given the limitations of these methods, it is quite likely a better model could be constructed using more sophisticated methods. A computer algorithm with the Australian weights is available from the author. It should be noted that the SF36V2 is copyright, and that researchers must have purchased a licence to use it from QualityMetric prior incorporating it into a study.

Population norms provide guidelines for interpreting SF36V2 scores. When available by age group, gender or instrument decile, these are particularly useful for describing populations, providing benchmarks for the proportion of cases returned to good or full health, or they may provide yardsticks against which the effectiveness of interventions can be assessed. The methods used in this paper for deriving local weights and population norms may be useful to other researchers, as may the populations norms presented.

5.8 Recommendations

This study has provided Australian weights for computing T-scores for the SF36V2 scales and for the two summary scales. This has been done in the interests of enabling researchers to make direct comparisons between their samples and the Australian population generally. To assist researchers, these estimates have been provided in considerable detail, including by age group, gender, score deciles and health status.

It is recommended that these estimates are used in studies involving Australian samples, and that the data are reported as being weighted with Australian values.

This study has also shown that there are small and important differences in the underlying health concept between Australians and Americans, as expressed through different factor loadings and consequent scale weights. For those researchers wishing to directly compare their data with US samples, tables providing US-weighted norms have been provided.

It is recommended that there be further research into the SF36V2 in Australia, including research into the reasons why different factor loadings were obtained and the meaning of these differences.

Finally, the findings show that the SF36V2 scales are sensitive to incontinence status, and it would appear that the SF36V2 would be an appropriate measure to use in incontinence studies for the assessment of health status.

6. Concluding Summary and Recommendations

Based on a population sample of South Australians, involving 3015 participants in the 2004 South Health Omnibus Survey weighted by Australian Bureau of Statistics population estimates to achieve representativeness, this report examined three aspects of incontinence in Australia. It provides:

- Population prevalences (section 3)
- Estimates of the impact of incontinence on peoples' quality of life (section 4)
- An estimate of the effect of incontinence on peoples' health using the SF-36 Version 2 (section 5).

Urinary incontinence was measured by the Incontinence Severity Index (ISI) and the Urogenital Distress Inventory – Short Form (UDI-6). Faecal incontinence was assessed by the Wexner Continence Grading Scale (Wexner). Soiling was measured by two additional questions.

Quality of life was assessed by utility, which is the value of quality of life to a person. Utility scales use 1.00 to represent the best possible quality of life, and 0.00 represents death-equivalent states. The utility scales used in this study were the Assessment of Quality of Life (AQoL), the EQ5D, the Health Utilities Index – 3 (HUI3), the 15D and the SF6D (derived from the SF36). The psychometric properties of each of these instruments was assessed, along with their sensitivity to incontinence.

Health status was assessed with the SF36V2 (SF-36 Version 2). Australian norms are provided, as are estimates of the association between incontinence and health status.

6.1 Results

Incontinence Prevalence

For urinary and faecal incontinence, the best estimate based on the ISI and Wexner measures was that the prevalence of any incontinence is 27% (95%CI: 26% – 29%). For females it is 40% (38% – 43%) and for males 14% (12% – 15%).

Based on self-report of any symptoms of urinary leakage, the ISI estimated prevalence of urinary incontinence was 24% (95%CI: 23% – 26%) overall. When broken down by gender, it was 38% (95%CI: 36% – 41%) for females and 10% (95%CI: 9% – 12%) for males. When measured by the UDI-6, which measures being bothered by symptoms, the overall prevalence of urinary incontinence was 47% (95%CI: 45% – 48%); for females it was 60% (95%CI: 58% – 63%) and for males 33% (30% – 35%). These estimates for the UDI-6 are confounded due to its poor psychometric properties; thus the ISI estimates are preferred.

For faecal incontinence, the standard Wexner Scale data suggested that the prevalence was 35% (95%CI: 33% – 36%). For females this was 38% (95%CI: 35% – 40%), and for males it was 32% (95%CI: 29% – 34%). However, the Wexner includes flatus, which is excluded from the current International Continence Society faecal incontinence definition. If the flatus question is excluded from the Wexner, the data show that the prevalence would be 8% (95%CI: 7% – 9%). For females this would be 10% (95%CI: 8% – 11%) and 6% (5% – 7%) for males. In the interests of consistency with international definitions, these modified prevalence estimates are preferred.

The impact of Incontinence on Quality of Life

The impact of incontinence on quality of life was assessed by the world's leading five utility instruments (the AQoL, EQ5D, HUI3, 15D and SF6D). First, population norms for all five measures were computed, and then the disutility (i.e. the loss of quality of life[10]) due to incontinence assessed.

Regarding population norms, for the AQoL the mean utility was 0.81 (SD = 0.20), for the EQ5D it was 0.82 (0.22), for the HUI3 it was 0.82 (0.21), for the 15D it was 0.93 (0.08) and for the SF6D it was 0.81 (0.14).

[10] Disutility is the difference between the norm and the quality of life state of interest. For example, if the norm is 0.75 and for those with urinary incontinence it is 0.68, then the disutility associated with urinary incontinence would be 0.75-0.68 = 0.07.

When the impact of incontinence on quality of life was assessed, the data showed that incontinence has a small to mild effect upon quality of life. The range in disutility for those with moderate urinary incontinence as measured by the ISI was between 0.08 (15D) and 0.14 (AQoL). For those with weekly faecal incontinence (measured by the Wexner) the range was from 0.07 (15D) to 0.15 (EQ5D).

When the utility instruments were assessed by responsiveness to incontinence, it was observed that the most sensitive instrument for urinary incontinence was the 15D, then the HUI3 and AQoL. The EQ5D and SF6D were less sensitive. For faecal incontinence the most sensitive instruments were the 15D, AQoL and EQ5D. The HUI3 and SF6D were less sensitive. Overall, urinary incontinence as measured by the ISI explained between 2-7% of the variance in utility scores, and faecal incontinence as measured by the Wexner between 5-13%.

These data suggest that the different utility instruments do not provide equivalent estimates of the impact of incontinence on quality of life. Interestingly, the most sensitive instrument (the 15D) was also the instrument which reported the smallest effect of incontinence on quality of life.

Consequently, the psychometric properties of the five utility instruments were examined using a combination of classic, modern and econometric test theory. The results suggested that there were particular measurement difficulties with the 15D, because it is not weighted with a preference-based technique, it uses an additive scale which prevents loss of utility for severe health states, and the data from respondents was found to provide a poor fit to the 15D utility model. There were also measurement difficulties with the SF6D due to the restricted scoring range. The lower boundary for the SF6D is 0.30, which implies that while scores are well reported for those with 'healthy' conditions, for those with severe health conditions there is an ever-increasing gap between the theoretical utility model (score range 0.00 to 1.00) and obtained scores. For the EQ5D two measurement problems were observed. Examination of its internal structure suggested that the 5 items were measuring two different constructs which led to difficulties with the underlying measurement model. A second issue concerned the obtained data distribution: the scores were 'lumpy' and clustered around certain values. This lumpiness is caused by the presence of an additional weight that comes into effect whenever a person endorses the worst health state level on any EQ5D item. The effect of this additional weight is to cause an increase/decrease of utility between 0.1 and 0.3. The impact of this additional EQ5D weight is to confer increased sensitivity on the EQ5D whenever a respondent moves from a level-3 endorsement to a level-2 endorsement. It also, however, has the effect of undermining the necessary interval property needed for use during cost-utility analysis.

The two better performing instruments were the AQoL and HUI3. Both possessed good psychometric properties, with the AQoL performing slightly better (e.g. it was the more reliable of the two and had the better data to model fit indices). No particular problems were identified for either of these two measures.

In conclusion, there were substantial differences in scores between the MAU-instruments such that utilities obtained from one measure cannot be assumed to be compatible with those from the other measures. These differences reflect different descriptive systems, assigned weights, and scoring mechanisms. That these deliver utilities that are statistically significantly different across a wide range of values, suggests the results for the different instruments cannot all be right, and that study results may depend upon the instrument chosen as much as actual treatment benefits.

Incontinence and Health Status

Examination of the psychometric properties of the SF36V2 suggested that there were important differences between the Australian and US versions, both in the descriptive systems and in the obtained scale scores. In addition, when Australian factor weights for the two summary scales (PCS (physical health) and MCS (mental health)) were computed using the identical methods used by the SF36V2 developers differences were also observed. Given the limitations of these methods, it is quite likely a better model could be constructed using more sophisticated methods. These findings, however, suggest that there are differences between the US samples used for the SF36V2 weights and the Australian sample reported in this study. Consequently, Australian weights were used in reporting the study findings. A feature of the SF36V2, when compared with the SF36V1, is that all scale scores are reported as T-scores. Based on the Australian weights, therefore, all of the eight sub-scales and the two summary scales have population norms of 50 and standard deviations of ±10 points.

When examined by age and gender, for physical function (PF), role physical (RP), bodily pain (BP) and general health (GH), although there are differences between males and females, in general there are progressive declines over the lifetime. For the other scales (vitality (VI), social function (SF), role emotion (RE) and mental health (ME)) there were small variations over the lifetime. On the physical summary scale (PCS) for both genders there was a progressive decline over the lifetime, but this was not evidenced for the mental summary scale (MCS).

In addition to these population norms, the proportion of cases within scale score deciles were examined. This revealed that for the role emotion (RE) scale 79% of all cases fell within the top decile, as did 64% for the role physical (RP) scale, 61% for the social function (SF) scale and 54% for the physical function (PF) scale. These findings are suggestive of extreme skew on these scales, and it is recommended that researchers should either transform their data prior to analysis or report medians rather than means.

When the association between incontinence status and health status as measured by the SF36V2 scales was examined, the results showed that as incontinence severity increased health status deteriorated. This was the case for all four measures of incontinence and for both males and females, although there were different patterns of decline in health status by gender. Generally, for those with severe urinary or faecal incontinence their health status was 1 standard deviation or more below the health status of those with no urinary incontinence symptoms. This finding was consistent with that of the utility instruments suggesting that severe urinary incontinence has a similar effect as severe faecal incontinence.

The SF36V2 was shown to be suitable for measuring health status in incontinence studies.

6.2 Summary of Recommendations

6.2.1 Incontinence Prevalence

The incontinence prevalence estimates reported in this study are consistent with the literature in general and suggest that urinary incontinence is a common condition, particularly among females.

- To adequately quantify this for medical decision-making and policy direction, there is need for an excess burden of disease study.
- These data would also suggest the need for trials evaluating the relative impacts of preventive programs (e.g. pelvic floor exercises, health literacy) and acute interventions (e.g. surgery).

There is, however, considerable uncertainty over the measurement of incontinence. As this study has shown, none of the existing measures – whether for urinary or faecal incontinence – could be used with a great deal of confidence. Depending upon which instrument was used, or which items were included or excluded, there were very different prevalence estimates. This implies that all of the measures, to some degree, provided misleading estimates.

Subject to the above, the results of this study suggest that the preferred urinary incontinence measure is the ISI. It was found to possess superior measurement properties than the UDI-6. Because the UDI-6 measures the impact of urinary incontinence on peoples' lives rather than incontinence per se, it may overstate incontinence prevalence and the impact of this on peoples' lives (defined as their health status and their quality of life). Given its poor psychometric properties, there is a prima facie case for major revision of the UDI-6. Although the ISI is the preferred measure, because it violates the assumptions of classic psychometric theory relating to scale stability, further research into its properties is also recommended.

- It is recommended, therefore, that based on the SAHOS dataset a full psychometric evaluation of the urinary incontinence measures used in this study is undertaken with the intent of developing better measures, and that these revised measures are then tested in future incontinence studies.

For faecal incontinence the current definition by the International Continence Society excludes flatus, yet this is included in the Wexner. In addition to this definitional inconsistency, the evidence from this study suggested that the inclusion of flatus led to overestimates of faecal incontinence prevalence.

- It is recommended that further work on the Wexner is undertaken to remove flatus and to improve its measurement properties.

6.2.2 The Impact of Incontinence on Quality of Life

This study has shown that there are substantial differences in manifest scores between five leading generic MAU-instruments (the AQoL, EQ5D, HUI3, 15D and SF6D). The differences are such that utilities obtained from one measure cannot be assumed to be compatible with those from the other measures. (The two measures which provided the most compatible scores are the EQ5D and HUI3.) This key finding provides empirical evidence supporting Thomas et al's (27) review, which came to the same conclusion based on examination of the published literature (see Appendix A).

The inconsistencies reported in this study reflect different descriptive systems, assigned weights, and scoring mechanisms. That these deliver utilities that are statistically significantly different across a wide range of values, suggests the results for the different instruments cannot all be right. When taken in conjunction with the differences in implied QALYs, effect sizes and relative efficiencies, they are suggestive that study results may depend upon the instrument chosen rather than actual treatment benefits.

Regarding recommendations, the results of this study support those reached by Thomas et al (27),.

- It is recommended that two utility measures should be included in any particular study and that both sets of results should be reported with appropriate sensitivity analyses.
- The preferred instrument would be the Australian AQoL since it performed at least as well as any of the other MAU-instruments and because it is weighted with Australian TTO-values. The instrument of second choice would be the HUI3.
- Where direct comparison between Australian and international data is required, the EQ5D could be used. Because of its measurement shortcomings it should not be used alone.

Given that all five utility instruments are contained within the SAHOS dataset, further research into similarities and differences between the utility measures could be undertaken with the objective of providing standardized algorithms for the development of a common scoring metric enabling imputation of scores from each instrument to each other instrument.

- It is recommended that further research is undertaken into providing standardized algorithms for the development of a common scoring metric enabling imputation of scores from each instrument to each other instrument.

6.2.3 The Impact on Incontinence on Health Status as measured by the SF36V2

Because the SF36V2 has not been previously reported in a large-scale Australian population study, considerable effort was made to understand its properties in this study's sample. The results suggested that the structure of Australian responses from the SAHOS participants was significantly different to that of the published US samples. The implication is that Australians conceptualize health differently to their US counterparts.

Consequent upon this finding, suitable Australian weights for scoring the SF36V2 are provided in this report for computing T-scores for the SF36V2 scales and for the two summary scales. These weights have been provided in the interests of enabling researchers make direct comparisons between their samples and the Australian population generally. To assist researchers, these estimates have been provided in considerable detail, including by age group, gender, score deciles and health status.

- It is recommended that these estimates are used in studies involving Australian samples, and that the data are reported as being weighted with Australian values. For those researchers wishing to directly compare their data with US samples, tables providing US-weighted norms have been provided.

The Australian weights were derived using the identical methods to those used by the SF36V2 developers. The shortcomings of these methods are acknowledged. It was also noted that SF36V2 scale scores were extremely skewed. Consequently it is recommended that:

- Further work on scoring the SF36V2 be undertaken, including research into the reasons why different factor loadings were obtained and the meaning of these differences.
- Researchers should either transform their data prior to analysis or report medians rather than means.

Finally, the findings show that the SF36V2 scales are sensitive to incontinence status, and it would appear that the SF36V2 would be an appropriate measure to use for the assessment of health status in incontinence studies.

Appendix A: Literature Review of Utility Instruments[11]

This section reviews multi-attribute utility (MAU) instruments in the context of future Australian epidemiological research on incontinence, including population screening or surveillance as well as clinical treatment trials. It was authored by A/Professor Graeme Hawthorne.

The instruments reviewed are, in order of publication, the Rosser Index, QWB, HUI3, 15D, EQ5D, AQoL and SF6D.

The Sources and Publications used

The literature used in this report comes from searches of Medline, Psychlit and Econolit. Additionally, citations such as reports were sought out where they were deemed relevant. The following shows the number of such references for each of the instruments: Rosser Index: 18 (18 journal articles); QWB: 94 (90 journal articles); EQ6D: 308 (299 journal articles); 15D: 19 (10 journal articles); HUI3: 23 (21 journal articles); AQoL: 36 (13 journal articles); SF6D: 6 (5 journal articles).

The low number of journal articles for the AQoL as a proportion of all references may reflect the author's familiarity with non-journal publications for the AQoL. Subject to this bias, it would appear there is publication bias by instrument publication date: in general the earlier instruments (QWB, EQ5D) have greater publication; this of course, does not hold true for the Rosser Index which has a more specialized market. Regarding the literature, then, use in studies should not be accepted as indicative of instrument validity as this would imply that popular usage confers known properties!

A.1 Economic Evaluation, Cost-Utility and the Axioms of Utility Measurement

The growing interest in the measurement of health-related quality of life (HRQoL) can be attributed to four interrelated health and health care changes (146). Health care technologies have reduced early mortality and prolonged the lives of those who would otherwise have died (147); there has been a shift in economically developed societies from exogenous to endogenous chronic diseases (148); many health services are now designed to prevent deterioration in quality of life (149); and there is increasing conflict between potentially useful interventions and the resources available to fund them (150).

These changes suggest health resources should be allocated in ways that best benefit communities (146, 147). This can be achieved through providing services which lead to benefits in people's life-length and HRQoL. Evaluative research (151) is needed to ensure that potential benefits are realised, including economic evaluation contributing to the decision-making processes associated with resource allocation.

A.2 HRQoL and Economic Evaluation

The World Health Organization's definition of quality of life (QoL) is an individual's perception of their position in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns (144).

Since it is the individual who experiences and values this state, the challenge is to find ways of measuring this individual perspective that are sensitive, valid and reliable and yields respondent information which can be considered alongside clinical and clinician data.

Health care evaluative research has traditionally been at the level of summative evaluation aimed at assessing the extent to which the intervention benefited people's health (151-153). Where comparative information across interventions is sought for resource allocation decisions,

[11] This Appendix is copied from the Thomas report (27). The numbering of sections, layout and referencing have been changed to make it stylistically consistent with the rest of this report.

economic evaluation is used. Useful introductions to economic evaluation can be found in Drummond (154), Singh et al (155), and Drummond et al (150). Economic evaluation offers three important mechanisms: it can describe the cost of the burden of incontinence, it can predict the level of resources that will be needed in the future for treating incontinence, and it can provide information about the best use of resources available for incontinence interventions. Although there are four major kinds of economic analysis, the most comprehensive are cost-utility and cost-benefit (150).

Multi-attribute utility (MAU) instruments are designed explicitly to be used in cost-utility evaluations. This review provides an overview of the leading MAU-instruments and assesses them against issues relevant to incontinence.

A.3 The Axioms of Utility Measurement

The basic axiom of cost-utility analysis is simple: life years are weighted by the value of a given health state in such a way that the values — referred to as 'utilities' — act as an exchange rate between the quantity and quality of life. In this context, 'utilities' are assumed to be preferences for a given health state. Regarding the measurement of utilities, Torrance (88) provides the classic text.

To understand utilities, consider the following. Most people would prefer to be healthy over a given time rather than suffer constant urinary or faecal incontinence. Utility measurement refers to valuing these preferences on a life-death scale with endpoints of 1.00 and 0.00, where 0.00 is death equivalent and 1.00 is perfect (very good) HRQoL. For example, the measured utility for urinary incontinence may be 0.60. If treatment improves this to 0.70, then the value of the treatment is 0.70 – 0.60 = 0.10. If this utility gain is maintained over time, say for 10 years, then the gain is 0.10 x 10 = 1.00 Quality adjusted life year (QALY). Because utilities fall on the life-death scale, they are (in theory) common across all health states and therefore can be used to compare the effect of interventions in different health fields, or different interventions within the same field. For example, the QALYs gained from Treatment A for incontinence could be compared with those gained from Treatment B for depression. Where treatment costs (including costs to the patient) are known, the treatment providing the lowest cost-per-QALY gained is preferred as this ensures society gains the greatest benefit from the health care dollar.

To allow for comparison, utility measures must be generic and must allow for respondents to report they have excellent HRQoL (full health equivalent state: 1.00); additionally they must allow those who have appalling HRQoL to report this (death equivalent state: 0.00). If an instrument does not permit this full range of responses, it cannot accurately measure the HRQoL of people who fall outside its range. For example, if an instrument only allows measurement between 0.50 to 1.00, then it is incapable of reporting the effect of treatment for people who are in a desperate health state (say, close to death). Under these circumstances, any claim to generalisability for the instrument is foregone.

The instrument must be applicable to HRQoL states deemed worse than death (i.e. the respondent indicates they would rather die now than continue living in their current HRQoL state). These negative health states are needed to allow for people who commit suicide or euthanasia; they have clearly made the decision that death is preferable to living in their current health state and any possible future health states. When determining negative utility boundaries, the developers of the EQ5D and HUI3 adopted Torrance's symmetry argument. This states that since a person can 'lose' HRQoL value from 1.00 (full health) to 0.00 (death equivalent), they must be able to 'gain' an equivalent amount from –1.00 to 0.00 (88). However, since negative utility values do not possess the same interval properties as positive utility scores (156, 157), there are difficulties. For example, improving the HRQoL of a person from –0.35 to –0.25 (i.e. bringing them closer to a HRQoL death-equivalent state) does not have the same meaning as improving their HRQoL state from 0.25 to 0.35; yet both these would have a utility gain of +0.10. This is implausible. It seems likely, then, that negative values should have lower boundaries close to 0.00 (death equivalent) (156).

Implicit in axioms and mathematical modelling of utilities is that utility measurement must be at the interval level, where interval level refers to measurement scales that have equal-intervals between the measurement points. There are two forms of interval measurement that MAU-instruments must have if they are to do their job correctly. One is known as the "weak" interval property and the other the "strong" interval property (87). The weak interval property is where a gain of 0.10 means the same thing across the range of instrument scores. For a person who has

severe faecal incontinence, their utility score might be 0.15; as a result of treatment this rises to 0.25; i.e. the value of the treatment is 0.25–0.15 = 0.10. Similarly, the value of the treatment is also 0.10 for a person with urinary incontinence with an initial utility of 0.60, and who gains a utility of 0.70 after treatment (0.70–0.60 = 0.10). The strong interval property is where there is a direct relationship between gains in utility and gains in life-length. Since QALY calculation represents the time spent in a given state multiplied by the quality of that state, this implies that a 0.20 utility gain multiplied by 5 years in the improved health state equals 1.00 QALY (from 0.20 X 5). But a gain of 1 QALY could also be the product of a 0.40 utility gain over 2.5 years (or any other combination).

A.3.1 Measuring Utilities using MAU-instruments

There are two steps to measuring utilities using MAU-instruments. First, the health state of interest is described. Second, the value or utility of the health state is assigned.

When a person completes a MAU-instrument, their numerical responses provide a description of their health. For example, consider two people completing an imaginary instrument with four dimensions each of which has four levels. This instrument's 'descriptive system' would be: physical, mental, social and cognitive health dimensions, and the response levels are: 1 = normal, 2 = some impairment, 3 = major impairment, 4 = gross impairment. Person A, who is in the best of health, selects the best response to each item (i.e. '1': normal,). Her health state would be described as '1,1,1,1'. Person B who reported major incontinence (level 3: major impairment on the physical dimension), normal mental heath (level 1), some social impairment (level 2), and normal cognitive function (level 1). Her health state would be '3,1,2,1'.

Valuing these health states is called 'scaling'. Five procedures have been used: time trade-off (TTO), standard gamble (SG), visual analog rating scale (VAS), magnitude estimation (ME) and person trade-off (PTO). Brief descriptions are given.

- ***Time trade-off (TTO)***. A person with severe incontinence can have a treatment which will restore her to full health; but a side effect is she will live a shorter life. She is asked to choose how many years of her life she would be willing to 'give up' in order to be in full health. If, in her untreated condition, her life expectancy was 10 years and after the treatment this was 5 years she may reject the treatment. If after the treatment it was 9 years, she may accept it; if her life expectancy was 6 years, she may not. Her choices would continue back-and-forth like this until she indicated that she was indifferent to whether she had the treatment or not. If the point of indifference was that 8 years of full health was the equivalent of 10 years with severe incontinence, then the quality of life value for her current health state is 8/10 or 0.80.
- ***Standard gamble (SG)***. A person with urinary incontinence is presented with a treatment option that has two possible outcomes: either full health for the remainder of his life, or death. He is free to choose either the treatment or to remain with lifelong urinary incontinence. If the probability of full health is 1.00 (i.e. his incontinence will be cured and there is no chance of death), then obviously he will choose to have the treatment. If the probability of full health is 0.90 and death 0.10, he may still choose the treatment. However there would be a point, for example at 0.80 for full health and 0.20 for death, where he is not clear as to whether he would want the treatment or would choose to remain in his current health state. This point of indifference is the 'value' of his health state.
- ***Visual analog scale (VAS)***. The respondent is asked to consider an incontinent health state and then to rate this on a scale, where the endpoints are typically 0.00 (death equivalent) and 1.00 (full health equivalent). Unlike the TTO or SG, with the VAS there is no uncertainty: the respondent is not asked to 'trade' anything. Consequently many consider that VAS scores do not represent utilities because they provide a simple ranking of health states. Where VAS scores are used, a transformation is generally required, based on TTO or SG (86, 158, 159).
- ***Magnitude estimation (ME)***. The respondent is asked to consider the distance of the health state of interest (eg. incontinence) from 1.00 (full health). Once several of these rating exercises have been carried out, the respondent is then asked to rank these in order (160). Because there is no uncertainty, it is uncertain if ME represents utility.
- ***Person trade off (PTO)***. The respondent is asked to estimate the number of people that would have to be treated to make an intervention worthwhile. For example, a respondent might be asked to choose between extending the life of 10,000 people who were in full health by 1 year against a treatment which extended the life of N people with incontinence, also for 1 year. The number of people with incontinence would be varied until the respondent indicated they were indifferent between the two choices (160).

When these techniques are used to obtain the utility weights used in an MAU-instrument, in theory each health state described by the descriptive system can be scaled (as was done with the original Rosser Index (161)), but this is impractical because MAU-instruments typically generate thousands of different health states. Instead, a limited number of health states are scaled and the values for other health states are then inferred using econometric or decision analytic techniques, typically either an additive or multiplicative model (45). During scoring, the health state descriptors (1,2,3, etc.) are replaced with the appropriate values. For example, if the value of suffering mild pain based on TTO is '0.70' and the response levels on an item measuring pain were '1' (no pain), '2' (mild pain), and '3' (severe pain), then a person who selected '2' would have this level replaced with the value '0.70' during scoring of the instrument.

Once item-level values have been assigned, these are combined into an index on a life-death scale. Three procedures have been used.

- ***Additive models.*** The substituted importance values are summed and the resulting score represents the utility index. The limitation is that for full health equivalent HRQoL states each instrument item or dimension must contribute a fixed amount. Under this model, a respondent can obtain a very poor utility score only if they report poor scores on all items or dimensions. Consider an instrument measuring two dimensions: physical and mental health. In an additive model, each may contribute 0.50 towards the utility score. In this model, appalling mental health (leading to suicide) could never, by itself, lead to a utility value lower than 0.50 because 0.50 (a person in good physical health) + 0.00 (mental health) = 0.50. Thus additive models cannot explain people who commit suicide if their physical health is good, or euthanasia if their mental health is good.
- ***Econometric models***. The items are treated as explanatory variables to derive a regression equation predicting utilities. This method, however, suffers the same limitation as the additive model.
- ***Multiplicative models***. These involve multiplying items or dimension scores together. This overcomes the limitation of the additive model as it allows any dimension to carry a person to a death equivalent value. Consider the case above. Here the person's value for mental health would be 0.00, and 0.50 (physical health) x 0.00 (mental health) = 0.00.

Given these assumptions, preference independence is required to avoid double-counting, which is where the same underlying health condition contributes more than once to the MAU-instrument utility index. For example, if a person is incontinent this should be counted in their utility score once, although the effect of this health state may be measured in several different aspects of their life; i.e. on several different scales. Where these effects are measured using unidimensional scales that are orthogonal to each other there is no difficulty. Where the scales, however, are correlated the effect of incontinence will be counted several times over thereby biasing the utility measurement. It is for this reason that MAU-instruments are required to possess structural independence (i.e. where the scales are unidimensional and orthogonal) (162). For example, if incontinence is counted on dimensions measuring social, physical and psychological dimensions as well as its effects being directly measured, then there is loss of preference independence as the scores on the social dimension may be a function of physical scores.

A.4 Description of MAU-instruments

In order of their development, MAU-instruments are the Rosser Index (161), the Quality of Well-Being (QWB) (163, 164), the Health Utility Index 3 (HUI3) (10, 50, 139), the 15D (3, 4, 52), the EQ5D (formerly the EuroQol) (7, 165), the Assessment of Quality of Life (AQoL) (5, 166, 167) and the SF6D (16, 17). Additionally, Fryback et al (168) have prepared an algorithm for mapping SF-36 scores onto the QWB.

This report considers the Rosser Index, QWB, HUI3, EQ5D, AQoL and SF6D. Although there are three HUI instruments, only the HUI3 is considered. The Fryback et al SF-36 algorithm is not a MAU measure in its own right.

The descriptions presented in this section are largely based on those given by Hawthorne & Richardson (45, 108).

A.4.1 Rosser Index

The British Rosser Index was designed for use in hospital settings. The original version had two dimensions measuring disability and distress, and measured 29 health states. Values were elicited using magnitude estimation from a convenience sample of 70 respondents (161). A revised version was released in the early 1990s based on SG procedures and included an additional dimension of discomfort (161) . Administration requires a trained interviewer. The upper boundary is 1.00, and the lower boundary –1.49; i.e. health states worse than death are permitted. The Rosser Index has given rise to two variants: the Health Measurement Questionnaire (HMG) (169) and the Utility-based Quality of Life-Heart Questionnaire (UBQ-H) (170). Permission must be obtained for using the instrument, however there are no costs for its use. No website was identified for the Rosser Index.

A.4.2 Quality of Well-Being (QWB or IWB)

The American QWB was designed to bridge the gap between clinical measurement, functional status and health planning policy (171) and was an adaptation of US health surveys (172). It has three dimensions (Mobility, Physical Activity, and Social Activity) with 3–5 levels each. There are an additional 27 illness symptoms. Combined, these provide an index of 'Well-life expectancy' of which there are 43 functioning levels (163, 171, 173). This would seem to support Anderson et al's (174) description of it as measuring dysfunction as mental and social health are not measured. The QWB was designed for interview administration (15–35 minutes), although a shorter version has been developed which takes about 15 minutes (164). Interviewer training is required (175). The preference weights were elicited using VAS scores which were obtained from a sample of the San Diego population. A linear transformation was then used to place these on a 0.00–1.00 scale (164, 173). An additive model is used to compute the index. Extensive efforts to validate that VAS provides interval properties led to the release of a revised version (163, 173, 176). The upper boundary is 1.00, and the lower boundary is 0.00 (death equivalent) and health states worse than death are not permitted. Permission must be obtained to use the QWB and there are no costs for its use. Further information on the QWB can be obtained at: http://medicine.ucsd.edu/fpm/hoap/instruments.html.

A.4.3 Health Utilities Index, Mark 3 (HUI3)

The Canadian Health Utilities Index (HUI3), for general population use, is based on the HUI2 which was designed for survivors of childhood cancer. To render it generic and overcome reported difficulties, it was revised into the HUI3 (139). The HUI1 has been superseded. The HUI3 measures 'within the skin' functional capacity (10), a perspective adopted to enhance its use in clinical studies (48). Social aspects of HRQoL are not measured. Items have 4–6 response levels. Twelve of the 15 items form 8 attributes (Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition and Pain). Designed for self-completion, Nord (177) reported it took 2 minutes to complete, although 5–10 minutes is more likely given it has 15 items. The utility weights were elicited using the VAS, and scores then transformed based on four 'corner' health states valued with the SG where a 60 year timeframe was used. These results were based on stratified sampling (n = 256; response rate 22%) of the Hamilton, Ontario, population (49). A multiplicative function combines the attributes into the utility score (49, 50). The upper boundary is 1.00, and the lower boundary is –0.36, permitting health states worse than death. Users must be registered and the instrument is available at a cost of CAN $4,000 per trial. Copies of the HUI3 and application forms can be found at: http://www.healthutilities.com/hui3.htm/.

A.4.4 15D

The Finnish 15D was defined by Finnish health concerns, the WHO definition of health and medical and patient feedback (3, 51). It is concerned with impairment and disability of 'within the skin' functions. There are 15 items, each with 5 levels, measuring Mobility, Vision, Hearing, Breathing, Sleeping, Eating, Speech, Elimination, Usual Activities, Mental Function, Discomfort & Symptoms, Depression, Distress, Vitality and Sexual Function (3). Nord (177) reported it took 5–10 minutes for self-completion. The weights came from five random samples of the Finnish population (n = 1290 respondents; response rate 51%) using VAS questions; responses were combined using a simple additive model (4, 52). The upper boundary is 1.00, and the lower boundary is +0.11: death-equivalent and worse than death health states are not allowed. Permission must be obtained to

use the instrument, however there are no costs for its use. The 15D has been translated into a number of European languages. Although there is no website devoted to the 15D, details can be obtained from http://195.101.204.50:443/public/15D.html.

A.4.5 EQ5D (formerly the EuroQol)

The EQ5D (formerly the EuroQoL), developed by a team from 7 European countries (7, 178), was based on the QWB (179), the Sickness Impact Profile (180), the Nottingham Health Profile (181), the Rosser Index (161), and group members' opinions. Designed for use in cross-cultural comparisons it has 5 items, each with 3 response levels, measuring Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. It takes 1-2 minutes to self-complete (177). The utility weights are from a British population random sample (n = 3395 respondents, response rate 56%) based on the TTO for 42 marker health states using a 10 year timeframe (47). Other utility values were regression modelled (47, 182, 183). The index is computed using an econometric regression model. The upper boundary is 1.00, and the lower boundary is –0.59: it permits health state values worse than death. Although the EQ5D is in the public domain for public health research, the EQ5D management group ask that researchers register their use of it. There are no costs for its use, unless it is used by commercial organisations. The EQ5D has been translated in many languages. Further information on the EQ5D can be obtained from: http://www.eur.nl/bmg/imta/eq-net/EQ5d.htm.

A.4.6 Assessment of Quality of Life (AQoL)

The Australian AQoL used the WHO's definition of health, and items describe 'handicap' as distinct from impairment and disability (46). The descriptive system has 15 items and 12 are used in computing the index (45). Each item has 4 levels. There are five dimensions: Illness (not used in utility computation), Independent Living, Social Relationships, Physical Senses and Psychological Well-being (5). Designed for self-completion, Nord (177) reported the AQoL took 5-10 minutes. A stratified sample (n = 350 respondents; response rate 72%) representative of the Australian adult population completed TTOs based on a 10 year timeframe to provide the utility weights (184). A multiplicative model is used to compute the utility index (108). The upper boundary is 1.00, and the lower boundary is –0.04: it permits health state values worse than death. Permission to use the AQoL must be obtained, but there is no cost for its use. Further information can be obtained at: http://chpe.buseco.monash.edu.au/aqol.html#1.

Due to a concern that the AQoL is insensitive at the upper end (i.e. for well health states), the AQoL research team are developing AQoL II for use in health promotion. As part of this development, AQoL II has been designed to enable the addition of disease-specific modules. One is currently being developed for the visually impaired.

A.4.7 SF6D

Two different algorithms have been published by for deriving preference-based values from the SF-36 (16, 17). They are referred to as the SF6D-1 and SF6D-2. The SF-36 descriptive system is American and the SF6D weights are British. The advantage of the SF6D procedures is that wherever SF-36 raw scores are available, the SF6D preference measure can be used.

Brazier et al's (16) SF6D-1 drew upon 20 of the 36 items; these were selected to avoid double-counting. During scoring items are combined into composites and each composite has 2–6 response levels. There are six sub-scales; Physical Function, Role Limitation, Social Function, Bodily Pain, Mental Health and Vitality. Utility weights were computed from VAS scores and modelled using SG values for three 'link' health states. These values were derived from a convenience sample of 165 British respondents. An additive model computes the utility index. The upper boundary is 1.00, the lower boundary is +0.46: it does not permit poor health states, death equivalent or worse than death health state values.

The SF6D-2 (17) uses 10 items from the SF-36: three from the physical functioning scale, one from physical role limitation, one from emotional role limitation, one from social functioning, two bodily pain items, two mental health items and one vitality item. These form 6 dimensions: Physical Functioning (PF: 6 levels), Role Limitation (RL: 4 levels), Social Functioning (SF: 5 levels), Pain (PA: 6 levels), Mental Health (MH: 5 levels) and Vitality (VI: 5 levels). Utility weights were computed from VAS scores, which were modelled using SG values for two link health

states. Values were obtained from a random sample (n = 611; response rate = 45%) of the British population. An additive econometric model is used to compute the utility index. The endpoints for the SF6D are 1.00, and 0.30 for the worst possible health state. No website for the SF6D was identified.

A.5 Comparison of Instruments

Hawthorne and Richardson (45) outlined the axioms of utility measurement which MAU-instruments should conform to in order to possess basic validity. These axioms can be used as a checklist in instrument selection. They are

- The use of a preference measurement to weight instrument items.
- Instruments must measure the dimensions of HRQoL deemed to be important. These are usually defined as physical, mental, social and somatic sensations (eg. pain).
- There must be coverage of the full spectrum of HRQoL values, from full health states to values representing states worse than death.
- The combination rule for the utility index must prevent double-counting.
- There must be evidence of both weak and strong interval measurement.
- Instruments must be sensitive to the health states of interest. This requirement is covered in the next section. For general sensitivity comparisons between the instruments the three validation papers published by Hawthorne et al should be consulted (45, 107, 166).

An additional requirement is that:

- There must be evidence of valid and reliable measurement.

A.5.1 Use of a Preference Measurement Technique to Weight Instrument Items

Instruments using the SG or TTO may be regarded as possessing preference weights since both involve decisions under uncertainty. In the SG, the life outcome is uncertain (the probability of full health versus death). In the TTO, life-length is uncertain (how many life-years a person is willing to sacrifice).

There are doubts over whether ME delivers preferences because the procedure requires the respondent to estimate the divergence of a given health state from the 'full' health state (which is assigned a value of 1.00). Once several given health states have been so assigned, the respondent is then asked to rank these in order (160).

As reported above, there is doubt whether the VAS delivers preference measurement. Consequently it has been argued that the VAS has no place in economic theory (86) and that untransformed VAS scores should not be used (185, 186). It is recommended that VAS data should always be transformed, based on TTO or SG (109, 158, 159); the transformation function that has been used was developed by Torrance et al (187). The preference measurement of instruments weighted with VAS scores therefore rests upon the validity of this transformation. For the EQ5D, Dolan et al (8) reported that the explanatory power of the transformations used was r^2= 0.46, which was considered to be very good. However Sintonen (52) reported that when applied to the 15D VAS data it assigned 12–25% of the adult population to values worse than death, a result he stated was 'implausible'. Bleichrodt & Johannesson (188) noted that individual transformations were unstable; Robinson et al (185) reported difficulties with the transformations; as did Torrance et al (186).

Instruments weighted with a preference measure are the EQ5D (which used the SG) and the AQoL (the TTO). The Rosser Index relies upon ME. The HUI3 and the SF6D both rely upon transformed VAS scores; the extent to which these can claim preference weighting is dependent upon the validity of the transformations. Nord (189) has questioned the validity of the linear transformations for the QWB, arguing that one of the primary reasons its use in Oregon was so heavily criticised was that it lacked cardinal values. Given that the 15D uses untransformed VAS ratings there are doubts that it meets this requirement, although Martin (190) argued that this gave the opportunity to quickly establish new weights for different populations — a procedure which Sintonen argued should be followed for each population from which study participants were drawn (3).

A.5.2 Instruments must Measure the Dimensions of HRQoL deemed to be Important

Important areas of HRQoL are usually defined as physical, mental, social and somatic sensations (eg. pain). Unless instruments measure all these they cannot claim to be 'generic'. It should be remembered that the measurement of utilities was explicitly developed to enable cross-condition, health state and health care comparisons; by definition MAU-instruments are supposed to be generic.

Generally there are no published formal tests of content validity (45). Where this is mentioned, instrument developers have reported 'face' validation, i.e. that instrument content as judged by the instrument developers 'looks about right'. For example, it has been argued the very restricted Rosser Index descriptive system makes it insensitive and provides a narrow band of responses (191-194). In a study of the EQ5D descriptive system it was reported that it only covers 39% of the concepts regarded by the public as salient to health (195). Feeny et al (10) reported that the HUI3 was valid because all levels of scores had been assigned at least once in population surveys. These various assertions do not engender confidence that the universe of utilities is actually measured by any of the instruments, a point which has been noted in the literature.

In three recent review articles Hawthorne et al (45, 107, 108) mapped the content of MAU-instruments against the dimensions of 14 HRQoL instruments published between 1971 and 1993. Table A.1 summarizes their work. This shows that even in the better instruments coverage of the universe is limited. Some instruments offer very narrow measurement (for example, the Rosser Index and EQ5D), others have in-depth or duplicated measurement in particular areas (for example, the QWB, 15D and HUI3), and some offer very broad but sketchy coverage (for example, the AQoL and SF6D). Duplicated measurement may bias the obtained utility values. Two examples illustrate the problems. Despite its broad coverage, the QWB primarily measures pain and physical disability (163) yet does not include either social or mental health (174), and analysis of the HUI3 showed it was a measure of physical impairment which did not adequately measure physical, social or mental dimensions (118).

A.5.3 There must be Coverage of the Full Spectrum of HRQoL Values

This refers to instruments providing utility values from full health states to values representing states worse than death. There are two issues here. First, instruments must have combination rules permitting very poor HRQoL, irrespective of how this is caused. Second, the range of utility scores must cover the full spectrum.

Regarding combination rules, multiplicative models are to be preferred for the reasons outlined above. Instruments with multiplicative models are the HUI3 and AQoL. The EQ5D and SF6D rely upon regression models which are essentially additive in nature, and the 15D is an additive instrument.

The Rosser Index, EQ5D and HUI3 allow large negative values. Given the difficulty with the symmetry argument, these values are problematic. Hawthorne & Richardson (196) calculated that the effect of restricting the lower boundary for the HUI3 and EQ5D to 0.00, in population studies, would raise mean utility values by 9% and 14% respectively. This suggests the net effect of the symmetry argument is to overstate the value of interventions where people are in very poor health states. This problem does not apply to the QWB and AQoL which have lower boundaries at or near to 0.00.

The lower endpoints for the 15D (+0.11) and SF6D (+0.46 and +0.30 for the SF6D-1 and SF6D-2 respectively) raise other questions. Hawthorne & Richardson (156) reported these boundaries resulted in very different QALY estimates: a 1 QALY gain from the AQoL, EQ5D or HUI3, where a person was returned from the lowest quartile to full health for 1 year, implied a 0.50 and 0.37 QALY gain on the 15D and SF6D respectively (45). These contradictory results suggest that at least one of these of instrument groups is wrong.

As they allow the full range of scores, the QWB or AQOL instruments would be preferred, as would the 15D.

Appendix A: Table 1: Content of MAU-instruments (a)

HRQoL dimensions (b)	*AQoL*	*EQ5D*	*HUI3*	*15D*	*QWB (c)*	*Rosser Kind (d)*	*SF6D*
Relative to the body							
Anxiety/depression/distress	*	*		**	**	*	**
Bodily care	*	*					*
Cognitive ability			*	*	*		
General health							
Memory			*				
Mobility	*	*	*	*	*		**
Pain	*	*	*	*	********		*
Physical ability/vitality/disability			*	*	*****	*	*
Rest and fatigue	*			*	**		**
Sensory functions	**		****	*****			
Social expression							
Activities of daily living	*	*		*			*
Communication	*		**	*	*		
Emotional fullfilment			*				
Environment					*		
Family role	*						
Intimacy/Isolation	*						
Medical aids use					**		
Medical treatment							
Sexual relationships				*	*		
Social function	*				*		*
Work function							*

Note: a = Table shows only those items used in calculation of utility scores. Each asterisk represents an item. Based on item content examination.

b = Dimensions of HRQoL defined by a review of 14 HRQoL instruments, 1971–1993.

c = Excludes intoxication.

d = Areas subsumed within the two items: mobility, employment, housework.

Source: Adapted from Hawthorne et al (44).

A.5.4 The Utility Combination Rule must prevent Double-counting

During construction of the Rosser Index, care was taken to ensure orthogonality between the dimensions (161). Brazier et al (109) reported that for QWB there is multicollinearity between the scales and symptoms. In papers describing the EQ5D there is no mention of this issue (7, 165). Based on clinicians' opinions, structural independence was claimed for the HUI3 (48); the factor analysis of the HUI3 published by Richardson & Zumbo (118), which revealed a lack of independence between the attributes, challenges this claim. Sintonen claimed independence for the 15D, although no evidence was provided (52).

For the SF6D-1 Brazier et al (197) used correlation analysis: where items were highly correlated only one was included. Brazier et al (17) noted that since an econometric model was used for the SF6D-2, preference independence, structural independence and double-counting were unimportant. Yet the form of the SF6D-2 for the prediction of SG scores is essentially an additive model. Therefore, this argument seems extraordinary given that orthogonality to prevent double-counting caused by multicollinearity has been axiomatic of both psychometric and decision-making theory for over 50 years (162, 198).

For the AQoL, exploratory factor analysis was used during construction to ensure orthogonality (5); the structure has since been confirmed by structural equation modelling (107).

A.5.5 There must be Evidence of both Weak and Strong Interval Measurement

For meeting this criterion, all MAU-instruments rely on the presumed interval properties of the TTO, SG, or VAS. No instrument construction or validation paper has reported any formal testing of these properties. It has not been convincingly demonstrated that these properties are embedded within the TTO, SG and VAS (86). Rosser (161) argued that the magnitude estimation procedure used with the Rosser Index produced cardinal values; thus, like the EQ5D and AQoL, the Rosser Index may meet the weak interval requirement.

The Weak Interval Property

VAS responses may be functions of adaptation, context, endpoints or anchorpoints, end-aversion and rating effects. These imply VAS may produce ordinal rather than interval data (87, 185, 186, 199). For the TTO and SG even less is known as these issues do not appear to have ever been properly investigated. Although Cook et al (199) challenged the claim of interval data for all three techniques, this was refuted by Hawthorne et al (89) on account of some major methodological difficulties.

Subject to these caveats, Hawthorne & Richardson (45) asserted it was likely the SG and TTO possessed interval properties given they allowed incremental probabilities (SG) or time fractions (TTO). On this basis, those instruments weighted with the SG or TTO should be preferred.

The Strong Interval Property

This means that any given incremental value in HRQoL utility was directly equivalent to the same incremental value in life-length or life-probability. There is no evidence available for any of the MAU-instruments that they meet this requirement.

A.5.6 Valid and Reliable Measurement

The validity and reliability of various MAU-instruments has been assessed through either test of concurrent validity where monotonic relationships are sought, or test-retest. Additionally, there are issues around the stability of the utility values used in the different instruments due to sample bias.

Monotonicity refers to a relationship in which the instrument of interest group or mean scores progressively increase in line with a criterion measure. For example, if a sample of people suffers incontinence from "a few drops" to "no bladder control at all", then an instrument measuring this underlying health condition should report manifest scores that systematically increase with the level of incontinence. This does not imply, of course, that there will always be a 1:1 relationship between the two measures, for there will be individual variation.

Hawthorne et al (45) examined monotonicity for the EQ5D, 15D, HUI3, AQoL and SF6D-1 against health status as defined by their sample strata of community random sample, outpatients and inpatients; they also examined the same instruments by combined utility quartile (45) and by instrument predictive power (107). In general their findings support monotonicity for all the instruments, although they did observe that the instruments formed two groups: those which correctly classified >50% of cases (AQoL, 15D and SF6D) and those which predicted <50% (EQ5D and HUI3).

Data on the Rosser Index are mixed. Although Rosser Index scores have been shown to match empirical and population general health data quite well when predicting the healthy/unhealthy dichotomies (169, 192), in a replication study it was shown that there are several health states where monotonicity is violated leading to difficulties with assigning logical QALY values (160).

For the QWB there is mixed evidence regarding monotonicity. Kaplan et al (200) reported very high correlations with a number of chronic conditions, where the average was r = 0.96. Based on the revised version, similar correlations with chronic conditions have been reported (163, 176). For example, Kaplan et al (201) reported a monotonic relationship between QWB scores and HIV-status; similarly monotonicity has been reported for functional status of children suffering

cancer (202). Against this the QWB has been criticised for producing QALY values that are non-monotonic. Thus a person wearing glasses is worse off than someone confined to a wheelchair, or curing five people with pimples would equate with saving one life (192, 203). In a study of heart disease, non-monotonicity was reported for half the QWB scales (204).

The Hawthorne et al results for the EQ5D (see above) were particularly interesting as they indicated that the EQ5D assigned too many cases to a utility value of 1.00, a finding consistent with earlier work by Brazier et al (205). Both research groups suggested this may have been due to the insensitivity of the EQ5D at the healthy end of the range and the consequent limited capacity to discriminate between those with full health and some health problems. At the other end of the range (very poor health states) Nord et al (192), in a study comparing Norwegian and Australian populations, reported that the EQ5D assigned excessively low values for some health states; a finding supported by Hawthorne et al (167) who found that the EQ5D assigned 4% of a population sample to health states worse than death. In a comparison with the SF-36, Brazier (205) pointed out that the EQ5D correlated poorly with physiological symptoms, and Andersen et al (206) reported that the EQ5D assigned non-monotonic values for people with fractures: a person with a fractured arm was assigned worse utility than someone with a fractured vertebrae.

Sintonen (52) tested the 15D for monotonicity in five population-based samples, reporting that up to 2.5% of respondents valued health states inconsistently, rising to 20% who valued 'death' higher than being 'unconscious'.

For the AQoL, other than Hawthorne et al's work there is as yet insufficient published material examining its properties for any conclusive assessment to be made. Hawthorne et al's papers (45, 107, 108) described above all report monotonicity. Monotonicity has also been reported for cochlear implants (94), the health status of those with long-term depression (207), suicidal ideation (208) and depression in a population sample (209), and stroke (210).

Test-retest reliability estimates have been reported for the QWB, 15D, EQ5D, HUI3 and AQoL. For the QWB, Kaplan et al (200) reported test-retest reliability at r = 0.93–0.98. In a study of chronic obstructive pulmonary disease, at 14-day separation, Stavem (85) reported that the EQ5D and 15D test-retest reliability using Spearman correlations were r = 0.73 and r = 0.90 respectively. This result for the 15D is more encouraging than that reported by Sintonen (4), who did not give a statistical estimate but stated that the agreement was not very good. In a study of stroke patients, Dorman et al (97) reported test-retest reliability estimates for the EQ5D of ICC = 0.83; and in a Dutch population study of the EQ5D where test-retest was carried out at 10-month intervals the correlation was r = 0.90 (98). Studies of the HUI3 (10, 101), based on a community random sample with telephone follow-up, reported test-retest reliability where r = 0.77. For the AQoL, Hawthorne (90), using random population sampling and mail/telephone comparisons reported the test-retest ICC = 0.83. An earlier study reported test-retest reliability for the AQoL descriptive system where = 0.80 (91).

Finally, and importantly, there are issues concerning the stability of the utility weights used in the various instruments. This concern stems from that fact that utility weights for most of the instruments — with the notable exception of the EQ6D where the sample size was 3395 — were obtained from a either small (e.g. 70 cases for the Rosser Index) or conveniences samples (e.g. the 1290 respondents for the 15D). In most cases, this was because of the cost of data collection: face-to-face interviews where SG or TTO questions are asked are costly. Because the SG or TTO is extremely tedious, all the instrument designers eroded their sample sizes further by breaking their health states up into sub-interview routines and then administering each sub-interview to a strata within the sample. This is commonly referred to as a 'sort' procedure. The extreme case where this occurred was with the SF6D-2 (17). The weights for the revised SF6D-2 were obtained from a representative sample of 836 Englishpersons of which 611 interviews were used. Based on a sort procedure, each respondent was asked to value 6 health states out of a possible 249 health states. Altogether 3,518 valuations were made: there was an average of 15 responses for each health state (the range was from 8 for health state 5,3,5,6,4,6 to 19 for health state 1,3,1,5,4,2) . Similar procedures were followed for the HUI3 (49), AQoL (211), and 15D (52), although in each case the numbers were greater than for the SF6D-2. For example, for the HUI3 the numbers for each health state varied from 19 to 246; for the AQoL the range was 70 through 225). These difficulties for each instrument were compounded by the relatively low response rates (typically about 50% although the AQoL's was higher).

These wafer-thin estimates raise fundamental questions concerning the transparency of utility scores, their stability and the generalisability of the instruments. In no case have instrument

developers reported validation of the obtained utility results or published an analysis of these data. Given this, it is highly likely the utility values for all instruments, other than the EQ5D, are biased and lack transparency. Because of the restricted response rates and small sample sizes utility weights may be less than stable; a problem compounded by the fact that all instrument weights have been derived using means rather than medians. Clearly, under these circumstances, claims for generalisability to many health conditions, including incontinence, should be interpreted cautiously.

A.6 A Review of MAU-instruments used in Incontinence Studies

The previous section examined the evidence for criterion validity where the criteria were the axioms of utility and psychometric measurement. This section reviews the performance of the MAU-instruments in incontinence studies where the criterion is the sensitivity of the instrument to detecting differences between those who are incontinent and continent.

To identify published studies a search of Medline and Econolit was undertaken using the terms 'utility' and 'incontinence', as well as the names of the instruments reviewed. Ninety articles were identified. Review of the abstracts revealed 16 articles using utility measures; all were retrieved. Reviews showed six did not contain any utility data so these were discarded.

No published papers were found for the AQoL or SF6D. For the AQoL and SF6D this was unsurprising given their recent development. Unpublished Australian data were available for both these instruments and these data have been reported here (for the SF6D the calculated values are from Brazier's second algorithm and the SF6D is described as the SF6D-2).

Of the studies reviewed, there were three population surveys, four trials (two non-randomised), and a modelling exercise.

A.6.1 Studies excluded

As part of a validation study of the uretal stent symptom questionnaire (USSQ), Joshi et al (212) reported values for the EQ5D. The treatment group were 85 patients, and the controls were 25 healthy volunteers. The EQ5D was administered to the patients 4 weeks after stent insertion and again at 4 weeks after stent removal. The results were reported as medians and inter-quartile ranges: for the stent group (with stent) the median scores was 0.76 (IQR: 0.62–0.90), after removal it was 1.00 (0.80–1.00), compared with the controls; 1.00 (0.76–1.00). A confounding factor was pain which was associated with the indwelling stents. The extent to which the differences in EQ5D scores were due to incontinence is therefore uncertain. Although the median EQ5D score at 4-weeks after stent removal was the same as that of the healthy controls, a proportion of the treatment group still reported incontinent, so, thus suggesting that the primary cause of the low EQ5D scores with stents may have been pain rather than incontinence. The implication is that the EQ5D may not be particularly sensitive to incontinence.

Krahn et al (213) elicited utility values from 141 older males (mean age = 72 years) who had treatment for prostate cancer. The utilities were elicited, on average, at 4 years post-diagnosis. The utility values were stratified by UCLA Prostate Cancer Index Scores. Urinary function was then reported for the HUI3 and the QWB. The values for the HUI3 were 1st quartile 0.85, 2nd 0.76, 3rd 0.80 and 4th 0.76; and for the QWB they were 1st 0.71, 2nd 0.64, 3rd 0.64, and 4th 0.57. If it is assumed that those in the 1st quartile had no urinary incontinence (something that is not stated in article), then the differences attributable to incontinence would be 0.09 for the HUI3 and 0.14 for the QWB. The HUI3 difference was reported as being non-significant, while the difference for the QWB was statistically significant ($p<0.001$). In the case of the HUI3, the lack of significance could be attributable to either large variation in scores within quartiles or to the lack of a monotonic relationship with increasing severity across the quartiles. A limitation of the paper is that no standard errors, standard deviations or confidence intervals were reported for these estimates, therefore this paper was excluded from further analysis.

Manca et al (214) used the EQ5D in a randomized study comparing tension-free vaginal tape (n = 117) with colposuspension (n = 97). EQ5D data were collected at baseline (means 0.78, 0.79 for the tension-free and colposuspension groups respectively), 6-weeks (0.79, 0.75) and 6-months (0.81, 0.79) follow-up. Unfortunately the data were not reported as means, medians

and interquartile ranges. The EQ5D values for women who were continent at the end of the study were not reported. Therefore this study was excluded from further analysis.

One Australian paper utilising the Rosser Index (215), was brought to the attention of the author although it was not identified in the literature search. This study compared five different treatments for female incontinence, reporting utility improvements between 1-2% across the five treatments, and the costs per QALY gained were between AUD$28,000—$134,000. Insufficient details of Rosser scores were included in the paper for it to be included in this review.

Several other papers were reviewed and excluded. One study which reported data for the EQ5D where scores were obtained on the EuroQoL (EQ5D) VAS which is not a utility measure (216) was excluded. (On the EQ5D VAS a respondent is asked to rate their health state on a 100-point scale.) The study by Ogawa et al (217) which reported a global utility rating was also excluded. Kobelt's (218) study of willingness-to-pay which included the EQ5D and correlated scores with micturition (Spearman $\rho = -0.25$) reported the mean EQ5D score was 0.68. Since no further information was given in the paper, it was also excluded from more detailed analysis.

A.6.2 Procedures

In the interests of comparability, Cohen's effect size (d) (d 112) has been calculated from the data in the various studies. Given the variability in treatment and comparator groups and how data have been reported, the full formula was used:

$$d = \frac{m_A - m_B}{\sqrt{\frac{\sigma_A^2 + \sigma_B^2}{2}}}$$

where:

m_A = the mean score of the incontinence group expressed in raw (original) measurement units;

m_B = the mean of the comparator group or comparator expressed in raw (original) measurement units; and

σ = the standard deviation of either population, on the assumption that the sample standard deviation equalled σ.

Cohen provided the following classification for interpreting d: 0.20 = a small effect, 0.50 = a moderate effect, and 0.80 = a large effect.

Based on the effect sizes, a modified version of the relative efficiency statistic (113, 219, 220) was computed, thus allowing the relative sensitivities to be examined:

$$RE_{Ins1vs.In2} = \frac{d_{In1}^2}{d_{In2}^2}$$

where:

d_{In2} = the instrument with the smallest effect size; and

d_{In1} = the effect size for the instrument of interest.

When interpreting d statistics, it should be borne in mind that the data were badly skewed in almost all the papers reviewed, which is normal for utility values (most people are healthy). Cohen argued for the robustness of the d measure and provided an example where the mean difference was 2.0 and the standard deviation 2.8 with respect to the number of trials rats required to learn a maze (112, p41).

The studies are reviewed in chronological order, and the effect sizes and relative efficiencies presented in Table A.2.

A.6.3 Results

The HUI3 was used in the Canadian National Population Health Survey in 1994/5. Mittmann et al (221) broke down the data by chronic conditions reported by the 17,626 participants (54% were women). When computing the HUI3 utility score, the researchers used HUI2 weights. To identify chronic health conditions respondents were asked: 'Do you have (chronic condition) that has been diagnosed by a health professional?'. Twenty-two persons reported urinary incontinence only (i.e. no comorbid chronic conditions); the number with incontinence and comorbidities was not reported. Although the researchers stated that those with urinary incontinence obtained the lowest HUI3 utility scores of any chronic condition (HUI3 utility score = 0.70, sd = 0.23; there were no differences by gender), this assessment was based on all cases including those where multiple chronic conditions were reported. When they reported condition specific cases, for those with 'incontinence only' the HUI3 utility was 0.85, sd = 0.12. That there was such a discrepancy in the HUI3 scores by 'incontinence only' versus 'incontinence only' and 'incontinence and comorbidities' suggests that most of the difference in HUI3 scores can be attributed to the effect of the comorbidities. For the purposes of reporting the sensitivity of the HUI3, comparison in this report was made between those with incontinence only and those without any reported chronic condition. This latter group are those reported as the comparator group in Table A.2.

Stach-Lempinen et al (222) reported data for the 15D on urinary incontinent women, where they were exploring the relationship between the Urinary Incontinence Severity Score, a visual analog scale describing the burden of incontinence ('How bothered are you by incontinence at this moment?') and the 15D. Treatment participants were 85 women consecutively presenting for symptomatic incontinence treatment; the comparators were 29 healthy women with urinary leakage who did not want any medical intervention. Two sets of results were reported: 15D scores comparing the two groups, and a comparison of pre-post scores for a sub-sample (n = 49) of women who improved as a result of treatment (d = 0.95) at follow-up (13 months); for women who did not statistically improve (n = 12), d = 55. Separately reported were the scores on the item within the 15D measuring elimination.

The EQ5D was used to examine the costs of incontinence in a study modelling the cost-effectiveness of tolterodine and oxybutynin (223). The authors defined four levels of incontinence, based on micturition (normal: M ≤ 8; mild: M = 9–12; moderate: M = 13–15; severe: M ≥ 16) and assigned women who participated in three 12-week trials comparing tolterodine with oxybutynin to these levels. The reported frequencies across all three trials were 1%, 38%, 33% and 28%. They then assigned EQ5D scores based on a sample of 455 women from a Swedish study examining the measurement of economic outcomes from incontinence (218). After assigning the Swedish women's data to the same four classification levels, the resulting EQ5D values were assigned to the women from tolterodine/oxybutynin trials in order to estimate costs.

In a randomised trial of colposuspension (laparoscopic versus abdominal incision) carried out in Melbourne, 45 women were administered at 6 month follow-up after surgery the AQoL and SF-36 as part of an investigation into patient satisfaction with medical care (224). Although the data have yet to be published, permission to use the data was given by the authors. The data were analysed by the incontinence status of the women at follow-up; of the 45 cases 16 were still incontinent at the time of data collection. The definition of incontinent was where the patient suffered both stress and urinary incontinence at follow-up as defined by self-report. The SF6D-2 scores were computed from the SF-36 data.

In the 1998 South Australian Health Omnibus Survey (HOS, n = 3010) both the AQoL and SF-36 were used. For this report the SF6D-2 utility scores were computed. There were three HOS questions on incontinence (faecal incontinence; loss of urine when coughing, sneezing or laughing; and urge incontinence involving urine loss prior to reaching a toilet). These data have not been previously published and are presented here with permission from Professor Alastair MacLennan, Department of Obstetrics and Gynaecology, Adelaide University. The HOS is a user-pays survey for health organizations, covering people aged 15+ years, involving random sampling from the SA population (2). For this report, urinary incontinence was defined as those cases meeting the two questions on urine loss (n = 194) and faecal incontinence referred to those who endorsed the faecal incontinence question (n = 87). There were 23 cases who reported both; these cases have been assigned to faecal incontinence. The comparator was all cases not assigned to incontinence. These data, of course, ignore the presence or absence of comorbidities.

Schultz and Kopec (225) re-analysed HUI3 data from the Canadian National Population Health Survey, 1996/7 (n = 73,402) with respect to 21 chronic conditions. The definition of a chronic

condition was where a condition had lasted or was expected to last for six months or more and had been diagnosed by a health professional. When HUI3 utility scores were assessed for those with no comorbid conditions, incontinence was the third most severe condition after Alzheimer's disease and stroke. For those with incontinence only the mean HUI3 utility was 0.82 (sd = 0.46), for those with incontinence and other comorbidities it was 0.61 (sd = 0.48). As with the Mittmann et al (221) study, the fact that there was such a discrepancy in the HUI3 scores by 'incontinence only' versus 'incontinence and comorbidities' suggests that most of the difference in HUI3 scores can be attributed to the effect of the comorbidities. For the purposes of reporting the sensitivity of the HUI3, comparison in this report was made between those with incontinence only and those without any reported chronic condition. This latter group are those reported as the comparator group in Table A.2.

A.6.4 Discussion of the Results presented in Table A.2

Examination of the studies shows that they involved very different populations and samples, and the definitions of incontinence varied. For example, although the AQoL South Australian and HUI3 Canadian data (221, 225) were both population surveys involving self-report, the AQoL estimates were derived from non-clinical questions whereas the HUI3 estimates were reports of clinical assessments. Additionally, the estimates for the HUI3 exclude those with comorbidities whilst those for the AQoL make no distinction between those with incontinence alone and those with comorbidities. This difference, however, does of itself invalidate the comparison since Mittmann et al (221) stated that those with incontinence alone reported the worst HUI3 utility value. The Schultz HUI3 study (225) reported the differences between those with incontinence alone and those with comorbidities, the effect of which was to reduce HUI3 values by 26%.

Appendix A: Table 2: Summary of Literature reporting use of MAU-instruments in Incontinence

	Study	*Year*	*Groups*	*Utility (sd) (a)*	*N*	*Effect size (b)*	*Relative efficiency*
AQoL	SA HOS (c)	2002	No incontinence	0.84 (0.19)	2729		
			Urinary incontinence	0.71 (0.26)	194	0.57	2.98
			Faecal incontinence	0.58 (0.29)	87	1.06	
AQoL	Hawthorne/Harmer (c)	1999	No incontinence	0.78 (0.23)	29		
			Incontinence	0.67 (0.23)	16	0.48	2.12
EQ5D	O'Brien et al	2001	No incontinence	0.74 (0.11)	6		
			Mild	0.72 (0.22)	209	0.12	
			Moderate	0.69 (0.27)	182	0.24	
			Severe	0.61 (0.38)	154	0.46	1.94
HUI3	Mittmann et al	1999	No incontinence	0.93 (0.08)	7509		
			Incontinence	0.85 (0.12)	22	0.78	5.59
HUI3	Schultz & Kopec	2003	No incontinence	0.95 (0.08)	71773		
			Incontinence	0.82 (0.46)	195	0.39	1.40
15D	Stach-Lempinen et al	2001	Comparator group (d)	0.91 (0.08)	85		
			Incontinence (baseline)	0.80 (0.09)	29	1.29	15.28
SF6D-2	SA HOS (c)	2002	No incontinence*	0.76 (0.13)	2729		
			Urinary incontinence	0.71 (0.13)	194	0.38	1.33
			Faecal incontinence	0.63 (0.15)	87	0.93	
SF6D-2	Hawthorne/Harmer (c)	1999	No incontinence	0.70 (0.10)	29		
			Incontinence	0.67 (0.08)	16	0.33	1.00

Notes: * = Median (IQR)

(a) = Where the study reported 95%CIs, these have been calculated.

(b) = Calculated from published study data. The comparator is each case is the No incontinence cohort.

(c) = Unpublished data. See the text for an explanation.

(d) = Women who were incontinent but who were not seeking treatment.

These differences must be kept in mind when interpreting the data in Table A.2. Although caution should be exercised, the analysis may be useful as a guide to instrument selection.

Generally, the obvious conclusion is that although incontinence has a significant effect on people's lives, it is not catastrophic. The table shows that all instruments report utility values for incontinence that are in the top half of the utility range. Indeed for urinary incontinence, the lowest utility value was 0.61 for those with severe incontinence on the EQ5D.

Regarding the effect of incontinence, under the axioms of utility theory it is possible to model the number of people who need to be treated to make a 1-QALY gain where it is assumed that the treatment effects last for one year

$$n = \frac{1.00}{U_{FH} - U_{IN}}$$

where U_{FH} is the mean utility value of non-incontinence and U_{IN} is the incontinence utility mean. The results show that a 1-QALY gain could be obtained by treating 7.69 people (EQ5D; 95%CI: 3.60–∞), 7.69 (AQoL, SA HOS study; 95%CIs: 5.56–12.50); 7.69 (HUI3, 95%CIs:5.12–15.42 (225)), 9.09 (AQoL, colcosuspension study; 95%CI: 2.13–∞), 9.09 (15D; 95%CI: 6.25–16.67), 12.50 (HUI3, 95%CI: 7.58–35.65 (221)), 20.00 (SF6D, HOS study; 95%CI: 14.29–50.00) and 33.33 (SF6D, colcosuspension study; 95%CI: 9.31–∞). Although the very broad confidence intervals are due to the small numbers in some of the studies, these findings suggest that among those with incontinence there are large variations in how it affects people's lives.

There is also considerable variation in the effect sizes due to the large standard deviations. The range from was 0.33 (SF6D) to 1.29 (15D). When interpreted using Cohen's criteria (112) the SF6D is capable of detecting a small effect, the EQ5D a moderate effect, and the HUI3, AQoL and 15D are capable of detecting large effects. That the 15D obtained the largest effect size is not surprising given its small standard deviations and its specific question on elimination.

The relative efficiencies also show marked differences. The data suggest that the least efficient instrument at detecting differences between incontinent and continent cases was the SF6D. Relative to the SF6D the efficiency of the other instruments was, in order, the EQ5D, the HUI3 and the AQoL, which were probably similar, and the 15D.

Finally, it should be noted that there are a number of inconsistencies between the utility values assigned across the different instruments for continent and incontinent cases. The mean incontinence score for the EQ5D (0.61) was approximately the same as faecal incontinence as reported by the SF6D-2 (0.63) when it might be expected that faecal incontinence would be worse than urinary incontinence (as shown on both the AQoL and SF6D-2). Another inconsistency, based on utility values that were derived from population estimates in Canada and Australia, is between the HUI3 (0.85 (221) and 0.82 (225) for incontinence)), the AQoL (0.84 for continent), the EQ5D (0.74 for continent) and the SF6D (0.76 (SA HOS) and 0.70 (colcosuspension study)).

There is also an inconsistency between the 15D (0.80 for incontinence), the SF6D-2 (0.76 and 0.70 for continent), the EQ5D (0.74 for continent) and the AQoL (0.78 for continent).

In addition to these findings from Table A.2, the percentage gains reported in the Foote & Moore (215) study using the Rosser Index, suggest that a 1-QALY gain could be obtained through the treatment of between 50-100 cases. This would suggest the Rosser Index is likely to be less sensitive in incontinence studies than the instruments above.

Although the inconsistencies reported above reflect the different descriptive systems, assigned weights, scoring mechanisms and study populations, they also suggest the different instruments deliver very different estimates: the results for the different instruments cannot all be right. When taken in conjunction with the differences in implied QALYs, effect sizes and relative efficiencies, they are suggestive that for similar studies the different instruments may give very different results when used in incontinence cost-utility analyses. The implications are extremely worrying as they suggest that study results may depend upon the instrument chosen rather than actual treatment benefits.

A.7 Recommendations

There has been so little work carried out in the area of incontinence and utility measurement that any recommendation is very speculative. The results of this review indicate that none of the existing MAU-instruments meet all the requirements of utility theory. Additionally, available data show that there is a great deal of inconsistency among those instruments that have been used in incontinence studies. In short, no instrument can be recommended as the 'gold standard' at this point.

The recommendations below have been framed by the fact that there are three levels at which utility instruments could be used in incontinence studies: (a) clinicians working in clinical practice, (b) specialists working in clinical practice, and (c) researchers or program evaluators.

At the clinical level, measurement is usually related to clinical management of individual patients and there are time and data collection issues which impact on recommended practice. Any instruments used at this level must possess sufficient nomological evidence to be used at the case level; i.e. for individual patient assessment. Additionally, at this level, data collection should be as brief and as possible and there should be few data analysis demands upon clinicians.

Under (b), data collected need to be sufficient to meet the needs of specialists. Whilst these include the requirements of clinical measurement, specialists need more information and are often involved in research or evaluation.

Researchers and program evaluators' needs centre round data that are useful for answering research questions where analyses are group-based; where data collection procedures may be remote; and where findings are aimed at demonstrating the effect of new treatments or at influencing policy decisions.

MAU-instruments, by definition, were designed for use by researchers undertaking economic evaluation. However, this does not necessarily imply that they have no role to play in clinical or specialist services. At the individual level, MAU-instruments may provide HRQoL profiles based on responses to individual questions or utility scores which may be compared to group or population norms. Additionally, in a health care system committed to evidence-based practice, basic data should be collected and held at the clinician level for local analysis as well as transference to research (eg. for inclusion in incontinence monitoring or surveillance).

A.8 Summary Comments

In general, conclusions drawn from this review should be placed in the following contexts which are germane to using MAU-instruments in incontinence studies.

- Instrument length. Given that the chosen MAU-instrument is likely to be used in an instrument battery and that longer batteries place higher cognitive demands on respondents who may be in frail health states, it would seem that short instruments should be considered. These would include the EQ5D, AQoL, HUI3 (both of which could be shortened to just the 12 items contributing to the utility score) and the 15D.

- **Coverage.** There is a clear difference between the utility instruments in relation to their coverage. If instruments providing 'within the skin' coverage are to be preferred, the choices would be between the 15D and HUI3. On the other hand, if the 'social expression' of HRQoL is desired, the AQoL or SF6D would be the instruments of choice.
- **Administration. Recommended** national instruments are likely to be used in a variety of settings, particularly in studies where data are collected through self-completion (whether in mail or interview situations) and where follow-up data are likely to be collected by mail or telephone. Instruments which require interviewer administration are therefore probably not recommended. Given the cognitive demand of telephone interviews, longer instruments with more complex item responses should also be avoided. This would preclude the use of the QWB and Rosser Index; and perhaps the 15D. Instruments that are more suitable for telephone administration are the EQ5D, AQoL and HUI3.
- **Ease of use.** Instruments which respondents find simple and easy to use are the EQ5D, HUI3, 15D, AQoL, and Rosser Index.
- **Time to complete.** To reduce the burden on participants and the costs associated with data collection this should be as short as possible. Those instruments taking less than ten minutes are the EQ5D, HUI3, 15D and AQoL.

- **Translation.** Translations will almost certainly be required for some sample sub-groups given the heterogenous Australian population. Because the SF6D relies upon the SF-36, the SF6D would be regarded as having been widely translated; i.e. SF6D scores can be obtained from any language into which the SF-36 has been translated. The only MAU-instrument, per se, that is readily available in a number of languages is the EQ5D. The 15D has been translated into several European languages.
- **Ease of scoring.** Although this does not directly impinge upon data collection, it does have some implications for data analysis where research groups may not have ready access to either a statistician or instrument technical support. The simpler the instrument the better. The simpler instruments regarding scoring are the 15D, EQ5D, AQoL and HUI3.
- **Sensitivity.** This is important given that in some situations the critical effect sizes for some incontinence interventions are likely to be small. Those instruments likely to be more sensitive are the 15D, AQoL and HUI3.
- **Reliability.** All the instruments reviewed are likely to possess similar reliability characteristics. However, this has not been fully investigated for all instruments.
- **Validity.** All the instruments reviewed have some questions about their validity. This has not been satisfactorily established and published for any of the instruments, particularly in relation to the generalisability of the utility weights — with the exception of the EQ5D — and the necessary strong interval property. Based on the available literature it would appear there are potential difficulties with the Rosser Index and QWB. There may also be some issues around sensitivity for the EQ5D, and some doubt as to whether the 15D actually measures utilities.
- **Utility axioms**. None of the instruments reviewed meet all the requirements for utility measurement at this time. However, the review suggests that those instruments with the better claims for meeting these axioms would be the HUI3 and AQoL, then the EQ5D and perhaps the 15D.

The following table provides a summary of the findings from this study. Each of the instruments reviewed was assessed against the descriptions and validity evidence presented in this report. For each of these criteria, the assessment was made on a 3-point scale where a low score indicated minimally meeting the criteria and a high score indicated mostly meeting the criteria. Additionally, following discussions with Jan Sansoni and A/Professor Shane Thomas, each of the criteria was weighted according to its perceived relevance to the Australian context (for example, it was decided that although having multiple language versions was advantageous, in the Australian context where English is almost universally spoken this was not an important criteria for instrument selection). The results suggest that the instruments of choice would be the AQoL, EQ5D and HUI3.

Of the tools studied in the Multi-Attribute Utility instruments category, three obtained the requisite 47-point score or higher. These were the Assessment of Quality of Life (AQoL), the European Quality of Life Measure-5D (EQ5D) and Health Utilities Index (HUI3). All three tools are recommended.

Appendix A: Table 3: Summary of Ratings for Multi-attribute Utility Instruments

Criteria	*Tool*						
	EQ5D	*AqoL*	*HUI3*	*15D*	*SF6D*	*QWB*	*Rosser*
Availability of comparison data/usage	3	2	2	1	1	2	2
Length, ease and time to complete	3	2	2	2	1	1	1
Method of administration	3	2	2	2	3	1	1
Translations available	3	1	2	1	3	1	1
Ease of scoring	3	3	3	3	3	2	1
Sensitivity to incontinence	1	2	2	3	1	1	1
Reliability evidence available	2	2	2	1	1	1	1
Validity evidence available	2	2	2	1	1	1	1
Adherence to psychometric axioms	2	3	3	1	1	1	1
Cost of using the instrument	2	3	1	3	3	3	3
Weighted Total	**55**	**54**	**51**	**42**	**38**	**33**	**31**

EQ5D European Quality of Life Measure – 5D (formerly the EUROQOL)
AQoL Assessment of Quality of Life
HUI3 Health Utilities Index – Version 3
15D Fifteen-Dimensional measure of health-related quality of life
SF6D No acronym available
QWB Quality of Well Being
Rosser Rosser Quality of Life Index

Recommendations

There are, therefore, a number of options which could be considered either individually or collectively.

1. A single MAU-instrument could be recommended as the preferred instrument of choice for routine use at the clinician- and specialist-levels. This instrument should be short, easy to administer and score, and population norms could be made available for easy reference. If such a policy was adopted, it would be in light of the limitations outlined in this report and there would be no guarantee that results obtained would be comparable with results obtained elsewhere using another instrument. Indeed, where QALYs were computed as the result of a treatment, it is likely these would reflect instrument choice as much as treatment effect. Where two MAU-instruments were recommended as the preferred measures, these difficulties would be compounded if some studies included one of the instruments and other studies opted for the other instrument.
2. To overcome this uncertainty, it could be recommended that two MAU-instruments be included in any particular research or evaluation study, and that researchers be encouraged to provide both sets of results. One of the recommended instruments should be that recommended for clinician use. This strategy would have the benefit of reducing the bias inherent in a one-instrument strategy, and it would produce a range of estimated benefits from interventions, thus acknowledging the limitations of relying upon any particular existing MAU-instrument. Given that, inevitably, comparisons will be made with incontinence studies overseas, this strategy would have the further benefit of enabling cross-cultural comparisons.
3. Several instruments could be trialled in 3–4 large incontinence studies for the explicit purpose of identifying the instrument to be recommended for future use. Whilst this would impose an immediate burden for, say, 3 to 5 years, it would enable many of the questions raised in this report regarding the validity of MAU-instruments to be thoroughly investigated in an Australian context. This would place Australia in a position of world leadership in incontinence and utility research; it would enable a fully informed decision to be made regarding instrument selection; and it is likely the Australian model would become the world standard in the immediate future given the paucity of current research in the field. Should this latter scenario eventuate, it is likely this would enhance international cooperation in the field.
4. As an alternative to #3, the multi-attribute utility instruments that were considered above could be included in the 2004 South Australian Health Omnibus Survey (HOS), together with suitable questions on incontinence and incontinence-related health sequelae. This would enable the rapid collection of data and its analysis leading to instrument selection and recommendation. Since the HOS involves drawing a weighted population sample, the findings could be used to establish population norms against which future work could be interpreted. (NB: This work is currently in progress.)
5. A specific study could be funded to develop an incontinence module for attachment to a generic MAU-instrument descriptive system. This recommendation arises from the consideration that there are HRQoL areas of concern to those with incontinence that are not addressed with fully generic instruments. If an incontinence module for an existing instrument were constructed, researchers would be in a position to report both incontinence-specific HRQoL effects and generic utility scores. This model has been followed by the SF-36, for which there are now many disease-specific modules, and it is being followed by the World Health Organization Quality of Life Group for the WHOQOL-OLD (being specifically developed for use with older adults)(see Murphy & Hawthorne 2001 (226)), and also in Australia in the area of visual impairment and the AQoL. The chief difficulty lies in selecting the base instrument.

References

1. WHO. *International Classification of Impairments, Disabilities and Handicaps*. Geneva: World Health Organization; 1980.
2. Wilson D, Wakefield M, Taylor A. *The South Australian Health Omnibus Survey*. Health Promotion Journal of Australia 1992;2(3):47-49.
3. Sintonen H, Pekurinen M. *A fifteen-dimensional measure of health-related quality of life (15D) and its applications*. In: Walker S, Rosser R, editors. *Quality of Life Assessment*. Dordrecht: Kluwer Academic Publishers; 1993.
4. Sintonen H. *The 15D measure of health-related quality of life: reliability, validity and sensitivity of its health state descriptive system*. Melbourne: National Centre for Health Program Evaluation; 1994. Report No.: Working Paper 41.
5. Hawthorne G, Richardson J, Osborne R. *The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health related quality of life*. Quality of Life Research 1999;8:209-224.
6. Hawthorne G, Osborne R. *Population norms and meaningful differences for the Assessment of Quality of Life (AQoL) measure*. Australian and New Zealand Journal of Public Health 2005;29(2):136-142.
7. EuroQol Group. *EuroQol: a new facility for measurement of health-related quality of life*. Health Policy 1990;16:199-208.
8. Dolan P, Gudex C, Kind P, Williams A. *Social tariff for EUROQoL: results from a UK general population survey*. York: Centre for Health Economics, University of York; 1995. Report No.: Discussion Paper 138.
9. Feeny D, Furlong W, Boyle M, Torrance GW. *Multi-attribute health status classification systems*. Health Utilities Index. Pharmacoeconomics 1995;7(6):490-502.
10. Feeny D, Torrance G, Furlong W. Health utilities index. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2nd ed. Philadelpha: Lippincott-Raven Publishers; 1996.
11. Sandvik H, Hunskaar S, Seim A, Hermstad R, Vanvik A, Bratt H. *Validation of a severity index in female urinary incontinence and its implementation in an epidemiological survey*. Journal of Epidemiology and Community Health 1993;47(6):497-9.
12. Sandvik H, Seim A, Vanvik A, Hunskaar S. *A severity index for epidemiological surveys of female urinary incontinence: comparison with 48-hour pad-weighing tests*. Neurourology and Urodynamics 2000;19(2):137-45.
13. Ware J, Sherbourne C. *The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection*. Medical Care 1992;30(6):473-483.
14. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 *Health Survey: Manual and Interpretation Guide*. Boston: The Health Institute, New England Medical Centre; 1993.
15. Ware JE, Kosinski MA, Dewey JE. *How to Score Version 2 of the SF-36 Health Survey*. Lincoln: Quality Metric Inc.; 2000.
16. Brazier J, Usherwood T, Harper R, Thomas K. *Deriving a preference-based single index from the UK SF-36 Health Survey*. Journal of Clinical Epidemiology 1998;51(11):1115-1128.
17. Brazier J, Roberts J, Deverill M. *The estimation of a preference-based measure of health from the SF-36*. Journal of Health Economics 2002;21:271-292.
18. Shumaker S, Wyman J, Uebersax J, McClish D, Fanti J. *Health-related quality of life measures for women with urinary incontinence: the incontinence impact questionnaire and the urogenital distress inventory*. Quality of Life Research 1994;3:291-306.
19. Uebersax J, Wyman J, McClish D, Shumaker F, McKlish J, Santl J. *Short forms to assess life-quality and symptoms distress for urinary incontinence in women; the Incontinence Impact Questionnaire and the Uro-Genital Distress Inventory*. Neurology and Uro-Dynamics 1995;14:131-139.

20. Jorge JM, Wexner SD. *Etiology and management of fecal incontinence. Diseases of the Colon and Rectum.* 1993;36(1):77-97.

21. Avery A, Taylor A, Gill T. *Incontinence in South Australia: Prevalence, Risks and Priorities*. Adelaide: Population Research and Outcomes Studies Unit, South Australian Department of Health; 2004.

22. Muscatello DJ, Rissel C, Szonyi G. *Urinary symptoms and incontinence in an urban community: prevalence and associated factors in older men and women*. Internal Medical Journal 2001;31(3):151-60.

23. Hughes AM, Sladden MJ, Hirst GH, Ward JE. *Community study of uncomplicated lower urinary tract symptoms among male Italian immigrants in Sydney, Australia*. European Urology 2000;37(2):191-8.

24. Chiarelli P, Brown W, McElduff P. *Leaking urine: prevalence and associated factors in Australian women*. Neurourology and Urodynamics 1999;18(6):567-77.

25. Millard RJ. *The incidence of urinary incontinence in Australia: a demographic survey conducted in the Sydney area in 1983*. Journal of Urology 1985;57:98-99.

26. Ouslander JG, Kane RL, Abrass IB. *Urinary incontinence in elderly nursing home patients*. Journal of the American Medical Association 1982;248(10):1194-8.

27. Thomas S, Moore K, Nay R, Fonda D, Marosszeky N, Sansoni J, et al. *Continence Outcomes Measurement Suite Project Final Report*. Melbourne: La Trobe University; InPress April.

28. Harrison Health Research. *Findings from Autumn 2004 Health Omnibus Survey*. Adelaide: Harrison Health Research; 2004.

29. Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U, et al. *The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society*. Neurourology and Urodynamics 2002;21(2):167-78.

30. Abrams A, Cardozo L, Khoury S, Wein A, editors. *Incontinence: Volume 1: Basics and Evaluation*. Paris: International Continence Society; 2005.

31. Abrams A, Cardozo L, Khoury S, Wein A, editors. *Incontinence: Volume 2: Management*. Paris: International Continence Society; 2005.

32. Bates P, Bradley WE, Glen E. *Standardization of terminology of lower urinary tract function*. Journal of Urology 1979;121:551-554.

33. Swithinbank LV, Donovan JL, du Heaume JC, Rogers CA, James MC, Yang Q, et al. *Urinary symptoms and incontinence in women: relationships between occurrence, age, and perceived impact*. British Journal of General Practice 1999;49(448):897-900.

34. Norton C, Whitehead WE, Bliss DZ, Metsola P, Tries J. *Conservative treatment and pharmacological management of faecal incontinence in adults*. In: Abrams A, Cardozo L, Khoury S, Wein A, editors. *Incontinence: Volume 2: Management*. Paris: International Continence Society; 2005. p. 1521-1564.

35. Nelson R, Norton N, Cautley E, Furner S. *Community-based prevalence of anal incontinence*. Journal of the American Medical Association 1995;274(7):559-61.

36. Whitehead WE, Wald A, Diamant NE, Enck P, Pemberton JH, Rao SS. *Functional disorders of the anus and rectum*. Gut 1999;45 Suppl 2:II55-9.

37. Royal College of Physicians. *Incontinence. Causes, management and provision of services*. A Working Party of the Royal College of Physicians. Journal of the Royal College of Physicians of London 1995;29(4):272-4.

38. Hanley J, Capewell A, Hagen S. *Validity study of the severity index, a simple measure of urinary incontinence in women*. British Medical Journal 2001;322(7294):1096-7.

39. Hannestad YS, Rortveit G, Sandvik H, Hunskaar S. *A community-based epidemiological survey of female urinary incontinence: the Norwegian EPINCONT study*. Epidemiology of Incontinence in the County of Nord-Trondelag. Journal of Clinical Epidemiology 2000;53(11):1150-7.

40. Lemack GE, Zimmern PE. *Predictability of urodynamic findings based on the Urogenital Distress Inventory-6 questionnaire*. Urology 1999;54(3):461-6.

41. Pang MW, Leung HY, Chan LW, Yip SK. *The impact of urinary incontinence on quality of life among women in Hong Kong*. Hong Kong Medical Journal 2005;11(3):158-63.

42. Vaizey CJ, Carapeti E, Cahill JA, Kamm MA. *Prospective comparison of faecal incontinence grading systems*. Gut 1999;44(1):77-80.

43. McCall WA. *How to measure in education*. New York: Macmillan; 1922.

44. Sansoni J, Costi J. SF-36: Version 1 or Version 2: the need for Australian normative data. In: *Proceedings of Health Outcomes 2001: The Odyssey Advances Conference; 2001 December 2001*; Canberra: Australian Health Outcomes Collaboration; 2001.

45. Hawthorne G, Richardson J. *Measuring the value of program outcomes: a review of utility measures*. Expert Review of Pharmacoeconomics Outcomes Research 2001;1(2):215-228.

46. Hawthorne G, Richardson J. *An Australian MAU/QALY Instrument: Rationale and Preliminary Results*. In. Melbourne: Centre for Health Program Evaluation; 1995. p. 29.

47. Dolan P. *Modeling valuations for EuroQol health states*. Medical Care 1997;35(11):1095-108.

48. Furlong WJ, Feeny DH, Torrance GW, Barr RD. *The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies*. Annals of Medicine 2001;33:375-384.

49. Furlong W, Feeny D, Torrance G, Goldsmith C, DePauw S, Zhu Z, et al. *Multiplicative Multi-attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report. Working Paper*. Hamilton: McMaster University, Centre for Health Economics and Policy Analysis; 1998. Report No.: 98-11.

50. Torrance GW, Furlong W, Feeny D, Boyle M. *Multi-attribute preference functions. Health Utilities Index*. Pharmacoeconomics 1995;7(6):503-20.

51. Sintonen H. *The 15D instrument of health-related quality of life: properties and applications*. Annals of Medicine 2001;33:328-336.

52. Sintonen H. *The 15D measure of health-related quality of life: feasibility, reliability and validity of its valuation system*. Melbourne: National Centre for Health Program Evaluation; 1995. Report No.: Working Paper 42.

53. Hawthorne G, Elliott P. *Imputing cross-sectional missing data: a comparison of common techniques*. Australian and New Zealand Journal of Psychiatry 2005;39:583-590.

54. Kalantar JS, Howell S, Talley NJ. *Prevalence of faecal incontinence and associated risk factors; an underdiagnosed problem in the Australian community?* Medical Journal of Australia 2002;176(2):54-7.

55. Avery JC, Gill TK, MacLennan AH, Chittleborough CR, Grant JF, Taylor AW. *The impact of incontinence on health-related quality of life in a South Australian population sample*. Australian and New Zealand Journal of Public Health 2004;28(2):173-9.

56. Tariq SH. *Geriatric fecal incontinence*. Clinics in Geriatric Medicine 2004;20(3):571-87, ix.

57. Peet SM, Castleden CM, McGrother CW. *Prevalence of urinary and faecal incontinence in hospitals and residential and nursing homes for older people*. British Medical Journal 1995;311(7012):1063-4.

58. Liu C, Andrews GR. *Prevalence and incidence of urinary incontinence in the elderly: a longitudinal study in South Australia*. Chinese Medical Journal 2002;115(1):119-22.

59. Sladden MJ, Hughes AM, Hirst GH, Ward JE. *A community study of lower urinary tract symptoms in older men in Sydney, Australia*. Australian and New Zealand Journal of Surgery 2000;70(5):322-8.

60. Boyle G. *Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? Personality and Individual Differences.* 1991;12(3):291-294.

61. Cortina J. *What is coefficient alpha? Examination of theory and applications*. Journal of Applied Psychology 1993;78(1):98-104.

62. Cronbach J, Meehl P. *Construct validity in psychological tests*. Psychological Bulletin 1955;52(4):281-302.

63. Landis JR, Koch GG. *The measurement of observer agreement for categorical data*. Biometrics 1977;33:159-174.

64. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 4th ed. New York: Harper Collins; 2001.

65. Nunnally J. *Psychometric Theory*. 2nd ed. New York: McGraw Hill; 1978.

66. Cundiff GW, Harris RL, Coates KW, Bump RC. *Clinical predictors of urinary incontinence in women*. American Journal of Obstetrics and Gynecology 1997;177(2):262-6; discussion 266-7.

67. Hair J, Anderson R, Tatham R, Black W. *Multivariate Data Analysis*. 5th ed. Upper Saddle River: Prentice Hall; 1998.

68. Moran LA, Guyatt GH, Norman GR. *Establishing the minimal number of items for a responsive, valid, health-related quality of life instrument*. Journal of Clinical Epidemiology 2001;54(6):571-9.

69. Chiarelli P. *Urinary incontinence: the last taboo?* Australian Journal of Rural Health 2004;12(6):277-8.

70. Bosch JL, Halpern EF, Gazelle GS. *Comparison of preference-based utilities of the Short-Form 36 Health Survey and Health Utilities Index before and after treatment of patients with intermittent claudication*. Medical Decision Making 2002;22(5):403-9.

71. Bosch JL, Hunink MG. *Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication*. Quality of Life Research 2000;9(6):591-601.

72. Brazier J, Roberts J, Tsuchiya A, Busschbach J. *A comparison of the EQ-5D and SF-6D across seven patient groups*. Health Economics 2004;13(9):873-84.

73. Hatoum HT, Brazier JE, Akhras KS. *Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting*. Value in Health 2004;7(5):602-9.

74. Oostenbrink R, HA AM, Essink-Bot ML. *The EQ-5D and the Health Utilities Index for permanent sequelae after meningitis: a head-to-head comparison*. Journal of Clinical Epidemiology 2002;55(8):791-9.

75. Gerard K, Nicholson T, Mullee M, Mehta R, Roderick P. *EQ-5D versus SF-6D in an older, chronically ill patient group*. Applied Health Economics and Health Policy 2004;3(2):91-102.

76. Holland R, Smith R, Harvey I, Swift L, Lenaghan E. *Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL*. Health Economics 2004;13(8):793-805.

77. Pickard AS, Johnson JA, Feeny DH. *Responsiveness of generic health-related quality of life measures in stroke*. Quality of Life Research 2005;14(1):207-19.

78. Stavem K, Froland SS, Hellum KB. *Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS*. Quality of Life Research 2005;14(4):971-980.

79. Conner-Spady B, Suarez-Almazor ME. *Variation in the estimation of quality-adjusted life-years by different preference-based instruments*. Medical Care 2003;41(7):791-801.

80. Kopec JA, Willison KD. *A comparative review of four preference-weighted measures of health-related quality of life*. Journal of Clinical Epidemiology 2003;56(4):317-25.

81. Longworth L, Bryan S. *An empirical comparison of EQ-5D and SF-6D in liver transplant patients*. Health Economics 2003;12(12):1061-7.

82. Bryan S, Longworth L. *Measuring health-related utility: Why the disparity between EQ-5D and SF-6D?* European Journal of Health Economics 2005.

83. Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, et al. *A comparison of four indirect methods of assessing utility values in rheumatoid arthritis*. Medical Care 2004;42(11):1125-31.

84. O'Brien BJ, Spath M, Blackhouse G, Severens JL, Dorian P, Brazier J. *A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index*. Health Economics 2003;12(11):975-81.

85. Stavem K. *Reliability, validity and responsiveness of two multiattribute utility measures in patients with chronic obstructive pulmonary disease*. Quality of Life Research 1999;8(1-2):45-54.

86. Brazier J, Deverill M. *A checklist for judging preference-based measures of health related quality of life: learning from psychometrics*. Health Economics 1999;8(1):41-51.

87. Richardson J. *Cost utility analysis: what should be measured?* Social Science and Medicine 1994;39(1):7-21.

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

89. Hawthorne G, Osborne RH, Elliott P. *Commentary on: a psychometric analysis of the measurement level of the rating scale, standard gamble and time-trade off*. by Cook et al. Social Science & Medicine 2003;56:895-897.

90. Hawthorne G. *The effect of different methods of collecting data: mail, telephone and filter data collection issues in utility measurement*. Quality of Life Research 2003;12:1081-1088.

91. Hawthorne G, Osborne R, McNeil H, Richardson J. *The Australian Multi-attribute Utility (AMAU): Construction and Initial Validation*. In. Melbourne: Centre for Health Program Evaluation; 1996. p. 45.

92. Richardson J, Hawthorne G. *The Australian quality of life (AQoL) instrument: psychometric properties of the descriptive system and initial validation*. Australian Studies in Health Services Administration 1998;85:315-342.

93. Osborne R, Hawthorne G, Papanicolaou M, Wegmuller Y. *Measurement of rapid changes in health outcomes in people with influenza symptoms*. Journal of Outcomes Research 2000;4:15-30.

94. Hogan A, Hawthorne G, Kethel L, Giles E, White K, Stewart M, et al. *Health-related quality-of-life outcomes from adult cochlear implantation: a cross-sectional study*. Cochlear Implants International 2001;2(2):115-128.

95. Osborne RH, Hawthorne G, Lew EA, Gray LC. *Quality of Life assessment in the community-dwelling elderly: validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36*. Journal of Clinical Epidemiology 2003;56(2):138-147.

96. Hawthorne G, Hogan A, Giles E, Stewart M, Kethel L, White K, et al. *Evaluating the health-related quality of life effects of cochlear implants: a prospective study of an adult cochlear implant program*. International Journal of Audiology 2004;43(4):183-192.

97. Dorman P, Slattery J, Farrell B, Dennis M, Sandercock P. *Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke*. United Kingdom Collaborators in the International Stroke Trial. Stroke 1998;29(1):63-68.

98. van Agt HM, Essink-Bot ML, Krabbe PF, Bonsel GJ. *Test-retest reliability of health state valuations collected with the EuroQol questionnaire*. Social Science and Medicine 1994;39(11):1537-1544.

99. Fransen M, Edmonds J. *Reliability and validity of the EuroQol in patients with osteoarthritis of the knee*. Rheumatology 1999;38(9):807-13.

100. Ruiz M, Rejas J, Soto J, Pardo A, Rebollo I. *Adaptation and validation of the Health Utilities Index Mark 3 into Spanish and correction norms for Spanish population*. Medicina Clinica 2003;120(3):89-96.

101. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. *Reliability of the Health Utilities Index – Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire*. Quality of Life Research 1995;4:249-257.

102. Barr RD, Simpson T, Whitton A, Rush B, Furlong W, Feeny DH. *Health-related quality of life in survivors of tumours of the central nervous system in childhood – a preference-*

based approach to measurement in a cross-sectional study. European Journal of Cancer 1999;35(2):248-55.

103. Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH, et al. *A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease*. Journal of Rheumatology 2003;30(10):2268-74.

104. Costet N, Le Gales C, Buron C, Kinkor F, Mesbah M, Chwalow J, et al. *French cross-cultural adaptation of the Health Utilities Indexes Mark 2 (HUI2) and 3 (HUI3) classification systems*. Clinical and Economic Working Groups. Quality of Life Research 1998;7(3):245-56.

105. Pickard AS, Johnson JA, Feeny DH, Shuaib A, Carriere KC, Nasser AM. *Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and Health Utilities Index*. Stroke 2004;35(2):607-12.

106. Verrips GH, Stuifbergen MC, den Ouden AL, Bonsel GJ, Gemke RJ, Paneth N, et al. *Measuring health status using the Health Utilities Index: agreement between raters and between modalities of administration*. Journal of Clinical Epidemiology 2001;54(5):475-81.

107. Hawthorne G, Richardson J, Day N. *A comparison of five multi-attribute utility instruments*. Australian Studies in Health Services Administration 2001;89:151-179.

108. Hawthorne G, Richardson J, Day NA. *A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments*. Annals of Medicine 2001;33(5):358–370.

109. Brazier J, Deverill M, Green C, Harper R, Booth A. *A review of the use of health status measures in economic evaluation*. Health Technology Assessment 1999;3(9):1-165.

110. Byrne BM. *Structural Equation Modeling with AMOS: Basic Concepts, Applications and Programming*. Mahwah, London: Lawrence Erlbaum; 2001.

111. Arbuckle J, Wothke W. Amos. In. 4.0 ed. Chicago: SmallWaters Corporation; 1999.

112. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.

113. Fayers P, Machin D. *Quality of Life: Assessment, Analysis and Interpretation*. Chichester: Wiley; 2000.

114. SPSS. SPSS for Windows, Version 13.1. In. 11.5 ed. Chicago: SPSS Inc.; 2004.

115. GraphPad. Prism 4. In. San Diego: GraphPad Software; 2003.

116. GraphPad. InStat. In. 3.02 ed. San Diego: GraphPad Software; 2000.

117. Streiner D, Norman G. Health Measurement Scales: *A Practical Guide to their Development and Use*. 2nd ed. Oxford: Oxford Medical Publications; 1995.

118. Richardson C, Zumbo B. *A statistical examination of the Health Utility Index-Mark III as a summary measure of health status for a general population survey*. Social Indicators Research 2000;51(2):171-191.

119. Sanson-Fisher RW, Perkins JJ. *Adaptation and validation of the SF-36 Health Survey for use in Australia*. Journal of Clinical Epidemiology 1998;51(11):961-7.

120. McCallum J. *The new 'SF-36' Health Status Measure: Australian Validity Tests. Working Paper*. Canberra: National Centre for Epidemiology and Population Health; 1994. Report No.: 2.

121. McCallum J. *The SF-36 in an Australian sample: validating a new, generic health status measure*. Australian Journal of Public Health 1995;19:160-6.

122. Perkins JJ, Sanson-Fisher RW. *An examination of self- and telephone-administered modes of administration for the Australian SF-36*. Journal of Clinical Epidemiology 1998;51(11):969-73.

123. Watson E, Firman D, Baade P, Ring I. *Telephone administration of the SF-36 health survey: validation studies and population norms for adults in Queensland*. Australian and New Zealand Journal of Public Health 1996;20(4):359-363.

124. ABS. *National Health Survey: SF-36 Population Norms, Australia*. Canberra: Australia Bureau of Statistics; 1997.

125. Andresen E, Patrick D, Carter W, Malmgren J. *Comparing the performance of health status measures for healthy older adults*. Journal of the American Geriatrics Society 1995;43(9):1030-1034.

126. Bullinger M, Alonso J, Apolone G, Leplège A, Sullivan M, Wood Dauphinee S, et al. *Translating health status questionnaires and evaluating their quality: the IQOLA Project approach.* International Quality of Life Assessment. Journal of Clinical Epidemiology 1998;51(11):913-23.

127. Hayes V, Morris J, Wolfe C, Morgan M. *The SF-36 health survey questionnaire: is it suitable for use with older adults?* Age and Ageing 1995;24(2):120-5.

128. O'Mahoney PG, Rodgers H, Thomson RG, Dobson R, James OFW. *Is the SF-36 suitable for assessing health status of older stroke patients?* Age and Ageing 1998;27(1, Jan):19-24.

129. Parker SG, Peet SM, Jagger C, Farhan M, Castleden CM. *Measuring health status in older patients. The SF-36 in practice*. Age and Ageing 1998;27(1, Jan):13-18.

130. Jenkinson C. *Evaluating the efficacy of medical treatment: possibilities and limitations*. Social Science and Medicine 1995;41(10):1395-401.

131. Keller SD, Ware JE, Jr., Gandek B, Aaronson NK, Alonso J, Apolone G, et al. *Testing the equivalence of translations of widely used response choice labels: results from the IQOLA Project*. International Quality of Life Assessment. Journal of Clinical Epidemiology 1998;51(11):933-44.

132. Bjorner JB, Damsgaard MT, Watt T, Groenvold M. *Tests of data quality, scaling assumptions, and reliability of the Danish SF-36*. Journal of Clinical Epidemiology 1998;51(11):1001-11.

133. Razavi D, Gandek B. *Testing Dutch and French translations of the SF-36 Health Survey among Belgian angina patients*. Journal of Clinical Epidemiology 1998;51(11):975-81.

134. Taft C, Karlsson J, Sullivan M. *Performance of the Swedish SF-36 version 2.0*. Quality of Life Research 2004;13(1):251-6.

135. Jenkinson C, Stewart-Brown S, Petersen S, Paice C. *Assessment of the SF-36 version 2 in the United Kingdom*. Journal of Epidemiology and Community Health 1999;53(1):46-50.

136. Gandek B, Ware JE, Jr. *Methods for validating and norming translations of health status questionnaires: the IQOLA Project approach*. International Quality of Life Assessment. Journal of Clinical Epidemiology 1998;51(11):953-9.

137. Barmes DE. *Public policy on oral health and old age: a global view*. Journal of Public Health Dentistry 2000;60(4):335-7.

138. Ware J, Kosinski M, Keller S. SF-36 *Physical and Mental Health Summary Scales: A User's Manual*. Boston: The Health Institute, New England Medical Centre; 1994.

139. Feeny D, Furlong W, Torrance G. *Health Utilities Index Mark 2 and Mark 3 (HUI2/3) 15-item questionnaire for self-administered, self-assessed usual health status*. Hamilton: Centre for Health Economics and Policy Analysis, McMaster University; 1996.

140. Rummel R. *Applied factor analysis*. Evanston: Northwestern University Press; 1970.

141. Ware J, Gandek B, Keller S. *Evaluating instruments used cross-nationally: methods from the IQOLA project*. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics*. Philadelphia: Lippincott-Raven Publishers; 1996. p. 681-692.

142. Skevington SM, Sartorius N, Amir M. *Developing methods for assessing quality of life in different cultural settings. The history of the WHOQOL instruments*. Social Psychiatry and Psychiatric Epidemiology 2004;39(1):1-8.

143. Szabo S, Orley J, Saxena S, WHOQoL Group. *An approach to response scale development for cross-cultural questionnaires*. European Psychologist 1997;2(3):270-276.

144. WHOQoL Group. *The World Health Organization Quality of Life Assessment (WHOQOL): position paper from the World Health Organization*. Social Science and Medicine 1995;41(10):1403-1409.

145. Hawthorne G, Herrman H, Murphy B. *Interpreting the WHOQOL-Brèf: preliminary population norms and effect sizes*. Social Indicators Research InPress;Accepted April 2005.

146. Imhof A. *The implications of increased life expectancy for family and social life*. In: Wear A, editor. *Medicine in Society: Historical Essays*. Cambridge: Cambridge University Press; 1992.

147. Nordenfelt L, editor. *Concepts and Measurement of Quality of Life in Health Care*. Dordrecht: Kluwer Academic; 1994.

148. Walker S, Rosser R. *Quality of life assessment: key issues in the 1990s*. Dordrecht: Kluwer Academic Publishers; 1993.

149. Bowling A. Measuring Health: *A Review of Quality of Life Measurement Scales*. Milton Keynes: Open University; 1991.

150. Drummond M, O'Brien B, Stoddart G, Torrance G. *Methods for the Economic Evaluation of Health Care Programmes*. 2nd ed. Oxford: Oxford University Press; 1998.

151. Suchman EA. *Evaluative Research: Principles and Practice in Public Service & Social Action Programs*. New York: Russell Sage Foundation; 1967.

152. Shortell R, Richardson W. *Health Program Evaluation*. St Louis: Mosby; 1978.

153. Ovretveit J. *Evaluating Health Interventions*. Buckingham: Open University Press; 1998.

154. Drummond M. *Introducing economic and quality of life measurements into clinical studies*. Annals of Medicine 2001;33:344-349.

155. Singh B, Hawthorne G, Vos T. *The role of economic evaluation in mental health care*. Australian and New Zealand Journal of Psychiatry 2001;35:104-117.

156. Richardson J, Hawthorne G. *Negative Utilities and the Evaluation of Complex Health States: Issues Arising from the Scaling of a Multiattribute Utility Instrument*. In. Melbourne: Centre for Health Program Evaluation; 2000. p. 38.

157. Hawthorne G, Richardson J, Day N, McNeil H. *Life and Death: Theoretical and Practical Issues in Utility Measurement*. In. Melbourne: Centre for Health Program Evaluation; 2000. p. 22.

158. Bennett K, Torrance G. *Measuring health state preferences and utilities: rating scale, time trade-off, and standard gamble techniques*. In: B S, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia: Lippincott-Raven Publishers; 1996. p. 253-265.

159. Robinson A, Dolan P, Williams A. *Valuing health states using VAS and TTO: what lies behind the numbers?* Social Science and Medicine 1997;45(8):1289-1297.

160. Gudex C, Kind P, van Dalen H, Durand M, Morris J, Williams A. Comparing Scaling Methods for Health State Valuations – Rosser Revisited. Discussion Paper. York: Centre for Health Economics, University of York; 1993. Report No.: 107.

161. Rosser R. *A health index and output measure*. In: Walker S, Rosser R, editors. *Quality of Life Assessment: Key Issues in the 1990s*. Dordrecht: Kluwer Academic Publishers; 1993.

162. von Winterfeldt D, Edwards W. *Decision analysis and behavioural research*. Cambridge: Cambridge University Press; 1986.

163. Kaplan R, Anderson J, Ganiats T. *The Quality of Well-Being Scale: rationale for a single quality of life index*. In: Walker S, Rosser R, editors. *Quality of Life Assessment: Key Issues in the 1990s*. Dordrecht: Kluwer Academic Publishers; 1993.

164. Kaplan R, Ganiats T, Sieber W, Anderson J. *The Quality of Well-being Scale*. Medical Outcomes Trust Bulletin 1996:2-3.

165. Kind P. *The EuroQoL instrument: an index of health-related quality of life*. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2nd ed. Philadelpha: Lippincott-Raven Publishers; 1996. p. 191-201.

166. Hawthorne G, Richardson J, Day N. *A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments*. In: Sintonen H, editor. XII Medical Symposium "Quality of Life Measurement in Clinical Studies"; 2000; Helsinki, Finland: Annals of Medicine; 2000. p. 358-3760.

167. Hawthorne G, Richardson J, Day N. *A comparison of five multi-attribute utility instruments*. In: Bridges J, editor. *Australian Health Economics Society Conference: Economics and Health: 2000 Proceedings of the Twenthieth-First Australian Conference of Health Economists; 2000; Gold Coast, Australia*: Australian Studies in Health Service Administration; 2000. p. 151-179.

168. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BE. *Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study*. Medical Decision Making 1997;17(1):1-9.

169. Kind P, Gudex C. *Measuring health status in the community: a comparison of methods*. Journal of Epidemiology and Community Health 1994;48:86-91.

170. Martin A, Glasziou P, Simes R. *A Utility-Based Quality of Life Questionnaire for Cardiovascular Patients: Reliability and Validity of the UBQ-H(eart) Items*. Sydney: NHMRC Clinical Trials Centre, University of Sydney; 1996.

171. McDowell I, Newell C, editors. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press; 1987.

172. Cadet B. *History of the construction of a health indicator integrating social preference: the Quality of Well-Being Scale*. In: 7th Meeting, International Network on Health Expectancy (REVES); 1994 23-25 February; Canberra; 1994.

173. Kaplan R, Bush J, Berry C. *Health status: types of validity and the Index of Well-being*. Health Services Research 1976;11(4):478-507.

174. Anderson J, Kaplan R, Berry C, Bush J, Ruben R. *Interday reliability of function assessment for a health status measure: The Quality of Well-Being scale*. Medical Care 1989;27(11):1076-1084.

175. Bombardier C, Raboud J. *A comparison of health-related quality-of-life measures for rheumatoid arthritis research*. The Auranofin Cooperating Group. Controlled Clinical Trials 1991;12(4 Suppl):243S-256S.

176. Coons S, Kaplan R. *Quality of life assessment: understanding its use as an outcome measure*. Hospital Formulary 1993;28:486-498.

177. Nord E. *A Review of Synthetic Health Indicators*. Oslo: National Institute of Public Health for the OECD Directorate for Education, Employment, Labour and Social Affairs; 1997.

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

179. Kaplan RM, Anderson JP. *A general health policy model: update and applications*. Health Services Research 1988;23(2):203-35.

180. Bergner M, Bobbitt RA, Carter WB, Gilson BS. *The Sickness Impact Profile: development and final revision of a health status measure*. Medical Care 1981;19(8):787-805.

181. Hunt S, McKenna S, McEwen J, Williams J, Papp E. *The Nottingham Health Profile: subjective health status and medical consultations*. Social Science and Medicine 1981;15A:221-229.

182. MVHGroup. *The Measurement and Valuation of Health: Final Modelling of Valuation Tariffs*. York: Centre for Health Economics; 1995.

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

184. Hawthorne G, Richardson J, Day N, McNeil H. *Using the 'Assessment of Quality of Life' (AQoL) instrument*. Technical Report. Melbourne: CHPE; 2000. Report No.: 12.

185. Robinson A, Loomes G, Jones-Lee M. *Visual analog scales, standard gambles and relative risk aversion*. Medical Decision Making 2001;21:17-27.

186. Torrance GW, Feeny D, Furlong W. *Visual analog scales: do they have a role in the measurement of preferences for health states?* Medical Decision Making 2001;21(4):329-34.

187. Torrance G, Boyle M, Horwood S. *Application of multi-attribute theory to measure social preferences for health states*. Operations Research 1982;30:1043-1069.

188. Bleichrodt H, Johannesson M. *An experimental test of a theoretical foundation for rating-scale valuations*. Medical Decision Making 1997;17(2):208-16.

189. Nord E. *Unjustified use of the Quality of Well-Being Scale in priority setting in Oregon*. Health Policy 1993;24(1):45-53.

190. Martin A. *Relative merits of the multi-dimensional health status indexes*. In: AIHW Workshop Expert Group Evaluation of Measures for the Coordinated Care Trials; 1996; Canberra; 1996.

191. Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. *Measuring changes in quality of life following magnetic resonance imaging of the knee: SF-36, EuroQol or Rosser index?* Quality of Life Research 1995;4(4):325-34.

192. Nord E, Richardson J, Macarounas-Kirchmann K. *Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys*. International Journal of Technology Assessment in Health Care 1993;9(4):463-78.

193. Mulkay M, Ashmore M, Pinch T. *Measuring the quality of life: a sociological intervention concerning the application of economics to health care*. Sociology 1987;21:541-564.

194. Elvik R. *The validity of using health state indexes in measuring the consequences of traffic injury for public health*. Social Science and Medicine 1995;40(10):1385-98.

195. DeptHealth. *Research Group on the measurement and valuation of health*. In: Methodology Workshop, EUROQoL Conference; 1995; London: Economics & Operational Research Division, Department of Health; 1995.

196. Hawthorne G, Richardson J. *Negative utility scores: theoretical and practical difficulties with an essential component of utility instruments*. In: 8th Annual Conference of the International Society for Quality of Life Research; 2001; Amsterdam, The Netherlands: Quality of Life Research; 2001. p. 207.

197. Brazier J, Rice N, Roberts J, South B. *Modelling health state values for the SF6D: a multilevel approach*. In: Health Economics Study Group Conference; 1998; Sheffield; 1998.

198. Cattell R. *Factor Analysis: an Introduction and Manual for the Psychologist and Social Scientist*. New York: Harper & Row; 1952.

199. Cook K, Ashton C, Byrne M, Brody B, Geraci J, Giesler R, et al. *A psychometric analysis of the measurement level of the rating scale, time trade-off, and standard gamble*. Social Science and Medicine 2001;53:1275-1285.

200. Kaplan R, Bush J, Berry C. *The reliability, stability and generalizability of a health status index*. In: Proceedings of the Social Statistics Section; 1978: American Statistical Association; 1978. p. 704-709.

201. Kaplan RM, Anderson JP, Patterson TL, McCutchan JA, Weinrich JD, Heaton RK, et al. *Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection*. HNRC Group. HIV Neurobehavioral Research Center. Psychosomatic Medicine 1995;57(2):138-47.

202. Bradlyn AS, Harris CV, Warner JE, Ritchey AK, Zaboy K. *An investigation of the validity of the quality of Well-Being Scale with pediatric oncology patients*. Health Psychology 1993;12(3):246-50.

203. O'Connor R. Issues in the *Measurement of Health-Related Quality of Life*. Melbourne: National Centre for Health Program Evaluation; 1993. Report No.: working Paper 30.

204. Visser M, Fletcher A, Parr G, Simpson A, Bulpitt C. *A comparison of three quality of life instruments in subjects with angina pectoris: the Sickness Impact Profile, the Nottingham Health Profile, and the Quality of Well Being Scale*. Journal of Clinical Epidemiology 1994;47(2):157-163.

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

206. Andersen L, Kristiansen I, Falch J, Aursnes I. *Cost-effectiveness of Alendronate for the prevention of osteoporotic fractures in Norwegian women*. Working Paper. Oslo: Folkehelsa; Statens Institutt for Folkehelsa; 1995. Report No.: 11/1995.

207. Herrman H, Hawthorne G, Thomas R. *Quality of life assessment in people living with psychosis*. Social Psychiatry and Psychiatric Epidemiology 2002;37(11):510-518.

208. Goldney RD, Fisher LJ, Wilson DH, Cheok F. *Suicidal ideation and health-related quality of life in the community*. Medical Journal of Australia 2001;175(10):546-549.

209. Hawthorne G, Cheok F, Goldney R, Fisher L. *The excess cost of depression in South Australia: a comparative study of two methods of calculating burden*. Australian and New Zealand Journal of Psychiatry 2003;37(3):362-373.

210. Sturm JW, Osborne RH, Dewey HM, Donnan GA, Macdonell RA, Thrift AG. *Brief comprehensive quality of life assessment after stroke: the Assessment of Quality of Life instrument in the north East melbourne stroke incidence study (NEMESIS)*. Stroke 2002;33(12):2888-94.

211. Hawthorne G, Richardson J, Osborne R, McNeil H. *The Assessment of Quality of Life (AQoL) Instrument: Construction, Initial Validation and Utility Scaling*. In. Melbourne: Centre for Health Program Evaluation; 1997. p. 23.

212. Joshi HB, Stainthorpe A, MacDonagh RP, Keeley FX, Jr., Timoney AG, Barry MJ. *Indwelling ureteral stents: evaluation of symptoms, quality of life and utility*. Journal of Urology 2003;169(3):1065-9; discussion 1069.

213. Krahn M, Ritvo P, Irvine J, Tomlinson G, Bremner KE, Bezjak A, et al. *Patient and community preferences for outcomes in prostate cancer: implications for clinical policy*. Medical Care 2003;41(1):153-64.

214. Manca A, Sculpher MJ, Ward K, Hilton P. *A cost-utility analysis of tension-free vaginal tape versus colposuspension for primary urodynamic stress incontinence*. British Journal of Obstetrics and Gynaecology 2003;110(3):255-62.

215. Foote A, Moore K. *A comparative cost utility analysis of five treatments for female urinary incontinence*. Australian Continence Journal 2001;7(4):108-109.

216. Fuertes ME, Garcia Matres MJ, Gonzalez Romojaro V, de la Rosa S, Anguera Vila A, de la Pena J, et al. *Clinical trial to evaluate trospium chloride (Uraplex) effectiveness and tolerance in patients with detrusor instability incontinence and its impact on quality of life*. Archivos Espanoles de Urologia 2000;53(2):125-136.

217. Ogawa A, Shimazaki J, Mitsuya H, Miyazaki S, Kurita T. *Clinical effects of terodiline hydrochloride on urinary frequency and sense of residual urine--a double blind clinical trial using flavoxate hydrochloride as a control*. Hinyokika Kiyo. Acta Urologica Japonica 1988;34(4):739-753.

218. Kobelt G. *Economic considerations and outcome measurement in urge incontinence*. Urology 1997;50(6A Suppl):100-107.

219. Liang M, Larson M, Cullen K, Schwartz J. *Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research*. Arthritis and Rheumatism 1985;28(5):542-547.

220. Wright J, Young N. *A comparison of different indices of responsiveness*. Journal of Clinical Epidemiology 1997;50(3):239-246.

221. Mittmann N, Trakas K, Risebrough N, Liu BA. *Utility scores for chronic conditions in a community-dwelling population*. Pharmacoeconomics 1999;15(4):369-376.

222. Stach-Lempinen B, Kujansuu E, Laippala P, Metsanoja R. *Visual analogue scale, urinary incontinence severity score and 15 D-psychometric testing of three different health-related quality-of-life instruments for urinary incontinent women*. Scandinavian Journal of Urology and Nephrology 2001;35(6):476-83.

223. O'Brien BJ, Goeree R, Bernard L, Rosner A, Williamson T. *Cost-effectiveness of tolterodine for patients with urge incontinence who discontinue initial therapy with oxybutynin: a Canadian perspective*. Clinical Therapeutics 2001;23(12):2038-49.

224. Hawthorne G, Harmer C. GUTSS: *The Genito-Urinary Treatment Satisfaction Scale Study*. In. Melbourne: Centre for Health Program Evaluation; 2000. p. 50.

225. Schultz SE, Kopec JA. *Impact of chronic conditions*. Health Reports 2003;14(4):41-53.

226. Murphy B, Hawthorne G. *Report of focus group research undertaken for the World Health Organization Quality of Life for Older Persons (WHOQOL-OLD) Study*. Melbourne: Melbourne WHOQOL Field Study Centre, Australian Centre for Posttraumatic Mental Health, Department of Psychiatry, University of Melbourne; 2001.