Health-related quality of life assessment in dementia
The psychology of health in economic evaluation

Chua, Kia-Chong

Awarding institution:
King's College London

Health-related quality of life assessment in dementia: The psychology of health in economic evaluation

Kia-Chong Chua

A thesis submitted for the degree of Doctor of Philosophy

King's College London, Institute of Psychiatry

July 2014

ABSTRACT

The health-related quality of life (HRQL) of people with dementia has received growing attention in formulating decisions about the provision and financing of health and social care. There is a need for measurement perspectives to determine whether HRQL assessment has captured what is important to the target population, to generate a coherent body of evidence to guide clinical and policy decisions. The thesis first investigates if HRQL in dementia is meaningfully interpreted as a general phenomenon in which the whole is greater than the sum of its parts. Overall total scores on DEMQOL and DEMQOL-Proxy were more sensitive to a general theme of individual differences in HRQL than subscale scores from multiple themes. Next, based on this measurement perspective, inconsistencies in self- and informant report behaviour were examined between geographical region, gender, and dementia severity. Items that demonstrated desirable measurement properties at this stage were selected for the short-form versions, DEMQOL-SF and DEMQOL-Proxy-SF, which included the preference-based items in DEMQOL-U and DEMQOL-Proxy-U. These provided the basis for a set of analyses exploring whether changes in subjective HRQL are influenced by response shift in meaning, priorities, or expectations over time. The thesis reports the findings that differences that emerge over repeated HRQL assessments could not be attributed to re-conceptualisation, re-prioritisation, or re-calibration of internal standards. Furthermore, differences in raw total scores over time were sensitive to HRQL improvement or deterioration. As such gains or losses in utility values from DEMQOL-U and DEMQOL-Proxy-U items would be

consistent with item responses that reflect longitudinal changes in HRQL. Taken together, the thesis suggests that the DEMQOL measurement system has tenable foundations for the clinical and economic evaluation of HRQL changes in dementia treatment interventions across clinical and social care settings.

ACKNOWLEDGEMENTS

I am indebted to my supervisors who helped me through a complex and difficult project, and spurred the development of its scope and meaning.

- Professor **Sube Banerjee**, Professor of Dementia & Associate Dean for Strategy, Brighton and Sussex Medical School (formerly Professor of Mental Health and Ageing, Institute of Psychiatry, King's College London)
- Dr **Renee Romeo**, Lecturer, Institute of Psychiatry, King's College London

Special thanks to two individuals who gave their time generously and provided technical insights at critical junctures of my analysis.

- Dr **Anna Brown**, Lecturer, School of Psychology, University of Kent
- Dr **George Ploubidis**, Reader, Department of Quantitative Social Science, Institute of Education, University of London (formerly Lecturer, Department of Population Studies, London School of Hygiene and Tropical Medicine)

This project would not have been completed without funding support after the first year of my study. I hereby express my gratitude for this lifeline from **King's Continuation Scholarship** of King's College London.

Special thanks also to friends, whose companionship helped me through the different chapters. This thesis is dedicated to Chua Peng-Kiat, and Lim Puay-Huan, my parents, as well as to Chua Kai-Leng, and Chua Kia-Heong, my siblings.

TABLE OF CONTENTS

Abstract 2

Acknowledgements 4

Table of Contents 5

List of Tables 13

List of Figures 16

Chapter 1 INTRODUCTION 18

1.1 Background 18

1.1.1 Drug treatment in dementia 18

1.1.2 Psychosocial intervention 19

1.1.3 Societal impact 19

1.1.4 Economic evaluation in dementia care 20

1.1.5 Health-related quality of life 21

1.1.6 Health utilities 22

1.2 Distinct concepts, common phenomenon 23

1.2.1 Valuation shift 24

1.2.2 Response shift 24

1.2.3 Illness adaptation 25

1.2.4 Response shift studies in utility assessment 27

1.3 Knowledge gaps 32

1.4 Methodological issues 33

1.4.1 Feasibility 34

1.4.2 Recall bias 34

1.4.3 Inadequate coverage 36

1.5 Scope and methodology 38

1.5.1 HRQL measurement system 38

1.5.2 Longitudinal analysis 43

1.5.3 Bifactor measurement perspective 44

1.5.4 Cross-validation 51

1.5.5 Short-form HRQL assessments 53

1.6 Thesis structure 54

Chapter 2 MEASUREMENT MODEL 57

2.1 Bifactor measurement model 57

2.2 Methods 60

2.2.1 Participants 60

2.2.2 Measures 60

2.3 Analysis 62

2.3.1 Multiple imputation 62

2.3.2 Bifactor EFA 62

2.3.3 Bifactor CFA and model comparisons 63

2.3.4 Factor strength 64

2.3.5 Model estimation 65

2.3.6 Model evaluation 66

2.4 Results 67

2.4.1 Sample characteristics 67

2.4.2 Bifactor EFA 69

2.4.3 Bifactor EFA with testlets 74

2.4.4 Bifactor CFA with testlets 76

2.4.5 Factor strength 86

2.5 Discussion 87

2.6 Limitations 91

Chapter 3 CROSS-VALIDATION 93

3.1 Measurement invariance between groups 93

3.2 Methods 98

3.2.1 Participants 98

3.2.2 Measures 99

3.3 Analysis 99

3.3.1 Covariates 99

3.3.2 DIF detection method 100

3.3.3 Model specifications and fit evaluation 103

3.3.4 Short-form derivation 104

3.4 Results 106

3.4.1 Sample characteristics 106

3.4.2 Measurement models 108

3.4.3 MIMIC models: magnitude of DIF 108

3.4.4 MIMIC models: Impact of DIF 116

3.4.5 Item selection for short-form versions 120

3.4.6 Measurement models for short-form versions 130

3.4.7 Short-form estimates of group differences 133

3.5 Discussion 134

3.5.1 DIF detection in DEMQOL 135

3.5.2 DIF detection in DEMQOL-Proxy 135

3.5.3 DEMQOL-SF and DEMQOL-Proxy-SF 136

3.5.4 Quasi-trait 137

3.5.5 Implications of HRQL as a quasi-trait 138

3.5.6 Content validity of DEMQOL-SF and DEMQOL-P-SF 139

3.6 Limitations 140

Chapter 4 RESPONSE SHIFT 144

4.1 Measurement invariance across time 144

4.2 A psychometric typology of change 145

4.3 Methods 148

4.3.1 Participants 148

4.3.2 Measures 149

4.4 Analysis 151

4.4.1 Model estimation 151

4.4.2 Model evaluation 151

4.4.3 Measurement model 152

4.4.4 Factor collapse in measurement model 153

4.4.5 Longitudinal SEM model 154

4.4.6 Modelling issues in Mplus 156

4.4.7 Longitudinal configural invariance 156

4.4.8 Longitudinal scalar invariance 157

4.4.9 Longitudinal invariance of measurement errors 159

4.4.10 Response shift and longitudinal estimates of change in HRQL 160

4.5 Results 161

4.5.1 Sample characteristics 161

4.5.2 Measurement model 165

4.5.3 Longitudinal configural invariance 166

4.5.4 Longitudinal scalar invariance 172

4.5.5 Longitudinal invariance of measurement errors 176

4.5.6 Response shift and longitudinal estimates of change in HRQL.... 183

4.6 Discussion 185

4.7 Limitations..... 189

Chapter 5 Research conclusions 193

5.1 Key findings 195

5.1.1 Forming conclusions about HRQL using DEMQOL measurement system 195

5.1.2 Conducting HRQL assessments in different settings and populations 197

5.1.3 Clinical and economic evaluation of longitudinal changes in HRQL 200

5.2 Limitations..... 202

5.3 Study implications 204

5.3.1 HRQL assessment in clinical research and practice 204

5.3.2 Illness adaptation in dementia 205

5.3.3 The value of life years in dementia 206

5.4 Future research directions 208

5.4.1 Importance of social network 208

5.4.2 Heterotypic continuity 208

5.4.3 QALY estimates in the presence of response shift 209

5.5 Concluding remarks 210

References 211

Appendices to Chapter 2 233

Appendices to Chapter 3 243

Appendices to Chapter 4 249

LIST OF TABLES

Table 1.1 DEMQOL question items. 41

Table 1.2 DEMQOL-Proxy question items. 42

Table 2.1 Demographic and clinical characteristics of study participants with complete/partial/missing HRQL assessment 68

Table 2.2 DEMQOL (28 items) bifactor EFA model standardised factor loadings 70

Table 2.3 DEMQOL-Proxy (31 items) bifactor EFA model standardised factor loadings 71

Table 2.4 Model fit evaluation 81

Table 2.5 DEMQOL bifactor CFA Model 1 standardised factor loadings for 20 items and 4 testlets 82

Table 2.6 DEMQOL-Proxy bifactor CFA Model 1 standardised factor loadings for 21 items and 5 testlets 83

Table 3.1 Demographic and clinical characteristics of study sample and participants who were excluded due to missing HRQL data 107

Table 3.2 DEMQOL bifactor CFA model standardised factor loadings 109

Table 3.3 DEMQOL-Proxy bifactor CFA model standardised factor loadings.. 110

Table 3.4 DEMQOL standardised factor loadings in baseline and final MIMIC model..... 111

Table 3.5 DEMQOL-Proxy standardised factor loadings in baseline and final MIMIC model 112

Table 3.6 Magnitude of DIF effects..... 115

Table 3.7 Group differences in HRQL and its domains for DEMQOL..... 117

Table 3.8 Group differences in HRQL and its domains for DEMQOL-Proxy... 118

Table 4.1 Croydon Memory Service observational cohort baseline data 163

Table 4.2 HTA-SADD trial sample baseline data..... 164

Table 4.3 Model fit information for longitudinal SEM models. 168

Table 4.4 DEMQOL-SF unstandardised factor loadings in multi-wave SEM model (Model 1)..... 169

Table 4.5 DEMQOL-P-SF unstandardised factor loadings in multi-wave SEM model (Model 1)..... 170

Table 4.6 Longitudinal scalar invariance in DEMQOL-SF bifactor CFA unstandardised factor loadings and thresholds..... 174

Table 4.7 Longitudinal scalar invariance in DEMQOL-Proxy-SF bifactor CFA unstandardised factor loadings and thresholds..... 175

Table 4.8 Longitudinal invariance of measurement error in DEMQOL-SF bifactor CFA unstandardised factor loadings and thresholds 177

Table 4.9 Longitudinal invariance of measurement error in DEMQOL-Proxy-SF bifactor CFA unstandardised factor loadings and thresholds 178

Table 4.10 Latent mean estimates (SE) reflecting changes from baseline assessment 182

Table 5.1 DEMQOL (28 items) factor analytic themes 195

Table 5.2 DEMQOL-Proxy (31 items) factor analytic themes 196

LIST OF FIGURES

Figure 1-1 An example of DEMQOL bifactor model. This is an 'incomplete' bifactor model as not all items load on both the general and a domain factor. Some load only on the general factor of HRQL..... 45

Figure 1-2 Factor analytic approaches that are predominant in the literature..... 48

Figure 2-1 DEMQOL 'incomplete' bifactor model (24 items and 4 testlets)..... 78

Figure 2-2 DEMQOL-Proxy 'incomplete' bifactor model (21 items and 5 testlets) 79

Figure 3-1 Item information curves for DEMQOL (POS, NEG, HRQL) 121

Figure 3-2 Item information curves for DEMQOL-Proxy (POS, NEG) 121

Figure 3-3 Item information curves for DEMQOL (COG) 122

Figure 3-4 Item information curves for DEMQOL-Proxy (COG)..... 122

Figure 3-5 Item information curves for DEMQOL (SOC) 123

Figure 3-6 Item information curves for DEMQOL-Proxy (SOC) 123

Figure 3-7 Item information curves for DEMQOL (HRQL) 124

Figure 3-8 Item information curves for DEMQOL-Proxy (HRQL) 124

Figure 3-9 Test information curve for DEMQOL and DEMQOL-SF. 129

Figure 3-10 Test information curve for DEMQOL-Proxy and DEMQOL-P-SF. 129

Figure 3-11 DEMQOL-SF bifactor CFA model standardised estimates 131

Figure 3-12 DEMQOL-P-SF bifactor CFA model standardised estimates 132

Figure 4-1 A longitudinal SEM model for two waves of HRQL assessment 155

CHAPTER 1 INTRODUCTION

This chapter considers background issues that have led to the growing use of economic evaluation to inform health policy on dementia care. It describes cost-utility analysis (CUA), a type of economic evaluation, and its relevance for dementia. Measurement issues are discussed to illustrate how utility measurement in CUA is linked to health-related quality of life (HRQL) assessment. There is a particular focus on response shift, a phenomenon that has been reported both in HRQL and utility measurement studies which is of particular relevance in dementia. Developments in the methodology for investigating response shift are reviewed to identify knowledge gaps and highlight those that form the basis of the present research. In closing, the scope and aims of this thesis are outlined.

1.1 Background

1.1.1 Drug treatment in dementia

Dementia brings about a decline in memory, reasoning and communication skills, and a gradual loss of skills needed in daily life for independent living (Knapp et al., 2007). At any stage of illness, individuals may also develop behavioural and psychological symptoms of dementia (BPSD) such as depression, psychosis (hallucinations and delusions), aggression and wandering. Available drug treatment may improve symptoms temporarily, but none has been shown to slow or stop the disease process (Thies, Bleiler, & Alzheimer's Association, 2013). Current standard treatments for BPSD continue to be the subject of clinical trials

due to long-standing concerns over drug efficacy and safety (Ballard et al., 2009; Banerjee et al., 2011).

1.1.2 Psychosocial intervention

Alongside pharmacological treatment, psychosocial interventions are growing in numbers as new intervention strategies evolve to support people with dementia and/or their carers. National guidelines developed in the UK (NICE SCIE, 2006) and across Europe (Vasse et al., 2012) draw attention to a range of psychosocial interventions that may be considered as part of health and social care in dementia. Owing in part to the complexity and/or intensity of these interventions, rigourous evidence of their effectiveness is available only from a few studies. Recent systematic reviews (Cooper et al., 2012; Knapp, Iemmi, & Romeo, 2013; Spijker et al., 2008) have gathered evidence to fill knowledge gaps about what components may be part of standard care in dementia. However, the paucity of well-conducted evaluations makes it difficult to draw conclusions with confidence.

1.1.3 Societal impact

Without effective intervention and support in place, complications in daily living can lead to the need for institutional care, where costs are dramatically higher than care in the community for all but the most complicated cases (Knapp et al., 2013). Notwithstanding the financial impact on individual families, these costs also have fiscal implications in countries where healthcare is financed by the state. In 2007 in the UK, costs of institutional care total £7 billion a year, of which two-thirds

are paid for by national health services (Knapp et al., 2007). This carries far-reaching implications for government spending on other national priorities (Banerjee, 2012). Even in countries where long-term care costs are paid for by families themselves, policy attention is drawn to the hidden costs that threaten economic growth as a segment of the population is at risk of being taken out of paid employment to care for a family member with dementia at home. Regardless of national contexts of health policy, the impact of dementia will only continue to grow. This is due to demographic trends worldwide which see a steady increase in the number of people who will be 65 years of age and older. According to the 2010 World Alzheimer Report (Wimo & Prince, 2010), the likelihood of developing dementia roughly doubles every five years after the age of 65. The number of people at risk is hence projected to rise sharply and the impact may outstrip the individual and societal resources that are available to meet the demands of this profoundly life-changing illness.

1.1.4 Economic evaluation in dementia care

Urgent and difficult decisions have to be made about the financing and provision of health and social care amidst conditions of uncertainty, conflicting objectives, and resource constraints. This brings into sharper focus the task of weighing alternative courses of action in meeting the needs of people with dementia in terms of costs and consequences. Cost-utility analysis (CUA) is one of a range of evaluative methods that is growing in use to inform health policy in dementia care. This method of evaluation is widely used as it focuses not just on the costs –

which are important in their own right – but on the amount of improvement in health outcomes associated with investing resources, and also on the strength of preferences for these target outcomes. For an understanding of its potential for informing about dementia care, two key elements which feature in CUA need to be considered. Issues associated with the measurement of health-related quality of life and health utilities are elaborated in the sections which follow.

1.1.5 Health-related quality of life

With prevailing challenges in the treatment of dementia, the goal of 'adding years to life' often includes explicit considerations of 'adding life to years' (Clark, 1995). While medications used for people with dementia target cognitive and psychiatric symptoms, these symptoms do not give a complete picture of the illness experience (Banerjee, 2007). An assessment of treatment effectiveness requires a full view of the range of domains in which impairments can occur in dementia, and the ways in which individuals can improve despite progressive illness (Rabins & Black, 2007). This is the objective of assessment of health-related quality of life (HRQL).

HRQL refers to a general state of well-being that depends on multiple aspects of physical and mental health. In dementia, these include aspects like memory, mood, and social behaviour (Lawton, 1994). No single aspect alone gives a full and accurate understanding of HRQL. There is wide variation in the domains considered important, or the ways in which good (or poor) HRQL in a domain are represented across assessment measures (Perales, Cosco, Stephan, Haro, &

Brayne, 2013). Nonetheless, consensus is clear that a broad focus is necessary to ensure that treatment benefits are not overlooked and potential harms not missed (Banerjee et al., 2006). In CUA, preferences about target outcomes are based on HRQL scenarios. This allows decision making to focus not only on the impact of clinically important symptoms but also on wider consequences of ill health when allocating scarce resources for health and social care across diagnostic groups.

1.1.6 Health utilities

With HRQL as the underpinning basis, CUA establishes the desirability of treatment benefits in the form of utility values. Direct elicitation of HRQL preferences is a necessary first step in constructing a mathematical algorithm to calculate utility values. A small (sub)set of HRQL question items (also referred to as a preference-based measure) is used to describe different scenarios of health states so that preferences can be investigated in population-based studies where people in the community are interviewed about their choices between various scenarios. Several methods exist for studying these preferences but they all aim to estimate the value of living in a set of circumstances if HRQL is impaired by an illness. A systematic overview of preference elicitation techniques is beyond the scope of this discussion and has been provided by Green, Brazier, and Deverill (2000). Given insights on the choices made by the general public, a preference-based algorithm is derived and thereafter applied across research studies to convert results of HRQL assessments into utility values for use in economic evaluation studies. The original HRQL assessments results are usually Likert-

scale responses (e.g. 1=not at all, 2=a little, 3=quite a bit, 4=a lot) to items that ask about the impact of a health condition on life circumstances (e.g. social relationship). Each level of item response is given by the algorithm a new value, termed as utility weights or tariffs, to convert the original result into preference-based ratings. These are then combined mathematically into a single estimate of utility value to estimate the overall perceived value of living with a scenario of HRQL outcomes (or health state).

Utility values range from 0 to 1, representing increasing levels of desirability from death to full health (i.e. absence of HRQL impairment). Treatment interventions that lead to more desirable outcomes show higher utilities. This preference-based account of quality of life also allows quantity of life to be re-considered in terms of quality-adjusted life years (QALYs) where one QALY represents a year lived in full health. The value of treatment interventions in this sense is compatible with the goal of 'adding life to years' in dementia care. Assessing whether an emerging treatment provides good value for money, in the context of CUA, refers to the potential to incur lower costs for each QALY gained as a result of intervention.

1.2 Distinct concepts, common phenomenon

CUA provides the potential for a unique source of insights on the clinical practice and the policy relevance of treatment benefits in dementia. In light of this emerging influence on decision making, measurement issues in CUA warrant further examination. The research described in this thesis addresses a measurement issue that has received attention both in HRQL and utility

assessment, but often under two different research agendas which diverge in methodological focus.

1.2.1 Valuation shift

In utility assessment, despite the undesirability of being in poor health, utility values given for scenarios of impaired health are often higher than expected. Individuals in poor health commonly value their scenarios of health states less negatively (i.e. they report higher utilities) than people in the general population (Sackett & Torrance, 1978). This has been reported even in individuals hospitalised with a serious illness (Tsevat et al., 1995). Such findings have fuelled a longstanding debate on whether societal benefit of healthcare interventions might be underestimated by patient perspectives or overestimated by the general public (Menzel, Dolan, Richardson, & Olsen, 2002; Ubel, Loewenstein, & Jepson, 2003). The debate is complicated by findings that the gap between patient and public perspectives is not found after a recovery of health in the former group (D. M. Smith, Sherriff, Damschroder, Loewenstein, & Ubel, 2006). This suggests that the same individuals can experience a 'valuation shift' (Dolan, 1996) that alters the perceived value of health state scenarios depending on when it was reported.

1.2.2 Response shift

Similar findings have been documented in HRQL literature. Despite living with severe chronic illness, individuals commonly report that they experience moderate to good HRQL (Kagawa-Singer, 1993; Padilla, Mishel, & Grant, 1992). Studies that compared chronically ill individuals (e.g. cancers, arthritis, diabetes) with the

general population found that the former do not show higher levels of anxiety and depression (Groenvold et al., 1999), or experience poorer HRQL (Breetvelt & Van Dam, 1991; Cassileth et al., 1984), even though they are aware of a decline in their health states (Andrykowski & Hunt, 1993). On the other hand, when healthcare providers and significant others (e.g. family carers) are asked for their proxy perspectives on HRQL of individuals in ill health, proxy reports are often not as positive as self-reports (Sneeuw, Sprangers, & Aaronson, 2002; Sprangers & Aaronson, 1992). As in the literature on utility assessment, the gap is not simply a difference in perspectives. Studies that asked individuals for a second evaluation of their initial HRQL found that their retrospective assessments in post-recovery tend to be more negative than their own initial assessments. The discrepancies resemble the gap between self- and proxy reports at time of initial assessment, indicating a 'response shift' (Sprangers & Schwartz, 1999) in which the criteria they used for judging their own HRQL have altered since the time of original report.

1.2.3 Illness adaptation

While HRQL assessment measures perceptions of health states, utility assessment measures the perceived value of living in a scenario of health states. As distinct concepts, both nonetheless share common influences as evident from reports of valuation shift and response shift.

One of the explanations for a gap between patient and public values is illness adaptation (Edelaar-Peeters et al., 2012). With the experience of illness,

individuals gain a better understanding of what life is like in states of impaired health (Menzel et al., 2002), and experience a change in the relationship between what happens and how one feels (Ubel, Loewenstein, & Jepson, 2005). Without such insights, it is difficult for the general public to anticipate illness adaptation (Kahneman & Snell, 2000; Loewenstein & Frederick, 1997), and they tend to estimate a more negative impact (i.e. lower utilities) than actual patients. This is corroborated by experimental evidence that has showed a valuation shift in individuals during the health state valuation process (Damschroder, Zikmund-Fisher, & Ubel, 2005; McTaggart-Cowan, Tsuchiya, O'Cathain, & Brazier, 2011; Ubel et al., 2005). Though at least one study did not produce a shift in values (Damschroder, Zikmund-Fisher, & Ubel, 2008), the majority to date show that study participants tend to re-estimate a less negative illness impact (i.e. higher utilities) after they have been informed by actual patient accounts or encouraged to reflect on one's ability to adapt.

A central preoccupation in understanding the impact of adaptation is to clarify considerations for and against the use of utility values that have shifted due to illness adaptation (Menzel et al., 2002). In contrast, the study of adaptation in HRQL literature is aimed at clarifying and predicting changes in HRQL as a result of illness or interventions (Menzel et al., 2002). Consequently, while the impact of adaptation is well studied in the utility assessment literature, the underlying processes have received relatively less attention (Edelaar-Peeters et al., 2012; Stiggelbout & de Vogel-Voogt, 2008). The theoretical framework of response shift (Sprangers & Schwartz, 1999), on the other hand, has stimulated

considerable HRQL research in this direction. First developed in other fields of psychology, response shift refers to a 'typology of change' in educational (Howard, Dailey, & Gulanick, 1979; Howard, Ralph, et al., 1979) and organisational change interventions (Golembiewski, Billingsley, & Yeager, 1976) which distinguishes between objective changes (alpha shift), changes in internal standards (beta shift), and reconceptualization (gamma shift). Their potential relevance as processes of adaptation have gained attention in HRQL research (Breetvelt & Van Dam, 1991; Sprangers, 1996) given that this understanding of processes underlying change may in turn inform more effective interventions (Norman & Parker, 1996). Relative to valuation shift, the response shift framework holds a finer explanation of illness adaptation and its impact on HRQL and utility measurement. It is for this reason, in the next section my focus will be on response shift in utility assessment.

1.2.4 Response shift studies in utility assessment

One of the earliest response shift investigations in utility assessment is a Dutch study (Postulart & Adang, 2000) that assessed HRQL in a group of insulin-dependent diabetes mellitus patients with end-stage renal disease. Using a visual analogue scale (VAS), HRQL of 22 patients was assessed before they received combined pancreas-kidney transplants and subsequently at 5, 12, and 18 months later. There was apparent improvement in HRQL at 5-months which was maintained at 12- and 18-month follow up. However, the authors had hypothesised that pre-transplant self-reports reflected HRQL of patients who have

already adapted to their illness. Hence, during each follow up assessment, they also asked patients to rate their pre-transplant HRQL again. Retrospectively, the same patients perceived their pre-transplant HRQL as having actually been worse than they had originally reported at initial assessment. As such, gains in HRQL after transplant would have been larger if compared against these retrospective reports of baseline HRQL. To augment their hypothesis about the impact of adaptation on self-reported HRQL at baseline, the study also used a convenience sample of 55 university students to imagine what it would mean to them to experience a similar scenario of health states and rate their HRQL on VAS. The assumption was that these proxy ratings would not have been influenced by circumstances that demanded illness adaptation. As hypothesised, they gave more negative HRQL evaluations than the patients. However, after a successful transplant, patient retrospective reports of their initial HRQL became as negative as proxy ratings. While the original study aim was to make these comparisons in terms of utility values, issues with statistical distributions in the sample data prevented the authors from using a power function to transform VAS ratings into utility values. Nonetheless, given that VAS was employed as a practical measure of preferences, the authors cautioned that CUA studies could produce different conclusions about the most cost-effective intervention, depending on whether HRQL reports had been influenced by response shift.

A similar investigation of response shift in utility values has been conducted in a Swiss clinical trial for patients with newly diagnosed colon cancer (Bernhard et al., 2001). With a larger sample (n= 122 to 132 patients, depending on specific

analysis), response shift was investigated for two clinical situations, first in a surgery phase and then in a post-discharge (adjuvant) treatment phase in which patients were randomised for chemotherapy or observation. HRQL was assessed using a linear analogue scale once before surgery, then at post-discharge, and approximately two months later. As in the Dutch study, retrospective ratings were obtained to detect response shift. A second evaluation of pre-surgery HRQL was obtained just before discharge and pre-adjuvant HRQL was also re-assessed about two months later. In both situations, patients rated their initial HRQL as worse than they had originally reported. Conventional analysis that compares pre- and post-intervention ratings showed no apparent improvement in HRQL after surgery and adjuvant phase. However, if retrospective ratings were used as the baseline HRQL for surgery and adjuvant phase, gains in HRQL were found. To explore the potential impact on CUA results, the authors used a power function to transform the linear analogue ratings into utility values. They found that utility values from retrospective reports were significantly lower than original estimates for adjuvant phase baseline (0.81 vs 0.88, $p < 0.01$) but not for surgery phase baseline (0.74 vs 0.79, $p > .05$). While this also implies a gain in QALYs at the end of adjuvant phase, longitudinal calculations were not reported. The authors concluded that patients with colon cancer change their internal standards of HRQL substantially and cautioned about the impact of this change on utility assessment.

While both the Dutch and Swiss study are among the earliest response shift investigations that paid explicit attention to utility assessment, their conclusions relied on power functions to calculate utility values rather than actually estimating

utility using preference-based algorithms. This has been attempted only recently. Using the EQ-5D (EuroQol Group, 1990), a widely employed preference-based HRQL measure that focusses on five core domains of general health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), an Australian study assessed HRQL in a convenience sample of 103 older adults who needed inpatient rehabilitation (McPhail & Haines, 2010). The study participants were asked to report their HRQL on the EQ-5D within three days of hospital admission and just before they were discharged (median length of stay = 38 days, interquartile range: 20 – 60). As in previous response shift studies, retrospective reports were obtained. In addition, the authors also tested the extent in which study participants could recall their initial EQ-5D reports. The preference-based algorithm developed by Dolan (1997) was used to assign utility weights to EQ-5D responses. The authors reported a pattern of findings that is similar to that in the Dutch and Swiss study. This sample of older adults experienced an average utility gain of 0.287 (95% CI: 0.216 – 0.359) by the time they were discharged. Taking into account response shift, a larger gain was found (0.441, 95% CI: 0.367 – 0.518), but this was less after adjusting for recall bias (0.303, 95% CI: 0.232 – 0.375).

One of the most recent response shift investigation in utility assessment was conducted in Singapore with 74 osteoarthritis patients undergoing total knee replacement surgery (Zhang et al., 2012). This study employed two widely used preference-based measures, the EQ-5D and SF-6D (Brazier, Usherwood, Harper, & Thomas, 1998), to assess HRQL before surgery and 18 months later. The SF-

6D was also administered at 6-month follow up. Both baseline (i.e. before surgery) and 6-month HRQL were subsequently re-evaluated at 18-month to detect response shift. With the SF-6D (physical functioning, role limitations, social functioning, pain, mental health, and vitality), the study was also able to investigate whether response shift detection had been affected by a slightly different emphasis on general health relative to the EQ-5D (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). A familiar pattern of response shift results was found on both measures. The magnitude of response shift in HRQL self-reports (SF-6D) was significantly larger at baseline (0.14, 95% CI: 0.08 – 0.20) than at 6-month follow up (-0.05, 95% CI: -0.14 – 0.00), consistent with clinical impressions of a large degree of post-operative recovery which subsequently plateaus off. This suggested that recall bias was not a major influence on the study results. Relative to the SF-6D results, the EQ-5D detected substantially larger magnitudes of response shift (0.72, 95% CI: 0.22 – 0.91) at baseline. Bland-Altman plots were used to show that this was the case particularly for patients who experienced larger response shift. Given that utility gains (or losses) are considered clinically significant if there is a difference of 0.04 on SF-6D and 0.07 on EQ-5D (Walters & Brazier, 2005), the impact of response shift on utility assessment in this study was noteworthy. At 18-month, the study participants experienced utility gains of 0.16 (95% CI: 0.02 – 0.26) on the SF-6D and this rose to 0.30 (95% CI: 0.18 – 0.39) after taking into account response shift. On the EQ-5D this was 0.27 (95% CI: 0.02 – 0.62) and 0.95 (95% CI: 0.65 – 1.16) respectively. Assigning a hypothetical cost of US$10,000 for the surgery

and other related expenses, cost-effectiveness was estimated at US$62,500 for each QALY gained as a result of the surgical intervention. When response shift in SF-6D self-reports was taken into account, the cost for each QALY gained was reduced to US$33,333 (i.e. the intervention was more cost-effective by US$29,167). The EQ-5D results showed a similar impact (US$26,511). Given a threshold of US$50,000 per QALY to fund interventions, the authors concluded that response shift could potentially change funding decisions for this surgical intervention.

1.3 Knowledge gaps

The studies highlighted in the previous section demonstrate a need to explore the impact of response shift on utility assessment. Insofar that estimates of treatment benefits may be attenuated or exaggerated, clinical and policy decisions may warrant reconsideration. The study of response shift in HRQL research has been conducted for a number of chronic illness conditions, usually motivated by initial observations of incongruence between objective decline in health and subjective experience of stable HRQL, or discrepancies between proxy- and self-report HRQL. While similar findings have been documented in dementia (Lyketsos et al., 2003; Novella et al., 2001), they are often investigated as consequences of impaired insight (e.g. Ready, Ott, & Grace, 2006; Trigg, Watts, Jones, & Tod, 2011). This may in part explain the relative lack of response shift studies in this population. Studies that have conducted in-depth interviews have shown that, despite significant cognitive deficits, people with dementia do hold meaningful

insights on HRQL (Mozley et al., 1999; S. C. Smith, Murray, et al., 2005) and other issues in life (Lawrence, Samsi, Banerjee, Morgan, & Murray, 2011; MacRae, 2011; Steeman, Tournoy, Grypdonck, Godderis, & De Casterle, 2011). Similar conclusions have been reached in psychometric studies of HRQL measures which show that self-reports from people with mild to moderate dementia do carry coherent themes and are reliable over time (Brod, Stewart, Sands, & Walton, 1999; Hoe, Katona, Roch, & Livingston, 2005; Trigg, Jones, & Skevington, 2007).

Very little is known about how people adapt to the chronic and challenging circumstances living with dementia. An understanding of their adaptation strategies and patterns may help inform intervention objectives and planning. It may also inform treatment evaluation by highlighting non-apparent implications. Based on the emerging HRQL literature, resumption of normative expectations may deflate conventional estimates of treatment effectiveness. Conversely, lowered expectations may exaggerate treatment benefits. This knowledge has the potential to add insights to prevailing criteria for determining value of treatment interventions in dementia.

1.4 Methodological issues

The conduct of response shift investigations in general faces methodological challenges that vary in their impact across studies. Some hold particular relevance for dementia.

1.4.1 Feasibility

Response shift investigations may be described as two broad classes of study methods. The first is a range of study design approaches that involve the use of evaluation exercises (e.g. ranking) to reveal changes in HRQL perceptions or preferences. They vary in empirical foundations as the majority are novel approaches or adapted from other assessment purposes (Schwartz & Sprangers, 1999). On top of the original HRQL assessment task, most involve additional assessments and the increase in respondent burden is considerable for the more complex ones. The feasibility of these methods is limited for the very old or very ill. The second, known as analytic approaches, involves the use of statistical methods to study patterns that emerged from item responses on HRQL assessments. Relative to study design approaches, they generally place heavier demands on sample size requirements. However, they focus on the original HRQL assessment data and do not require additional assessment tasks. In this way, analytic approaches obviate some of the methodological challenges of study design approaches.

1.4.2 Recall bias

The use of retrospective self-report, also referred to as 'then-test' (Schwartz & Sprangers, 1999), is the study design approach used in significant majority of HRQL response shift studies (Schwartz et al., 2006). In the original conception of this approach (Howard, Ralph, et al., 1979), discrepancies between the initial and then-test baseline scores reflect changes in internal standards only if intervention

groups show larger discrepancies than controls, since impetus for change is theoretically absent in the latter. Notably, the inclusion of control groups is rare in HRQL studies (Schwartz & Sprangers, 2010). Interpretational challenges arise especially in studies where there may be difficulties or bias in memory recall.

In a Dutch study of patients receiving a combined pancreas-kidney transplant (Postulart & Adang, 2000), a convenience sample of university students served as a comparison group so that the influence of illness adaptation was absent. Nonetheless, it is worth noting that these proxy ratings were not retrospective reports. A direct comparison with patient assessments might show differences that were due to illness adaptation, as well as inaccurate recall and response bias. Inaccurate recall is plausible since fallible memory has been documented even for intense experiences like pain (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). Response bias on the other hand might arise due to effort justification or social desirability (Conway & Ross, 1984; Ross, 1989). The authors acknowledged that such recall bias in patient retrospective reports had not been ruled out. While the Swiss clinical trial study had a control group among patients with colon cancers (Bernhard et al., 2001), differences between control and treatment arms were not statistically significant. The analysis was based on the overall sample and similar study limitations were acknowledged. In the Australian study of hospitalised older adults (McPhail & Haines, 2010), recall accuracy was explicitly tested. Response shift was nonetheless detected, but the impact was much weaker after adjusting for inaccurate recall. The proposed method of adjustment has however not yet seen wide application in the literature. A control

group was absent from the Singapore study with osteoarthritis patients undergoing knee replacement surgery (Zhang et al., 2012). When asked at the end of the survey, most patients actually thought they provided similar ratings in both initial and then-test reports of their baseline HRQL. Nonetheless, as response shift was detected when patients were at the end of a period of rapid recovery but not after they were clinically stable, the authors concluded that recall bias was minimal.

While the then-test approach is implicitly assumed to be a more valid assessment of utility in these studies, challenges in ruling out recall bias may result in misleading conclusions about outcomes and by extension cost-effectiveness. Furthermore, its application in dementia is challenging given that memory impairment is necessarily part of clinical presentation at time of diagnosis and will predictably increase over time.

1.4.3 Inadequate coverage

The implicit validity of then-test may also be challenged by another study design issue. To guide systematic inquiry on this phenomenon in HRQL research, Sprangers and Schwartz (1999) operationalised response shift as a change in HRQL appraisal due to:

(a) a change of internal standards (**re-calibration**);

(b) a change in perceived value or importance (**re-prioritisation**); or

(c) a change in perceived definition or meaning (**re-conceptualisation**).

This working definition departs from earlier frameworks in which re-prioritisation is considered an inherent part of re-conceptualisation (Sprangers & Schwartz, 1999). Despite the current distinction, it has been acknowledged that these processes are likely to be intertwined (Schwartz & Sprangers, 1999). Notably a significant number of studies, including those reviewed in section 1.2, have employed only the then-test approach in their investigations. In these studies, while then-test findings are reported as changes in internal standards, discrepancies between initial and retrospective assessment may reflect more than just re-calibration. In the event that re-prioritisation or re-conceptualisation may have also taken place in certain HRQL domains, conclusions about higher cost-effectiveness estimates may need to be re-considered. Concurrent investigation of all forms of response shift would therefore be valuable. On the controversy about patient utility values, Menzel et al. (2002) also pointed out that illness adaptation encompasses more than just a change in expectations (i.e. re-calibration of internal standards). Knowledge gaps about other underlying processes have to be filled so that there is clarity in what could be normatively regarded as adapting well whilst living with an illness.

Notwithstanding such implications, a broader scope of investigation also carries substantial clinical significance. As individuals consistently report moderate to good HRQL despite deterioration in health, this may obscure treatment impact in ways that imply an apparent lack of benefit. However, an apparent lack of change in HRQL may hide a shift in expectations. While one's HRQL ratings may be similarly optimistic before and after an intervention, the former may have been

based on a lower set of expectations, which has since risen to higher standards due to significant improvement in HRQL. Similarly, change may not be apparent in HRQL ratings despite a shift in perceived value or importance of HRQL domains, in ways that are adaptive to illness-related circumstances. These arguments may also be made for a change in perceived definition or meaning of HRQL that potentially has enduring implications over the course of a chronic illness. In this light, response shift is not merely a confounding influence to be ruled out so that real changes (alpha shift) can be accurately assessed. This phenomenon deserves study in its own right (Armenakis, 1988), for insights on adaptation processes that may be associated with an intervention. Response shift has also been held as a meaningful intervention target (Golembiewski et al., 1976), in order to maintain change (Norman & Parker, 1996).

1.5 Scope and methodology

The research reported in this thesis investigates the phenomenon of response shift in dementia. Due to interactions between substantive and analytic issues in this investigation, an account of both is provided here.

1.5.1 HRQL measurement system

The scope of the investigation reported here is underpinned by DEMQOL, a HRQL measurement system in dementia that can be employed for both clinical and economic evaluation. The conceptual foundations of this system were based on in-depth interviews with patients and their carers to explore what constitutes HRQL for people with dementia(S. C. Smith, Murray, et al., 2005). This

generated an initial pool of 70 candidate items which subsequently underwent two rounds of field testing with exploratory factor analysis (EFA), leading to a self- and informant-report measure, the DEMQOL (28 items) and DEMQOL-Proxy (31 items) respectively (S. C. Smith, Lamping, et al., 2005). Both are interviewer-administered measures with question items that inquire about the 'feelings', 'memory', and 'everyday life' of the person with dementia in the 'last week'. All items have a four-point Likert scale (a lot / quite a bit / a little / not at all) and the responses are coded so that higher total scores reflect better HRQL. Further econometric development work produced two preference-based algorithms that calculate utility values using five items (DEMQOL-U) from the DEMQOL and four items (DEMQOL-Proxy-U) from the DEMQOL-Proxy respectively (Mulhern et al., 2013). A list of DEMQOL and DEMQOL-Proxy items are presented in Tables 1.1 and 1.2 respectively.

Given their substantive emphasis on HRQL in dementia, utility values generated by DEMQOL-U and DEMQOL-Proxy-U stand to provide insights on top of those based on generic HRQL measures like the EQ-5D (EuroQol Group, 1990). The latter purports a focus on 'core' domains of HRQL which carry general relevance to permit comparisons across all health conditions and treatments. Despite being the dominant paradigm for policy decisions, concerns over the non-specificity of the health descriptive system of generic HRQL measures have persisted (McTaggart-Cowan et al., 2008). Without reference to health states of a specific illness condition, the generic content may have too distal a focus on aspects of health of which neither the patient nor the doctor are necessarily anticipating a

treatment impact (Brazier & Fitzpatrick, 2002). Consequently, generic measures may lack relevance (Guyatt, King, Feeny, Stubbing, & Goldstein, 1999), or may be insensitive to small but important changes in illness-specific health states (Guyatt, 2002; Jenkinson et al., 1997). In this light, the DEMQOL measurement system presents a unique opportunity for exploring the impact of response shift on HRQL reports in clinical and economic evaluation of dementia interventions.

Table 1.1 DEMQOL question items.

	First I'm going to ask about your feelings. In the last week, have you felt…
1	cheerful?
2	worried or anxious?
3	that you are enjoying life?
4	frustrated?
5	confident?
6	full of energy?
7	sad?
8	lonely?
9	distressed?
10	lively?
11	irritable?
12	fed-up?
13	that there are things that you wanted to do but couldn't?
	Next, I'm going to ask you about your memory. In the last week, how worried have you been about…
14	forgetting things that happened recently?
15	forgetting who people are?
16	forgetting what day it is?
17	your thoughts being muddled?
18	difficulty making decisions?
19	poor concentration?
	Now, I'm going to ask you about your everyday life. In the last week, how worried have you been about…
20	not having enough company?
21	how you get on with people close to you?
22	getting the affection that you want?
23	people not listening to you?
24	making yourself understood?
25	getting help when you need it?
26	getting to the toilet in time?
27	how you feel in yourself?
28	your health overall?

Response options (all items): 1 = a lot, 2 = quite a bit, 3 = a little, 4 = not at all
Item 1, 3, 5, 6, 10 reversed coded. Higher overall total score reflect better HRQL.
DEMQOL-U preference-based algorithm for item 1, 4, 8, 14, 24 available from
http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol

Table 1.2 DEMQOL-Proxy question items.

	First I'm going to ask you about (your relative's) feelings. In the last week, would you say that (your relative) has felt…
1	cheerful?
2	worried or anxious?
3	frustrated?
4	full of energy?
5	sad?
6	content?
7	distressed?
8	lively?
9	irritable?
10	fed-up
11	that he/she has things to look forward to?
	Next, I'm going to ask you about (your relative's) memory. In the last week, how worried would you say (your relative) has been about…
12	his/her memory in general?
13	forgetting things that happened a long time ago?
14	forgetting things that happened recently?
15	forgetting people's names?
16	forgetting where he/she is?
17	forgetting what day it is?
18	his/her thoughts being muddled?
19	difficulty making decisions?
20	making him/herself understood?
	Now, I'm going to ask about (your relative's) everyday life. In the last week, how worried would you say (your relative) has been about…
21	keeping him/herself clean (eg washing and bathing)?
22	keeping him/herself looking nice?
23	getting what he/she wants from the shops?
24	using money to pay for things?
25	looking after his/her finances?
26	things taking longer than they used to?
27	getting in touch with people?
28	not having enough company?
29	not being able to help other people?
30	not playing a useful part in things?
31	his/her physical health?

Response options (all items): 1 = a lot, 2 = quite a bit, 3 = a little, 4 = not at all
Item 1, 4, 6, 8, 11 reversed coded. Higher overall total score reflect better HRQL.
DEMQOL-Proxy-U preference-based algorithm for item 3, 8, 17, 22 available from http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol

1.5.2 Longitudinal analysis

In this thesis, the potential impact of response shift on HRQL and utility assessment is explored using latent variable modelling methods. This analytic approach allows for a concurrent detection of re-calibration, re-prioritisation, and re-conceptualisation. As additional retrospective assessments are not required, validity threats from recall bias do not pose a major issue. Alongside any response shift findings, the results also provide an estimate of changes in HRQL that could be attributed to the putative intervention. Response shift (if any) was examined with particular interests on preference-based items as changes in item response behaviour also affect the utility weights assigned for calculating the eventual estimate of utility values.

The use of latent variable modelling to investigate response shift falls under the structural equation modelling (SEM) framework of measurement invariance across multi-wave factor models (Oort, 2005). In each wave of HRQL assessment, a factor analytic model shows HRQL themes that are used to understand patterns in the responses on DEMQOL and DEMQOL-Proxy. While the themes hypothesised for each assessment occasion may be identical, this would not be the case if HRQL has been re-conceptualised. A change of this nature affects item response patterns such that the concept of HRQL shows a change in meaning by exhibiting different themes across different assessment occasions. The lack of longitudinal measurement invariance may also reflect re-calibration or re-

prioritisation depending on which other aspects of the model differ across time (methodological details in Chapter 4).

1.5.3 Bifactor measurement perspective

Hypothesising an appropriate factor analytic model for the observed item response patterns is crucial for all latent variable modelling purposes. The present research builds on the foundations of initial development work (Mulhern et al., 2013) in which exploratory factor analysis (EFA) showed that the DEMQOL measurement system captured individual differences in five domains. From self-report perspectives, the HRQL domain factors in DEMQOL carry the theme of 'positive emotion' (POS), 'negative emotion' (NEG), 'loneliness' (LON), 'worry about social relationship' (SOC), and 'worry about cognition' (COG). From informant perspectives, the HRQL domain factors in DEMQOL-Proxy carry the theme of 'positive emotion' (POS), 'negative emotion' (NEG), 'worry about cognition' (COG), 'worry about financial-related tasks' (FIN), and 'worry about appearance' (APP). An EFA study has also been conducted on Spanish versions of the DEMQOL and DEMQOL-Proxy (Lucas-Carrasco et al., 2010). Similar findings were reported but fewer HRQL themes were found in both self- and proxy-reports. This investigation began with a bifactor model perspective of the themes that can be used to understand the HRQL concept in DEMQOL measurement system. Alongside domain factors suggested by previous EFA studies, a general factor was also hypothesised in the factor analytic models for DEMQOL and DEMQOL-Proxy (see Figure 1.1 for an example).

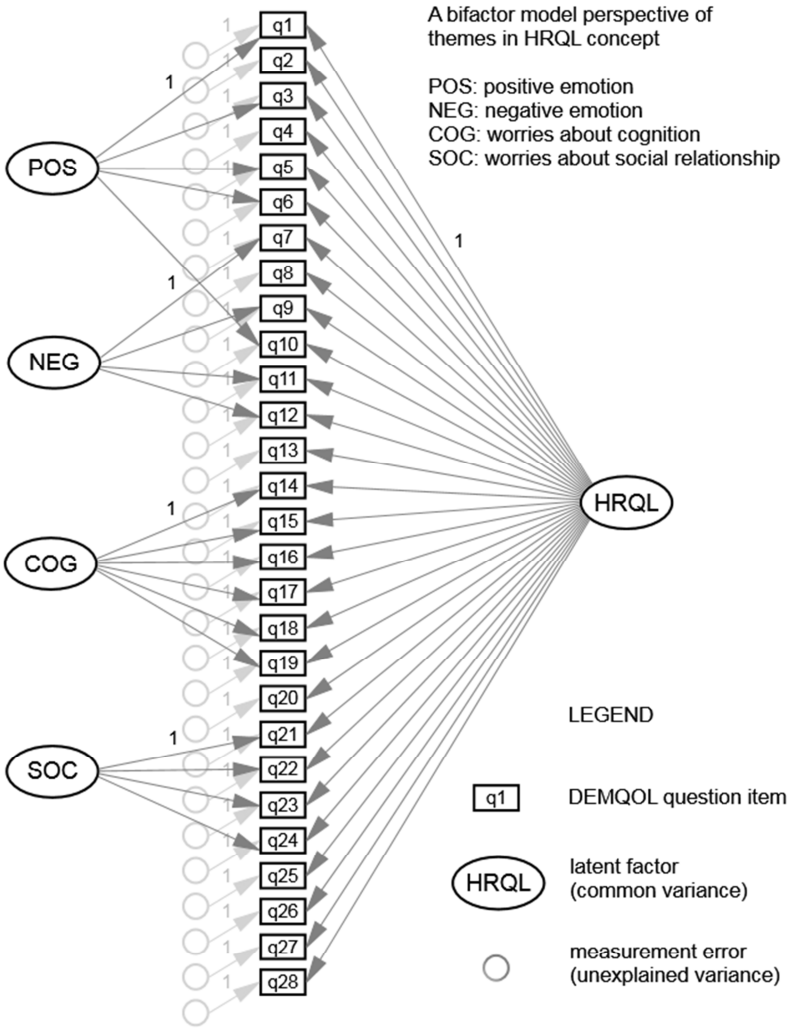

Figure 1-1 An example of DEMQOL bifactor model. This is an 'incomplete' bifactor model as not all items load on both the general and a domain factor. Some load only on the general factor of HRQL.

To highlight its potential significance for understanding HRQL, a brief review of the origins of bifactor model is given. First developed in the field of cognitive psychology, bifactor model framework refers to a perspective that views general intelligence as a construct with broad influence across multiple domains of cognitive abilities (e.g. verbal comprehension, perceptual reasoning, working memory, processing speed). The complex nature of this construct is reflected in the premise that no single domain of cognitive ability gives an adequate assessment of general intelligence. This measurement model has a broad latent factor representing a general theme that explains why test results in multiple cognitive domains share something in common (i.e. general intelligence). On top of this general source of common variance, a bifactor model perspective recognises that test performance on a cognitive ability domain may also give other information that is unrelated to general intelligence (e.g. prior experience or practice). In the measurement model, this additional information is represented as multiple sources of common variance in subsets of cognitive tests which carry a narrower theme. As these latent factors reflect influences that are independent of general intelligence, they are orthogonal to (or uncorrelated with) the broad latent factor of general intelligence. Their substantive nature and hence a meaningful label for narrower themes in the measurement model may be clear only in a wider context of SEM models that include other explanatory variables.

Like general intelligence, HRQL is commonly articulated as a complex phenomenon that can only be understood in terms of multiple domains (or dimensions) of life. For content and hence construct validity, HRQL measures

usually include a broad array of question items so as to achieve exhaustive coverage of the diversity entailed in a complex phenomenon (Reise, Morizot, & Hays, 2007). This content diversity often leads to findings of multidimensionality in factor analytic models (Reise, Moore, & Haviland, 2010). A bifactor model recognises that it is not realistic to expect HRQL measures to have strong content validity and yet be strictly unidimensional. Out of the complexities (or multidimensionality) of HRQL, a coherent overall impression of a general phenomenon can be constructed. With this perspective, the thesis maintains a strategic focus on the main assessment objective by retaining an enduring notion of 'essential unidimensionality' while simultaneously recognising multidimensionality in a complex construct like HRQL (Reise et al., 2010).

Such a model configuration holds a unique set of heuristics for a theoretical and empirical understanding of HRQL. In a bifactor model, the putative broad influence of a general HRQL factor is tested, together with an examination of whether item responses have additional sources of common influence of a narrower theme (i.e. domains) over and above the general influence of HRQL. The plausibility of a complex general phenomenon is supported by the presence of sizable factor loadings on the general factor. Insights on how well each item loads on HRQL are also useful given that a decision between competing results of multidimensionality can be guided by empirical insights on how well individual items measure this target construct which is also the main assessment objective in practical applications. Among predominant factor analytic approaches in the

literature (see Figure 1.2), only bifactor measurement models confer this level of clarity (Chen, West, & Sousa, 2006; Reise, 2012).

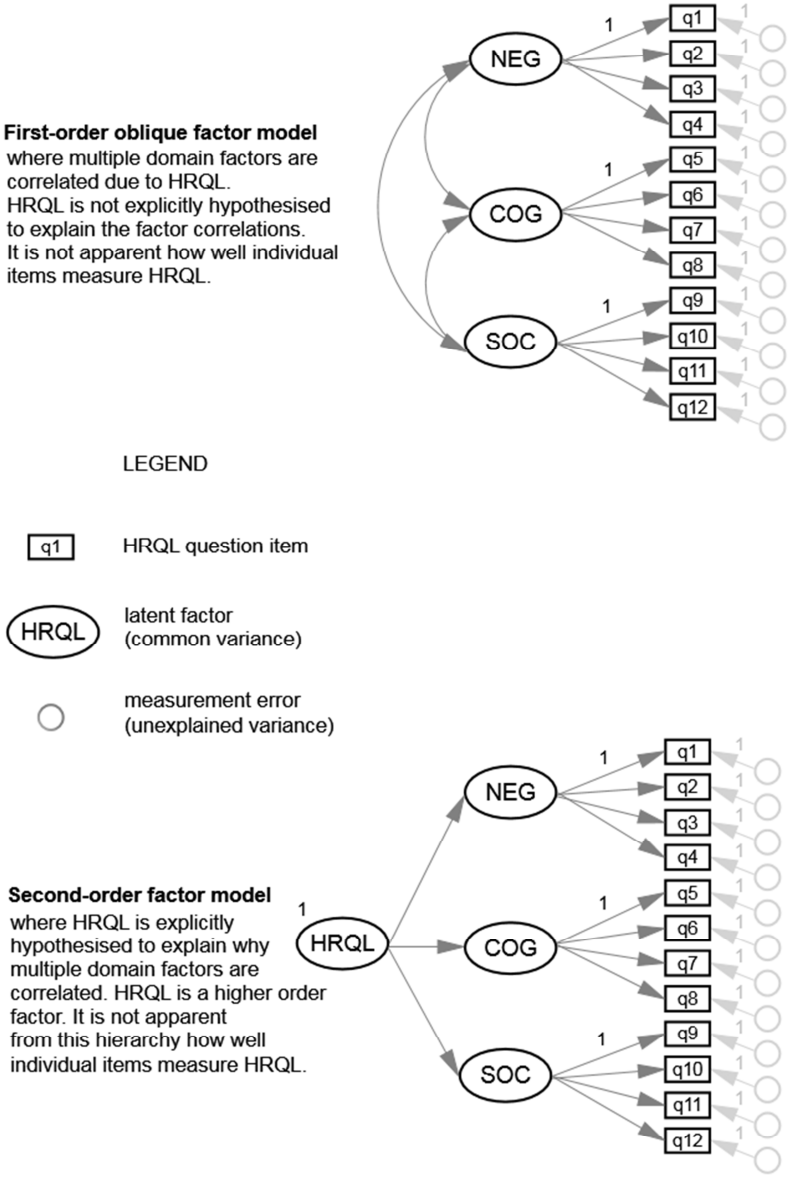

Figure 1-2 Factor analytic approaches that are predominant in the literature.

The plausibility of additional sources of common influence that carry a narrower theme is supported by the presence of sizable factor loadings on the domain factors. This aspect holds potential theoretical importance for the conceptual definition of HRQL. While apparent themes in item responses are logically expected to emerge as domain factors (i.e. sources of common variance), this may not be the case alongside a general factor in bifactor measurement models. An independent HRQL domain may fail to emerge in a bifactor EFA or exhibit model anomalies (e.g. factor variance not statistically significant) in a bifactor confirmatory factor analysis (CFA). Termed as 'factor collapse' in bifactor CFA models, this event is a statistical indication that responses to items of this 'domain' do not share any additional common variance (i.e. 'common theme') that is on top of the theme they have in common due to HRQL. In other words, a hypothesised HRQL domain may actually not exist and the items have non-zero factor loadings only on the general HRQL factor. Substantively, this implies that such narrower themes do not convey insights on another theme of individual differences other than that of the general theme of individual differences in HRQL. This is analogous to cognitive research findings that 'reasoning ability' does not convey additional information (i.e. does not exist as an independent domain alongside a broad factor in a bifactor model) beyond what it conveys about individual differences in general intelligence because performance on this ability test essentially reflects only general intelligence (Gottfredson, 1997; R. E. Snow, Kyllonen, & Marshalek, 1984). Factor collapse as such provides a potential

indication that the 'non-existent' domain lies at the heart of the conceptual definition of HRQL (Brunner, Nagy, & Wilhelm, 2012).

In addition to these insights, an evaluation of 'factor strength' can also be made from the factor loadings. Factor strength, also termed factor saturation, refers to the amount of variance in (total or subscale) scores that could be attributed to the target construct (i.e. general HRQL and its domains). When HRQL is the only source of common influence on item responses (i.e. strictly unidimensional), the same information is conveyed by cronbach's alpha (Reise et al., 2010). When there are multiple sources of common variance (i.e. multidimensional), cronbach's alpha is no longer an appropriate indication of score reliability (Cortina, 1993; Sijtsma, 2009a, 2009b). More specifically, when the target construct is a general factor alongside multiple sources of influence, an examination of factor strength recognises the bifactor hierarchy and provides a valid estimate of reliability (methodological details in Chapter 2). Low factor saturation in a domain factor suggests that variation in the scores of that HRQL subscale has poor reliability. Conversely, high factor saturation in a general factor suggests that variation in overall total scores is mainly due to individual differences in HRQL.

This knowledge can help steer debates about competing results of multidimensionality which imply different ways of calculating item scores in a HRQL assessment measure (e.g. different sets of subscales, or multiple subscales vs an overall total). It has been argued that subscale scores should be calculated because HRQL by definition is a multidimensional concept and respective domain

scores can help clarify treatment impact (Ettema, Droes, de Lange, Mellenbergh, & Ribbe, 2005; Perales et al., 2013). Furthermore, unless HRQL is a unidimensional construct, scaling individual differences with a HRQL total score can lead to inaccurate estimates (Reise, Bonifay, & Haviland, 2013). The use of total scores however is important for many practical applications of HRQL assessment (e.g. in randomised trials) where the goal is to capture the overall balance of the impacts of diverse domains (Kifley et al., 2012), especially in treatment interventions that target broad outcomes (Ebesutani, Reise, et al., 2012), so that treatment benefits are not overlooked and potential harms are not missed (Banerjee et al., 2006). This investigation of factor strength aims to clarify the feasibility of calculating subscale and/or overall total scores for DEMQOL and DEMQOL-Proxy. Of particular interest, overall total scores may still afford reliable estimates of general HRQL despite its inherent multidimensionality (Gustafsson & Aberg-Bengtsson, 2010; Reise et al., 2010). This insight also informs whether an unexpected result (positive or null findings) could exclude methodological concerns like low reliability (i.e. poor precision) in the assessment of individual differences (Brunner et al., 2012).

1.5.4 Cross-validation

Results that show a meaningful factor structure (or measurement model) only reveal HRQL themes that may be relevant in a specific study sample. If these themes carry strong validity as reflections of a general HRQL construct, the same measurement model should also emerge in different samples and across time.

This thesis therefore contains a first stage establishing whether there is a plausible bifactor CFA model for the DEMQOL and DEMQOL-Proxy, the second stage cross-validates these measurement models with an independent sample from another geographical region (Latin America), so as to provide firm empirical foundations for a final stage of response shift investigations. Besides hypothesising the same measurement model for the independent sample, cross-validation can employ the SEM framework of measurement invariance between groups (UK vs Latin America), so that a direct comparison can be made and evaluated statistically. Given a complex interplay of socioeconomic disparities between these two geographical regions, responses on HRQL assessments may differ in the absence of genuine differences. If such disparities influence responses for a DEMQOL or DEMQOL-Proxy item, the item is said to display differential item functioning (DIF). To detect DIF effects a type of SEM model that is known as a multiple indicators, multiple causes (MIMIC) model can be used. In a MIMIC model, a bifactor measurement model for HRQL is hypothesised alongside multiple causes to explain group differences in general HRQL and its domain factors. This flexibility in MIMIC models makes it possible to extend the DIF investigations beyond the focus on geographical disparities. Measurement invariance can be investigated to see if the same HRQL conceptual models could be used for both gender and across stages of dementia severity. Gender disparities are commonly found in HRQL reports (Fryback et al., 2007; Hanmer, Lawrence, Anderson, Kaplan, & Fryback, 2006), even after taking into account age, ethnicity, marital status, education, and income (Cherepanov, Palta, Fryback, &

Robert, 2010). It is plausible that a DIF investigation might surface similar disparities in DEMQOL and DEMQOL-Proxy responses. Disparities might also emerge due to dementia severity despite the absence of genuine differences in HRQL. Such disparities might obscure treatment impact in clinical trial studies where a central concern often lie in whether treatment interventions are effective only for people with mild dementia or that it also works for people with more advanced illness.

1.5.5 Short-form HRQL assessments

This cross-validation stage also has the potential to exploit a well-demonstrated correspondence between the family of factor analytic models in SEM and another family of statistical methodology known as item response theory (IRT) models (Kamata & Bauer, 2008; Reise, Widaman, & Pugh, 1993; Takane & Deleeuw, 1987). While DIF detection with MIMIC models essentially provides statistical adjustment for confounding influences such that the CFA measurement models attain a more sensitive discrimination of individual differences in HRQL, this analytic approach is not feasible for practical applications in clinical settings. Consequently, this thesis describes the development of shortened versions of the DEMQOL and DEMQOL-Proxy in which items that displayed DIF effects were considered for omission, thereby obviating the need for statistical adjustment in applied settings. To this end, IRT model results were derived from the CFA models to further study item response patterns in terms of the amount of information each item provided for discriminating individual differences and the

level of HRQL at which they were the most informative. Based on this knowledge, a smaller set of items was selected for short form versions of DEMQOL and DEMQOL-Proxy (DEMQOL-SF and DEMQOL-Proxy-SF), retaining similar levels and range of sensitivity as their parent versions. The development of short-form versions potentially serves a wider purpose of enhancing feasibility of HRQL assessments in dementia across clinical and social care settings, as well as that of repeated assessments in longitudinal studies. In the final part of the thesis bifactor CFA models using DEMQOL-SF and DEMQOL-Proxy-SF items are used to investigate response shift.

1.6 Thesis structure

Chapter 1 introduces background issues, highlights key concepts, and outlines the scope of the research.

Chapter 2 reports on the first stage of empirical investigation in which the goal was to establish an appropriate measurement model for investigating response shift in HRQL. Alongside factors that were suggested by previous exploratory factor analysis (EFA) studies, a general HRQL factor was also hypothesised in confirmatory factor analysis (CFA) models for DEMQOL and DEMQOL-Proxy, giving rise to a bifactor model. The research question was whether a coherent overall impression of general HRQL can emerge out of the complexity of multiple HRQL themes. Following recommended practice, bifactor EFAs were first conducted to surface potential modelling problems that might arise in CFAs. The final bifactor CFA models were examined for insights to the research question.

Secondary insights were obtained on whether particular HRQL domains hold core relevance to the general HRQL concept. Reliability of total and subscale scores was also examined in the bifactor models to inform scoring practices.

Chapter 3 reports on the second stage of empirical investigations in which the goal was to cross-validate the bifactor CFA models of DEMQOL and DEMQOL-Proxy in an independent sample, so that their empirical foundations are firm for subsequent purposes. The research question was whether DEMQOL and DEMQOL-Proxy permit identical interpretations about HRQL in other groups that differ by geographical region, gender, and dementia severity. The structural equation modelling (SEM) framework for measurement invariance between groups provided a statistical basis for making direct comparisons between pertinent groups. All group differences were investigated simultaneously using a type of SEM model that is known as MIMIC models. In this context, items that display DIF effects undermine measurement invariance. Item response theory models were used to add insights on the amount of information each item provided for discriminating individual differences and the level of HRQL at which they were the most informative. Taken together, this knowledge was used to develop short-form versions of DEMQOL and DEMQOL-Proxy. They formed the basis for the final stage of investigations.

Chapter 4 reports on the final stage in which response shift was investigated in longitudinal assessments of HRQL using DEMQOL-SF and DEMQOL-Proxy-SF items. The SEM framework of measurement invariance across multi-wave factor

models provided a statistical basis for comparing the bifactor CFA models across time points. Differences were examined for concurrent indications of re-calibration, re-prioritisation, and re-conceptualisation of HRQL at follow up assessment occasions. Response shift (if any) was examined with particular focus on preference-based items as changes in item response behaviour also affect the utility weights assigned for calculating the eventual estimate of utility values.

Chapter 5 summarises the research findings and discusses insights derived in the wider context of health assessment in economic evaluations.

CHAPTER 2 MEASUREMENT MODEL

This chapter reports on the first stage of empirical investigation in which the primary goal was to establish an appropriate measurement model for investigating response shift in HRQL. A secondary objective was to obtain insights for informing scoring practices when DEMQOL and DEMQOL-Proxy are employed for HRQL assessments. The research questions are:

(a) Do item responses on DEMQOL and DEMQOL-Proxy show a general theme of individual differences that supersedes the inherent complexities of multiple themes in a HRQL concept?

(b) Are overall total scores and/or multiple subscale scores from DEMQOL and DEMQOL-Proxy sensitive to individual differences in HRQL?

2.1 Bifactor measurement model

Bifactor model framework originated in the field of cognitive psychology for inquiry into whether a coherent overall impression of general intelligence can be constructed out of the complexities from multiple cognitive ability domains. This perspective retains a focus on the main assessment objective while recognising its inherent multidimensionality. The substantive emphasis on a complex general phenomenon is consistent with many assessment objectives with broad target constructs like depression (Brouwer, Meijer, & Zevalkink, 2013; Norton, Cosco, Doyle, Done, & Sacker, 2013), burnout (Meszaros, Adam, Szabo, Szigeti, & Urban, 2014), and quality of life (Chen et al., 2006).

In this thesis, HRQL is hypothesised as a general factor representing common variance across all items in DEMQOL and DEMQOL-Proxy respectively. This allows for a direct inquiry on how item responses are influenced by a complex general phenomenon. Besides the general influence of HRQL, responses on some items are more closely related to one another than they are with the rest of the item pool. This content similarity gives rise to narrower sources of common variance represented by domain factors. Given that the target construct is hypothesised as a complex general phenomenon, only the broad general factor provides a theoretical reflection of individual differences in HRQL. The domain factors reflect narrower sources of individual differences that are unrelated to HRQL in a bifactor model perspective. Furthermore, as there is no theoretical reason to expect domain factors to have logical relations other than because of HRQL, they no longer share another common source of influence once the general influence of HRQL is accounted for. As such, all latent factors (or sources of common variance) are orthogonal to one another in a canonical bifactor model.

Sizable factor loadings on the general factor would support the hypothesis that HRQL is a complex general phenomenon. Sizable factor loadings on domain factors suggest the presence of additional sources of common influence that carry narrower themes which are unrelated (i.e. orthogonal) to HRQL. As these themes are often labelled as substantive domains of the target construct, potential confusion arises as to how these domains are held to be part of HRQL and yet are unrelated to HRQL. With a bifactor model perspective, the reason why smaller groups of items share a narrow theme is unrelated to the reason why they also

share a general theme with items from other domains. Given the prospect of orthogonality, it is possible that an independent HRQL domain may not emerge in a bifactor EFA and the items have non-zero factor loadings only on the general HRQL factor. In a bifactor CFA, hypothesising a 'non-existent' HRQL domain may lead to model identification issues that surface as anomalies like: (i) domain factor variance that is not statistically significant; and/or (ii) domain factor loadings that are weak or not statistically significant. The CFA model may also simply fail to converge due to factor over-extraction (i.e. hypothesising more latent factors than there really are). Termed as 'factor collapse', these model anomalies suggest the domain factor in question should not be hypothesised. Individual differences in responses on items of a 'non-existent' domain convey information that essentially reflects only individual differences in HRQL. Substantively, this also suggests that such a domain lies at the heart of HRQL's conceptual definition.

A systematic evaluation of factor strength (also termed as factor saturation) can also be made with factor loadings on the general HRQL and domain factors. The extent to which an overall total and multiple subscale scores are sensitive to individual differences depends on factor saturation levels. In a bifactor model framework, this is determined from the omegaH coefficient (McDonald, 1999; Zinbarg, 2006; Zinbarg, Revelle, Yovel, & Li, 2005) which shows the percentage of variance in summed scores (overall total / subscale) that can be attributed to their target construct (general HRQL / HRQL domain). A high omegaH value for the general factor indicates that variation in overall total scores is mainly due to

HRQL. High levels of factor strength as such assure measurement reliability (or precision) for discriminating individual differences in HRQL. By modifying the mathematics of omegaH, described in detail by Brunner et al. (2012) and Reise et al. (2013), the reliability of subscale scores can also be determined. These insights inform on the feasibility of calculating an overall total and/or multiple subscale scores for HRQL assessments with DEMQOL and DEMQOL-Proxy.

2.2 Methods

2.2.1 Participants

The sample comprised community-dwelling participants, and their carers, who were referred to the Croydon Memory Service a service for early assessment and intervention in dementia based in South London. Study participants include those referred between December 2002 and June 2010 who, after a full multidisciplinary assessment, were given a formal clinical diagnosis of dementia using ICD-10 criteria (Banerjee et al., 2007). This sample therefore represents assessments of HRQL made at the time of diagnosis. No ethical committee approval was needed as this study was a secondary analysis on de-identified archival data.

2.2.2 Measures

The DEMQOL (28 items) and DEMQOL-Proxy (31 items) are interviewer-administered measures for obtaining self- and informant-reports of HRQL in people with dementia (S. C. Smith et al., 2007). The question items on both

measures inquire about the 'feelings', 'memory', and 'everyday life' of the person with dementia in the 'last week'. All items have a four-point likert scale (a lot / quite a bit / a little / not at all) and the responses are coded so that higher total scores reflect better HRQL. The full content of the measures and scoring instructions are available at http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol/.

While the primary focus was on the HRQL data, other assessment data were employed for conducting multiple imputation of DEMQOL and DEMQOL-Proxy data. These assessments included an evaluation of cognitive functioning as assessed by the Mini-Mental State Examination (MMSE, Folstein, Folstein, & McHugh, 1975), depression as assessed by a shortened version of the Geriatric Depression Scale (GDS, Yesavage & Sheikh, 1986), behavioural and psychological symptoms in dementia as assessed by the Neuropsychiatric Inventory (NPI, Cummings et al., 1994), and problems with daily life activities as assessed by the Bristol Activities of Daily Living Scale (BADL, Bucks, Ashworth, Wilcock, & Siegfried, 1996). Carers evaluations of their carer burden, using the Zarit Burden Interview (Zarit, Zarit, Reever, & Bach-Peterson, 1980), and of their general health, on the General Health Questionnaire (GHQ, Goldberg & Williams, 1988), were also included.

2.3 Analysis

2.3.1 Multiple imputation

Relative to those with complete or partial HRQL data, participants and their carers tended to have poorer health if this archival data were missing. This trend was more apparent for the NPI, BADL, and Zarit, but group differences were generally small. On this basis, we assumed that the data were missing at random (MAR) and conducted multiple imputation with auxiliary variables (Collins, Schafer, & Kam, 2001) to gain precision in the imputation. Specifically, missing DEMQOL and DEMQOL-Proxy data were imputed as ordered-categorical values using Bayesian estimation of the unrestricted variance covariance model (termed 'H1 model') as implemented in Mplus version 7 (Asparouhov & Muthén, 2010a). As an 'inclusive' strategy is recommended (Collins et al., 2001; Yoo, 2009), the MMSE, NPI, GDS, BADL, Zarit, and GHQ were employed as auxiliary variables for imputing the HRQL data. A total of 100 data sets were generated for DEMQOL and DEMQOL-Proxy respectively (Mplus syntax in Appendices p. 233-234).

2.3.2 Bifactor EFA

The investigations began with bifactor EFAs (Jennrich & Bentler, 2011) on imputed data sets of DEMQOL and DEMQOL-Proxy. This was conducted under exploratory structural equation modelling (ESEM) framework (Asparouhov & Muthén, 2009) in Mplus version 7. An orthogonal bifactor Geomin rotation (L. Muthén & Muthén, 1998-2012, pp. 103-104) was implemented so that all latent factors were orthogonal and items were free to load on a general factor as well as

any domain factors (Mplus syntax in Appendices p. 235-236). Alongside a general HRQL factor, five domain factors were first hypothesised for DEMQOL and DEMQOL-Proxy respectively. Bifactor EFA models with fewer domain factors were also estimated to see if they offer improved interpretability.

The bifactor EFA results were obtained primarily to surface potential modelling problems (e.g. sizable cross-loadings) instead of only looking for them (e.g. via modification indices) after imposing even more stringent assumptions in confirmatory bifactor models (Reise, 2012). This stage of analysis also provided early insights on whether previously reported themes (i.e. HRQL domains) could be replicated with bifactor model perspectives of multidimensionality. Of particular interest, the absence of a previously reported DEMQOL or DEMQOL-Proxy domain from a bifactor EFA model might signal the prospect of factor collapse if this domain was hypothesised in bifactor CFA context.

2.3.3 Bifactor CFA and model comparisons

Bifactor CFAs were conducted to address the main research question of whether a general theme of HRQL would emerge out of the diversity of multiple themes in DEMQOL and DEMQOL-Proxy. This substantive focus included an interest in the prospect of factor collapse of a HRQL domain as this event would shed light on what lies at the heart of HRQL concept in DEMQOL measurement system. Given that direct evidence of factor collapse might not emerge in bifactor CFA models, indirect evidence might be found by comparing tenability of models with and without factor collapse.

Compared to its original model, a model with factor collapse would have fewer latent factors and hence fewer parameters that were freely estimated from the data. In this way, the latter is said to be nested in the original model. A decline in model fit was expected in the nested models since they afforded a less complex explanation of the data (or more degrees of freedom) than the original model. Model comparisons were made to determine whether this decline in fit might be statistically significant. Given inconsequential differences, a less complex explanation (i.e. model with factor collapse) would be preferred. In Mplus version 7, model comparison cannot be implemented with imputed data sets. Hence, this stage of bifactor CFA was based on the original non-imputed data (Mplus syntax in Appendices p. 237-240).

2.3.4 Factor strength

Given tenable bifactor CFA models, the final stage of analysis examined factor strength (omegaH or ω_h) to inform on the feasibility of using an overall total and subscale scores in practical applications. For HRQL total scores, variance attributable to the general factor (VAR_g) can be obtained by first adding up all standardised factor loadings on the target construct HRQL, then squaring this total sum. The same calculation was made for each domain factor (VAR_{d1}, VAR_{d2}, and so on). Having accounted for these sources of explained variance, the unique variance of each item was obtained by subtracting their communalities (i.e. explained variance in an item, h^2) from the value of one. For a bifactor model

with three group factors, factor saturation due to the HRQL construct was hence obtained as follows:

$$\omega_h = VAR_g / [VAR_g + VAR_{d1} + VAR_{d2} + VAR_{d3} + \Sigma(1 - h^2)]$$

The mathematics can be extended to examine reliability of subscale scores by treating each domain factor (e.g. d1) as the target construct in the numerator:

$$\omega_h = VAR_{d1} / [VAR_g + VAR_{d1} + \Sigma(1 - h^2)]$$

As the focus was only on the subset of items represented by the domain factor, the denominator terms require slight modification. Given the equation above, only items of domain d1 were involved in the calculation of VAR_g and $\Sigma(1 - h^2)$. Didactic accounts are available in Brunner et al. (2012, p. 821 and 825) and Reise et al. (2013, p. 6).

2.3.5 Model estimation

Since there are four categories on the Likert response scale, it would be appropriate to treat DEMQOL and DEMQOL-Proxy item responses as ordered-categorical data (Rhemtulla, Brosseau-Liard, & Savalei, 2012). All modelling analysis were hence based on polychoric correlations rather than Pearson's correlations (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2008), and model parameters were estimated using robust weighted least squares with means and variances adjustment (WLSMV) as is recommended (Flora & Curran, 2004; B. Muthén, du Toit, & Spisic, 1997; Savalei & Rhemtulla, 2013). This approach uses a multivariate probit regression model to predict how

probabilities of (ordered-categorical) item responses are related to latent variables that represent (continuous) levels of HRQL and its domains. For models estimated with WLSMV, the DIFFTEST option in Mplus was required for model comparisons so as to obtain the correct chi square difference test between models (L. Muthén & Muthén, 1998-2012, pp. 451-452).

2.3.6 Model evaluation

As a fundamental basis for making interpretations, empirical fit between model predictions and the observed data must be adequate. Overall (i.e. omnibus) model fit was evaluated statistically using robust model Chi square (χ^2_{M}). An exact fit between model predictions and sample data, within bounds of sampling error, would result in a non-statistically significant χ^2_{M} value. In the absence of exact fit, the extent of approximate fit remains of interest. For this evaluation, Mplus provides four descriptive indices that offer a non-statistical summary of model fit for CFAs on ordered-categorical data. Based on commonly adopted standards, root mean square error of approximation (RMSEA, Steiger, 1990) values should be low (<0.10 for acceptable fit, <0.05 for very good fit), while comparative fit index (CFI, Bentler, 1990), Tucker Lewis index (TLI, Tucker & Lewis, 1973) values should be high (>0.90 for acceptable fit, >0.95 for very good fit) when approximate fit is adequate. These standards were drawn from extensive simulation studies with continuous data and their relevance for ordered-categorical data remains an area of active inquiry (Cook, Kallen, & Amtmann, 2009; Marsh, 2004; Marsh, Ludtke, Nagengast, Morin, & Von Davier, 2013;

West, Finch, & Curran, 1995). The weighted root mean square residual (WRMR, Yu, 2002) has been developed for ordered-categorical data and while a value of less than one has been recommended, it remains to be established for a wider range of simulations (e.g. bifactor model).

2.4 Results

2.4.1 Sample characteristics

A total of 1240 study participants were included in the analyses. Table 2.1 presents the demographic and clinical characteristics of all study participants with details for groups with complete, partial, or missing DEMQOL and DEMQOL-Proxy data.

Table 2.1 Demographic and clinical characteristics of study participants with complete/partial/missing HRQL assessment

		DEMQOL			DEMQOL-Proxy		
	Overall	Complete	Partial	Missing	Complete	Partial	Missing
Participants	1240	756	112	372	679	230	331
Age *	79.0 (8.4) n=1236	78.7 (8.5) n=753	77.9 (8.2) n=112	80.1 (8.2) n=371	78.8 (8.1) n=675	79.3 (9.0) n=230	79.3 (8.5) n=331
Gender							
Male	457	269	44	144	253	87	117
Female	783	487	68	228	426	143	214
Ethnicity							
White	1042	657	86	299	580	191	271
Black	90	43	13	34	38	21	31
Asian	83	41	11	31	48	12	23
Unknown	25	15	2	8	13	6	6
ICD-10 **							
AD	694	425	54	215	369	119	206
AD mixed	316	192	33	91	175	67	74
Vascular	147	84	15	48	84	30	33
Others	37	15	5	17	22	5	10
Unknown	46	40	5	1	29	9	8
MMSE *	20.4 (5.4) n=1239	21.1 (5.1) n=756	20.4 (5.3) n=112	18.9 (5.8) n=371	20.8 (5.3) n=679	19.5 (5.6) n=230	20.1 (5.5) n=330
GDS *	3.1 (2.7) n=1113	3.0 (2.6) n=692	3.1 (2.6) n=105	3.1 (3.1) n=316	2.9 (2.7) n=619	3.2 (2.6) n=198	3.2 (2.9) n=296
NPI *	13.4 (13.5) n=1113	12.7 (13.0) n=684	12.3 (14.1) n=102	15.3 (14.0) n=327	12.1 (12.1) n=668	15.2 (16.9) n=221	15.6 (13.0) n=224
BADL *	10.5 (9.6) n=1124	9.5 (9.2) n=691	10.3 (9.4) n=105	12.8 (10.2) n=328	9.5 (9.2) n=671	11.6 (9.9) n=225	12.4 (10.4) n=228
Zarit *	24.8 (17.3) n=914	23.3 (16.6) n=566	26.2 (19.2) n=88	27.6 (17.5) n=260	24.4 (17.0) n=565	24.6 (17.5) n=192	26.6 (17.9) n=157
GHQ *	4.5 (5.7) n=895	4.1 (5.4) n=548	4.8 (6.5) n=85	5.0 (6.0) n=262	4.3 (5.5) n=552	4.5 (6.1) n=186	4.9 (5.7) n=157

* Sample average with standard deviation in parentheses. As rate of missing data varies across variables, valid sample size (n) is reported.

** ICD-10 diagnosis: Alzheimer's Disease, late/early onset (AD), Alzheimer's Disease, mixed type (AD mixed), Vascular dementia (Vascular), Others / Unspecified (Others), ICD code not known (Unknown).

2.4.2 Bifactor EFA

Table 2.2 and 2.3 display the bifactor EFA results for the DEMQOL and DEMQOL-Proxy respectively. Most items loaded well (≥ 0.3) on the general factor, providing support for the putative broad scope of influence in HRQL as a complex phenomenon. With a general HRQL factor, both DEMQOL and DEMQOL-Proxy had 'incomplete' bifactor models (Chen et al., 2006) in which not all items loaded on both the general and a domain factor. Some items loaded only on the general factor.

Five HRQL themes had been reported for DEMQOL in previous validation work (Mulhern et al., 2013). In this study, the most interpretable bifactor model for DEMQOL provided preliminary support for four of these domains: positive emotion (POS), negative emotion (NEG), worries about cognition (COG), and loneliness (LON). Items that were previously reported for the domain of 'worries about social relationship' (SOC) did not emerge as a theme (i.e. did not share another source of common variance) after accounting for what they have in common due to the general theme of HRQL.

Table 2.2 DEMQOL (28 items) bifactor EFA model standardised factor loadings

Item	Question	GEN	DOM 1	DOM 2	DOM 3	DOM 4
10*	lively	.27	**.77**			
6*	full of energy	.36	**.72**			
3*	that you are enjoying life	.42	**.59**			
5*	confident	.41	**.51**			
1*	cheerful	.51	**.45**			
4	frustrated	.56		**.62**		
12	fed-up	.66	.22	**.39**		
11	irritable	.59		**.38**		
13	things that you wanted to do but couldn't	.46		**.29**		
17	your thoughts being muddled	.65			**.55**	
14	forgetting things that happened recently	.60			**.48**	
16	forgetting what day it is	.52			**.43**	
19	poor concentration	.67			**.40**	
15	forgetting who people are	.57	-.22		**.37**	
18	difficulty making decisions	.73			**.30**	
9	distressed	.72			.23	
8	lonely	.54				**.70**
20	not having enough company	.55				**.67**
7	sad	.63	.21			.21
2	worried or anxious	.65				
21	how you get on with people close to you	.72	-.26		-.21	
22	getting the affection that you want	.74	-.35		-.34	
23	people not listening to you	.71	-.31			
24	making yourself understood	.65	-.24			
25	getting help when you need it	.74	-.23			
26	getting to the toilet in time	.56				
27	how you feel in yourself	.78		-.23		-.20
28	your health overall	.65				-.29

All displayed factor loadings are statistically significant over 100 replications. For domain factors (DOM), only loadings of magnitude ≥ 0.2 are displayed. Loadings are in bold to clarify item assignment for each domain factor. Provisional labels for GEN: general HRQL, DOM1: positive emotion, DOM2: negative emotion, DOM3: worries about cognition, DOM4: loneliness

* For higher total HRQL score to reflect better HRQL, item 1, 3, 5, 6, 10 were reverse-scored (see: http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol/)

Sample size in imputed data, n = 1240. Model fit: χ^2_M = 1187.680 (df = 248, standard deviation over 100 replications, SD_{100} = 52.574), RMSEA = .055 (SD_{100} = .002), CFI = .953 (SD_{100} = .003), TLI = .929 (SD_{100} = .005), WRMR = 1.092 (SD_{100} = .030).

Table 2.3 DEMQOL-Proxy (31 items) bifactor EFA model standardised factor loadings

Item	Content	GEN	DOM 1	DOM 2	DOM 3	DOM 4	DOM 5
5	sad	.51	**.61**				
7	distressed	.57	**.56**				
10	fed-up	.55	**.54**				
2	worried or anxious	.50	**.53**				
9	irritable	.34	**.51**				
3	frustrated	.51	**.47**				
4*	full of energy	.18		**.86**			
8*	lively	.19		**.84**			
11*	that he/she has things to look forward to	.19	.22	**.52**			
1*	cheerful	.26	.42	**.51**			
6*	content	.30	.48	**.42**			
21	keeping him/herself clean	.58			**.71**		
22	keeping him/herself looking nice	.58			**.62**		
24	using money to pay for things	.58				**.70**	
25	looking after his/her finances	.56				**.59**	
23	getting what he/she wants from the shops	.60				**.44**	
29	not being able to help other people	.50					**.72**
30	not playing a useful part in things	.53					**.58**
27	getting in touch with people	.63				.24	**.36**
28	not having enough company	.56					**.28**
26	things taking longer than they used to	.63					.21
31	his/her physical health	.44					.21
15	forgetting people's names	.68					
13	forget things that happened a long time ago	.58					
19	difficulty making decisions	.78					
14	forgetting things that happened recently	.80		-.21	-.35		
12	his/her memory in general	.66		-.21	-.37		
16	forgetting where he/she is	.60					
17	forgetting what day it is	.74					
18	his/her thoughts being muddled	.82					
20	making him/herself understood	.70					

All displayed factor loadings are statistically significant over 100 replications. For domain factors (DOM), only loadings of magnitude ≥ 0.2 are displayed. Loadings are in bold to clarify item assignment for each domain factor. Provisional labels for GEN: general HRQL, DOM1: negative emotion, DOM2: positive emotion, DOM3: worries about appearance, DOM4: worries about financial-related tasks, DOM5: worries about social relationship

* For higher total HRQL score to reflect better HRQL, item 1, 4, 6, 8, 11 were reverse-scored (see: http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol/)

Sample size in imputed data, n = 1240. Model fit: χ^2_M = 1174.214 (df = 294, standard deviation over 100 replications, SD_{100} = 45.48), RMSEA = .049 (SD_{100} =.001), CFI = .965 (SD_{100} =.002), TLI = .944 (SD_{100} =.003), WRMR = .967 (SD_{100} =.022).

Five HRQL themes had also been reported for DEMQOL-Proxy in previous validation work (Mulhern et al., 2013). In this study, there was preliminary support for four of these domains: 'positive emotion' (POS), 'negative emotion' (NEG), 'worries about appearance' (APP), 'worries about financial-related tasks' (FIN). Items that were previously reported for the domain of 'worries about cognition' (COG) did not emerge as a theme after accounting for the general theme of HRQL. Instead, a theme provisionally labelled as 'worries about social relationship' (SOC) was found.

These early results suggested that the perspective of what was at the heart of HRQL concept differed between respondents and informants. From self-report perspectives, responses to questions on 'worries about social relationship' in DEMQOL conveyed only information about individual differences in HRQL. They had no other source of influence to provide further insights on individual differences. From informant perspectives, responses to questions on 'worries about cognition' in DEMQOL-Proxy essentially reflected individual differences in HRQL and had no other source of influence. While 'worries about social relationship' (SOC) held core relevance in self-report HRQL, 'worries about cognition' (COG) held core relevance in informant-rated HRQL.

Alongside a general theme of HRQL, the additional themes in DEMQOL (POS, NEG, COG, and LON) and DEMQOL-Proxy (NEG, POS, APP, FIN, and SOC) item responses suggested that individual differences unrelated to HRQL also had an influence on the HRQL assessment results. Within the context of the

measurement model, clarifying their substantive significance was challenging without the aid of other explanatory variables. However, since correlations reflected by domain factors were not relevant to the construct of HRQL, it was plausible that the two-item domains in DEMQOL (LON) and DEMQOL-Proxy (APP) reflected methods effects.

DEMQOL item 8 (*lonely*) and item 20 (*not having enough company*) in the LON domain had an association that was considerably stronger than all others in the correlation matrix. Besides a high level of redundancy in the information they convey (r = 0.8), the gap between this and the remaining inter-item correlations might be an artefact of their highly similar content that was unique in the pool of 28 items. This 'excess' similarity could have been 'inflated' by bloated specifics content (Cattell, 1996) and so reflected additional information that was not relevant to the HRQL construct (i.e. common variance that could not be attributed to the general factor). Similar observations were noted with the association between DEMQOL-Proxy item 21 (*keeping him/herself clean*) and 22 (*keeping him/herself looking nice*) in terms of their strength (r = 0.8) and relative magnitude in the correlation matrix. They share similarity in phrasing as well as proximity in item sequence (i.e. order effects). As with bloated specifics content, such 'inflated' similarity (i.e. additional common variance represented by APP) had no theoretical relevance to individual differences in HRQL.

2.4.3 Bifactor EFA with testlets

A decision was made at this juncture to employ testlets for DEMQOL item 8 and 20, as well as DEMQOL-Proxy item 21 and 22. Specifically, scores of these item-pairs were added up so that they were treated as a single item. As the 'excess' correlations in these item-pairs were not of theoretical interest, the use of testlets offered a practical strategy for removing their idiosyncratic impact on the modelling analysis. While this could also be achieved by omitting an item in each pair from the analysis, there would be less information loss with testlets. As a result of aggregation, testlets also have higher reliability than each individual item (Bandalos & Finney, 2001). Given interests at subsequent stage to study factor saturation for insights on scoring practices, it is worth noting that forming testlets by simple addition does not alter the basis of calculating a total score (Steinberg, Sharp, Stanford, & Tharp, 2013).

The bifactor EFA models were re-estimated for DEMQOL with 26 items and one testlet (item 8 and 20), and for DEMQOL-Proxy with 29 items and one testlet (item 21 and 22). In this series of bifactor EFAs, three to five domain factors were hypothesised alongside a general HRQL factor for DEMQOL and DEMQOL-Proxy. The objective was to screen for item-pairs that exhibit similar forms of 'local dependencies' (LD) in which their elevated correlations might be attributed to individual differences in HRQL and other independent causes that were not of theoretical interests (Steinberg et al., 2013).

While the fit with sample data was generally good after a bi-Geomin rotation, a few item-pairs consistently exhibited an anomalous impact on bifactor EFA results. In DEMQOL, they were:

- item 6 (*full of energy*) and 10 (*lively*),
- item 21 (*how you get on with people close to you*) and 22 (*getting the affection you want*),
- item 27 (*how you feel in yourself*) and 28 (*your health overall*).

In DEMQOL-Proxy, they were:

- item 4 (*full of energy*) and 8 (*lively*),
- item 12 (*his/her memory in general*) and 14 (*forgetting things that happened recently*),
- item 24 (*using money to pay for things*) and 25 (*looking after his/her finances*),
- item 29 (*not being able to help other people*) and 30 (*not playing a useful part in things*).

These item-pairs tended to undermine model interpretability either by (a) being embedded in a domain factor as an aberrant pair of negative loadings; and/or (b) emerging as a domain factor with only two items that had strong loadings; and/or (c) loading weakly on the general HRQL factor. In the matrices of pairwise item correlations of DEMQOL and DEMQOL-Proxy, these item-pairs displayed elevated correlations relative to other inter-item correlations that were of the same

putative domains (enclosed in Appendices p. 241-242). On screening the item content, their atypical correlations might have been due to bloated specifics content, wording and/or order effects that inflate similarity between two items. As these LD item-pairs reflected influences that have no theoretical relevance for individual differences in HRQL, testlets were employed to obviate their idiosyncratic influence on study results.

These modelling decisions reduced the number of indicators (items or testlets) that shared the theme of 'loneliness' (LON) in DEMQOL, and 'worries about appearance' (APP) and 'worries about finance-related tasks' (FIN) in DEMQOL-Proxy. The themes were hence omitted from bifactor CFA models as their domain content would not be well-represented by only one or two indicators. Their indicators load only on the general factor, without additional loadings on an independent domain factor. In other words, only the role they played in discriminating individual differences in HRQL were of interest at the bifactor CFA stage.

2.4.4 Bifactor CFA with testlets

Given the bifactor EFA insights, 'incomplete' bifactor CFA models were hypothesised for DEMQOL with 24 items and four LD testlets and for DEMQOL-Proxy with 26 items and five LD testlets. Having introduced testlets, the theme of loneliness (LON) in DEMQOL could no longer be hypothesised with a single testlet item. A domain factor for 'worries about social relationship' (SOC) was hypothesised instead. Items for this domain were selected based on previous

validation work (Mulhern et al., 2013). The theme of SOC did not emerge in the bifactor EFA. It is hence of interest whether this domain factor would show signs of factor collapse in the bifactor CFA. The bifactor CFA model hypothesised that responses on DEMQOL (Figure 2.1) were influenced by individual differences in general HRQL and four independent domains provisionally labelled as 'positive emotion' (POS), 'negative emotion' (NEG), 'worries about cognition' (COG), and 'worries about social relationship' (SOC).

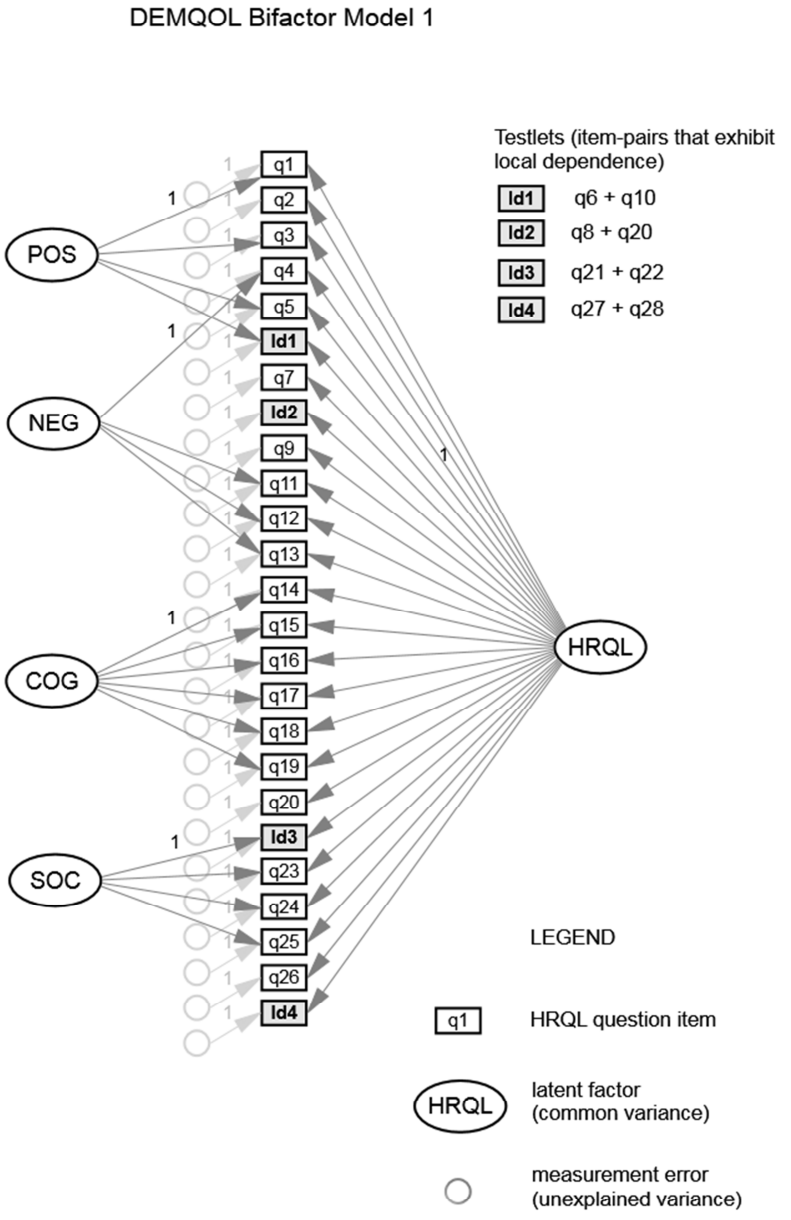

Figure 2-1 DEMQOL 'incomplete' bifactor model (24 items and 4 testlets)

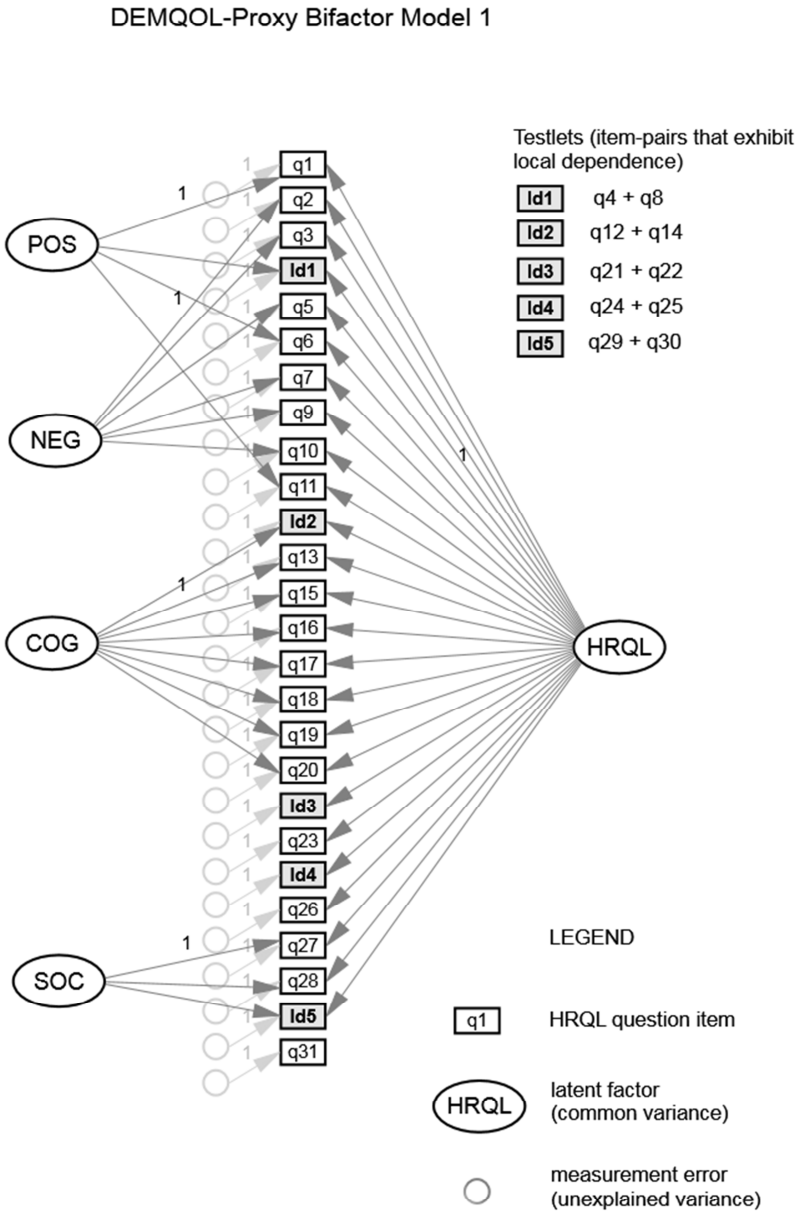

Figure 2-2 DEMQOL-Proxy 'incomplete' bifactor model (21 items and 5 testlets)

The introduction of LD testlets also led to a change in the themes hypothesised for DEMQOL-Proxy. Instead of 'worries about appearance' (APP) and 'worries about financial-related tasks' (FIN), a domain factor for 'worries about cognition' (COG) was hypothesised. Items for this domain were selected based on previous validation work (Mulhern et al., 2013). The theme of COG did not emerge in the bifactor EFA. It is hence of interest whether this domain factor would show signs of factor collapse in the bifactor CFA. The bifactor CFA model hypothesised that responses on DEMQOL-Proxy (Figure 2.2) were influenced by individual differences in general HRQL and four independent domains provisionally labelled as 'positive emotion' (POS), 'negative emotion' (NEG), 'worries about social relationship' (SOC), and 'worries about cognition' (COG).

The 'incomplete' bifactor models were specified: (a) 18 (DEMQOL) / 11 (DEMQOL-Proxy) items have a non-zero loading on the general factor and a domain factor, and zero loadings on the other domain factors; (b) six (DEMQOL) / five (DEMQOL-Proxy) items loaded on the general factor only (c) all latent (general / domain) factors are uncorrelated with one another; (c) measurement errors (or residual variance) of each item were uncorrelated with one another. To identify the model, one of the factor loadings on the general factor and one on each domain factor were fixed at 1, and factor variances were freely estimated from the sample data. This configuration, labelled as Model 1 for DEMQOL and DEMQOL-Proxy, formed the basis for direct evidence of factor collapse.

Table 2.4 Model fit evaluation

DEMQOL (n=868)	df	χ^2_M	RMSEA (90%CI)	CFI	TLI	WRMR
Model 1 GEN HRQL, POS, NEG, COG, SOC	234	1041.049	.063 (.059 - .067)	.927	.914	1.544
Model 2 GEN HRQL, POS, NEG, COG	238	1167.718	.067 (.063 - .071)	.916	.903	1.665
Model 3 GEN HRQL, POS, NEG, SOC	240	1371.269	.074 (.070 - .078)	.898	.883	1.829
DEMQOL-Proxy (n=909)	df	χ^2_M	RMSEA (90%CI)	CFI	TLI	WRMR
Model 1 GEN HRQL, NEG, POS, SOC, COG	278	1410.655	.067 (.064 - .070)	.921	.907	1.680
Model 2 GEN HRQL, NEG, POS, COG	281	1463.001	.068 (.065 - .071)	.917	.904	1.720
Model 3 GEN HRQL, NEG, POS, SOC	286	1808.373	.077 (.073 - .080)	.893	.879	1.961

Based on commonly adopted standards, RMSEA values should be low (<0.10 for acceptable fit, <0.05 for very good fit), while CFI and TLI values should be high (>0.90 for acceptable fit, >0.95 for very good fit) when approximate fit is adequate.

Table 2.5 DEMQOL bifactor CFA Model 1 standardised factor loadings for 20 items and 4 testlets

Item	Question	HRQL	POS	NEG	COG	SOC	ξ
LD4*	[testlet]: item 6 + 10 (r = .68)	.31	.69				.43
3*	that you are enjoying life	.36	.68				.41
1*	cheerful	.46	.56				.48
5*	confident	.40	.54				.54
4	frustrated	.57		.76			.10
12	fed-up	.69		.32			.42
13	things that you wanted to do but couldn't	.46		.29			.70
11	irritable	.60		.28			.56
7	sad	.70					.51
2	worried or anxious	.70					.51
9	distressed	.80					.36
17	your thoughts being muddled	.63			.59		.26
14	forgetting things that happened recently	.58			.51		.40
16	forgetting what day it is	.49			.48		.53
15	forgetting who people are	.52			.46		.51
19	poor concentration	.64			.45		.39
18	difficulty making decisions	.72			.31		.39
23	people not listening to you	.60				.70	.15
24	making yourself understood	.61				.39	.48
LD1	[testlet]: item 21 + 22 (r = .75)	.63				.37	.46
25	getting help when you need it	.69				.37	.40
26	getting to the toilet in time	.54					.71
LD2	[testlet]: item 8 + 20 (r = .70)	.56					.69
LD3	[testlet]: item 27 + 28 (r = .70)	.69					.52
	omegaH	.86	.59	.28	.34	.30	

*Item 1, 3, 5, 6, 10 reverse-scored, so that higher total scores reflect better HRQL.
r: polychoric correlation between testlet items; ξ: item uniqueness = 1 – communality

Table 2.6 DEMQOL-Proxy bifactor CFA Model 1 standardised factor loadings for 21 items and 5 testlets

Item	Question	HRQL	NEG	POS	SOC	COG	ξ
9	irritable	.41	.49				.59
7	distressed	.64	.48				.35
5	sad	.62	.46				.40
3	frustrated	.56	.44				.49
2	worried or anxious	.58	.43				.49
10	fed-up	.66	.43				.37
LD3*	[testlet]: item 4 + 8 (r = .77)	.23		.67			.50
1*	cheerful	.40		.62			.46
11*	that he/she has things to look forward to	.26		.58			.60
6*	content	.48		.54			.48
LD5	[testlet]: item 29 + 30 (r = .71)	.57			.45		.47
27	getting in touch with people	.65			.45		.37
28	not having enough company	.62			.32		.52
15	forgetting people's names	.46				.60	.43
LD2	[testlet]: item 12 + 14 (r = .77)	.57				.52	.41
17	forgetting what day it is	.58				.49	.42
18	his/her thoughts being muddled	.70				.47	.29
13	forget things that happened a long time ago	.45				.43	.61
19	difficulty making decisions	.68				.42	.37
20	making him/herself understood	.58				.42	.49
16	forgetting where he/she is	.55				.33	.59
LD1	[testlet]: item 21 + 22 (r = .82)	.56					.68
23	getting what he/she wants from the shops	.67					.56
LD4	[testlet]: item 24 + 25 (r = .76)	.66					.57
26	things taking longer than they used to	.66					.56
31	his/her physical health	.50					.75
	omegaH	.82	.34	.60	.24	.36	

*Item 1, 4, 6, 8, 11 reverse-scored, so that higher total scores reflect better HRQL.
r: polychoric correlation between testlet items; ξ: item uniqueness = 1 – communality

Model 1 for DEMQOL and DEMQOL-Proxy did not predict a pattern of inter-item correlations that had an exact match with those found in the sample within bounds of sampling error (Table 2.4). In terms of approximate fit, RMSEA values suggested that the average discrepancy between actual and predicted covariances per degree of freedom was in an acceptable range. Both CFI and TLI values also indicated that they offered an acceptable amount of improvement in empirical fit when compared to the 'worst model' (Miles & Shevlin, 2007) in which none of the elements is correlated. However, WRMR values indicated that the average discrepancy between the observed and predicted correlation matrices was not at acceptable levels. Taken together, Model 1 for DEMQOL and DEMQOL-Proxy held satisfactory tenability for making substantive interpretations.

Most items loaded well on the general factor of DEMQOL (Table 2.5) and DEMQOL-Proxy (Table 2.6), a necessary condition for reliable assessment of individual differences in HRQL (Reise et al., 2010). Furthermore, among items that also loaded on a HRQL domain, their domain factor loadings tended to be weaker than their general factor loadings. This suggests that responses on these items conveyed more information about a broad phenomenon than about the theme of their narrower domain. Items that carried the theme of 'positive emotion' (POS) presented a notable exception. As these were also the only items that required reverse-scoring, there was considerable 'excess' similarity (i.e. additional common variance represented by POS factor) that could not be attributed to the general HRQL factor. A growing body of research has recommended against the use of reverse-scored items and/or employed other types of factor models to

address the undue influence of such method effects on item responses (Brown, 2003; Carlson et al., 2011; Ebesutani, Drescher, et al., 2012; Lindwall et al., 2012; Marsh, 1986, 1996; Tomás, Oliver, Galiana, Sancho, & Lila, 2013; van Sonderen, Sanderman, & Coyne, 2013). With a bifactor model perspective, this result provided preliminary evidence that 'positive emotion' constituted an integral part of HRQL (i.e. substantial loadings on general HRQL factor) despite the potential presence of method effects which would not be relevant to individual differences in HRQL.

Model 1 for DEMQOL and DEMQOL-Proxy did not exhibit anomalies in factor variances or loadings. Direct evidence of factor collapse was absent. However the plausibility of factor collapse remained to be ruled out. For this purpose, two alternative bifactor models were examined in turn for DEMQOL and DEMQOL-Proxy respectively. Model 2 and 3 were hypothesised with only three domain factors alongside a general HRQL factor. Compared to the original model which had four domain factors, these alternative models were nested in Model 1. The specifications for Model 2 were identical to that for Model 1 but Model 2 did not have a domain factor for 'worries about social relationship' (SOC). This implied the hypothesis of factor collapse for SOC. The specifications for Model 3 were identical to that for Model 1 but Model 3 did not have a domain factor for 'worries about cognition' (COG). This implied the hypothesis of factor collapse for COG. Factor collapse for 'positive emotion' (POS) or 'negative emotion' (NEG) was not investigated because these domain factors consistently emerged at the bifactor EFA stage.

Model 2 for DEMQOL and DEMQOL-Proxy did not predict a pattern of inter-item correlations that had an exact match with those found in the sample within bounds of sampling error (Table 2.4). In terms of approximate fit, RMSEA, CFI and TLI values were favourable. The same evaluation did not support the tenability of Model 3 for DEMQOL and DEMQOL-Proxy. It is hence plausible that 'worries about social relationship' (SOC), but not 'worries about cognition' (COG), was at the heart of HRQL concept in DEMQOL measurement system.

The factor loadings in Model 2 were not presented as they were very similar to those in Model 1. With fewer domain factors, Model 2 afforded a less complex explanation of the data (or more degrees of freedom) than Model 1 and showed poorer exact fit with the sample data (i.e. larger χ^2_M values in Table 2.4). Model comparisons were made and the DIFFTEST results showed that Model 2 had statistically significant poorer model fit than Model 1 for DEMQOL ($\Delta\chi^2$ = 107.05, Δdf = 4, p < .001) and DEMQOL-Proxy ($\Delta\chi^2$ = 59.88, Δdf = 3, p < .001). These results did not favour the hypothesis of factor collapse of SOC.

2.4.5 Factor strength

Based on Model 1 (bottom row of Table 2.5 and 2.6), the general HRQL factor was clearly a dominant influence on variation in overall total scores in the DEMQOL (omegaH = 0.9) and DEMQOL-Proxy (omegaH = 0.8). In contrast, only 24 – 36% of the variation in subscale scores could be attributed to their domain factors after the general influence of HRQL has been accounted for. As before, the POS domain was an exception. While this domain had more factor

strength, its omegaH estimate of 0.6 (for both DEMQOL and DEMQOL-Proxy) indicated that this subscale score would not afford adequate reliability for making interpretations about individual differences.

2.5 Discussion

This study provides evidence to support the notion that HRQL is a general phenomenon for which a coherent overall impression of diverse life circumstances can be formed in dementia. Alongside a general HRQL factor, there were four other independent sources of influences on item responses for DEMQOL and DEMQOL-Proxy. The themes of these influences were provisionally labelled as 'positive emotion' (POS), 'negative emotion' (NEG), 'worries about cognition' (COG), 'worries about social relationship' (SOC). A SEM model (i.e. CFA measurement model with other explanatory variables) is required to clarify the substantive significance of these themes. As they are unrelated to general HRQL in a bifactor model perspective, the 'incremental prediction' (Ozer & Benet-Martinez, 2006) offered by each of these domain factors may not be in the expected direction in relation to other explanatory variables (e.g. Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Y. Yang et al., 2013).

Within the context of a bifactor measurement model, a theme like 'positive emotions' (POS) may reflect the influence of methods effects since POS items are the only ones that required reverse coding. Reporting whether one had more 'positive emotions' might be cognitively more demanding than reporting whether one had less 'worries'. A similar instance of such influences had been reported in

young children (Marsh, 1986). While such influences on item responses are not theoretically relevant to individual differences in HRQL, 'positive emotion' (POS) items generally had an integral contribution to the assessment of overall HRQL as evidenced by their sizable loadings (≥0.3) on the general factor.

In retaining the concept of an 'essentially unidimensional' target construct, bifactor EFA results raise the prospect that 'worries about social relationships' may hold core influence on how people with dementia evaluate their own HRQL. From informant perspectives, 'worries about cognitive functioning' may hold core influence on how they evaluate the HRQL of people with dementia. This resonates with the conclusions of other studies which reported that self- and informant-report HRQL are influenced by different things (Black et al., 2012; Moyle, Murfield, Griffiths, & Venturato, 2012; Novella et al., 2001; Vogel, Mortensen, Hasselbalch, Andersen, & Waldemar, 2006).

When these insights are re-examined in bifactor CFA context (Model 1), direct evidence of factor collapse is absent. Neither the domain factor in DEMQOL for 'worries about social relationships' (SOC), nor the domain factor in DEMQOL-Proxy for 'worries about cognitive functioning' (COG) exhibits model anomalies that suggest that they should not be hypothesised. When this is further examined in bifactor CFA models that implied factor collapse of SOC (Model 2) or COG domain factor (Model 3), model fit evaluation suggests that it is tenable that the 'worries about social relationship' (SOC), but not 'worries about cognition' (COG), is at the heart of HRQL concept in DEMQOL and DEMQOL-Proxy.

Despite affording a less complex explanation of the data, Model 2 (i.e. factor collapse of SOC) exhibits only slightly poorer model fit than Model 1 (i.e. no factor collapse). When compared statistically, DIFFTEST results favour Model 1 over its nested alternative for both DEMQOL and DEMQOL-Proxy. Nonetheless, rejecting a less complex explanation (Model 2) solely on statistical grounds is premature against a background of empirical literature demonstrating that social functioning plays a pivotal role in the illness experience (Frick, Irving, & Rehm, 2012; Hughes, Flatt, Fu, Chang, & Ganguli, 2013; Lou, Chi, Kwan, & Leung, 2013; MacRae, 2011) as well as healthy aging in general (Coyle & Dugan, 2012; Huxhold, Fiori, & Windsor, 2013; Ichida et al., 2013; Rook, Luong, Sorkin, Newsom, & Krause, 2012). Between better empirical fit (Model 1) and greater theoretical parsimony (Model 2), this study has equivocal results for a decision.

In item responses from self-report (DEMQOL) and informant-report (DEMQOL-Proxy), 'worries about social relationship' may be at the heart of individual differences in HRQL (Model 2). The potential concordance between self-report and informant perspectives is noteworthy in light of the body of literature that suggested otherwise. A possible explanation for the current findings may be that this study sample comprised people with dementia in the early stages of their life with a diagnosis of dementia. HRQL perceptions may change as the illness develops and as the person with dementia and their family carer cope and adapt to the daily life circumstances. While concordance may be affected by the progressive nature of this long term condition, there may be greatest agreement early in the illness. An important caveat at this juncture is that the provisional

label of ‘worries about social relationships’ (SOC) refers to a domain whose items differ between DEMQOL and DEMQOL-Proxy even though they focus on similar concerns around social relationships. The extent to which they therefore constitute ‘similar conceptions’ as implied by the same labels is open to debate.

While the primary goal of this stage was to identify plausible themes for making substantive interpretations with DEMQOL and DEMQOL-Proxy, a secondary objective was to obtain empirical insights for informing scoring practices. The study results show that the broad factor of general HRQL has sufficient factor strength for DEMQOL and DEMQOL-Proxy total scores to be sensitive to individual differences in HRQL. At least 80% of variation in overall total scores from DEMQOL and DEMQOL-Proxy is due to the general construct of HRQL. This is of importance as many practical applications of HRQL assessment (e.g. in RCTs) aim to capture the overall balance of the impacts of diverse domains (Kifley et al., 2012). The use of total scores is also concordant with treatment goals that target broad outcomes (Ebesutani, Reise, et al., 2012), so that treatment benefits are not overlooked and potential harms are not missed (Banerjee et al., 2006).

As multidimensionality has been reported for the DEMQOL and DEMQOL-Proxy, subscale scores may offer insights on how interventions influence HRQL. The use of subscale scores however poses interpretive challenges (Brouwer et al., 2013). Since HRQL domains are essentially different themes of the same target phenomenon, subscale scores are logically related and this multicollinearity

makes it problematic to analyse them as if they were independent themes with distinct implications for policy and clinical decisions (Brouwer et al., 2013; Brunner et al., 2012; Chen et al., 2012). With a bifactor model perspective, this study showed that if a subscale score is calculated among POS items, only 61-64% of the variance can be attributed to this theme of individual differences. This is even lower for the other subscales. In other words, subscale scores do not have adequate measurement reliability for discriminating individual differences in the respective HRQL domains.

2.6 Limitations

A number of study limitations must be noted alongside our conclusions. Firstly, we observed that the HRQL data were available from those who had less impairment from neuropsychiatric symptoms and were more independent in daily life activities. While this sample bias is not severe, the bifactor CFA findings may be less relevant for those with more impairment. Nonetheless, this limitation does not pose a major concern for interpretations about HRQL as a general phenomenon since the same conclusion was reached with data sets that were imputed with auxiliary variables to mitigate the sample bias.

Another study limitation is the use of testlets in the CFAs but not in the EFAs. The use of testlets, on its own, is controversial, with well-grounded arguments on both sides (Little, Rhemtulla, Gibson, & Schoemann, 2013; Marsh et al., 2013). While a discussion of the issues is beyond the scope of this paper, ‘distributive parceling strategies’ may be more problematic than with ‘homogenous parceling

strategies'. The testlets in this study are more consistent with the principles of homogenous parcels. Also, forming testlets by simple addition does not alter the basis of calculating a total score (Steinberg et al., 2013). Nonetheless, the present study is not immune to the criticisms on parcelling items. In employing testlets to capture the 'quirks' of item responses that add no meaningful theoretical information, the parcels may 'camouflage' model misfit which remained hidden as a potential source of bias in broader investigations that examine associations between HRQL and other variables. In light of this threat, sensitivity analyses would be useful to compare findings from a HRQL model with and without testlets (Little et al., 2013). The conclusions we draw must be treated as preliminary.

The present findings provide a starting point for further work to continue with bifactor model investigations of the DEMQOL and DEMQOL-Proxy, as well as other HRQL measures in dementia. As demonstrated by Ebesutani et al. (2011), bifactor models are well-suited for investigating 'heterotypic continuity' (Holmbeck, Devine, & Bruno, 2010) in which the same underlying phenomenon may be expressed differently at different stages of development. In the context of dementia, themes that hold a core influence in HRQL evaluations may differ between self- and informant-report, community and residential home samples, as well as stages of illness and diagnosis. Factor collapse in bifactor models may hence illuminate what lies at the heart of HRQL in people with dementia at different times of need.

CHAPTER 3 CROSS-VALIDATION

This chapter reports on the second stage of empirical investigation in which the goal was to determine whether the DEMQOL and DEMQOL-Proxy bifactor models that were reported in Chapter 2 can be replicated in an independent sample. The cross-validation process also provided a suitable basis for identifying a smaller set of items so that shorter HRQL assessments can be employed with similar levels and range of sensitivity as full-length versions of DEMQOL and DEMQOL-Proxy.

The research questions are:

(a) Do DEMQOL and DEMQOL-Proxy permit identical interpretations about HRQL across groups that differ by geographical region, gender, and dementia severity?

(b) Which items in DEMQOL and DEMQOL-Proxy should be used so that shorter HRQL assessments can be made and yet retain a similar level and range of sensitivity in full-length versions?

3.1 Measurement invariance between groups

Measurement invariance studies originated in the field of educational psychology in which the aim is to identify test items that may give an unfair advantage or disadvantage in certain groups of students such that test results do not reflect their true abilities. In the context of health psychology, ‘abilities’ refer to attributes like physical functioning, depression, or general HRQL. As the assessment of many

such attributes often include self-report, the evaluation process is likely to be influenced by factors like age, gender, or education. When these influences are found in the absence of genuine differences, assessment items are said to display differential item functioning (DIF). Depending on the strength of DIF effects, assessment scores may have lower validity in certain groups and findings of group differences may be questionable. As a result, clinical and policy decisions may be ineffective or suboptimal. For instance, Gallo, Rabins, Lyketsos, Tien, and Anthony (1997) have reported that older adults with clinical depression were often not treated because they did not display dysphoria and anhedonia, both of which are symptoms required by DSM-IV criteria for a diagnosis of major depression.

An examination of group differences in a DIF investigation differs from conventional group comparisons in that comparisons are matched so that group differences emerge only if there is measurement bias rather than genuine differences (Teresi & Fleishman, 2007). Among potential causes of DIF in health assessment, gender is one of the most ubiquitous. Gender differences in mental health have been documented in epidemiological studies and population surveys in several countries (Drapeau et al., 2010). Females generally report higher levels of psychological distress (Cockerham, Hinote, & Abbott, 2006; Drapeau et al., 2010) and depression (Inaba et al., 2005), even in late life (Djernes, 2006). These are central components in the concept of HRQL, and similar gender disparities are also apparent when HRQL assessments were conducted in population health studies using an array of commonly employed measures, the SF12, SF6D, EQ5D,

and HUI (Fryback et al., 2007; Hanmer et al., 2006). Investigations of plausible causes (e.g. age, ethnicity, marital status, education, and income) have provided support that these gender differences are genuine (Drapeau et al., 2010; Matud, 2004; Mirowsky, 1996). Despite this evidence, DIF cannot be excluded as a potential source of influence that heightens or masks true differences. In general, the content and wording of items that assess psychological functioning may be more consistent with the perception, expression, and interpretation of emotions among female respondents (Drapeau et al., 2010). Consequently, cultural norms may result in a tendency among male respondents to understate their emotional experiences, giving rise to higher scores among female respondents even though they actually may not differ in their levels of distress/depression. In a more recent population-based study that reported lower scores in female respondents on five HRQL measures, gender disparity was reduced but not removed after accounting for age, ethnicity, marital status, education, and income (Cherepanov et al., 2010). It is hence plausible that DIF due to gender is a potential source of measurement bias.

Two other potential sources of DIF warrant particular attention when assessing HRQL in dementia. First DIF due to dementia severity is a potential validity threat to a wide range of clinical research in dementia. A common concern in randomised trials is whether an intervention is effective only for people with mild dementia or that it also works for people with more advanced illness. While a treatment or intervention may confer similar benefits for people with mild and moderate dementia, HRQL may be evaluated differently in each group,

confounding overall estimates of treatment efficacy derived from group comparisons or even masking the real treatment impact within a group. We currently do not know whether people with mild dementia or more advanced illness evaluate HRQL in a similar way. Differences may arise due to impaired insight and/or distinct sets of values and expectations at progressive stages of illness. Studies have reported that accuracy of self-reported depression is associated with impaired insight, but not stage of illness (Horning, Melrose, & Sultzer, 2014; A. L. Snow et al., 2005). While loss of insight is linked to dementia severity, studies have demonstrated that significant cognitive deficits did not hinder meaningful self-report on HRQL (Mozley et al., 1999; Trigg et al., 2007; Trigg et al., 2011).

Second, cross-national differences may also give rise to different sets of values and expectations among people with dementia. As treatment innovations for dementia are still emerging, they will be tested globally for their effectiveness in different settings and populations. HRQL measures are likely to be employed in countries that differ importantly in culture, language, and health care systems. The complex interplay of these contextual factors may affect the way HRQL is perceived and reported by people from different countries. Given a potentially effective treatment, HRQL data may show inconsistent evidence if it is affected by such cross-national differences on top of the true treatment impact, particularly in non-randomised trials. As two-thirds of the world population of people with dementia reside in low and middle income countries (World Health Organization & Alzheimer's Disease International, 2012), this is a potentially important source

of variation and error to explore when using HRQL measures that have been developed mostly in high income countries.

Very few data exist with which to form a priori hypotheses about DIF effects in HRQL assessment for populations of people with dementia (e.g. Revell, Caskie, Willis, & Schaie, 2009). This chapter focuses on HRQL assessments using DEMQOL and DEMQOL-Proxy in the UK and Latin America. The primary objective was to detect any potential DIF due to gender, illness severity, and geographical region in samples of community-dwelling older adults with dementia. While the presence of DIF effects implies bias in item responses, an assessment of individual differences in HRQL is not problematic given that the detection process actually makes statistical adjustments so that estimates from the measurement model are no longer confounded. In other words, model estimates of individual differences in HRQL are based on all item responses regardless of extent or severity of DIF effects (Teresi & Fleishman, 2007). This allows for an assessment of the impact of DIF effects on HRQL assessment with and without an active investigation of such confounding influences. For instance, Fleishman and Lawrence (2003) showed that differences in mental health between African-Americans and European-Americans were rendered non-significant after DIF effects have been accounted for.

This analytic option is however not feasible in routine clinical practice and removing DIF items from HRQL measures may be considered in conjunction with reviews by knowledge experts. To this end, our secondary objective is to derive

short-form versions of DEMQOL and DEMQOL-Proxy in which only items that offer optimal discrimination of individual differences without displaying potential DIF are retained. With these short-form measures, HRQL assessment would also have greater feasibility across diverse clinical and social care settings, and for the very old or very ill.

3.2 Methods

3.2.1 Participants

The HRQL data for this study came from two samples. The first comprised community-dwelling elderly individuals, and their carers, referred to the Croydon Memory Service a service for early assessment and intervention in dementia based in South London. The sample comprised those referrals made between December 2002 and June 2010 who, after a full multidisciplinary assessment, were given a formal clinical diagnosis of dementia using ICD-10 criteria (Banerjee et al., 2007). The sample therefore represents assessments of HRQL made at the time of diagnosis. No ethical committee approval was needed as this study was a secondary analysis on de-identified data.

The second study sample comprised community-dwelling elderly individuals, and their carers, who took part in the second wave of population-based surveys conducted by the 10/66 Dementia Research Group (DRG) in Cuba, Dominican Republic, Peru, Venezuela, and Mexico (Prince et al., 2007). As part of 10/66 DRG's broader aims to capture the impact of dementia, HRQL was assessed at follow up using DEMQOL and DEMQOL-Proxy for people who received a

dementia diagnosis at baseline or follow up based on the 10/66 diagnostic protocol. No ethical committee approval was needed as this was an analysis of de-identified data which was publicly accessible.

3.2.2 Measures

The DEMQOL (28 items) and DEMQOL-Proxy (31 items) are interviewer-administered measures for obtaining self- and informant-reports of HRQL in people with dementia (S. C. Smith et al., 2007). The question items on both measures inquire about the 'feelings', 'memory', and 'everyday life' of the person with dementia in the 'last week'. All items have a four-point likert scale (a lot / quite a bit / a little / not at all) and responses are coded so that higher total scores reflect better HRQL (http://www.bsms.ac.uk/research/our-researchers/sube-banerjee/demqol/). The DEMQOL and DEMQOL-Proxy had been translated into Spanish by the 10/66 Dementia Research Group for use in Latin America.

3.3 Analysis

3.3.1 Covariates

DIF due to gender, dementia severity, and geographical region was investigated. For gender, female respondents were treated as the reference group. For dementia severity, people with moderate to severe dementia were treated as a single focal group and compared against mild dementia serving as the reference group. In the UK sample, mild dementia was defined by scores of 21-30 on the Mini-Mental State Examination (Folstein et al., 1975). In the Latin America samples, mild

dementia was defined by scores of 0.5-1.0 on the Clinical Dementia Rating scale (Morris, 1993). For geographical region, data from the five Latin America countries were treated as a single focal group and compared against the UK sample serving as reference group.

3.3.2 DIF detection method

DIF detection was conducted under a structural equation modelling (SEM) framework using the multiple indicators, multiple causes (MIMIC) model. This approach permits a straightforward specification of a multidimensional model, unlike most other DIF detection methods for which unidimensional models are required (Woods, 2009; F. M. Yang, Tommet, & Jones, 2009). When the assumption of unidimensionality is not adequately met, these methods can result in false DIF detection (Mazor, Hambleton, & Clauser, 1998). The MIMIC approach is therefore appropriate for the present study given that emerging psychometric literature (e.g. Reise, 2012), as well as bifactor model results in Chapter 2, support the use of a multidimensional measurement model with a general HRQL and four domain factors: 'positive emotion' (POS), 'negative emotion' (NEG), 'worries about cognition' (COG), and 'worries about social relationship' (SOC). The same flexibility in SEM framework also permits MIMIC models to include multiple sources of DIF (e.g. gender, stage of illness, geographical region) in a single investigation, as well as a concurrent examination of group differences adjusted for the impact of DIF.

Operationally, DIF detection with MIMIC models is a model building process in SEM that begins with a baseline structural model in which we hypothesise a HRQL bifactor measurement model alongside multiple causes to explain group differences in HRQL and its domains. Gender, dementia severity, and geographical region were three causes hypothesised to have an impact (i.e. structural path) on the general HRQL and four domain factors. To reflect the hypothesis that differences in item response probabilities were due only to group differences in HRQL and its domains, the three covariates were assumed to have an impact on item responses only via the general HRQL and four domain factors (Mplus syntax enclosed in Appendices p. 243-246).

The baseline model was then compared with a new model in which a direct path from a covariate to an item was added. This augmented model reflected the hypothesis that there were group differences in item response probabilities, beyond those that were explained by group differences in HRQL and its domains. This direct path was interpreted as a DIF effect in which a specific focal group (e.g. males) had higher / lower response probabilities on the item, despite being matched to the reference group (i.e. females) in terms of their levels of HRQL estimated by the measurement model. The decision on which direct paths to add was based on modification indices. These are derivatives of the model chi square which show an expected improvement in model fit if direct paths between the covariates and items are freely estimated (i.e. parameters no longer fixed at 0 to reflect the absence of DIF). Large modification indices values for these path parameters suggest that a significant amount of group differences in item response

probabilities remain unaccounted for (i.e. baseline model was mis-specified) when we assumed no DIF in the initial group comparisons. Multiple interim models were estimated in a forward stepwise manner in which these direct paths were added one at a time (based on largest modification indices value) to form a new model with an increasingly smaller set of DIF-free (i.e. anchor) items. The statistical significance of the added path constitutes a test for DIF for an individual item. Each augmented model was also compared against their preceding (i.e. nested) alternative and the iterations were stopped when adding a path parameter no longer led to a statistically significant improvement in model fit.

While statistical significance indicates the presence of DIF, the magnitude of individual DIF effects is also of practical interest. Following the analytic strategy of (F. M. Yang & Jones, 2007), model estimation through the iterative stages employed WLSMV estimation which uses multivariate probit regression model to describe how (ordered-categorical) item responses were related to (continuous) latent variables that represented HRQL and its domains. With this estimation method, commonly reported model fit statistics are available to support model fit evaluation and comparison procedures both of which are tasks that are integral for DIF detection with MIMIC models. The final model was then re-estimated using the maximum likelihood estimator with robust standard errors (MLR) which used a multivariate logistic regression model for the same purpose. Due to the large number of latent factors, MLR was implemented with Monte Carlo integration to circumvent computational limitations. This estimation method expresses path

parameters as logistic regression coefficients which can be mathematically transformed to obtain odds ratios (ORs). As such, the magnitude of DIF effects could be judged in terms of the proportional difference between the reference and focal group in their respective odds of responding to a symptom at any level of HRQL. Cole, Kawachi, Maller, and Berkman (2000) have proposed that proportional ORs > 2.0 or < 0.5 are to be considered 'relatively large' and meaningful measurement bias. Given the complexities that can arise from multiple DIF effects varying in magnitude and direction, the impact of DIF was evaluated by comparing original estimates of group differences from the baseline MIMIC model (i.e. unadjusted for DIF) with those from the final model that adjusted for DIF (Jones & Gallo, 2002; Reininghaus, McCabe, Burns, Croudace, & Priebe, 2012; Teresi & Fleishman, 2007).

3.3.3 Model specifications and fit evaluation

Both DEMQOL and DEMQOL-Proxy bifactor models were shown in Chapter 2 to have a general HRQL factor and four substantive domains (POS, NEG, COG, SOC). All items had non-zero direct loading on the general HRQL factor and the domain it was designed to measure, but zero loadings on the other domains. Some items loaded on the general factor only. All latent factors are uncorrelated (i.e. orthogonal). The error terms (i.e. unexplained variance) of all items are also orthogonal.

We anticipated a need in the current analysis to specify correlated residuals or testlets due to local dependence (LD) between some item-pairs that had highly

similar content and/or strongly correlated item responses due to close proximity in item sequence. As MLR estimation cannot accommodate correlated residuals between ordered-categorical items, latent factors were employed as an alternative representation of these correlations (Mplus syntax enclosed in Appendices p. 243-246). Modification indices were inspected to prioritise the need to hypothesise these additional factors. The actual number of LD domain factors hypothesised depended on model fit evaluation.

When WLSMV estimation was employed, model identification was achieved by having one of the factor loadings on the general factor fixed at one. In addition, one of the factor loadings on each domain was also fixed at one. For LD domain factors with only two items, the factor loadings were fixed to be equal and the factor variance was fixed at one. The variances of the other factors were freely estimated. When MLR estimation was employed, model identification was achieved by fixing all factor variances at one. All factor loadings were freely estimated except for equality constraints on item-pairs of LD domain factors.

Across the iterative stages, model fit was assessed using RMSEA, CFI and TLI and model comparisons employed the DIFFTEST option in Mplus, so as to obtain the correct chi square difference test between models that were estimated with WLSMV.

3.3.4 Short-form derivation

Items that displayed DIF were potential candidates for exclusion from short form versions of the DEMQOL and DEMQOL-Proxy. However reviews by content

experts are also necessary. To aid decision-making, we used item response theory (IRT) to further study the items in terms of the amount of information they provided for discriminating individual differences and the level of HRQL at which they were the most informative. For this, two-parameter logistic (2PL) IRT models were used to investigate item response probabilities in terms of discrimination and difficulty parameters.

In the present study, discrimination parameters captured the relation between item responses and HRQL construct. This is conceptually equivalent to factor loadings of measurement models in SEM. The closer the relationship, the more informative an item is for differentiating individuals in terms of HRQL. In IRT framework, this also equates to lower levels of measurement error. Difficulty parameters define the level of HRQL that must be present in individuals before they are likely to achieve successively higher levels of response for an item (this likelihood was defined as having 50% probability of achieving each successive level of response). This is conceptually equivalent to threshold parameters when measurement models in SEM are estimated using multivariate probit (WLSMV) or logistic (MLR) regression for ordered-categorical item data. The higher the HRQL threshold, the more 'difficult' an item is for individuals to achieve higher levels of responses. Only individuals with high levels of HRQL are likely to report higher levels of functioning when asked a 'difficult' item. With an 'easy' item, even individuals with low HRQL levels are likely report high levels of functioning.

Given the correspondence between these IRT and SEM models, both discrimination and difficulty parameters could be derived from CFA parameters (in the MIMIC models) for generating IRT plots in Mplus. We used item information curves to show how much information an item provided (calculated from discrimination parameters) across different HRQL levels. These are typically bell-shape curves with peaks located at the difficulty of each item. In general, preference was given to items that were more informative (i.e. higher peaks). Having a set of peaks that were located along different points on the HRQL continuum (i.e. different thresholds / difficulty) is crucial to assure adequate measurement precision such that changes or differences over a broad range of HRQL levels can be assessed reliably.

3.4 Results

3.4.1 Sample characteristics

Table 3.1 presents the demographic and clinical characteristics of all study participants; all had a diagnosis of dementia (ICD-10 or 10/66 diagnostic algorithm). Only those with complete or partial DEMQOL and DEMQOL-Proxy data were included in the analyses.

Table 3.1 Demographic and clinical characteristics of study sample and participants who were excluded due to missing HRQL data

	UK n=1240		Latin America n=498	
DEMQOL	sample	miss	sample	miss
Age (SD)	78.6 (8.5) n=865	80.1 (8.2) n=317	79.7 (7.6) n=366	80.3 (8.0) n=76
Gender				
Male	313	144	126	13
Female	555	228	287	72
Dementia severity				
Mild	517	171	265	21
Moderate	325	174	132	42
Severe	26	26	12	21
Unknown	0	1	4	1
Dementia type				
AD	479	215	122	38
AD mixed	225	91	31	11
Vascular	99	48	54	14
Others	20	17	45	7
Unknown	45	1	161	15
DEMQOL-Proxy	sample	miss	sample	miss
Age (SD)	78.9 (8.4) n=905	79.3 (8.5) n=331	79.7 (7.7) n=436	85.8 (8.4) n=6
Gender				
Male	340	117	138	1
Female	569	214	353	6
Dementia severity				
Mild	509	179	282	4
Moderate	358	141	172	2
Severe	42	10	32	1
Unknown	0	1	5	0
Dementia type				
AD	488	206	155	5
AD mixed	242	74	57	0
Vascular	114	33	67	1
Others	27	10	42	0
Unknown	38	8	175	1

sample: study participants with complete/partial HRQL data
miss: study participants without HRQL data

Latin America: Cuba (n=115), Dominican Republic (n=124), Mexico (n=104), Peru (n=91), Venezuela (n=64)

3.4.2 Measurement models

The DEMQOL bifactor model (Table 3.2) attained acceptable model fit when a general HRQL factor was hypothesised with four substantive domain factors (POS, NEG, COG, SOC), and an additional factor (LD1) for item 8 and 20. The DEMQOL-Proxy bifactor model (Table 3.3) attained acceptable model fit when a general HRQL factor was hypothesised with four substantive domain factors (POS, NEG, COG, SOC), and two additional factors for LD item-pairs (LD1 for item 21 and 22; LD2 for item 24 and 25).

Three POS items from DEMQOL-Proxy did not have statistically significant loadings on the general HRQL factor. This might be a consequence of combining the UK and Latin America samples for the current analysis since these items had loaded marginally well on the general HRQL factor in the UK sample (Chapter 2). It is hence of particular interest in the subsequent stage of analysis to see if geographical region might be a source of DIF effects among these items.

3.4.3 MIMIC models: magnitude of DIF

The factor loadings in DEMQOL (Table 3.4) and DEMQOL-Proxy (Table 3.5) measurement models remained fairly stable when three covariates (gender, dementia severity, geographical region) were introduced to form the baseline MIMIC models, as well as after DIF effects were accounted for in the final MIMIC models.

Table 3.2 DEMQOL bifactor CFA model standardised factor loadings

Item	Question	GEN	POS	NEG	COG	SOC	LD1
1	cheerful **	.36	.59				
2	worried or anxious	.66					
3	that you are enjoying life **	.29	.61				
4	frustrated	.55		.57			
5	confident **	.22	.65				
6	full of energy **	.18	.77				
7	sad	.68					
8	lonely	.58					.62
9	distressed	.71					
10	lively **	.13	.77				
11	irritable	.59		.34			
12	fed-up	.66		.43			
13	things that you wanted to do but couldn't	.49		.36			
14	forgetting things that happened recently	.57			.54		
15	forgetting who people are	.59			.49		
16	forgetting what day it is	.50			.56		
17	your thoughts being muddled	.65			.57		
18	difficulty making decisions	.68			.42		
19	poor concentration	.62			.49		
20	not having enough company	.63					.62
21	how you get on with people close to you	.71				.36	
22	getting the affection that you want	.69				.47	
23	people not listening to you	.69				.56	
24	making yourself understood	.66				.45	
25	getting help when you need it	.73				.37	
26	getting to the toilet in time	.63					
27	how you feel in yourself	.74					
28	your health overall	.64					

Sample n= 1281, Model fit: χ^2_M = 1980.455 (df = 329), RMSEA = .063 (90%CI: .060 - .065), CFI = .925, TLI = .914, WRMR = 1.936

Table 3.3 DEMQOL-Proxy bifactor CFA model standardised factor loadings

Item	Question	GEN	POS	NEG	COG	SOC	LD1	LD2
1	cheerful **	.11	.74					
2	worried or anxious	.50		.53				
3	frustrated	.44		.56				
4	full of energy **	.03 #	.74					
5	sad	.45		.56				
6	content **	.18	.72					
7	distressed	.58		.47				
8	lively **	.01 #	.84					
9	irritable	.29		.51				
10	fed-up	.50		.47				
11	that he/she has things to look forward to **	-.03 #	.49					
12	his/her memory in general	.46			.65			
13	forget things that happened a long time ago	.51			.45			
14	forgetting things that happened recently	.51			.76			
15	forgetting people's names	.54			.57			
16	forgetting where he/she is	.60			.39			
17	forgetting what day it is	.60			.50			
18	his/her thoughts being muddled	.65			.49			
19	difficulty making decisions	.63			.44			
20	making him/herself understood	.63			.31			
21	keeping him/herself clean	.47					.82	
22	keeping him/herself looking nice	.51					.82	
23	getting what he/she wants from the shops	.73						
24	using money to pay for things	.67						.63
25	looking after his/her finances	.63						.63
26	things taking longer than they used to	.69						
27	getting in touch with people	.68				.30		
28	not having enough company	.60				.24		
29	not being able to help other people	.58				.75		
30	not playing a useful part in things	.63				.49		
31	his/her physical health	.59						

Sample n= 1400, Model fit: χ^2_M = 3228.482 (df = 408), RMSEA = .070 (90%CI: .068 - .073), CFI = .911, TLI = .899, WRMR = 2.509

not statistically significant

Table 3.4 DEMQOL standardised factor loadings in baseline and final MIMIC model

	Baseline model						Final model					
Item	GEN	POS	NEG	COG	SOC	LD1	GEN	POS	NEG	COG	SOC	LD1
1	.35	.59					.34	.56				
2	.67						.70					
3	.28	.61					.27	.62				
4	.60		.57				.62		.61			
5	.22	.66					.23	.63				
6	.19	.76					.20	.77				
7	.68						.69					
8	.57					.62	.59					.62
9	.72						.72					
10	.13	.79					.14	.79				
11	.62		.34				.63		.39			
12	.70		.44				.71		.51			
13	.54		.45				.56		.34			
14	.60			.56			.62			.59		
15	.61			.51			.63			.53		
16	.53			.58			.55			.60		
17	.68			.57			.70			.60		
18	.70			.43			.72			.45		
19	.68			.51			.70			.47		
20	.62					.62	.64					.62
21	.70				.37		.72				.38	
22	.68				.48		.71				.47	
23	.68				.56		.70				.57	
24	.66				.45		.68				.46	
25	.74				.35		.76				.34	
26	.63						.64					
27	.76						.80					
28	.65						.66					

Baseline MIMIC model: Sample n= 1277, Model fit: χ^2_M = 2279.836 (df = 395), RMSEA = .061 (90%CI: .059 - .064), CFI = .919, TLI = .906, WRMR = 1.890
Final MIMIC model: Sample n= 1277, Model fit: χ^2_M = 2009.69 (df = 389), RMSEA = .057 (90%CI: .055 - .060), CFI = .931, TLI = .918, WRMR = 1.768

Table 3.5 DEMQOL-Proxy standardised factor loadings in baseline and final MIMIC model

	Baseline model							Final model						
Item	GEN	POS	NEG	COG	SOC	LD1	LD2	GEN	POS	NEG	COG	SOC	LD1	LD2
1	.13	.74						.12	.75					
2	.51		.56					.50		.50				
3	.46		.65					.45		.52				
4	.04 #	.75						.04 #	.74					
5	.47		.57					.45		.60				
6	.19	.73						.19	.73					
7	.59		.48					.58		.47				
8	.02 #	.84						.03 #	.84					
9	.32		.51					.31		.52				
10	.51		.49					.50		.47				
11	-.02 #	.51						-.03 #	.50					
12	.50			.66				.49			.63			
13	.53			.50				.49			.58			
14	.56			.77				.56			.74			
15	.58			.60				.56			.58			
16	.63			.41				.60			.45			
17	.63			.52				.62			.50			
18	.69			.51				.67			.50			
19	.66			.46				.65			.45			
20	.66			.34				.62			.40			
21	.48					.77		.47					.80	
22	.50					.77		.50					.80	
23	.72							.72						
24	.67						.63	.66						.63
25	.63						.63	.63						.63
26	.69							.69						
27	.69				.28			.67				.27		
28	.61				.25			.59				.24		
29	.57				.73			.57				.77		
30	.62				.53			.62				.50		
31	.59							.58						

Baseline MIMIC model: Sample n= 1359, Model fit: χ^2_M = 3226.965 (df = 480), RMSEA = .064 (90%CI: .062 - .066), CFI = .913, TLI = .899, WRMR = 2.198
Final MIMIC model: Sample n= 1359, Model fit: χ^2_M = 2771.408 (df = 467), RMSEA = .059 (90%CI: .057 - .062), CFI = .927, TLI = .913, WRMR = 2.208
not statistically significant

Potential DIF effects were flagged for six DEMQOL items and 11 DEMQOL-Proxy items (Table 3.6). Geographical region accounted for the majority and was the only source of DIF effects in DEMQOL. Item 27 in DEMQOL displayed the most severe DIF effects due to region (OR=7.2, 95% CI: 4.7 – 11.1). Compared to the UK sample of people with dementia, the Latin American group had much higher odds of reporting higher levels of functioning when asked if they worry about '*how you feel in yourself*'. This means that taking people with the same levels of HRQL, those in Latin America would give more positive evaluations on this item compared with those from the UK. All other DIF effects due to geographical region were also of substantial, but smaller, magnitudes going by the criteria (OR<0.5 or OR>2.0) proposed by Cole et al. (2000).

Dementia severity and gender evoked DIF effects only in DEMQOL-Proxy. Compared to informants of people with mild dementia, informants of people with more advanced illness had smaller odds of reporting higher levels of functioning on item 16 ('*worry about forgetting where he/she is*': OR=0.3, 95% CI: 0.2 – 0.5) and item 20 ('*worry about making him/herself understood*': OR=0.4, 95% CI: 0.3 – 0.6). This means that when HRQL levels do not differ between people with mild or moderate to severe dementia, informants for those with more advanced illness report less positive evaluations on these two items.

For gender, informants had larger odds of reporting higher levels of functioning on item 28 ('*worry about not having enough company*': OR=2.6, 95% CI: 1.9 – 3.5), but smaller odds of doing so on item 9 ('*irritable*': OR=0.6, 95% CI: 0.5 –

0.8). This means that among people of either gender who do not differ in their levels of HRQL, informant reports would indicate less worry about '*not having enough company*', but more feelings of '*irritable*' in males. There was however uncertainty in whether the magnitude of gender DIF effects was clinically meaningful. This was the case for one (out of six) DIF effects in DEMQOL and six (out of 13) DIF effects in DEMQOL-Proxy as they had confidence intervals that included ORs which were not always within the proposed range of relatively large and meaningful bias (last column of Table 3.6).

Table 3.6 Magnitude of DIF effects

Item	Fac	DEMQOL	DIF due to	Ustd	SE	Std	ORs (95% CI)	Cole
27	G	how you feel in yourself	Region	.71	.08	.33	7.2 (4.7 – 11.1)	Yes
19	C	poor concentration	Region	.54	.07	.24	4.7 (2.9 – 7.7)	Yes
2	G	worried or anxious	Region	.49	.06	.23	3.0 (2.2 – 4.1)	Yes
13	N	things to do but couldn't	Region	.38	.09	.18	2.1 (1.3 – 3.3)	Maybe
1	P	cheerful	Region	-.48	.06	-.21	0.3 (0.2 – 0.4)	Yes
3	P	enjoying life	Region	-.57	.06	-.25	0.2 (0.2 – 0.3)	Yes
		DEMQOL-Proxy						
3	N	frustrated	Region	.69	.06	.31	5.6 (4.0 – 7.7)	Yes
8	P	lively	Region	.43	.06	.20	4.9 (2.9 – 8.1)	Yes
4	P	full of energy	Region	.54	.06	.25	4.8 (3.2 – 7.4)	Yes
28	S	not enough company	Gender	.42	.06	.19	2.6 (1.9 – 3.5)	Maybe
2	N	worried or anxious	Region	.39	.05	.18	2.4 (1.8 – 3.3)	Maybe
9	N	irritable	Gender	-.24	.06	-.11	0.6 (0.5 – 0.8)	Maybe
11	P	things to look forward to	Region	-.46	.06	-.21	0.4 (0.3 - 0.6)	Maybe
27	S	get in touch with people	Region	-.32	.06	-.15	0.4 (0.3 – 0.6)	Maybe
20	C	making self understood	Severity	-.31	.06	-.15	0.4 (0.3 – 0.6)	Maybe
16	C	forgetting where	Severity	-.36	.06	-.18	0.4 (0.3 – 0.5)	Yes
16	C	forgetting where	Region	-.44	.07	-.21	0.3 (0.2 – 0.5)	Yes
20	C	making self understood	Region	-.55	.07	-.26	0.3 (0.2 – 0.4)	Yes
13	C	forget things that happened a long time ago	Region	-.73	.07	-.34	0.2 (0.1 – 0.3)	Yes

Fac: Latent factors, where G = general HRQL; P = positive emotion; N = negative emotion, C = worries about cognition; S = worries about social relationship
Unstd: unstandardised probit coefficients (i.e. WLSMV estimation); SE: standard errors of unstandardised coefficients; Std: standardised coefficients (STDYX metric)
ORs: Odds ratios based on standardised logistic coefficients (i.e. MLR estimation)
Cole and colleagues: clinically meaningful if OR $\leq$ 0.5 or OR $\geq$ 2.0

3.4.4 MIMIC models: Impact of DIF

With the complexities arising from DIF effects of varying magnitude and direction, the impact of DIF on group comparisons remained to be determined. For this purpose, we examined standardised coefficients (Table 3.7 for DEMQOL and Table 3.8 for DEMQOL-Proxy) that captured the associations between covariates (gender, dementia severity, region) and latent constructs (HRQL, POS, NEG, COG, SOC, LD factors) before and after adjustment for DIF (Jones & Gallo, 2002; Reininghaus et al., 2012). The continuous latent constructs had a standardised mean of zero with unit variance in the reference groups. As all three covariates were dichotomous dummy variables, the standardised coefficients represented the difference between a reference (e.g. UK) and focal group (e.g. Latin America) for each latent construct (e.g. HRQL) in terms of standard deviation units.

Based on unadjusted estimates in baseline MIMIC model for DEMQOL (Table 3.7), the Latin America sample had lower HRQL levels (standardised estimate = -0.23) than the UK sample. There were no statistically significant differences in HRQL for gender and dementia severity. The same conclusions were reached (standardised estimate = -0.34) after DIF effects were accounted for in the final MIMIC model. Ignoring DIF would have underestimated HRQL differences between the two regions by about 0.1 of a standard deviation.

Table 3.7 Group differences in HRQL and its domains for DEMQOL

	Baseline model			Final model			Short-form model		
	Unstd	SE	Std	Unstd	SE	Std	Unstd	SE	Std
HRQL									
Gender	.02	.05	.01	.02	.04	.01	.07	.05	.05
Severity	.02	.05	.02	.02	.04	.01	.03	.05	.02
Region	-.36 **	.05	-.23	-.54 **	.04	-.34	-.43 **	.05	-.28
POS									
Gender	.05	.05	.03	.05	.04	.03			
Severity	-.20 **	.05	-.12	-.20 **	.04	-.12			
Region	-.18 **	.05	-.11	.10	.04	.06			
COG									
Gender	-.07	.04	-.06	-.06	.03	-.06	-.11	.06	-.08
Severity	.08 *	.04	.08	.07 *	.03	.08	.15 *	.06	.11
Region	.39 **	.05	.35	.41 **	.04	.41	.67 **	.07	.48
NEG									
Gender	-.04	.04	-.05	-.03	.03	-.04	-.09 *	.04	-.09
Severity	-.03	.04	-.03	-.02	.03	-.03	-.04	.04	-.04
Region	.47 **	.05	.49	.39 **	.05	.53	.59 **	.06	.57
SOC									
Gender	-.07	.04	-.09	-.07	.04	-.10	-.13 *	.05	-.14
Severity	.01	.04	.01	.01	.04	.01	.02	.05	.02
Region	-.05	.04	-.07	.09 *	.05	.12	.03	.06	.03
LD1									
Gender	.39 **	.10	.18	.39 **	.09	.18			
Severity	-.08	.10	-.04	-.08	.08	-.04			
Region	-.09	.10	-.04	.16	.09	.07			

Reference groups: Female (Gender), Mild dementia (Severity), UK (Region)
Unstd: unstandardised coefficients, SE: standard errors, Std: standardised coefficients.
LD1: item 8 and 20
* $p < .05$; ** $p < .001$.

Table 3.8 Group differences in HRQL and its domains for DEMQOL-Proxy

	Baseline model			Final model			Short-form model		
	Unstd	SE	Std	Unstd	SE	Std	Unstd	SE	Std
HRQL									
Gender	.13 **	.04	.10	.08	.04	.07	.11 *	.03	.11
Severity	.05	.04	.04	.10 *	.04	.08	.06 *	.03	.07
Region	-.20 **	.04	-.16	-.07	.04	-.06	-.19 **	.04	-.20
POS									
Gender	.02	.03	.02	.03	.03	.02			
Severity	-.20 **	.03	-.19	-.20 **	.03	-.19			
Region	.01	.03	.01	-.11 **	.04	-.10			
COG									
Gender	.00	.03	-.01	.03	.03	.03	.01	.03	.01
Severity	.02	.02	.03	.02	.03	.02	.06 *	.03	.07
Region	.33 **	.04	.46	.38 **	.05	.45	.58 **	.05	.58
NEG									
Gender	-.02	.04	-.02	.05	.04	.05	-.01	.04	-.01
Severity	-.09 *	.04	-.09	-.13 **	.04	-.13	-.11 *	.04	-.12
Region	.36 **	.04	.35	.09 *	.04	.09	.39 **	.04	.40
SOC									
Gender	.04	.05	.03	.03	.04	.03	.02	.05	.02
Severity	.13 **	.05	.12	.07	.04	.07	.10 *	.05	.09
Region	-.01	.05	-.01	-.10 *	.05	-.09	.05	.05	.04
LD1									
Gender	.22 *	.09	.09	.27 **	.09	.11			
Severity	.23 **	.08	.10	.18 *	.08	.07			
Region	-1.10 **	.09	-.46	-1.26 **	.09	-.51			
LD2									
Gender	-.21 *	.10	-.10	-.13	.10	-.06			
Severity	.06	.09	.03	-.02	.09	-.01			
Region	-.07	.10	-.03	-.29 **	.10	-.14			

Reference groups: Female (Gender), Mild dementia (Severity), UK (Region)
Unstd: unstandardised coefficients, SE: standard errors, Std: standardised coefficients.
LD1: item 21 and 22; LD2: item 24 and 25
* $p < .05$; ** $p < .001$.

In the case of DEMQOL-Proxy, initial estimates (Table 3.8) showed that HRQL levels were slightly higher in males (standardised estimate = 0.10), and slightly lower in the Latin America sample (standardised estimate = -0.16). These differences were not statistically significant in the final model. Instead a small but statistically significant difference (standardised estimate = 0.08) emerged in favour of the group with more advanced illness. These small but statistically significant changes have to be interpreted with caution given the level of statistical power afforded by the sample size in the current study. Given that group differences were not large before and after DIF adjustment, the impact of DIF was limited.

DIF adjustment also led to different conclusions about statistically significant group differences in two HRQL domains in DEMQOL (POS and SOC) and three in DEMQOL-Proxy (POS, SOC, and LD2). For the majority that remained consistent, the impact of DIF generally resulted in a small bias in standardised estimates (< 0.1 standard deviation difference). A notable exception was the NEG domain of DEMQOL-Proxy. Initial estimates showed that the level of functioning was higher in the Latin American sample by 0.35 standard deviations. After DIF adjustment, this group difference became much smaller (standardised estimate = 0.08) despite remaining statistically significant. In other words, this group difference was substantially overestimated due to DIF.

3.4.5 Item selection for short-form versions

Item information curves for each item (28 for DEMQOL, 31 for DEMQOL-Proxy) were generated based on IRT discrimination and difficulty parameters that were converted from the CFA parameters in the MIMC models.

The vertical axis of these graphical plots depicts the level of information provided by an item across HRQL levels. On the horizontal axis, latent estimates of HRQL (from the measurement model) are standardised so that sample average is located at the mean of 0 with a standard deviation of 1.

The majority of DEMQOL and DEMQOL-Proxy items had information peaks that were located at 0.5 – 1.0 standard deviation below the sample average. There was most measurement precision for people with poorer HRQL relative to the average in community-dwelling samples. To maintain content coverage during item selection, we examined the item information curves in sets that corresponded to each HRQL domain.

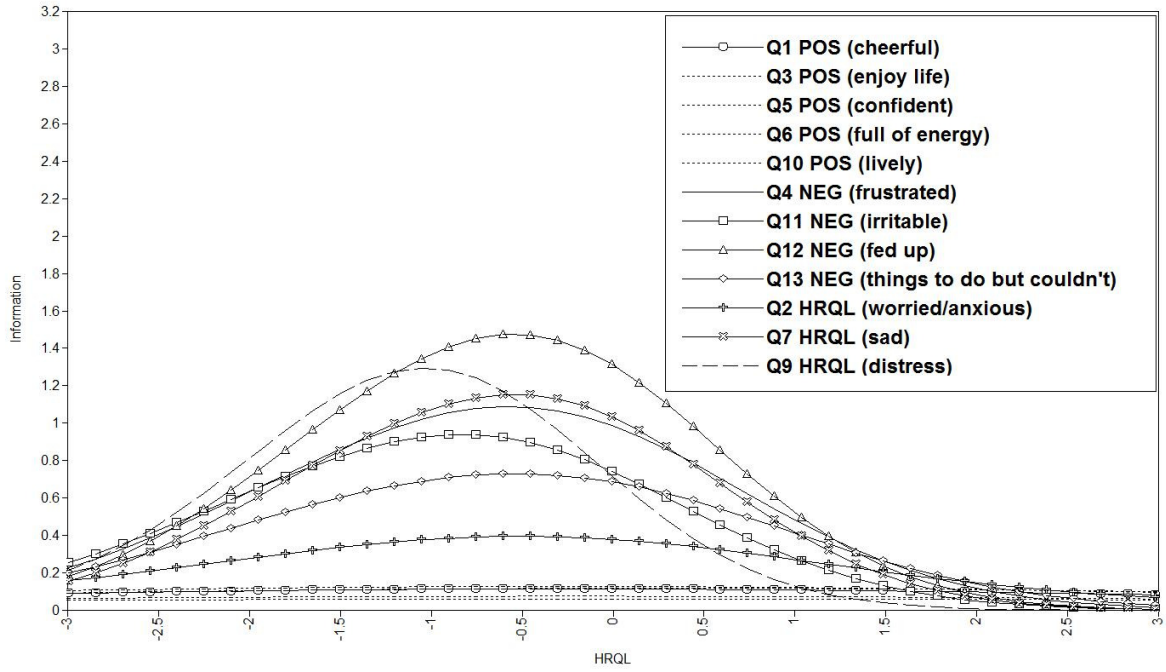

Figure 3-1 Item information curves for DEMQOL (POS, NEG, HRQL)

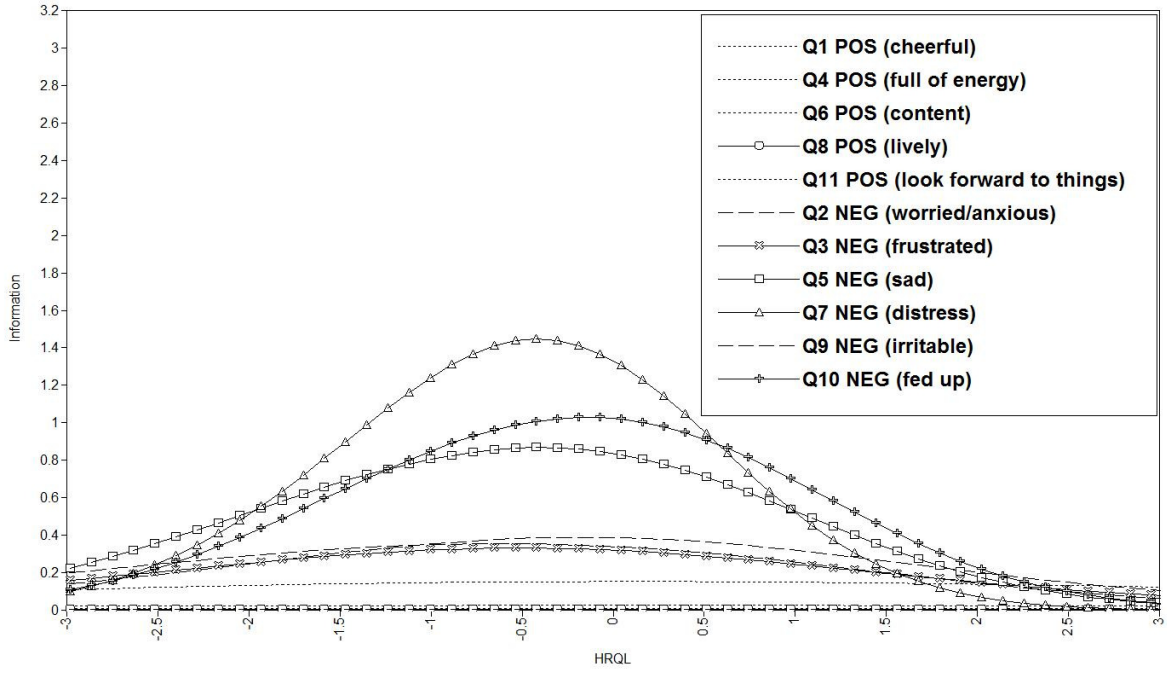

Figure 3-2 Item information curves for DEMQOL-Proxy (POS, NEG)

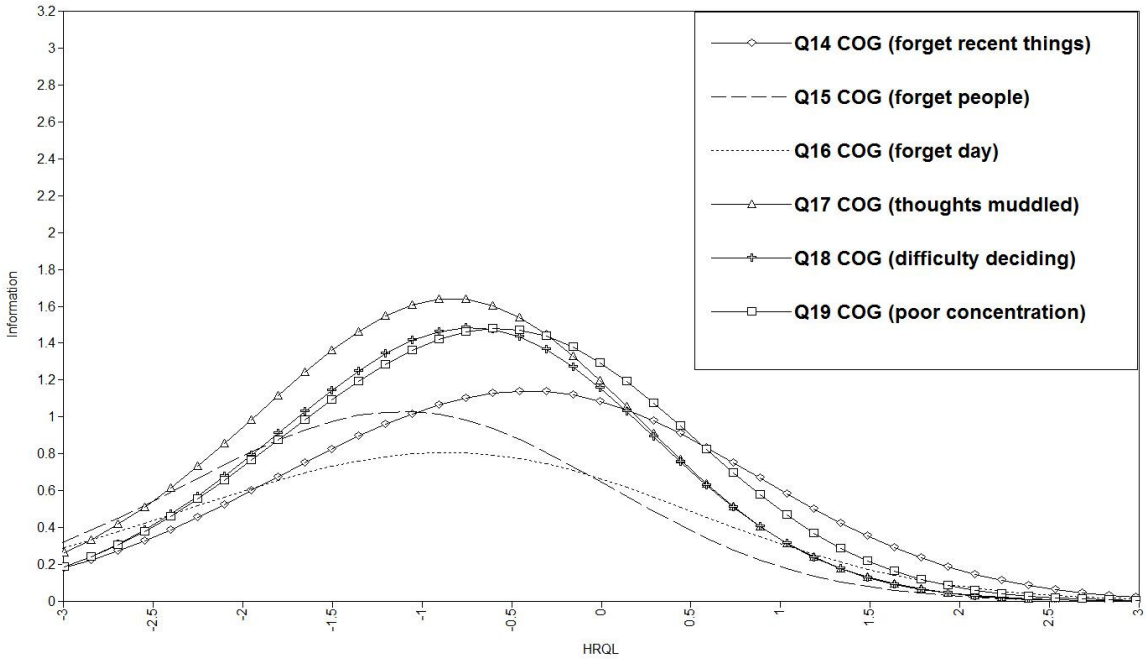

Figure 3-3 Item information curves for DEMQOL (COG)

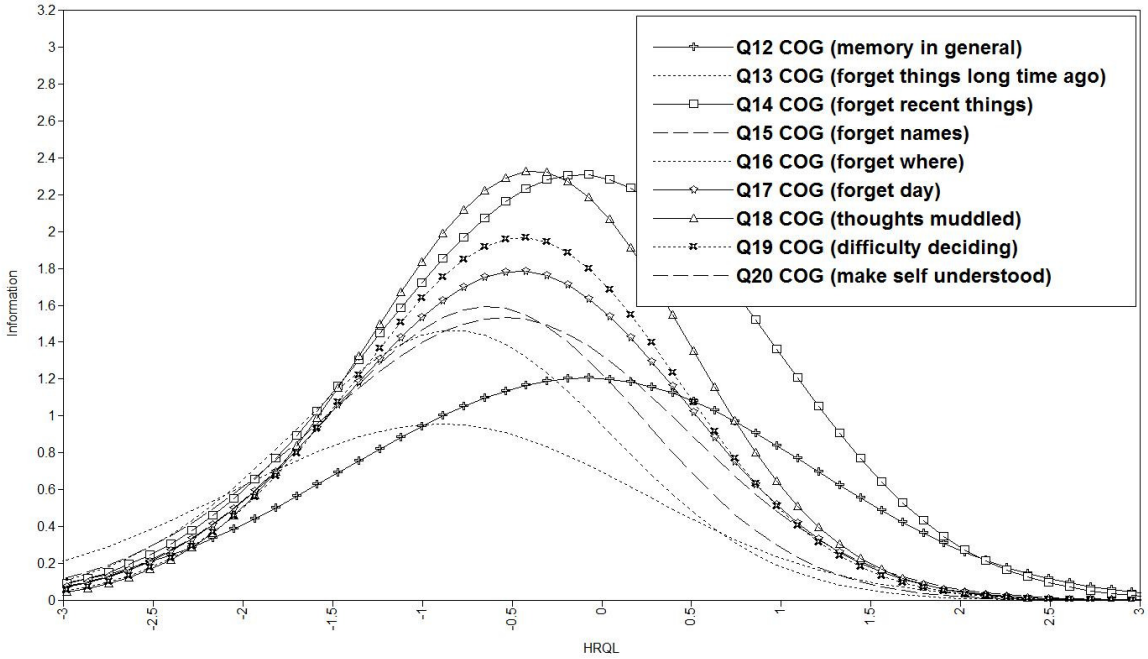

Figure 3-4 Item information curves for DEMQOL-Proxy (COG)

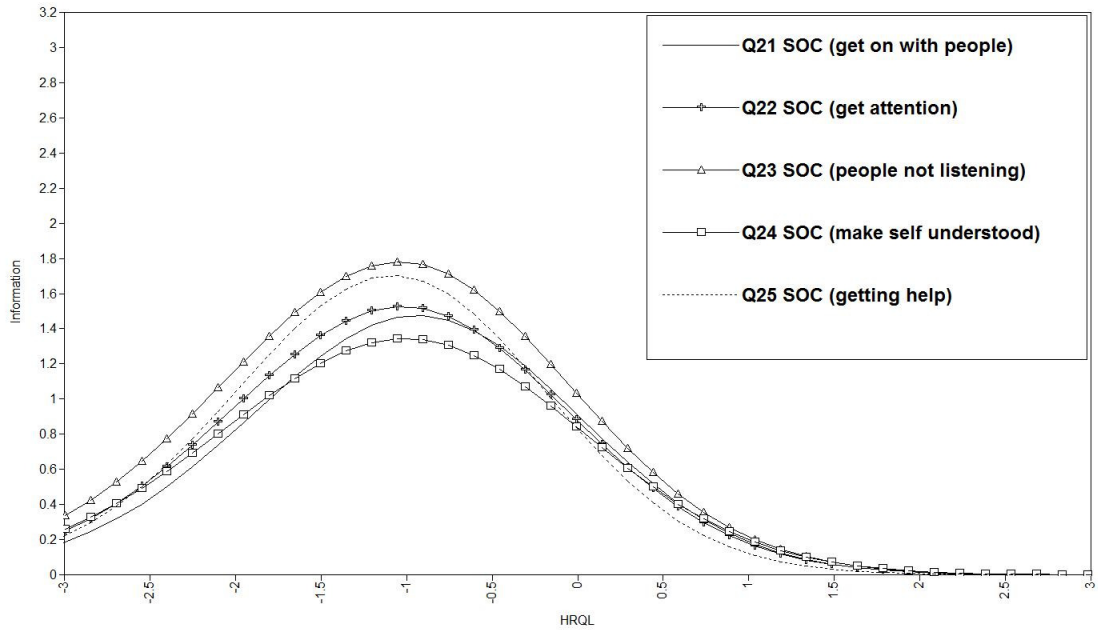

Figure 3-5 Item information curves for DEMQOL (SOC)

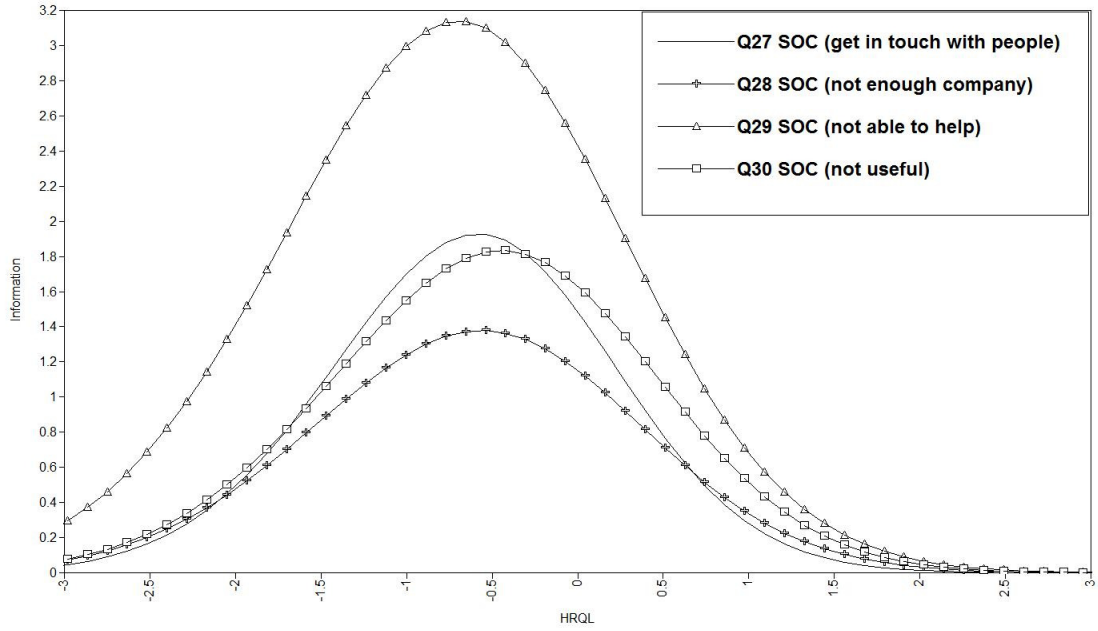

Figure 3-6 Item information curves for DEMQOL-Proxy (SOC)

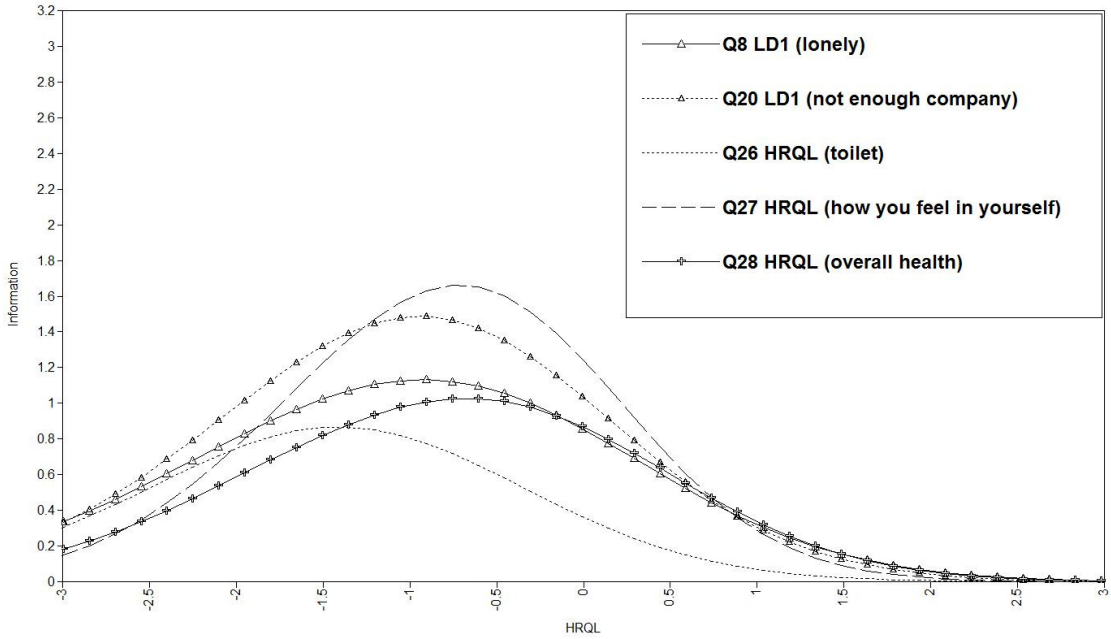

Figure 3-7 Item information curves for DEMQOL (HRQL)

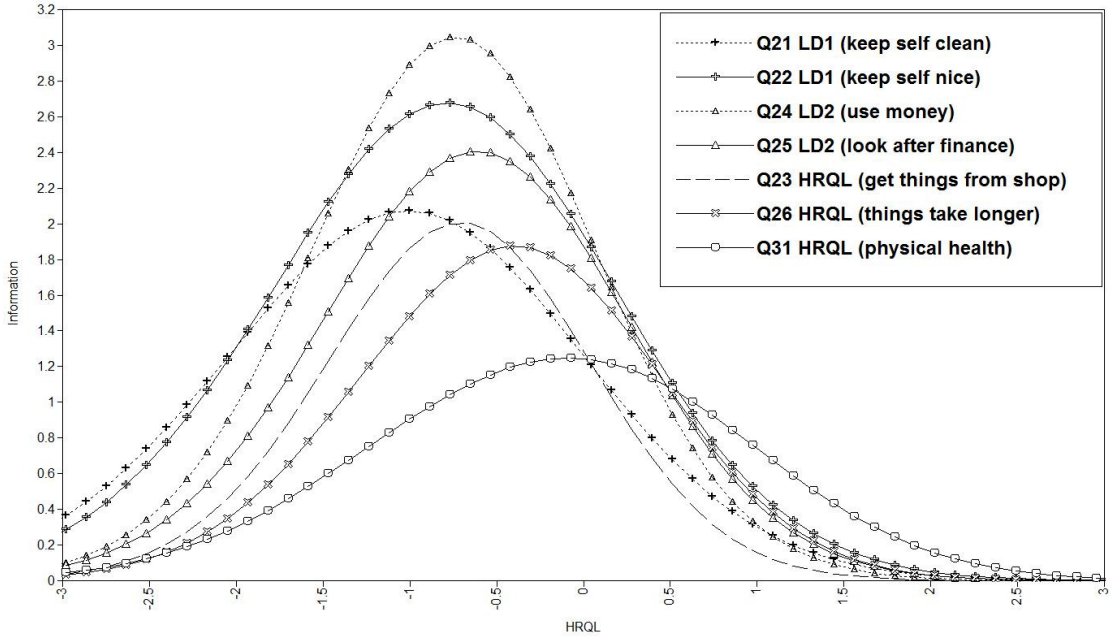

Figure 3-8 Item information curves for DEMQOL-Proxy (HRQL)

For POS domain, the information curves of five items in DEMQOL (Figure 3.1) and another five in DEMQOL-Proxy (Figure 3.2) showed that they were not discriminative of individual differences across the HRQL continuum. Therefore only one POS item was retained for DEMQOL (item 1: *cheerful*) and DEMQOL-Proxy (item 8: *lively*) in the short-form versions. The inclusion of these items was to maintain content correspondence with DEMQOL-U and DEMQOL-Proxy-U (Mulhern et al., 2013), which are preference-based versions that had been developed for economic evaluation. Both POS items displayed DIF due to region (Table 3.6) and item 8 in DEMQOL-Proxy also did not load on the general HRQL factor in the present study (Table 3.3 and 3.5).

For NEG domain, we retained all four items (4, 11, 12, 13) in DEMQOL (Figure 3.3) and another four (item 3, 5, 7, 10) out of the original six in DEMQOL-Proxy (Figure 3.4). This is in line with recommendations from simulation studies that showed having more items per latent factor helps insure adequate construct representation and leads to more stable estimation in SEM (Little, Lindenberger, & Nesselroade, 1999; Marsh, Hau, Balla, & Grayson, 1998). Item 13 (*things to do but couldn't*) in DEMQOL and item 3 (*frustrated*) in DEMQOL-Proxy displayed DIF due to region (Table 3.6). However, there was uncertainty in the clinical significance of DIF in the former (last column of Table 3.6). Despite having larger DIF effects due to region, the inclusion of item 3 was to maintain content correspondence between DEMQOL-Proxy and DEMQOL-Proxy-U.

Three additional DEMQOL items (2, 7, and 9) were considered in Figure 3.1 as they also related to negative emotions even though they did not load on NEG (but only on the general HRQL factor). Among them, item 2 and 7 were retained for the short-form. The current selection strategy did not focus on achieving maximum amount of information (e.g. item 9 had high information levels but was not selected). Instead, items were chosen to maximise information coverage over the region where HRQL levels are slightly above the sample average. With current indications that DEMQOL and DEMQOL-Proxy are well-suited for assessing HRQL impairment (or when HRQL levels are 0.5 – 1.0 standard deviation *below* sample average), priority was given to maintaining optimal levels of measurement precision when assessing treatment benefits (or when HRQL levels are 0.5 – 1.0 standard deviation *above* sample average).

For COG domain, we retained four (item 14, 17, 18, 19) out the original six items in DEMQOL (Figure 3.3) and another four (item 12, 14, 17, 18) out of the original nine in DEMQOL-Proxy (Figure 3.4). Item 19 (poor concentration) in DEMQOL displayed DIF due to region (Table 3.6) but offered relatively more information for assessing treatment benefits. All four COG items selected for DEMQOL-Proxy short form were DIF-free. Item 14 (forget recent things) in DEMQOL and item 17 (forget day) in DEMQOL-Proxy were also selected to maintain content correspondence with preference-based versions.

For SOC domain, we retained four (item 21, 22, 23, 24) out of the original five in DEMQOL (Figure 3.5) and all of the original four (item 27, 28, 29, 30) in

DEMQOL-Proxy (Figure 3.6). None of the selected four in DEMQOL displayed DIF. Item 27 (*get in touch with people*) and 28 (*not enough company*) in DEMQOL-Proxy displayed DIF due to region and gender respectively. However, there was uncertainty in the clinical significance of DIF effects in both cases (last column of Table 3.6). Item 24 (*make self understood*) in DEMQOL was also selected to maintain content correspondence with DEMQOL-U.

Both DEMQOL and DEMQOL-Proxy had domain factors that represented correlations between LD item-pairs (LD factors in Figure 3.7 and 3.8). Due to their highly similar content and/or order effects, the 'excess' associations highlighted potential redundancies which could be eliminated from the measurement models of short-form versions. Choosing between item 8 (*lonely*) and 20 (*not enough company*) in DEMQOL, the former was retained to maintain content correspondence with DEMQOL-U. Although its counterpart had a higher information peak, both were similarly discriminative in the target region most relevant for the assessment of treatment benefits (i.e. 0.5 – 1.0 standard deviation above the sample average). Item 21 (*keep self clean*) was more discriminative (ie had a higher information peak) than 22 (*keep self nice*) in DEMQOL-Proxy and offered better coverage in the target region (i.e. slightly higher than average HRQL levels). This item also corresponded with the content of DEMQOL-Proxy-U. In DEMQOL-Proxy, item 24 (*use money*) was more discriminative than 25 (*look after finance*) but the latter provided slightly better coverage in the target region and was therefore retained. The remaining items in DEMQOL (Figure 3.7) and DEMQOL-Proxy (Figure 3.8) loaded only on the general HRQL factor of

their respective measurement models. Of these, item 28 (*overall health*) in DEMQOL and item 26 (*things take longer*) and 31 (*physical health*) in DEMQOL-Proxy were retained to bolster information coverage over the target region.

In sum, a total of 17 items were retained for DEMQOL (out of 28 items) and DEMQOL-Proxy (out of 31 items) respectively. The item information curves can be added up to give an overall test information curve which summarises the level of information coverage that is offered by the entire set of items across the HRQL continuum. Figures 3.9 and 3.10 allow for a comparison of test information curves for the original and short-form versions of DEMQOL and DEMQOL-Proxy respectively. At a cost of less information (i.e. lower peaks) due to having fewer items, the short-form versions retained very similar coverage over the HRQL continuum to their parent versions. Of note, DEMQOL-SF had a reliability of at least 0.8 (= Information / Information +1) for assessing HRQL impairment (up to 2.5 SD below sample average) and treatment benefits (up to 1 SD above sample average). DEMQOL-Proxy-SF had a reliability of at least 0.8 over a slightly broader range (from -2.5 to +1.5 SD). Both short forms offered the highest reliability (>0.9) in the region of slightly poorer than average HRQL levels (-0.5 SD below sample average).

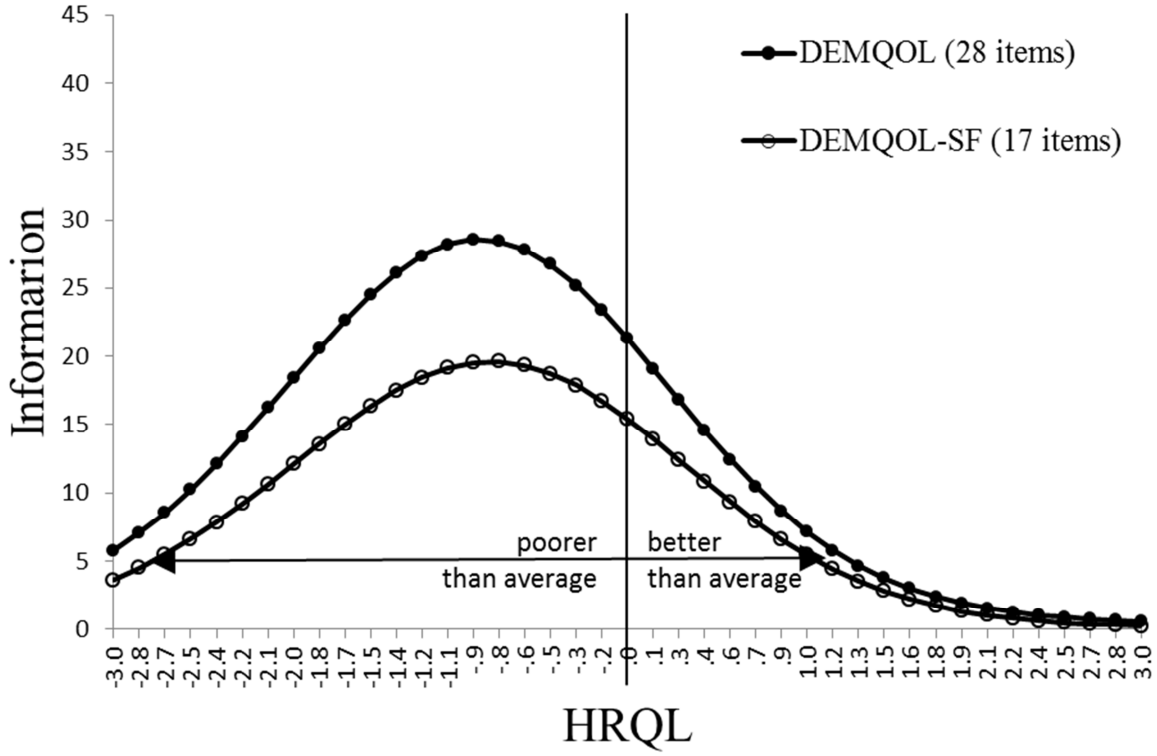

Figure 3-9 Test information curve for DEMQOL and DEMQOL-SF.
Horizontal line where Information = 5 indicates HRQL continuum where measurement reliability = 0.8 or more

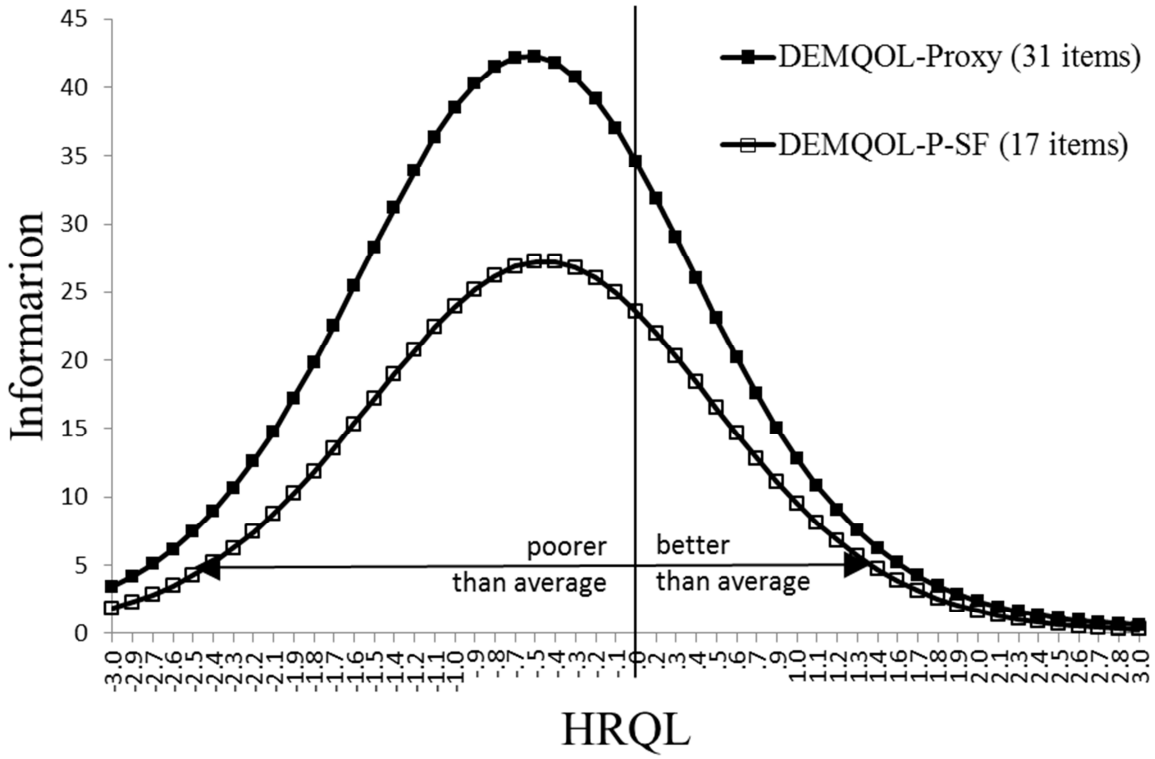

Figure 3-10 Test information curve for DEMQOL-Proxy and DEMQOL-P-SF.
Horizontal line where Information = 5 indicates HRQL continuum where measurement reliability = 0.8 or more.

3.4.6 Measurement models for short-form versions

To determine how short-form derivation might have affected the fidelity of the theoretical construct of HRQL in DEMQOL and DEMQOL-Proxy, bifactor CFAs were conducted with their short-form versions, DEMQOL-SF (Figure 3.11) and DEMQOL-Proxy-SF (Figure 3.12), to see if identical themes might be used to understand the response patterns in short-form versions. We hypothesised that responses on DEMQOL-SF and DEMQOL-Proxy-SF reflected the theme of general HRQL, as well as that of negative emotions (NEG), cognitive functioning worries (COG), and social functioning worries (SOC). The POS domain was no longer included in the measurement models due to the retention of only one POS item in both short forms. Concerns about any decline in content validity are addressed later in the discussion of this chapter. LD factors were also not included since only one item from each item-pair was retained. These items joined the few that loaded only on the general HRQL factor in the parent versions. Model fit evaluation suggested that model-data correspondence was acceptable. Nonetheless, DEMQOL-Proxy item 8 (lively) did not load on the general HRQL factor for DEMQOL-Proxy-SF. Both DEMQOL-SF and DEMQOL-P-SF retained a similar HRQL concept as in their parent versions in the absence of a POS domain.

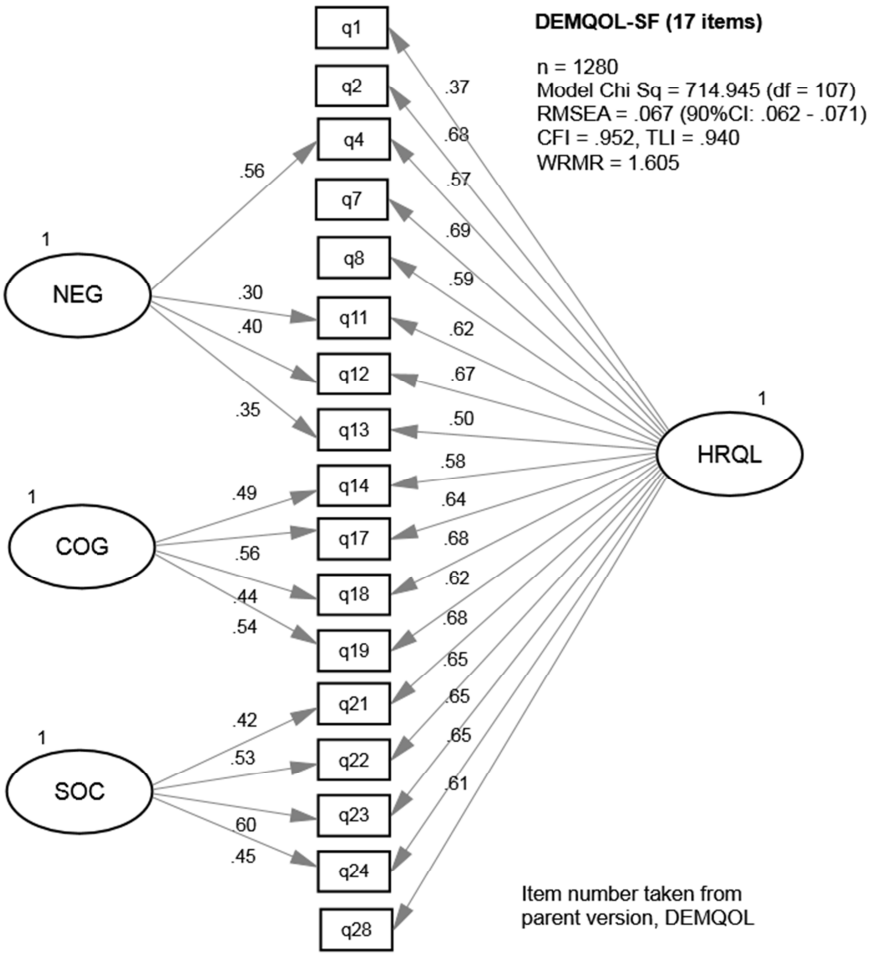

Figure 3-11 DEMQOL-SF bifactor CFA model standardised estimates

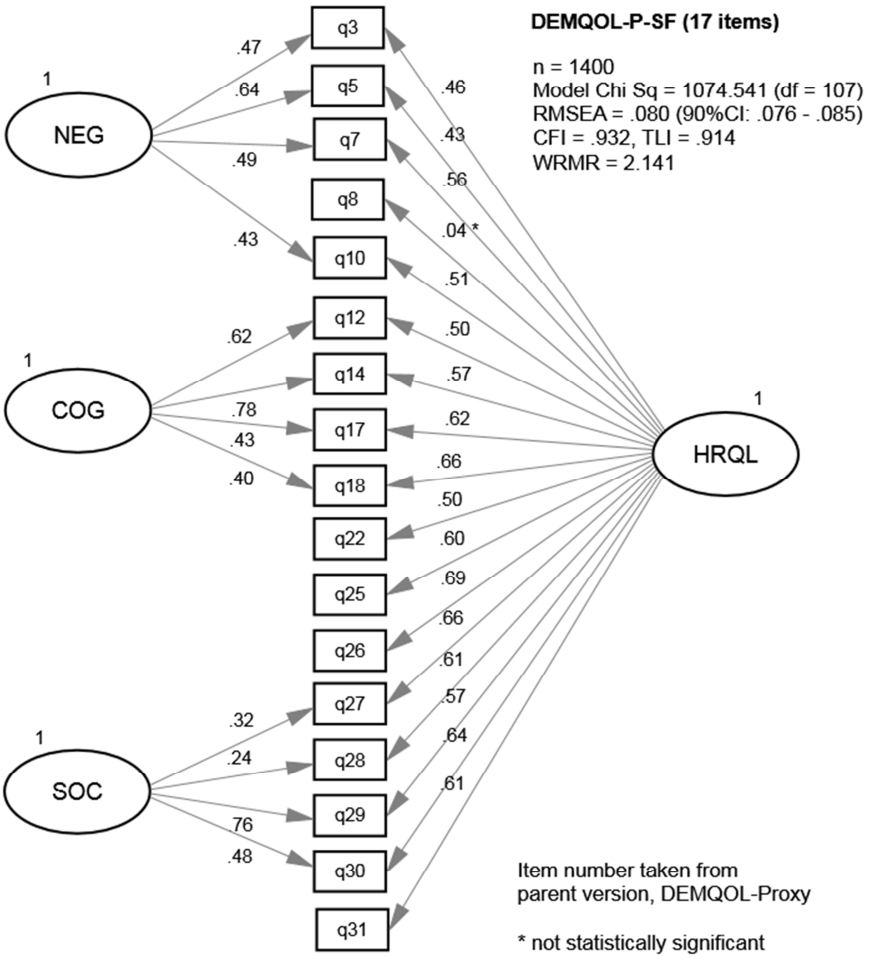

Figure 3-12 DEMQOL-P-SF bifactor CFA model standardised estimates

3.4.7 Short-form estimates of group differences

To examine whether the short forms could reproduce the same conclusions about group differences in HRQL, we introduced the three covariates (gender, dementia severity, region) into the measurement models for DEMQOL-SF and DEMQOL-Proxy-SF. No statistical adjustment was made in these MIMIC models despite the presence of some items that had displayed DIF in their parent versions. This was so the group comparisons resembled that of clinical research and practice where DIF adjustment is either absent or not feasible.

Results from both DEMQOL-SF and DEMQOL indicated that HRQL differences between gender or dementia severity were not statistically significant in this community-dwelling study sample (Table 3.7). Based on DEMQOL-SF, the Latin America sample had lower HRQL levels than the UK sample (standardised estimate = -0.28). The parent version showed a similar difference (baseline model: -0.23), but of a larger magnitude after DIF effects were accounted for (final model: -0.34). These group differences were hence underestimated (when unadjusted for DIF) by DEMQOL and DEMQOL-SF, though it appeared slightly less severe in the short form version.

Results from DEMQOL-Proxy-SF (Table 3.8) indicated that males had higher HRQL levels than females (standardised estimate = 0.11). The parent version (baseline model) showed a similar difference but this was no longer statistically significant after adjusting for DIF effects (final model). Gender differences were

therefore overestimated (when unadjusted for DIF) by DEMQOL-Proxy and DEMQOL-Proxy-SF to a similar extent.

Based on DEMQOL-Proxy-SF (Table 3.8), people with more severe dementia had slightly higher HRQL levels than those with mild dementia (standardised estimate =0.07). The parent version gave the same conclusion but only after DIF effects were taken into account. DEMQOL-Proxy-SF seemed less affected by DIF due to dementia severity and hence potentially more appropriate than its parent version for comparing HRQL levels between individuals at different stages of illness.

When comparison was made between UK and Latin American samples, both DEMQOL-Proxy-SF and its parent version (baseline model) showed higher levels of HRQL in the UK sample (Table 3.8). However, after accounting for DIF effects, the parent version (final model) showed that this difference was not statistically significant. Both DEMQOL-Proxy-SF and its parent version overestimated the difference (when unadjusted for DIF), and this appeared slightly more severe with the short form.

3.5 Discussion

These data offer insights into measurement invariance in HRQL assessment for people with dementia. Item response probabilities were examined to test the assumption that people with similar levels of HRQL would have similar responses on DEMQOL and DEMQOL-Proxy despite gender, dementia severity, and region. Items for which these assumptions were not met were identified as exhibiting DIF effects.

3.5.1 DIF detection in DEMQOL

We found no evidence of DIF effects due to gender and dementia severity in DEMQOL. Six items displayed DIF due to region, but there was no clear preponderance of particular DIF effects in a HRQL domain. Relative to the UK sample, the Latin American group tended to give less positive evaluations when responding to two positively-worded items (e.g. cheerful). For the other four items that had negative undertones (e.g. worried about 'poor concentration'), the same focal group tended to give more positive evaluations. The implications were investigated by comparing results before and after adjustment for these DIF effects. The initial results showed that differences in HRQL levels were not statistically significant for gender and dementia severity; the Latin American sample had lower HRQL levels than the UK sample. The same conclusions were reached with adjusted results but the difference between UK and Latin American samples became larger. The bias in standardised estimates was not large (difference of 0.1). The absence of DIF due to gender and dementia severity, supports the use of DEMQOL within each geographical region for assessing HRQL differences in gender or individuals at different stages of illness.

3.5.2 DIF detection in DEMQOL-Proxy

Eleven DEMQOL-Proxy items exhibited DIF effects mostly due to region. There was no clear preponderance of particular DIF effects in HRQL domain. When assessing three aspects of worry about cognitive function (worried about '*make self understood*', '*forget where*', '*forget things that happened long ago*'),

informant evaluations tended to be less positive in the Latin American sample (relative to UK), or when the assessment was made for people with more advanced illness (relative to mild dementia). While there was some uncertainty over clinical significance of DIF effects due to gender, DEMQOL-Proxy reports tended to show less worry about 'not having enough company', but more feelings of 'irritable' in males. When these DIF effects were ignored, HRQL differences were overestimated for gender and region, but underestimated for dementia severity. The bias in standardised estimates was not severe for gender and dementia severity (difference of 0.1 or less), but slightly more problematic for region. Using DEMQOL-Proxy to assess HRQL differences in gender or individuals at different stages of illness may be less problematic when results are compared within each geographical region.

3.5.3 DEMQOL-SF and DEMQOL-Proxy-SF

Unadjusted estimates of group differences based on short-form versions led to similar conclusions as those based on their parent versions. For DEMQOL-SF, HRQL differences were underestimated for region but this was less severe than that in its parent version. Measurement invariance for region appeared to be slightly stronger in DEMQOL-SF despite the retention of some items that displayed DIF in the parent version. For DEMQOL-Proxy-SF, HRQL differences were overestimated for gender and region, as with its parent version. The estimate for HRQL differences in dementia severity coincided closely with the DIF-adjusted estimate from its parent version. This suggested that measurement

invariance for dementia severity was stronger in the short-form version of DEMQOL-Proxy.

While DIF effects were not entirely eliminated from the short-form versions, they retained a high level of measurement precision for assessing HRQL impairment and treatment benefits. It is important to note that DEMQOL-SF and DEMQOL-Proxy-SF, like their parent versions, were more sensitive to change (i.e. higher measurement reliability) across the continuum of HRQL impairment (as low as 2 SD below sample average) than across the continuum of treatment benefits (as high as 1 SD above sample average). For instance, in individuals with relatively poor HRQL (e.g. -1.0 SD), the magnitude of HRQL impairment / improvement must exceed a 95% confidence interval of -1.5 to -0.6 (calculated from standard error, SE = 1 / √ Information), for change to be considered statistically significant. In contrast, for individuals with relatively good HRQL (e.g. +1.0 SD), the magnitude of HRQL impairment / improvement must exceed a larger 95% confidence interval of 0.2 to 1.9, for change to be considered statistically significant.

3.5.4 Quasi-trait

The apparent limitations in assessing treatment benefits is more likely to be a reflection of the theoretical nature of HRQL constructs, rather than of a theoretical deficiency in the construction of HRQL measures (Reise & Waller, 2009). In the clinical literature, Reise and Waller (2009) noted that scale scores from assessment measures tend to be skewed, with the majority having no/low levels of

psychopathology, and assessment items tend to have item difficulty locations that are more informative about presence of psychopathology rather than its absence. In the same way, item responses on DEMQOL-SF and DEMQOL-Proxy-SF were more informative about poor HRQL (i.e. presence of impairment) rather than about good HRQL (i.e. absence of impairment). Termed as a 'quasi-trait' (Reise & Waller, 2009), HRQL can be considered a unipolar construct in which trait levels are relevant only in one direction. While HRQL impairment can be readily identified in assessments, there is interpretive ambiguity at high levels of HRQL (i.e. absence of impairment). This is consistent with the observation made by (Lawton, 1994), whose conceptual model has been a major influence on HRQL measures in dementia, that HRQL is a construct 'concerned primarily with decrements from the average', and that good HRQL is related to but not exactly the reverse of poor HRQL. Such an understanding of HRQL is also consistent with the findings of a recent population study in UK which showed an asymmetry between strong adverse reactions to deteriorations in health, alongside weak increases in well-being after health improvements (Binder & Coad, 2013).

3.5.5 Implications of HRQL as a quasi-trait

Interpretive ambiguity at high levels of HRQL presents a challenge in constructing assessment items that are informative about the relative absence of impairment (Reise & Waller, 2009). This may explain why DEMQOL-SF and DEMQOL-Proxy-SF items offered less measurement precision at HRQL levels that exceed the average by a lot (e.g. > 1 SD). While this information gap would

logically be filled by reports of high levels of positive emotions, interpretive ambiguity in the region where HRQL impairment is mostly absent might have undermined the ability of POS items to load on the general HRQL factor in bifactor measurement models. Similar findings were reported in a study that recommended the elimination of positively-worded items on a well-known depression measure (Stansbury, Ried, & Velozo, 2006). A growing body of research has also suggested that reverse-scored/worded items should be avoided and that other types of factor models should be used to address the influence of these method effects on item responses (Brown, 2003; Carlson et al., 2011; Ebesutani, Drescher, et al., 2012; Lindwall et al., 2012; Marsh, 1986, 1996; Tomás et al., 2013; van Sonderen et al., 2013).

3.5.6 Content validity of DEMQOL-SF and DEMQOL-P-SF

The omission of 'positive emotion' (POS) items from HRQL assessment in short-form versions does not imply that positive states have no relevance in the health of people with dementia. Based on clinical observations of people with dementia in residential care, Lawton (1994) proposed that indicators of positive states may be found in both positive affect states and positive behaviours, such as behaviours that exemplify social engagement. When such positive behavioural states are undermined, it is plausible that people with dementia may express worries about how they get on with people or people not listening (item 21 and 23 in DEMQOL), or not being able to help or play a useful part in things (item 29 and 30 in DEMQOL-Proxy). The assessment of HRQL provided by DEMQOL-SF

and DEMQOL-Proxy-SF includes a consideration of positive states in terms of 'worries about social relationship' (SOC). Any decline in content validity with the omission of POS domain might not be pivotal. Maintaining a focus on SOC is also fundamental for the clinical relevance of HRQL assessment. As suggested by Lawton (1994), this is 'a treatment goal that seems appropriate for an illness whose manifestations in general appear to represent estrangement from the external world'.

3.6 Limitations

The conclusions drawn must be considered in light of a number of study limitations. Firstly, while the findings were based on fairly large samples of community-dwelling elderly, the extent to which they are representative of the general population of people with dementia in their respective countries is not clear. The generalisability of study findings was based primarily on an assessment of missing data rates, which were generally low for the Latin America countries as a group. While the rates were higher in the UK sample, there were only minor differences in clinical characteristics between study participants with complete/partial HRQL data and those for whom HRQL data was missing (see Chapter 2). Nonetheless, given that missing data often reflects the challenging nature of the phenomenon under study (X. Yang, Li, & Shoptaw, 2008), HRQL data may have been missing for those with more HRQL impairment.

The investigation of DIF due to dementia severity was based mostly on people with mild to moderate severity of dementia. Only a small minority had a diagnosis

of severe dementia in these community-dwelling samples. Though reference is made to people with more advanced illness, the results pertain largely to a comparison between people with mild or moderate dementia. As DIF effects due to dementia severity were largely absent in this study, it implies that illness progression, or different stages of illness, does not change item response behaviour in HRQL evaluations. However, the findings are based on cross-sectional data and more conclusive knowledge would require longitudinal studies with individuals who experienced illness progression.

Geographical region presented the most impediments to measurement invariance in DEMQOL and DEMQOL-Proxy. This may be due to differences between countries in availability of formal care and other intervention services, which may also explain the higher HRQL levels in the UK sample. It may also be a consequence of translation between English and Spanish version of the HRQL assessments (Teresi, 2006). The research presented here does not allow for further investigation of the complex interplay between socioeconomic factors (e.g. ethnicity, language, education). Nonetheless, these preliminary results highlighted a potential need to consider revisions in translation and/or statistical adjustment in research studies.

It has been shown that selecting a non-invariant item to be the reference indicator (i.e. factor loading on HRQL fixed at 1 for one item) can result in erroneous results in DIF detection (Johnson, Meade, & DuVernet, 2009). Here, item 9 (distress) in DEMQOL was selected to be the reference indicator for HRQL

because earlier analyses indicated that it loaded only on the general HRQL factor without showing additional covariation with other items. For consistency, item 7 (distress) in DEMQOL-Proxy was also selected to be the reference indicator for HRQL. As noted in the literature review of this chapter, gender differences are commonly reported for 'distress' and DIF due to gender may have a role in these differences. The sensitivity analyses employed a different reference indicator for HRQL. The item 'forget what day it is' was chosen in DEMQOL and DEMQOL-Proxy because it did not exhibit DIF effects in the primary analysis. The set of DIF effects detected in the re-analysis remained remarkably similar. No DIF effects emerged for the item 'distress' in DEMQOL and DEMQOL-Proxy. The original choice of reference indicator for HRQL had little impact on the reported results.

A MIMIC model approach with a forward stepwise model building strategy was used. Alternative strategies within the MIMIC model family (Wang & Shih, 2010; Woods, 2009) are available and may generate different conclusions. An important caveat is that MIMIC models can detect only uniform DIF. This refers to a scenario in which the focal group consistently reported higher (or lower) levels of a symptom despite being matched to their counterparts in the reference group at any level of HRQL. If the probability of response is higher/lower in the focal group only when HRQL is severely impaired, this would be an instance of non-uniform DIF since it varies according to the level of HRQL. While non-uniform DIF cannot be detected with MIMIC models, it may still surface as (uniform) DIF so long as it results in a shift in conditional probabilities of item responses

(Reininghaus et al., 2012; F. M. Yang et al., 2009). The correct identification of the specific type of DIF is challenging even with methods that can detect non-uniform DIF (Finch & French, 2008). In terms of uniform DIF, simulation studies have demonstrated that a MIMIC model approach compares favourably with established methods of DIF detection (Finch, 2005; Willse & Goodman, 2008; Woods, 2009). As such, MIMIC models are potentially useful as a first-stage detection when the initial concern is presence of any DIF (Finch, 2005; Reininghaus et al., 2012; F. M. Yang et al., 2009).

In this chapter, both DEMQOL-SF and DEMQOL-Proxy-SF measurement models showed acceptable fit with the sample data and stronger measurement invariance relative to their parent versions. Nonetheless, validation studies based on the actual short-form versions (see Appendices p. 247-248) , rather than extracting item data from their parent versions, are necessary (G. T. Smith, McCarthy, & Anderson, 2000).

CHAPTER 4 RESPONSE SHIFT

This chapter addresses the primary objective of investigating response shift in order to understand whether changes captured in longitudinal HRQL assessments have been influenced by adaptation processes as individuals cope with dementia. The research questions are:

(a) What was the impact of response shift (if any) on HRQL changes captured by longitudinal assessments with DEMQOL-SF or DEMQOL-Proxy-SF?

(b) What are the implications for utility assessment?

4.1 Measurement invariance across time

As people with dementia cope with a gradual loss in capacity for independent living, due to impairments in memory, reasoning and communication skills, their expectations, values, or definitions of HRQL may change over time. These psychological processes may be adaptive (or maladaptive). To explore such implications, an understanding of the underlying processes is required. With foundations in the field of educational intervention (Howard, Ralph, et al., 1979) and organisational change (Golembiewski et al., 1976), as noted above, response shift in health research has been operationalised by (Sprangers & Schwartz, 1999) as a change in appraisal responses due to three underlying processes:

(a) a change of internal standards (**re-calibration**);

(b) a change in perceived value or importance (**re-prioritisation**); or

(c) a change in perceived definition or meaning (**re-conceptualisation**).

Despite this distinction in the working definition, it has been acknowledged that these processes are likely to be intertwined (Schwartz & Sprangers, 1999). This implies that concurrent investigation of these processes is necessary. The methods proposed for investigating response shift vary in empirical substantiation as the majority are novel approaches or adapted from other assessment purposes (Schwartz & Sprangers, 1999). Most involve additional tasks on top of the usual HRQL assessment (e.g. then-test) and the more complex ones (e.g. card sort approach) have limited feasibility for the very old or very ill. Furthermore, multiple additional tasks are generally required as none of the approaches alone offers complete coverage of response shift. Given these challenges, investigations of response shift processes often rely on statistical approaches.

4.2 A psychometric typology of change

Expanding on early statistical paradigms for investigating response shift phenomenon (Millsap & Hartog, 1988; Schmitt, 1982), Oort (2005) re-expressed the working definition from Sprangers and Schwartz (1999) in the context of latent variable modelling using the structural equation modelling (SEM) framework of measurement invariance across multi-wave factor models.

HRQL at each assessment occasion (or wave) is first represented by measurement models with identical themes. This is reflected by an invariant pattern of factor loadings which shows that the same DEMQOL-SF or DEMQOL-Proxy-SF items can be grouped under identical themes of HRQL regardless of assessment

occasion. Termed as 'configural invariance' (Horn & McArdle, 1992), this form of measurement invariance is an indication that the HRQL concept is stable over time. Re-conceptualisation would result in longitudinal differences in the pattern of factor loadings (Oort, 2005).

Given a stable HRQL concept, inquiry can proceed to examine the relative importance of HRQL elements at each assessment occasion. This can be inferred from the size of factor loadings which show how 'indicative' DEMQOL-SF or DEMQOL-Proxy-SF items are of their designated themes (Oort, 2005). An invariant set of factor loadings across measurement occasions shows that no item has become more, or less, important in discriminating individual differences in HRQL. Termed as 'metric invariance' (Horn & McArdle, 1992), this stricter form of measurement invariance is an indication that the unit of measurement is identical over time, due to a stable order of priorities. Otherwise, re-prioritisation would result in longitudinal differences in the size of factor loadings (Oort, 2005).

Within the same set of priorities, inquiry can proceed to examine the expectations with which HRQL is evaluated at each assessment occasion. This is inferred from item intercepts, which are conceptually analogous to the level of item difficulty in item response theory (IRT) framework. A 'difficult' item is one that requires individuals to have high levels of HRQL in order to achieve a high score. An 'easy' item is one where individuals are likely to get a high score even at low levels of HRQL. An invariant set of item intercepts shows that no element of HRQL has become 'easier' or more 'difficult' over time. Termed as 'scalar

invariance' (Meredith, 1993), this form of measurement invariance is an indication that the measurement origins are identical over time, due to a stable set of internal standards. Otherwise, re-calibration of internal standards would result in longitudinal differences in item intercepts (Oort, 2005).

Taken together, even in the presence of re-prioritisation and/or re-calibration, these longitudinal SEM models provide an estimate of change in HRQL that could be attributed to the putative intervention. This is referred to as 'true change' in the statistical model, based on a difference between the means of latent variables that represent HRQL at each wave. Unlike raw score differences, latent score differences offer an adjusted estimate of HRQL change in the presence of response shift. It is nonetheless worth noting that, in the absence of response shift (i.e. if there is scalar invariance), raw score differences can provide reasonable estimates of HRQL changes if score reliabilities in each wave are also invariant. Stated differently, when the amount of measurement error is identical across assessment occasions, the assessment of change can be based on raw score differences (Marsh, Scalas, & Nagengast, 2010). In latent variable models, measurement error is captured by residual variances that are unique to each DEMQOL-SF or DEMQOL-Proxy-SF item. Consequently, the assessment of latent score changes has been adjusted for measurement error, regardless of whether they are invariant across assessment occasions. The assessment of raw score changes, on the other hand, would require longitudinal invariance of factor loadings (i.e. concept and priorities), intercepts (i.e. internal standards), and residuals (i.e. measurement error).

This study is an exploration of the response shift phenomenon in dementia. Very little is known about how people adapt to the chronic and challenging circumstances of living with dementia. An observational cohort of memory clinic patients provided the first dataset for this investigation. This was repeated in a randomised controlled trial study sample. Response shift (if any) was examined with a particular focus on preference-based items since changes in item response behaviour also affect the utility weights assigned for calculating the eventual estimate of utility values.

4.3 Methods

4.3.1 Participants

The HRQL data for this study came from two samples. The first comprised community-dwelling individuals, and their carers, referred to the Croydon Memory Service for early assessment and intervention in dementia based in South London. The sample consisted of referrals made between December 2002 and June 2010 who, after a full multidisciplinary assessment, were given a formal clinical diagnosis of dementia using ICD-10 criteria (Banerjee et al., 2007). HRQL assessments were obtained at an initial visit from self- and proxy reports. Follow ups were scheduled at 6- and 12-months after the initial visit, and annually thereafter. Only individuals who attended the clinic for at least a year after initial diagnosis were included. The current investigation focussed on assessment data from the first three waves (baseline, 6- and 12-month). No ethical committee approval was needed as this study was a secondary analysis on de-identified data.

The second study sample comprised community-dwelling elderly individuals, and their carers, referred by old age psychiatry services in UK into HTA-SADD, a randomised, multicentre, double-blind, placebo-controlled trial, funded by the UK National Institute of Health Research (NIHR), on the use of antidepressants for depression in dementia (Banerjee et al., 2011). Trial participants met National Institute of Neurological and Communicative Diseases and Stroke (NINCDS)–Alzheimer's Disease and Related Disorders Association (ADRDA) criteria for probable or possible Alzheimer's disease, and had co-existing depression that was assessed as potentially needing antidepressants. None were clinically critical (e.g. suicide risk), contraindicated to study drugs, on antidepressants, in another trial, all had carer to provide data. HRQL assessments obtained from self- and proxy reports at baseline, 13-, and 39-weeks formed the basis of the response shift investigations. The primary results of the RCT have been published (Banerjee et al., 2011), the current study is secondary data analysis and did not require separate ethical committee approval.

4.3.2 Measures

The DEMQOL-SF (17 items) and DEMQOL-Proxy-SF (17 items) are short-form versions of DEMQOL (28 items) and DEMQOL-Proxy (31 items) respectively. They were developed using DEMQOL and DEMQOL-Proxy CFA models from which IRT results were obtained to further study item response patterns in terms of the amount of information each item provided for discriminating individual differences, and the level of HRQL at which they were the most informative

(Chapter 3). Based on this, a smaller set of items was selected for DEMQOL-SF and DEMQOL-Proxy-SF, retaining similar levels and range of sensitivity to their parent versions. As the current investigation extracted the short-form versions from DEMQOL and DEMQOL-Proxy respectively, the original item numbers are retained in this thesis for ease of reference.

From the 17 questions on general HRQL, both DEMQOL-SF and DEMQOL-Proxy-SF have 12 items that cover the theme of negative emotions (NEG: 4 items), worries about cognitive functioning (COG: 4 items), or worries about social functioning (SOC: 4 items). Preference-based algorithms have been developed for 5 items (DEMQOL-U) from DEMQOL and 4 items (DEMQOL-Proxy-U) from DEMQOL-Proxy for economic evaluation purposes. These preference-based items are also present in DEMQOL-SF and DEMQOL-Proxy-SF. Like their parent versions, both are interviewer-administered assessments with responses obtained on a four-point Likert scale (1 = a lot; 2 = quite a bit; 3 = a little; 4 = not at all) and coded so that higher total scores reflect better HRQL.

While HRQL data was the focus, a number of other clinical assessments were also available for a broader sample description. These included: Mini-Mental State Examination (MMSE, Folstein et al., 1975), Geriatric Depression Scale (GDS, Yesavage & Sheikh, 1986), Neuropsychiatric Inventory (NPI, Cummings et al., 1994), Bristol Activities of Daily Living Scale (BADL, Bucks et al., 1996), Zarit Burden Interview (Zarit, Zarit et al., 1980), and the General Health Questionnaire (GHQ, Goldberg & Williams, 1988). The MMSE, GDS, NPI, and BADL

assessments focussed on the health of people with dementia, whereas Zarit and GHQ were targeted at the health of family carers who were also the source of proxy HRQL reports. The HTA-SADD trial conducted similar assessments with the MMSE, NPI, and BADL. In addition, the Cornell Scale for Depression in Dementia (CSDD, Alexopoulos, Abrams, Young, & Shamoian, 1988) was used to screen for trial eligibility and as the primary outcome.

4.4 Analysis

4.4.1 Model estimation

Latent variable modelling was completed using Mplus version 7. Since there are four categories on the Likert response scale, it was appropriate to treat DEMQOL-SF and DEMQOL-Proxy-SF item responses as ordered-categorical data (Rhemtulla et al., 2012). Model estimation employed robust weighted least squares with means and variances adjustment (WLSMV) as is recommended (Flora & Curran, 2004; B. Muthén et al., 1997; Savalei & Rhemtulla, 2013). This approach uses a multivariate probit regression model to predict how probabilities of (ordered-categorical) item responses are related to (continuous) latent variables that represent levels of HRQL and its domains.

4.4.2 Model evaluation

As a fundamental basis for making interpretations, empirical fit between model predictions and the observed data must be adequate. Overall (i.e. omnibus) model fit was evaluated statistically using robust model Chi square (χ^2_{M}). An exact fit

between model predictions and sample data, within bounds of sampling error, would result in a non-statistically significant χ^2_M value. In the absence of exact fit, the extent of approximate fit remains of interest. For this evaluation, Mplus provides four descriptive indices that offer a non-statistical summary of model fit for CFAs on ordered-categorical data. Based on commonly adopted standards, root mean square error of approximation (RMSEA, Steiger, 1990) values should be low (<0.10 for acceptable fit, <0.05 for very good fit), while comparative fit index (CFI, Bentler, 1990), Tucker Lewis index (TLI, Tucker & Lewis, 1973) values should be high (>0.90 for acceptable fit, >0.95 for very good fit) when approximate fit is adequate. These standards were drawn from extensive simulation studies with continuous data and their relevance for ordered-categorical data remains an area of active inquiry (Cook et al., 2009; Marsh, 2004; Marsh et al., 2013; West et al., 1995). The weighted root mean square residual (WRMR, Yu, 2002) has been developed for ordered-categorical data and while a value of less than one has been recommended, it remains to be established for a wider range of simulations (e.g. bifactor model).

4.4.3 Measurement model

Using the findings from Chapter 3 (see Figure 3.11 and 3.12), bifactor CFA models were hypothesised to explain the response patterns on DEMQOL-SF and DEMQOL-Proxy-SF respectively: (a) 17 items had a non-zero loading on the general factor (HRQL); (b) each item, except for five, also had non-zero loading on a group factor representing the domain of negative emotion (NEG, 4 items),

worries about cognitive functioning (COG, 4 items), or worries about social relationship (SOC, 4 items), but zero loading on the other domain factors; (c) all latent (general / domain) factors were orthogonal to one another; (d) residual variance of each item were orthogonal to one another. To identify the model, one of the factor loadings on the general factor and one on each domain factor were fixed at the value of one. All latent factor variances were freely estimated from the sample data.

4.4.4 Factor collapse in measurement model

Before testing for longitudinal configural invariance in a multi-wave SEM model, the measurement model at each wave was explored individually. As noted in Chapter 2, alongside a general HRQL latent factor, factor collapse of a HRQL domain is a plausible event. In other words, a hypothesised HRQL domain might not exist and the items have non-zero factor loadings only on the general HRQL factor. The prospect of factor collapse meant that measurement models of each wave might not be identical due to the absence of certain HRQL domains at a particular assessment occasion. Hypothesising an identical bifactor CFA model for every wave in a longitudinal SEM model might then result in poor model fit. In this instance, a lack of longitudinal configural invariance is not a result of HRQL being re-conceptualised. Factor collapse of a HRQL domain is an indication that this 'non-existent' domain factor is at the heart of the general HRQL factor. If the domain that held core relevance to general HRQL was different between assessment occasions, this would be re-prioritisation rather than

re-conceptualisation. This is consistent with Oort (2005)'s psychometric typology as factor collapse would result in stronger factor loadings on the general HRQL factor since the items in question would no longer load on another HRQL domain factor. In the bifactor models, the lack of configural invariance might be due to re-conceptualisation or re-prioritisation. Exploring the measurement models individually would guard against mis-attribution.

4.4.5 Longitudinal SEM model

The longitudinal SEM models for DEMQOL-SF and DEMQOL-Proxy-SF each had three waves of bifactor CFA models. In the observational cohort from Croydon Memory Service, these waves referred to the baseline, 6 and 12 months HRQL assessment time points. In the HTA-SADD trial sample, they referred to baseline HRQL assessment, and 13 and 39 weeks (or approximately 3 and 9 months) later. To account for plausible associations between HRQL reports provided by the same individual, correlations were hypothesised between general HRQL of each wave. The respective HRQL domains were correlated in the same fashion. Besides these across-occasion latent factor correlations, the residual variance (or measurement error) of each item was also correlated with that of the same item from the other two assessment occasions (see Figure 4.1).

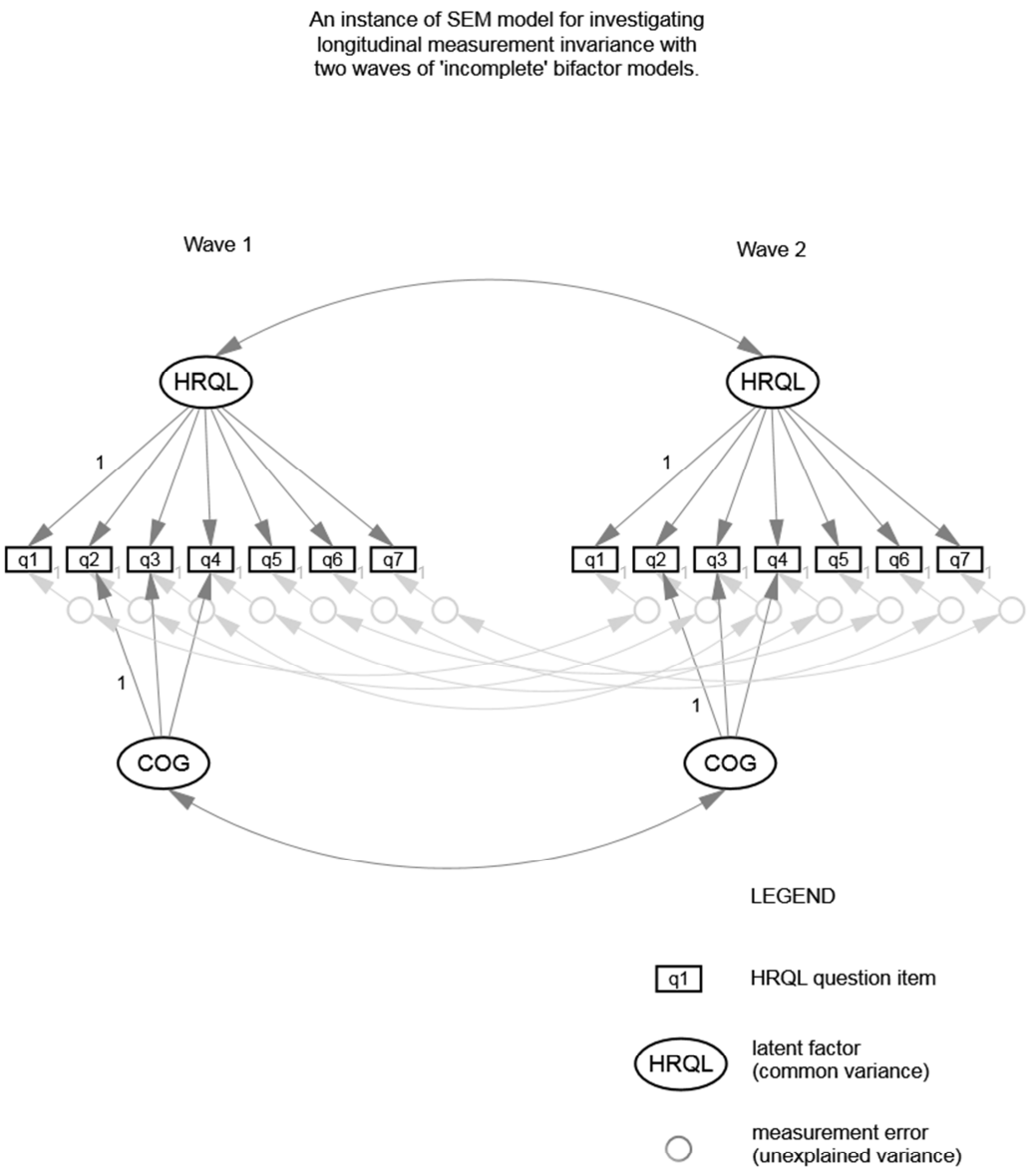

Figure 4-1 A longitudinal SEM model for two waves of HRQL assessment

4.4.6 Modelling issues in Mplus

Before an operational account of the longitudinal modelling, two issues deserve mention. First, with ordered-categorical item data, the concept of item intercept (in continuous data) is replaced by item thresholds which refer to the level of difficulty of achieving the next higher score on a Likert scale (e.g. an item with four response categories has three thresholds). In this context, item response probabilities are influenced concurrently by factor loadings and item thresholds (L. Muthén & Muthén, 1998-2012, p. 485). While it is technically possible in some cases to tease apart their influences, it is not so when items load on more than one latent factor (B. Muthén, 2013, pp. 9-10), as in the case of bifactor models in this investigation. Consequently, both factor loadings and item thresholds need to be studied in tandem (Rosen, Beron, & Underwood, 2013).

Second, given the focus on modelling residual variance and covariance in the longitudinal SEM models, the default parameterisation setting in Mplus was switched from Delta to Theta (B. Muthén & Asparouhov, 2002; L. Muthén & Muthén, 1998-2012, p. 605). In Theta parameterisation, all residual variances are fixed at one by default (L. Muthén & Muthén, 1998-2012, p. 461). The implications of both of these issues on model specifications will be discussed in the following sections.

4.4.7 Longitudinal configural invariance

Response shift investigation began with a SEM model in which an identical CFA measurement model was hypothesised for every wave of HRQL assessments. This

was labelled as Model 1 for DEMQOL-SF and DEMQOL-Proxy-SF respectively and investigated with data from both the memory service and clinical trial. All factor loadings and thresholds were freely estimated from the sample data to see whether they exhibited an invariant pattern for every wave. At this stage, longitudinal changes were not investigated since response shift might have altered the meaning, values, or expectations of HRQL. All latent factor means were therefore fixed at zero and their variances fixed at one (Mplus syntax enclosed in Appendices p. 249 - 252). Given tenable model fit (see model evaluation in section 4.5.2), further support for the hypothesis of longitudinal configural invariance was determined from the presence of sizable factor loadings on the general HRQL factor in every wave. Otherwise, differences might emerge from the pattern of factor loadings that attained statistical significance only on some assessment occasions, depending on how HRQL might have been re-conceptualised over time.

4.4.8 Longitudinal scalar invariance

When a stable factor loading pattern was found, the next stage proceeded to investigate a stricter form of invariance by further constraining factor loadings to be equal across assessment occasions (i.e. metric invariance). However, as both factor loadings and item thresholds had to be studied in tandem, metric invariance could not be investigated directly. Instead the longitudinal SEM model was modified by constraining both factor loadings and item thresholds to be equal across assessment occasions to test the hypothesis of scalar invariance (Model 2).

In Theta parameterisation, all residual variances are fixed at one by default (L. Muthén & Muthén, 1998-2012, p. 461). Hence the amount of measurement error was identical across assessment occasions, which constituted an even more stringent hypothesis of measurement invariance than was intended in Model 2. To retain focus on scalar invariance hypothesis, the constraints on residual variances were relaxed so that they were freely estimated from the data. Specifically, measurement errors in second and third assessment occasion were allowed to vary from baseline (where residual variances remained fixed at the value of one to achieve model identification). This decision was also in line with the logic that measurement invariance hypotheses should be tested in an increasingly stringent order (Vandenberg & Lance, 2000).

Given that the concept, unit and origin of measurement were hypothesised to be invariant across assessment occasions, Model 2 provided a basis for comparing latent factor means to assess longitudinal changes. To reflect these plausible longitudinal changes, factor means and variances were freely estimated in Model 2 for the second and third assessment occasions. Factor means and variances at baseline remained fixed at the value of 0 and 1 respectively (as in Model 1), so as to achieve model identification. Taken together, these constraints in Model 2 allowed the latent factor means (e.g. general HRQL) at second and third assessment occasion to be interpreted as longitudinal changes relative to baseline (Mplus syntax enclosed in Appendices p. 253 – 260).

Given tenable model fit (see model evaluation in section 4.5.2), further support for the hypothesis of scalar invariance was drawn from model comparisons. Based on the set of assumptions in Model 1 (e.g. pattern of factor loadings), Model 2 imposed further restrictions (e.g. across-occasion equality constraints on factor loadings) and hence had fewer model parameters that were freely estimated from the sample data. Model 2 as such was nested in Model 1. A decline in fit with the sample data was expected for Model 2 because it offered a less complex (or more parsimonious) explanation of the data, and hence held more scientific falsifiability (i.e. more degrees of freedom) than Model 1. Model comparisons were made to determine whether this decline in model fit might be statistically significant. For models estimated with WLSMV, the DIFFTEST option in Mplus was required for model comparisons so as to obtain the correct chi square difference test between models (L. Muthén & Muthén, 1998-2012, pp. 451-452). Given inconsequential decline in model fit (i.e. DIFFTEST not statistically significant), a less complex explanation of the data (Model 2) was favoured.

4.4.9 Longitudinal invariance of measurement errors

When scalar invariance was found to be tenable, the next model imposed further restrictions by re-invoking the constraints on residual variances as in the Theta parameterisation defaults such that factor loadings, thresholds and residual variances were hypothesised to be equal across assessment occasions (Model 3). Given tenable model fit (see model evaluation in section 4.5.2), further support for this hypothesis was drawn from comparisons with Model 1. As in the

preceding instance, Model 3 was nested in Model 1. The DIFFTEST results showed that the costs of less model complexity in Model 3 (i.e. poorer empirical fit relative to Model 1) were mostly inconsequential.

When this was not the case, modification indices were examined for guidance on how to re-specify Model 3 to improve model fit. Modification indices are derivatives of the model chi square (χ^2_M) which show an expected improvement in model fit if Model 3 assumptions were relaxed. The removal of untenable restrictions (i.e. allowing parameters to be freely estimated instead of being fixed equal across assessment waves) proceeded progressively with further model comparisons made for each re-specification of Model 3. At the conclusion of these iterative steps, the extent of longitudinal measurement invariance was determined based on the final model (Model 4).

4.4.10 Response shift and longitudinal estimates of change in HRQL

The impact of response shift was assessed by comparing conclusions about HRQL changes based on models where scalar invariance was held tenable (Model 2 and 3) and models where measurement invariance was undermined by response shift. In the presence of response shift, Model 4 would offer a statistical estimate of 'true' change adjusted for any re-prioritisation, re-calibration, and/or time-varying measurement error. However, as no response shift was detected in this study, Model 4 was expected to produce similar substantive conclusions about longitudinal HRQL changes as Model 2 and 3.

Particular attention was paid to items 1 (cheerful), 4 (frustrated), 8 (lonely), 14 (forget things that happened recently), and 24 (making yourself understood) in DEMQOL-SF, as well as item 3 (frustrated), 8 (lively), 17 (forget what day it is), and 22 (keeping him/herself looking nice) in DEMQOL-Proxy-SF. These are the preference-based items that make up DEMQOL-U and DEMQOL-Proxy-U respectively (Mulhern et al., 2013). In addressing validity threats to the measurement of change in utility values, the absence of response shift raised no concerns.

4.5 Results

4.5.1 Sample characteristics

Table 4.1 and 4.2 present the demographic and clinical characteristics of study participants from the Croydon Memory Service cohort (n=468) and HTA-SADD trial sample (n=326) respectively. The number of HRQL reports differed between each wave of assessment. However, as the estimation algorithm in Mplus used all available data, individuals were included in the analysis if they had HRQL data for at least one assessment occasion. Consequently, the observational cohort had a sample size of 432 (36 excluded) for DEMQOL-SF and 407 (61 excluded) for DEMQOL-Proxy-SF. In the clinical trial sample, this was 306 (20 excluded) and 324 (2 excluded) respectively.

Based on clinical assessments that were common in both study samples, individuals from the HTA-SADD trial generally had more impaired health than those attending the Croydon Memory Service clinic. Given the population, this

was to be expected. Within each study sample, individuals who did not have a HRQL assessment for any wave (labelled as 'Miss' in Table 4.1 and 4.2) tended to have more impaired health due to neuropsychiatric symptoms (as assessed by NPI) and loss of skills needed in activities of daily living (as assessed by BADL). This trend was also observed in subsequent assessment waves. The data from both samples were not missing completely at random (MCAR).

Table 4.1 Croydon Memory Service observational cohort baseline data

		DEMQOL			DEMQOL-Proxy		
	Overall	Full	Partial	Miss	Full	Partial	Miss
Participants	468	133	299	36	98	309	61
Age	76.9 (9.2)	75.2 (9.4)	77.8 (8.9)	75.8 (9.4)	77.4 (8.0)	76.8 (9.2)	76.4 (10.6)
Gender							
Female	305	85	196	24	67	194	44
Male	163	48	103	12	31	115	17
Ethnicity							
White	408	115	264	29	92	267	49
Others	60	18	35	7	6	42	12
ICD-10 **							
AD	292	73	201	18	57	199	36
AD mixed	110	33	64	13	23	71	16
Vascular	18	4	10	4	5	12	1
Others	6	1	4	1	0	6	0
Unknown	42	22	20	0	0	1	0
MMSE	22.0 (4.7)	22.4 (4.7)	21.9 (4.7)	21.3 (5.4)	21.6 (5.3)	22.0 (4.5)	22.4 (4.9)
GDS *	2.9 (2.7) n=406	3.1 (2.7) n=113	2.8 (2.6) n=266	3.3 (3.8) n=27	3.0 (2.6) n=80	2.9 (2.7) n=276	3.0 (2.7) n=50
NPI *	11.2 (11.8) n=417	10.4 (13.0) n=118	11.2 (11.4) n=272	14.7 (10.1) n=27	11.1 (12.1) n=98	11.3 (12.0) n=293	10.9 (8.4) n=26
BADL *	7.2 (7.3) n=421	6.7 (7.2) n=117	7.1 (7.3) n=277	10.7 (7.2) n=27	7.6 (7.7) n=97	7.0 (7.0) n=297	8.8 (8.1) n=27
Zarit *	21.4 (16.1) n=351	19.8 (16.6) n=93	21.0 (15.3) n=234	31.5 (18.3) n=24	19.6 (15.6) n=79	21.9 (16.5) n=252	22.1 (12.2) n=20
GHQ *	3.7 (5.1) n=342	3.7 (5.2) n=87	3.6 (5.1) n=232	4.9 (4.9) n=23	2.9 (3.8) n=75	4.0 (5.4) n=249	3.1 (4.3) n=18

* Sample average with standard deviation in parentheses. As rate of missing data varies across variables, valid sample size (n) is reported.

** ICD-10 diagnosis: Alzheimer's Disease, late/early onset (AD), Alzheimer's Disease, mixed type (AD mixed), Vascular dementia (Vascular), Others / Unspecified (Others), ICD code not known (Unknown).

Table 4.2 HTA-SADD trial sample baseline data

		DEMQOL			DEMQOL-Proxy		
	Overall	Full	Partial	Miss	Full	Partial	Miss
Participants	326	171	135	20	209	115	2
Age	79.4 (8.5)	78.8 (8.5)	80.1 (8.5)	79.2 (11.6)	78.6 (8.4)	80.9 (8.6)	74.2 (1.8)
Gender							
Female	221	119	89	13	143	77	1
Male	105	52	46	7	66	38	1
Ethnicity							
White	303	160	127	16	191	110	2
Others	31	11	8	4	18	5	0
CSDD	13.0 (4.2)	12.8 (4.1)	13.2 (4.5)	13.4 (3.6)	12.8 (4.2)	13.3 (4.4)	9.5 (0.7)
MMSE *	18.1 (6.7) n=251	19.6 (6.1) n=156	16.2 (6.7) n=90	5.4 (4.8) n=5	18.1 (6.9) n=171	18.2 (6.1) n=79	-
NPI *	29.0 (18.6) n=318	26.9 (16.9) n=168	30.2 (19.7) n=130	39.4 (21.3) n=20	28.5 (17.4) n=205	29.8 (20.7) n=111	33.5 (13.4) n=2
BADL *	17.7 (11.0) n=324	14.5 (9.8) n=171	20.4 (10.7) n=134	28.2 (13.0) n=19	16.5 (10.7) n=208	19.8 (11.2) n=115	41 (-) n=1

***** Sample average with standard deviation in parentheses. As rate of missing data varies across variables, valid sample size (n) is reported.

4.5.2 Measurement model

Among the six bifactor CFA models (3 waves x 2 data sets) for DEMQOL-SF and DEMQOL-Proxy-SF respectively, there were indications of factor collapse across all 12 models. Groups of items hypothesised for particular HRQL domains did not share additional common variance above that due to general HRQL factor. This is a statistical indication that the hypothesised domains did not carry additional information beyond what they conveyed about individual differences in general HRQL.

Among the six DEMQOL-SF bifactor models, the domain of 'negative emotion' (NEG) and 'worries about social functioning' (SOC) had factor variances and/or factor loadings that were not statistically significant or only marginally significant, indicative of factor collapse. These model anomalies were observed across all waves in both the Croydon and HTA-SADD trial sample. The consistency of these indications provided a plausible basis for hypothesising that NEG and SOC domains were at the heart of HRQL concept in DEMQOL-SF.

Among the six DEMQOL-Proxy-SF bifactor models, the domain of 'worries about social functioning' (SOC) displayed similar sets of model anomalies or led to an under-identified model. These indications were also consistent across all waves in both study samples, so providing a plausible basis for concluding that SOC was at the heart of HRQL concept in DEMQOL-Proxy-SF. An issue was noted with item 14 of the domain of 'worries about cognitive function' (COG). In at least one assessment occasion for both study samples, the factor loading of this

item had unusually large values and standard errors and/or was not statistically significant on both the general HRQL and COG domain factor. The other three items (12, 17, 18) in this domain did not display this problem and factor variance results provided no apparent indication of factor collapse. Given that identical models at other assessment occasions converged with admissible values, this might be the data-related problem known as 'empirical under-identification' (Kline, 2011, pp. 146-147). Measurement models where items load on more than one latent factor (e.g. bifactor models) are known to be susceptible to this modelling problem, even if the model is apparently over-identified and model fit adequate (Marsh, 1989; Marsh & Bailey, 1991). Given that these data were from DEMQOL-Proxy, this issue might be present only in the item response patterns of the parent version. When the actual DEMQOL-Proxy-SF is administered, the item response patterns may emerge differently. As the specific causes can be difficult to diagnose, a practical solution is to fix the factor loadings at some plausible values. This strategy was implemented by taking factor loadings from models that converged with admissible values, thereby circumventing the issue with empirical under-identification.

4.5.3 Longitudinal configural invariance

Having explored the measurement models individually, three waves of 'incomplete' bifactor CFA models were hypothesised in a longitudinal SEM model. For the DEMQOL-SF, each wave of 'incomplete' bifactor CFA model had 17 items loading on a general HRQL factor and only four (item 14, 17, 18, 19)

had additional loadings on a domain factor for worries about cognitive functioning (COG). For the DEMQOL-Proxy-SF, a similar 'incomplete' bifactor CFA model was hypothesised with a general HRQL factor and two additional domain factors. Besides loading on the general HRQL factor, four items (item 12, 14, 17, 18) had additional loadings on the domain factor of 'worries about cognitive function' (COG), and another four (item 3, 5, 7, 10) had additional loadings on the domain of 'negative emotions' (NEG). The pattern of factor loadings was hypothesised for each assessment occasion was identical (i.e. longitudinal configural invariance).

Table 4.3 summarises model fit information for the longitudinal SEM models. Neither DEMQOL-SF nor DEMQOL-Proxy-SF models of longitudinal configural invariance (Model 1) predicted a pattern of inter-item correlations that had an exact match (within bounds of sampling error) with those found in the data of the memory service and clinical trial sample. In terms of approximate fit, RMSEA values indicated that the discrepancy between actual and predicted covariances per degree of freedom was not unacceptably large. Both CFI and TLI values also indicated that there were acceptable amounts of improvement in empirical fit when compared to the 'worst model' (Miles & Shevlin, 2007) in which none of the elements is correlated. Similarly, WRMR values indicated that the average discrepancy between the observed and predicted correlation matrices was at acceptable levels. Across the array of approximate fit indices, model-data correspondence at large was adequate for drawing conclusions with the model results.

Table 4.3 Model fit information for longitudinal SEM models.

DEMQOL-SF Croydon (n=432)	df	χ^2_M	RMSEA (90% CI)	CFI	TLI	WRMR	DIFFTEST
Model 1 (par = 273) Configural invariance	1155	1702.659	.033 (.030 – .036)	.924	.916	1.211	-
Model 2 (par = 171) Scalar invariance	1257	1767.239	.031 (.027 – .034)	.929	.928	1.245	105.749 df=102, p>.05
Model 3 (par = 137) Strict invariance	1291	1742.639	.028 (.025 – .032)	.937	.938	1.267	130.930 df=136, p>.05
DEMQOL-SF HTA-SADD (n=306)	df	χ^2_M	RMSEA (90% CI)	CFI	TLI	WRMR	DIFFTEST
Model 1 (par = 273) Configural invariance	1155	1566.929	.034 (.030 – .038)	.935	.928	1.117	-
Model 2 (par = 171) Scalar invariance	1257	1655.475	.032 (.028 – .036)	.937	.936	1.162	122.362 df=102, p>.05
Model 3 (par = 137) Strict invariance	1291	1665.534	.031 (.026 – .035)	.941	.942	1.214	165.673 df=136, p=.04
Model 4 (par = 138) Item 23 λ	1290	1656.304	.030 (.026 – .035)	.942	.943	1.202	155.621 df=135, p>.05
Model 4 (par = 138) Item 12 ξ	1290	1652.838	.030 (.026 – .035)	.943	.943	1.203	153.337 df=135, p>.05
DEMQOL-Proxy-SF Croydon (n=407)	df	χ^2_M	RMSEA (90% CI)	CFI	TLI	WRMR	DIFFTEST
Model 1 (par = 286) Configural invariance	1142	1365.314	.022 (.017 – .026)	.965	.961	.987	-
Model 2 (par = 182) Scalar invariance	1246	1461.466	.021 (.016 – .025)	.966	.965	1.030	111.008 df=104, p>.05
Model 3 (par = 148) Strict invariance	1280	1484.493	.020 (.015 – .024)	.968	.968	1.061	147.477 df=138, p>.05
DEMQOL-Proxy-SF HTA-SADD (n=324)	df	χ^2_M	RMSEA (90% CI)	CFI	TLI	WRMR	DIFFTEST
Model 1 (par = 286) Configural invariance	1142	1352.476	.024 (.018 – .029)	.972	.969	.925	-
Model 2 (par = 182) Scalar invariance	1246	1447.454	.022 (.017 – .027)	.973	.972	.972	109.458 df=104, p>.05
Model 3 (par = 148) Strict invariance	1280	1476.224	.022 (.016 – .027)	.974	.974	1.009	150.382 df=138, p>.05

Par: number of freely estimated parameters.

All χ^2_M were statistically significant. Based on commonly adopted standards, RMSEA values should be low (<0.10 for acceptable fit, <0.05 for very good fit), while CFI and TLI values should be high (>0.90 for acceptable fit, >0.95 for very good fit) when approximate fit is adequate.

All DIFFTEST refer to model comparison with Model 1. For DEMQOL-SF in HTA-SADD trial, Model 4 had either:
(i) item 23 factor loading (λ) freely estimated at third occasion; or
(ii) item 12 residual variance (ξ) freely estimated at baseline occasion.

Table 4.4 DEMQOL-SF unstandardised factor loadings in multi-wave SEM model (Model 1)

Time	CMS 1		CMS 2		CMS 3		HTA 1		HTA 2		HTA 3	
Item	GEN	DM	GEN	DM	GEN	DM	GEN	DM	GEN	DM	GEN	DM
1	.41		.36		.50		.51		.39		.59	
2	.82		.83		.94		.84		.88		1.04	
4	.91		.80		1.12		.71		.99		.85	
7	.93		1.12		.91		.90		.97		.86	
8	.59		.63		.57		.70		.71		.83	
11	.96		.90		1.09		.74		.78		.90	
12	1.03		1.00		1.04		.79		1.15		1.40	
13	.66		.53		.75		.59		.60		.96	
14	.83	.80 $^{\Delta}$	.99	.82 $^{\Delta}$	1.00	.89 $^{\Delta}$	.72	.66 $^{\Delta}$	.99	.89 $^{\Delta}$	.93	1.12 $^{\Delta}$
17	1.04	1.08 $^{\Delta}$	1.07	1.20 $^{\Delta}$	1.17	1.00 $^{\Delta}$	1.28	1.24 $^{\Delta}$	1.30	1.14 $^{\Delta}$	1.15	1.08 $^{\Delta}$
18	1.08	.49 $^{\Delta}$	1.00	.80 $^{\Delta}$	1.25	.48 $^{\Delta}$	1.26	.99 $^{\Delta}$	.96	.76 $^{\Delta}$	.71	.71 $^{\Delta}$
19	1.12	1.05 $^{\Delta}$	.94	.97 $^{\Delta}$	1.12	.81 $^{\Delta}$	1.33	1.25 $^{\Delta}$	1.28	1.22 $^{\Delta}$	1.27	1.18 $^{\Delta}$
21	.94		.89		1.28		1.13		.92		.72	
22	1.12		1.25		1.00		1.03		.85		.73	
23	1.02		.82		.95		.74		.67		.36	
24	1.06		1.00		.95		.95		.86		.62	
28	.77		.68		.76		.98		1.08		.79	

CMS: Croydon Memory Service, HTA: HTA SADD trial

GEN: general HRQL; DM: Domain factor, where Δ marks the items for 'worries about cognitive function' (COG).

Table 4.5 DEMQOL-P-SF unstandardised factor loadings in multi-wave SEM model (Model 1)

Time	CMS 1		CMS 2		CMS 3		HTA 1		HTA 2		HTA 3	
Item	GEN	DM	GEN	DM	GEN	DM	GEN	DM	GEN	DM	GEN	DM
3	.65	.60 #	.77	.96 #	1.01	1.03 #	.40	.73 #	.57	.70 #	.90	1.04 #
5	1.21	1.01#	1.36	1.04 #	.98	.92 #	.65	1.35 #	.52	1.06 #	.56	1.12 #
7	1.03	.67 #	1.07	.60 #	1.29	1.12 #	.50	1.21 #	.58	.94 #	.77	.98 #
8	.28		.29		.34		-.10 ns		-.01 ns		-.14 ns	
10	1.20	.77 #	.96	.81 #	1.21	1.42 #	.42	.96 #	.88	1.48 #	.49	1.38 #
12	.77	1.01 Δ	.88	.98 Δ	.77	.70 Δ	1.16	.75 Δ	1.08	.58 Δ	1.16	.82 Δ
14	1.50	1.80 Δ	1.68	1.96 Δ	1.28	1.76 Δ	1.35	1.26 Δ	1.47	1.15 Δ	1.40	1.30 Δ
17	.71	.34 Δ	.81	.62 Δ	.63	.51 Δ	.99	1.10 Δ	1.03	.98 Δ	.84	.86 Δ
18	1.09	.50 Δ	1.39	.94 Δ	1.05	.60 Δ	1.32	.93 Δ	1.59	1.66 Δ	1.16	.91 Δ
22	.60		.85		.94		.52		.51		.56	
25	.83		.91		.66		.69		.63		.69	
26	.91		.96		.68		.84		.93		1.00	
27	.90		1.22		1.08		.73		.79		.81	
28	.90		.76		.81		.71		.85		1.02	
29	.94		.88		1.21		.98		1.09		1.27	
30	.92		1.20		.99		1.27		1.45		1.53	
31	.51		.74		.54		.72		.84		1.01	

CMS: Croydon Memory Service, HTA: HTA SADD trial

GEN: general HRQL; DM: Domain factor, where # marks the items for 'negative emotion' (NEG), and Δ marks the items for 'worries about cognitive function' (COG)

ns: factor loading not statistically significant.

Item 14 factor loadings were fixed for GEN (1.50) and COG (1.80) at baseline of Croydon Memory Service (CMS 1); and for GEN (1.40) and COG (1.30) at third wave of HRQL assessment in HTA-SADD trial (HTA 3).

In both Croydon Memory Service and HTA-SADD trial samples, DEMQOL-SF items loaded well on a general HRQL factor, as well as a domain factor labelled 'worries about cognitive function' (COG), in every wave of HRQL assessments (Table 4.4). Similarly, DEMQOL-Proxy-SF items loaded well on a general HRQL factor, as well as two domain factors labelled 'negative emotion' (NEG) and 'worries about cognitive function' (COG), in every wave of HRQL assessments for both study samples (Table 4.5). The factor loadings were generally similar across repeated assessments, with no apparent sign of re-conceptualisation which would have altered the pattern of sizable factor loadings in an assessment wave.

Two issues merited attention before the investigations could proceed. For DEMQOL-Proxy-SF, item 8 (*lively*) did not load on the general HRQL factor in the longitudinal SEM model for the HTA-SADD trial sample. This was noted previously in the short-form development phase (Chapter 3). As before, this investigation was based on data extracted from the parent versions. Consequently, the issue with item 8 might be specific to DEMQOL-Proxy item response patterns, rather than that of DEMQOL-Proxy-SF. When the actual 17-item short-form version is administered, the observed response patterns might differ from those in the 31-item full version. This was also plausible given that the same issue did not persist in the DEMQOL-Proxy-SF model for Croydon Memory Service cohort.

The second issue concerned empirical under-identification. Consistent with initial exploration of individual measurement models, item 14 for DEMQOL-Proxy-SF displayed unusually large values and standard errors for Croydon Memory Service

baseline assessment as well as the third assessment occasion in HTA-SADD trial. They were also not statistically significant. The longitudinal SEM models were re-estimated after fixing these factor loadings with plausible values based on approximate averages of the values from the other two assessment occasions within each study sample. These DEMQOL-Proxy-SF models did not alter the original substantive conclusions and they are reported as Model 1 (Table 4.3 and 4.5) for subsequent investigations.

4.5.4 Longitudinal scalar invariance

Given tenability of longitudinal configural invariance, Model 1 was modified such that both factor loadings and item thresholds were identical across the assessment occasions for DEMQOL-SF and DEMQOL-Proxy-SF respectively. With these across-occasion constraints, the occasion-specific constraints that were imposed on DEMQOL-Proxy-SF item 14 to address empirical under-identification were no longer necessary.

Model fit evaluation (Table 4.3) indicated that neither DEMQOL-SF nor DEMQOL-Proxy-SF model of scalar invariance (Model 2) predicted a pattern of inter-item correlations that had an exact match (within bounds of sampling error) with those found in the data of both the memory service and clinical trial cohorts. In terms of approximate fit, all four instances of Model 2 (i.e. DEMQOL-SF / DEMQOL-Proxy-SF model for Croydon / HTA study sample) had acceptable RMSEA, CFI, TLI and WRMR values that were comparable to Model 1. Model-

data correspondence at large was adequate for drawing conclusions with the model results.

Table 4.6 and 4.7 present the unstandardized factor loadings and thresholds for DEMQOL-SF and DEMQOL-Proxy-SF respectively. As they were identical in every wave of assessment, only the results from a single wave bifactor CFA model was presented. With across-occasion equality constraints, Model 2 had fewer freely estimated parameters (or less model complexity) than Model 1 resulting in poorer fit with the sample data (i.e. larger χ^2_{M} values). When compared statistically, DIFFTEST results indicated that the poorer empirical fit of Model 2 relative to Model 1 was inconsequential for both DEMQOL-SF and DEMQOL-Proxy-SF (Table 4.3). This result reinforced the tenability of Model 2, strengthening support for the plausibility of scalar invariance in DEMQOL-SF and DEMQOL-Proxy-SF assessments. With longitudinal invariance in factor loadings and thresholds, this study found no evidence of re-prioritisation and re-calibration over repeated HRQL assessments in both the memory service and clinical trial cohort.

Table 4.6 Longitudinal scalar invariance in DEMQOL-SF bifactor CFA unstandardised factor loadings and thresholds

	CMS Model 2		Thresholds			HTA Model 2		Thresholds		
Item	GEN	COG	τ_1	τ_2	τ_3	GEN	COG	τ_1	τ_2	τ_3
1	.41		-1.98	-.74	1.16	.51		-1.03	.13	1.55
2	.82		-2.35	-1.25	.15	.85		-1.48	-.68	.33
4	.92		-2.04	-1.18	.07	.72		-1.16	-.42	.47
7	.99		-2.75	-1.83	-.23	.91		-1.55	-.60	.60
8	.58		-2.22	-1.50	-.56	.72		-1.39	-.88	-.01
11	.97		-2.81	-1.76	-.24	.76		-1.77	-.90	.47
12	1.03		-2.52	-1.45	.07	.82		-1.14	-.32	.73
13	.68		-1.63	-.78	.50	.58		-1.12	-.22	.59
14	.82	.76	-2.16	-.84	.46	.76	.75	-1.45	-.49	.40
17	1.04	1.06	-3.10	-1.78	-.35	1.25	1.17	-2.32	-1.08	.27
18	1.11	.57	-2.83	-1.90	-.39	1.20	.99	-2.56	-1.37	-.18
19	1.10	1.00	-2.72	-1.58	.34	1.32	1.23	-2.35	-1.11	.35
21	.94		-2.75	-2.06	-1.07	1.07		-2.31	-1.58	-.87
22	1.07		-2.93	-2.39	-1.44	.99		-2.51	-1.83	-1.08
23	1.00		-3.23	-2.35	-1.05	.71		-2.34	-1.52	-.70
24	1.01		-2.97	-2.13	-.84	.92		-2.15	-1.34	-.35
28	.74		-2.61	-1.70	-.36	1.00		-1.73	-.99	-.13

Item thresholds refer to the level of difficulty of achieving the next higher score on Likert scale with four response categories. Individuals with poorer than average HRQL were more likely to have a high score on items with response categories that were ‘easy’ (or small threshold values), than on items with response categories that were ‘difficult’ (or large threshold values).

Table 4.7 Longitudinal scalar invariance in DEMQOL-Proxy-SF bifactor CFA unstandardised factor loadings and thresholds

	CMS Model 2			Thresholds			HTA Model 2			Thresholds		
Item	GEN	NEG	COG	τ_1	τ_2	τ_3	GEN	NEG	COG	τ_1	τ_2	τ_3
3	.67	.62		-1.86	-.75	.62	.45	.67		-.94	.07	1.17
5	1.15	.85		-3.16	-1.96	.07	.65	1.48		-1.92	-.30	1.85
7	1.01	.63		-3.08	-1.61	-.28	.55	1.04		-1.77	-.52	.85
8	.30			-.58	.43	1.74	-.09 ns			.04	.99	2.06
10	1.22	.94		-2.84	-1.41	.73	.45	1.05		-.84	.27	1.60
12	.73		.87	-1.69	-.37	1.05	1.14		.74	-1.50	-.15	.90
14	1.62		2.15	-3.12	-.51	2.28	1.40		1.42	-1.38	.10	1.21
17	.70		.47	-1.94	-1.07	.00	.97		.99	-1.33	-.51	.38
18	1.05		.62	-2.61	-1.27	.24	1.24		1.03	-1.71	-.42	.98
22	.68			-2.67	-1.76	-1.04	.51			-1.87	-1.08	-.47
25	.77			-2.14	-1.59	-.75	.70			-1.66	-1.07	-.58
26	.91			-2.84	-1.66	-.13	.84			-1.92	-1.06	-.19
27	.94			-2.78	-1.97	-.92	.73			-1.82	-1.24	-.49
28	.86			-2.29	-1.66	-.49	.70			-1.31	-.72	-.12
29	.98			-3.03	-1.98	-.83	1.01			-2.33	-1.39	-.45
30	.97			-2.79	-1.54	-.35	1.22			-2.04	-.97	.00
31	.54			-1.64	-.79	.23	.73			-1.34	-.59	.19

Item thresholds refer to the level of difficulty of achieving the next higher score on Likert scale with four response categories. Individuals with poorer than average HRQL were more likely to have a high score on items with response categories that were ‘easy’ (or small threshold values), than on items with response categories that were ‘difficult’ (or large threshold values).

4.5.5 Longitudinal invariance of measurement errors

Given tenability of scalar invariance, Model 2 was modified such that all item residual variances (or measurement error) were fixed at one and so were identical across the assessment occasions. Model fit evaluation (Table 4.3) indicated that neither DEMQOL-SF nor the DEMQOL-Proxy-SF Model 3 predicted a pattern of inter-item correlations that had an exact match with those found in the sample data within bounds of sampling error. In terms of approximate fit, all four instances of Model 2 (i.e. DEMQOL-SF / DEMQOL-Proxy-SF model for Croydon / HTA-SADD study sample) had acceptable RMSEA, CFI, TLI and WRMR values that were comparable to Model 1. Model-data correspondence at large was adequate for drawing conclusions with the model results.

Table 4.8 and 4.9 present the unstandardized factor loadings and thresholds for DEMQOL-SF and DEMQOL-Proxy-SF respectively. As a nested (or more restricted) version of Model 1, Model 3 had a poorer exact fit with the sample data. When compared statistically, DIFFTEST results indicated that the poorer empirical fit of Model 3 relative to Model 1 was inconsequential for DEMQOL-SF in the memory service sample and DEMQOL-Proxy-SF in both study samples (Table 4.3). DEMQOL-SF Model 3 showed a statistically significant poorer fit than Model 1 in the HTA-SADD trial sample ($\Delta\chi^2_{\mathrm{M}}$ = 165.673, df = 136, p = 0.04).

Table 4.8 Longitudinal invariance of measurement error in DEMQOL-SF bifactor CFA unstandardised factor loadings and thresholds

	CMS Model 3		Thresholds			HTA Model 3		Thresholds		
Item	GEN	COG	τ_1	τ_2	τ_3	GEN	COG	τ_1	τ_2	τ_3
1	.42		-2.01	-.76	1.16	.48		-1.01	.12	1.48
2	.85		-2.42	-1.30	.14	.91		-1.58	-.73	.34
4	.93		-2.04	-1.19	.06	.83		-1.34	-.49	.52
7	.96		-2.65	-1.76	-.23	.91		-1.57	-.61	.59
8	.58		-2.21	-1.51	-.57	.75		-1.45	-.93	-.02
11	.97		-2.80	-1.75	-.25	.79		-1.86	-.95	.48
12	1.01		-2.44	-1.41	.06	1.06		-1.47	-.43	.93
13	.64		-1.51	-.72	.46	.59		-1.20	-.24	.61
14	.92	.86	-2.37	-.93	.50	.88	.89	-1.68	-.57	.45
17	1.07	1.09	-3.13	-1.81	-.36	1.25	1.18	-2.34	-1.09	.27
18	1.08	.56	-2.75	-1.85	-.39	.99	.84	-2.18	-1.16	-.14
19	1.05	.94	-2.58	-1.50	.32	1.28	1.20	-2.31	-1.09	.33
21	1.00		-2.89	-2.14	-1.13	.95		-2.11	-1.42	-.78
22	1.10		-3.00	-2.45	-1.48	.89		-2.31	-1.68	-.99
23	.92		-2.99	-2.19	-.98	.61		-2.04	-1.35	-.61
24	1.00		-2.94	-2.11	-.83	.82		-1.97	-1.22	-.32
28	.73		-2.55	-1.66	-.37	.94		-1.65	-.94	-.12

Item thresholds refer to the level of difficulty of achieving the next higher score on Likert scale with four response categories. Individuals with poorer than average HRQL were more likely to have a high score on items with response categories that were ‘easy’ (or small threshold values), than on items with response categories that were ‘difficult’ (or large threshold values).

Table 4.9 Longitudinal invariance of measurement error in DEMQOL-Proxy-SF bifactor CFA unstandardised factor loadings and thresholds

	CMS Model 3			Thresholds			HTA Model 3			Thresholds		
Item	GEN	NEG	COG	τ_1	τ_2	τ_3	GEN	NEG	COG	τ_1	τ_2	τ_3
3	.73	.67		-2.05	-.83	.69	.53	.78		-1.15	.07	1.39
5	1.11	.83		-3.11	-1.92	.09	.55	1.23		-1.66	-.26	1.60
7	1.07	.66		-3.25	-1.69	-.27	.57	1.06		-1.87	-.55	.89
8	.29			-.57	.44	1.75	-.07 ns			.02	.87	1.81
10	1.12	.88		-2.66	-1.31	.71	.53	1.24		-1.03	.30	1.92
12	.77		.92	-1.80	-.39	1.12	1.04		.70	-1.40	-.13	.86
14	1.42		1.87	-2.74	-.45	2.02	1.42		1.47	-1.45	.09	1.24
17	.69		.47	-1.92	-1.06	.01	.88		.96	-1.26	-.48	.35
18	1.10		.65	-2.77	-1.32	.26	1.23		1.10	-1.74	-.43	1.01
22	.74			-2.89	-1.92	-1.13	.49			-1.85	-1.07	-.46
25	.78			-2.14	-1.61	-.76	.61			-1.48	-.94	-.50
26	.83			-2.60	-1.52	-.10	.85			-1.97	-1.09	-.18
27	.99			-3.00	-2.09	-.96	.72			-1.83	-1.23	-.48
28	.81			-2.20	-1.58	-.45	.78			-1.45	-.80	-.12
29	.96			-3.01	-1.96	-.81	1.01			-2.36	-1.39	-.44
30	.97			-2.85	-1.54	-.33	1.31			-2.18	-1.04	.03
31	.56			-1.73	-.82	.26	.78			-1.45	-.63	.23

Item thresholds refer to the level of difficulty of achieving the next higher score on Likert scale with four response categories. Individuals with poorer than average HRQL were more likely to have a high score on items with response categories that were ‘easy’ (or small threshold values), than on items with response categories that were ‘difficult’ (or large threshold values).

Modification indices were inspected for guidance on how to re-specify Model 3 so that it could have better fit with DEMQOL-SF data from HTA-SADD trial. The two largest modification indices (28.1 and 18.5) flagged a need to consider hypothesising correlated residuals between item 21 and 22 in the first and last assessment occasion of Model 3. A decision was made against increasing model complexity in this way due to issues noted in Chapter 2 (section 2.4.3). At that stage of investigation when bifactor EFAs were conducted with its parent version (i.e. DEMQOL), the results highlighted that item 21 (*how you get on with people close to you*) and 22 (*getting the affection you want*) might have 'excess' association due to highly similar content and close item proximity. As this source of common variance was not theoretically relevant to individual differences in HRQL, the gain in empirical fit from specifying correlated residuals might not outweigh the loss in model parsimony when a more complex SEM model was estimated. Moreover, since the current short-form data was extracted from their parent versions, the need for correlated residuals might also be more appropriate for DEMQOL than for DEMQOL-SF.

Further inspection of smaller modification indices that exceeded the value of 10 (minimum set by Mplus defaults) revealed only four out of the remaining 12 had substantively plausible implications. The largest of these flagged a need to consider relaxing the equality constraint on item 23 at the third assessment occasion (modification indices = 14.0). This suggested that conclusions about scalar invariance based on DEMQOL-SF Model 2 might need to be reconsidered in HTA-SADD trial sample. On the other hand, given that Model 2 tenability was

supported by both approximate fit indices and model comparison (Model 1 vs 2), it might also be argued that the current stage of hypothesis testing (Model 3) should restrict an examination of modification indices to focus only on the tenability of item residual variances constraints. Going by this argument, only the constraint on residual variance of item 12 at the first assessment occasion warranted attention (modification indices = 13.9).

To weigh both substantive and statistical considerations, Model 3 was modified to explore the implications of relaxing either constraint. In Model 4, either the factor loading of item 23 at the third assessment occasion, or the residual variance of item 12 at the first assessment occasion, was freely estimated from HTA-SADD trial data. In both cases, Model 4 exhibited adequate approximate fit with the sample data (Table 4.3). When item 23 factor loading was allowed to differ from the other two occasions (i.e. re-prioritisation had taken place), Model 4 showed a poorer empirical fit than Model 1 that was inconsequential (i.e. DIFFTEST results not statistically significant). Similarly, when item 12 residual variance was allowed to differ from the other two occasions (i.e. measurement error vary over time), the model showed no significant decline in tenability relative to Model 1.

Since neither was significantly 'worse' than Model 1, attention turned to whether any of the two showed substantial improvement from DEMQOL-SF Model 3 in HTA-SADD trial sample, where tenability was satisfactory (i.e. adequate approximate fit, but DIFFTEST showed poorer exact fit than Model 1 that was marginally significant). When item 12 residual variance was freely estimated in

Model 4, there was significant improvement from Model 3 in empirical fit ($\Delta\chi^2_M$ = 14.066, df = 1, p < 0.001). This was also the case when the factor loading of item 23 was allowed to differ between assessment occasions in Model 4 ($\Delta\chi^2_M$ = 6.354, df = 1, p = 0.01). However, the magnitude of improvement was notably smaller. Given these indications, measurement error for item 12 was freely estimated at baseline in Model 4 and no further model re-specification was pursued with DEMQOL-SF data from HTA-SADD trial.

Table 4.10 Latent mean estimates (SE) reflecting changes from baseline assessment

Croydon	6m			12m		
DEMQOL -SF	HRQL	COG		HRQL	COG	
Model 2	0.06 (0.08)	0.26 (0.13)		0.28** (0.09)	0.16 (0.13)	
Model 3	0.04 (0.07)	0.27* (0.11)		0.24** (0.07)	0.19 (0.11)	
DEMQOL-Proxy-SF	HRQL	COG	NEG	HRQL	COG	NEG
Model 2	0.11 (0.10)	0.29* (0.11)	0.11 (0.16)	0.11 (0.11)	0.41** (0.13)	0.02 (0.19)
Model 3	0.23* (0.10)	0.24* (0.11)	0.02 (0.16)	0.10 (0.09)	0.40*** (0.11)	0.09 (0.16)
HTA-SADD	3m			9m		
DEMQOL -SF	HRQL	COG		HRQL	COG	
Model 2	0.33*** (0.08)	-0.07 (0.12)		0.47*** (0.09)	0.13 (0.18)	
Model 3	0.32*** (0.07)	-0.07 (0.11)		0.45*** (0.08)	0.12 (0.15)	
Model 4	0.30*** (0.07)	-0.06 (0.11)		0.42*** (0.08)	0.13 (0.14)	
DEMQOL-Proxy-SF	HRQL	COG	NEG	HRQL	COG	NEG
Model 2	0.35*** (0.10)	0.13 (0.13)	0.57*** (0.11)	0.50*** (0.13)	0.12 (0.15)	0.49*** (0.12)
Model 3	0.42*** (0.08)	0.09 (0.11)	0.56*** (0.09)	0.54*** (0.11)	0.05 (0.12)	0.49*** (0.11)

* $p < .05$, ** $p < .01$, *** $p < .001$

4.5.6 Response shift and longitudinal estimates of change in HRQL

Table 4.10 presents the latent mean estimates of longitudinal changes in general HRQL and its domains since baseline assessment. Scalar invariance held in all models and they differed only in the number of invariant item residuals. As latent mean estimates were adjusted for measurement error regardless of the number of invariant item residuals, there were generally only small differences between model results.

In the Croydon Memory Service cohort, self-reports on DEMQOL-SF showed gains in HRQL at 12-month but not in the earlier waves. Based on Model 3 results, the latent estimate of this change was three times the size of its observed variability (standardised response mean, SRM = 0.24/0.07 = 3.43). The gains in HRQL were substantial using conventional criteria in which SRM values of 0.20, 0.50, and 0.80 represent small, moderate, and large effect sizes respectively (Husted, Cook, Farewell, & Gladman, 2000). Informant reports on DEMQOL-Proxy-SF present a slightly different picture. From their perspectives, the individuals with dementia had substantial HRQL gains (SRM = 0.23/0.10 = 2.30) at 6-month but not in later follow up. Interpretations about changes in COG were difficult at this stage. In the context of bifactor models, this latent factor was orthogonal to general HRQL, indicating that it reflected additional information that was on top of individual differences in general HRQL.

In the HTA-SADD trial cohort, DEMQOL-SF showed substantial gains in HRQL at 3-month (SRM = 0.30/0.07 = 4.29) and this improvement was maintained at 9-

month (SRM = 0.42/0.08 = 5.25). Informant reports on DEMQOL-Proxy-SF gave a similar picture. From their perspectives, the individuals with dementia had substantial HRQL gains at 3-month (SRM = 0.42/0.08 = 5.25) and this was maintained at 9-month (SRM = 0.54/0.11 = 4.91). These results were consistent with primary findings from the trial which showed a significant decline in depression symptoms at 3-month that was maintained at 9-month follow up (Banerjee et al., 2011), if not attributable to the antidepressants compared with placebo.

Model 2 (scalar invariance) reflected changes based on a stable HRQL concept (factor loading patterns), priorities (factor loadings), and expectations (thresholds). With Model 3, both DEMQOL-SF and DEMQOL-Proxy-SF demonstrated an even stricter form of invariance in which the amount of measurement error (item residual variances) was not different between each assessment wave. While DEMQOL-SF model results in HTA-SADD trial data (Model 4) weakened this claim, the impact on substantive conclusions was negligible in the current study. By and large, there was sufficient measurement invariance to support the use of raw score differences for assessing longitudinal changes in HRQL. Given that no response shift was detected, the assignment of utility weights in preference-based items was consistent with item response probabilities that reflected longitudinal gains in HRQL.

4.6 Discussion

This study found no evidence that people with dementia had changed the meanings, priorities, or expectations that they held about HRQL when re-interviewed within a 12-month period. HRQL reports provided by their carers showed no indication of response shift from their perspectives. Differences that emerged in repeated HRQL assessments could not be attributed to re-conceptualisation, re-prioritisation, or re-calibration of internal standards. It is hence plausible that the observed differences reflected real changes in subjective HRQL over time. On top of scalar invariance, the amount of measurement error in DEMQOL-SF and DEMQOL-Proxy-SF data was also stable across repeated assessments. Taken together, these psychometric properties provide a robust basis for employing raw score differences as a practical measure of HRQL changes in observational studies or randomised trials. This also implies that, in economic evaluations, utility weights assigned by the preference-based algorithms of DEMQOL-U and DEMQOL-Proxy-U would be consistent with responses that mainly reflect longitudinal gains (or losses) in HRQL.

Null or weak findings of response shift have been documented in similar studies with other chronic illness populations. In an American research registry of multiple sclerosis patients (n=1767), response shift was investigated for an understanding of how HRQL, as assessed by the SF-12, was associated with relapse and symptom change over two 6-month intervals (King-Kallimanis, Oort, Nolte, Schwartz, & Sprangers, 2011). While meaningful associations between

HRQL and health states were found, there was little response shift. The authors speculated that since the patients were not subjected to a planned intervention, there was no clear catalyst of health state changes to trigger response shift.

A similar issue may explain the lack of response shift in the samples of the current study. In the Croydon Memory Service cohort, after the initial referral for an early diagnosis, follow ups were made so that multidisciplinary care planning could be provided (Banerjee et al., 2007). The course of care, tailored to individual circumstances, is characterised by regular contact with a myriad of health and social care services as a result of the memory service intervention. In this context, the catalyst of health state changes is not readily discernible amidst the complexities of treatment and care planning. With the HTA-SADD trial, though this was a planned intervention, the trial results showed a decline in depression symptoms for all, with or without anti-depressant treatment (Banerjee et al., 2011). The active treatment arms as such might not be considered as a catalyst of health state changes. However, as these trial participants were recruited from old age psychiatry services, there are also substantial levels of 'treatment-as-usual' involved in their health and social care. Health state changes in this context may not have a clearly discernible catalyst to trigger response shift in a salient way. Amidst notable improvements in HRQL in both study samples, illness adaptation may result from only small changes that do not surface as response shift within a 12-month period.

Another plausible explanation for the findings may be that response shift might have occurred before a dementia diagnosis was made. In the UK, only 44% of those with dementia received a formal diagnosis (Alzheimer's Society, 2013), usually given late in the illness, and often initiated only after a crisis (NAO, 2007). By the time a diagnosis is made, illness adaptation might have already resulted from substantial changes in the meanings, priorities, and expectations that one held about HRQL.

The plausibility of this explanation has been demonstrated by a Canadian study of stroke patients in which response shift was investigated for an understanding of how HRQL, as assessed by the SF-36, was affected by stroke (n=238 patients) relative to the impact of natural aging (n=468 controls) over four 6-month intervals (Ahmed, Mayo, Corbiere, et al., 2005). While the impact of stroke was hypothesised to trigger more response shift than natural aging, none was found for the patient or control group. Longitudinal invariance was demonstrated within each group. However, when multi-group invariance was investigated by cross-sectional comparisons of baseline measurement models in patient and control group, the results suggested that their perceptions of HRQL were not identical. In other words, the onset of stroke might have triggered re-conceptualisation. The authors speculated that if stroke patients did experience response shift, it might be evident at onset but not after.

This speculation gained further support from another study by the same group of investigators who re-focused the inquiry on the immediate post-recovery period

where there is usually a period of salient health improvement which plateaus by about three months (Ahmed, Mayo, Wood-Dauphinee, Hanley, & Cohen, 2005). While these clinically significant transitions might trigger a change in how HRQL is evaluated, no response shift was found in HRQL reports (SF-36) of stroke patients (n=190) through their transitions in recovery (within 1 week post-stroke, 6 and 24 weeks after).

The need to consider earlier timeframes has also been underscored by a more recent study with stroke patients (Barclay & Tate, 2014). Unlike previous response shift studies where pre-stroke HRQL was inferred from baseline assessments made soon after stroke onset, this Canadian study had HRQL data that were obtained from a group of older men (n=168, mean age = 80.1 years) on average 1.3 years before an incident of stroke. When HRQL assessments using the SF-36 were repeated on average 1.5 years after stroke, the study found that role limitations due to emotional problems mattered more (i.e. re-prioritisation), whereas expectations about physical function have been lowered (i.e. re-calibration).

While there is a need to explore response shift earlier, the current study does not imply that illness adaptation is not relevant at mild to moderate stages of dementia. With the modelling of means and covariances in SEM, response shift and true change would be apparent as aggregate estimates only if a substantial number of individuals in the study sample showed the same type of change (Oort, 2005). It is plausible that some individuals have experienced response shift at later

HRQL assessments but the consequences were not salient at group level. In demonstrating longitudinal scalar invariance in DEMQOL-SF and DEMQOL-Proxy-SF, this study suggests that any response shift due to illness adaptation may not have a major impact on group estimates of HRQL changes from observational or clinical trial research in dementia.

4.7 Limitations

The findings reported here have to be interpreted in light of study limitations. First, there was data loss at each assessment occasion that could not be classified as missing completely at random (MCAR). In both the memory service cohort and clinical trial sample, individuals with missing HRQL data had more impairment due to neuropsychiatric symptoms (NPI) and loss of skills needed in daily life for independent living (BADL). As some information loss could be recovered if auxiliary variables like NPI and BADL had been used to augment the modelling (Collins et al., 2001; Yoo, 2009), such data loss might be classified as missing at random (MAR). However, given that missing data often reflects the challenging nature of the phenomenon under study (X. Yang et al., 2008), it was also plausible that HRQL data might be missing for those with more HRQL impairment. Such data loss would be classified as missing not at random (MNAR).

The full-information maximum likelihood (FIML) estimator is the ML estimation method in SEM for dealing with missing data that are MCAR or MAR (Allison, 2003; Yoo, 2009). In Mplus, both ML and WLSMV algorithms use all available data for model estimation but the latter is suited only when missing data is MCAR

(Asparouhov & Muthén, 2010b). The theoretical advantage that FIML has over WLSMV is clear when missing data is MCAR or MAR. However, given that the current data might also be MNAR, both FIML and WLSMV would yield biased estimates unless special modelling techniques (e.g. pattern-mixture modelling) were employed (L. Muthén & Muthén, 1998-2012, pp. 393-396). As this option adds considerable complexities on top of the multi-wave bifactor CFA models, it was not implemented for the present study.

The choice between ML and WLSMV was also considered in light of findings from recent simulation studies. Rhemtulla et al. (2012) demonstrated that for ordinal data with fewer than five response categories, factor loadings and robust standard errors were generally most accurately estimated by robust categorical least squares estimators (e.g. WLSMV) relative to robust ML estimation. Moshagen and Musch (2014) showed that WLSMV has strong convergence properties with good recovery of population parameters (e.g. factor loadings and standard errors) even when the model is large and sample size is small. The simulation results also suggested that there is little reason to prefer ML over WLSMV when the data are ordinal. Taken together, model estimation proceeded with WLSMV despite potential issues with missing data that were MAR or MNAR. It is worth noting that WLSMV estimation is not problematic with missing data that is MAR if the missing data modelling technique of multiple imputation is used to recover missing information (Asparouhov & Muthén, 2010b). This analytic strategy would require the current SEM investigations to be based on several imputed data sets. However, when implemented (see Chapter 2),

modification indices would not be available and model comparisons could not be made with imputed data sets in Mplus (version 7.11). For this reason, this option was not used.

While missing data that might be MNAR remains an issue, this has to be considered in light of the characteristics of the samples in the two data sources. Based on clinical assessments that were common between the memory service and clinical trial (Table 1), the HTA-SADD trial study sample had more severe impairment in cognition (MMSE), neuropsychiatric symptoms (NPI), and skills needed in daily life for independent living (BADL). This however did not have an apparent impact on study results given that both Croydon Memory Service and HTA-SADD trial samples had identical CFA and SEM models that showed tenable fit with the DEMQOL-SF and DEMQOL-Proxy-SF data. Missing data that were MNAR raises concern with whether model results would be different if subsequent assessment occasions included individuals with more severe HRQL impairment. In terms of cognition, neuropsychiatric symptoms and daily functioning, severity of impairment did not demonstrate an impact on model results. It is plausible that severity of HRQL impairment may have only a weak impact on the current findings of longitudinal measurement invariance in DEMQOL-SF and DEMQOL-Proxy-SF.

This study may be under-powered due to model complexity (e.g. bifactor CFA where items load on more than one factor), model size (e.g. multi-wave SEM), and missing data (e.g. MAR). Each of these conditions is known generally to

inflate demands for statistical power in SEM. Nonetheless, a recent simulation study showed that several rules-of-thumb are problematic because they are based on a narrow range of model configuration and may lead to grossly over- or underestimated sample size requirements in other contexts (Wolf, Harrington, Clark, & Miller, 2013). Evidence of great variability in sample size requirements was also reported in another simulation study that focused on WLSMV estimation (Moshagen & Musch, 2014). After studying a range of conditions, the authors concluded that models estimated with WLSMV algorithm were most likely to have proper convergence, accurate recovery of factor loadings and covariances, as well as satisfactory approximation of standard errors and the model chi square, when sample size is greater than 300. This sample size is considerably larger than the practical criteria ($n \geq 200$) used by a meta-analysis to rate the quality of studies in response shift research (Schwartz et al., 2006). While sample sizes in the current investigation exceeded 300 (Table 4.3), the study findings need to be replicated with larger sample sizes.

Finally, the study of response shift relied solely on the DEMQOL measurement system. Given that other HRQL measures in dementia differ in content coverage, they may show more themes that also carry core relevance in HRQL.

CHAPTER 5 RESEARCH CONCLUSIONS

This thesis examined HRQL assessment in dementia, with a focus on the measurement of change in clinical and economic evaluation of treatment interventions. The inquiry was conducted in three stages. First, the conceptual definition of HRQL in DEMQOL and DEMQOL-Proxy was validated from a bifactor model perspective. This measurement model was subsequently cross-validated with an independent sample in three aspects (geographical region, gender, dementia severity). DEMQOL and DEMQOL-Proxy items that demonstrated desirable psychometric properties at this stage were selected for short-form (SF) versions. In the final stage, DEMQOL-SF and DEMQOL-Proxy-SF item responses formed the basis for exploring whether improvement (or deterioration) in subjective HRQL had been influenced by a shift in meaning, priorities, or expectations over time.

The thesis exploited three recent methodological developments in tandem for investigating response shift. First, DEMQOL is the only condition-specific HRQL measurement system to date that also has a preference-based algorithm for cost-utility analysis in dementia. In addressing the knowledge gap of whether response shift is a concern for HRQL assessment in dementia, an inquiry on utility assessment is also part of the investigation. Relative to other HRQL measures that were developed specifically for dementia, the DEMQOL measurement system holds a strategic advantage for exploring both clinical and policy implications of response shift in HRQL assessments.

Second, against a background of response shift investigations in which a significant majority employed the then-test method, the detection of response shift in this thesis leveraged on emerging applications of latent variable modelling under the SEM framework of longitudinal measurement invariance. Compared to the then-test method which carries an exclusive focus on re-calibration, the SEM framework aligns with the view that response shift processes are fundamentally intertwined, and provides a psychometric typology of change that translates into a concurrent examination of re-conceptualisation, re-prioritisation, and re-calibration.

Third, for understanding item response patterns and their themes in self- and proxy reports (i.e. DEMQOL and DEMQOL-Proxy), the current investigations conducted EFAs and CFAs from a bifactor model perspective. This latent variable measurement model views HRQL as a general theme (or source of common variance) that supersedes the complexities of a myriad of narrower themes of item response patterns, in the form of a broad latent factor that is independent of multiple domain factors. This focus on HRQL as the target construct differs from commonly employed factor analysis approaches (see Figure 1.2 on page 48) in that it yields direct insights on how well every DEMQOL / DEMQOL-Proxy item discriminates individual differences on the main assessment objective. Furthermore, bifactor CFA models may exhibit 'factor collapse' as an indication that a HRQL domain in question lies at the heart of HRQL concept. With this perspective, the thesis made use of conceptual and empirical foundations that maintain a focus on the measurement of a complex general phenomenon, while

holding a degree of versatility for exploring potential variation in what lies at the heart of HRQL in people with dementia at different times of need.

5.1 Key findings

5.1.1 Forming conclusions about HRQL using DEMQOL measurement system

The complex nature of HRQL in dementia is apparent from previous factor analytic studies (Mulhern et al., 2013), which shed light on multiple themes of individual differences in item response patterns of DEMQOL and DEMQOL-Proxy (Table 5.1 and 5.2 respectively).

Table 5.1 DEMQOL (28 items) factor analytic themes

PCA (Varimax rotation)	Bifactor EFA (Bi-geomin orthogonal rotation)	Bifactor CFA (24 items + 4 testlets)
POS: positive emotion Item 1, 3, 5, 6, 10	POS 1, 3, 5, 6, 10	POS 1, 3, 5, [6 + 10]
NEG: negative emotion Item 2, 4, 7, 11, 12	NEG 4, 11, 12, 13	NEG 4, 11, 12, 13
COG: worries about cognition Item 14, 15, 16, 17, 18, 19	COG 14, 15, 16, 17, 18, 19	COG 14, 15, 16, 17, 18, 19
SOC: worries about social relationship Item 21, 22, 23, 24, 25, 26		SOC [21 + 22], 23, 24, 25
LON: loneliness Item 8, 20	LON 8, 20	
Non/cross loading Item 9, 13, 27, 28	HRQL only 2, 7, 9, 21, 22, 23, 24, 25, 26, 27, 28	HRQL only 2, 7, 9, [8 + 20], 26 [27 + 28]

Results of principal component analysis (PCA) with varimax rotation were taken from Mulhern et al. (2013). In bifactor EFA and CFA models, all items also load on a general factor of HRQL. Some items load only on the general factor (i.e. HRQL only). Item pairs in testlets were denoted in square brackets e.g. [6 + 10].

Table 5.2 DEMQOL-Proxy (31 items) factor analytic themes

PCA (Varimax rotation)	Bifactor EFA (Bi-geomin orthogonal rotation)	Bifactor CFA (27 items + 5 testlets)
POS: positive emotion Item 4, 8, 11	POS 1, 4, 6, 8, 11	POS 1, [4 + 8], 6, 11
NEG: negative emotion Item 2, 3, 5, 7, 9, 10	NEG 2, 3, 5, 7, 9, 10	NEG 2, 3, 5, 7, 9, 10
COG: worries about cognition Item 12, 13, 14, 15, 17, 18, 19, 20, 26		COG [12 + 14], 13, 15, 17, 18, 19, 20
FIN: worries about financial tasks Item 23, 24, 25	FIN 23, 24, 25	
APP: worries about appearance Item 21, 22	APP 21, 22	
	SOC 27, 28, 29, 30	SOC 27, 28, [29 + 30]
Non/cross loading Item 1, 6, 16, 27, 28, 29, 30, 31	HRQL only 12, 13, 14, 15, 17, 18, 19, 20, 26, 31	HRQL only 18, 19, 20, [21 + 22], 23, [24 + 25], 26, 31

Results of principal component analysis (PCA) with varimax rotation were taken from Mulhern et al. (2013). The domain of 'worries about financial tasks' (FIN) was labelled as 'daily activities' in the original report. In bifactor EFA and CFA models, all items also load on a general factor of HRQL. Some items load only on the general factor (i.e. HRQL only). Item pairs in testlets were denoted in square brackets e.g. [4 + 8].

Using bifactor model perspectives, this investigation found that it was tenable to consolidate insights from multiple health-related domains of life to form a coherent overall conclusion about HRQL in dementia from DEMQOL and DEMQOL-Proxy assessments. In other words HRQL is most appropriately explained as a general phenomenon that supersedes the complexities of individual differences in a myriad of health-related domains.

Along this line of inquiry, bifactor EFA results suggested that 'worries about social relationship' (SOC) was a core theme in HRQL assessments with DEMQOL, whereas 'worries about cognition' (COG) might be a core theme in HRQL reports provided by informants on DEMQOL-Proxy. These findings were partially replicated in bifactor CFAs. There was also tentative evidence to suggest that 'worries about social relationship' (SOC) held core relevance in self-report HRQL. In informant report, instead of 'worries about cognition' (COG), bifactor CFA had tentative evidence that 'worries about social relationship' (SOC) was a core theme. The discrepancy between bifactor EFA and CFA precluded firm conclusions at this juncture.

To understand the practical implications of bifactor model findings, an evaluation of measurement reliability was made for overall total scores and multiple subscale scores from DEMQOL and DEMQOL-Proxy. In HRQL assessments provided by self- and informant report, subscale scores consistently showed inadequate measurement reliability for discriminating individual differences in their putative HRQL domains. On the other hand, if conclusions about HRQL were formed using overall total scores, both DEMQOL and DEMQOL-Proxy demonstrated a high level of sensitivity to individual differences.

5.1.2 Conducting HRQL assessments in different settings and populations

Having established a meaningful basis for interpreting DEMQOL and DEMQOL-Proxy assessments, cross-validation was performed to see if conclusions about HRQL could be made in the same manner across different settings and

populations. Identical bifactor CFA models were employed to see if an independent sample had the same HRQL perceptions and response behaviour for DEMQOL and DEMQOL-Proxy assessments. Direct comparisons were made and evaluated statistically for geographical region (UK vs Latin America), gender (of people with dementia), and dementia severity (mild vs moderate to severe).

The findings of this investigation suggested that conclusions about HRQL could be made in terms of a general theme that was independent of four narrower themes (i.e. POS, NEG, COG, SOC), regardless of geographical region, gender, and dementia severity. However, inconsistencies in response behaviour were found on both DEMQOL and DEMQOL-Proxy. For instance, compared to the UK sample of people with dementia, respondents in the Latin American sample had much higher odds of reporting higher levels of functioning when asked if they worry about '*how you feel in yourself*' (DEMQOL item 27). This means that taking people with the same levels of HRQL, those in Latin America would give more positive evaluations on this item compared with those from the UK.

Geographical region was the main source of inconsistencies in response behaviour for DEMQOL and DEMQOL-Proxy. Neither gender of people with dementia nor dementia severity resulted in major inconsistencies in HRQL self- and informant report behaviour. There was no preponderance in magnitude and direction of inconsistent response behaviour on any HRQL theme in DEMQOL and DEMQOL-Proxy. To understand the practical implications of these findings, estimates of group differences (e.g. UK vs Latin America) in HRQL were

compared before and after statistical adjustments were made for response bias. This evaluation found only a minor impact on DEMQOL estimates of group differences. This was more of a concern with DEMQOL-Proxy as statistical significance was altered for some group differences. However, trivial bias in group difference estimates might also have an impact on statistical significance given the statistical power afforded by the combined sample size (i.e. UK and Latin America). Of note, group differences were generally small before and after statistical adjustment for response bias in DEMQOL-Proxy.

As statistical adjustment is not feasible for HRQL assessments in clinical settings, this investigation presented an opportunity to develop short-form (SF) versions so as to reduce the number of items that were prone to inconsistencies in response behaviour. Based on item response patterns of DEMQOL and DEMQOL-Proxy, 17 items were selected for DEMQOL-SF and DEMQOL-Proxy-SF respectively such that they also retain similar levels and range of sensitivity as their parent versions for discriminating individual differences in HRQL. The selection process led to the retention of only one item from the domain of 'positive emotion' (POS) in DEMQOL and DEMOQL-Proxy. Bifactor CFAs showed that conclusions about HRQL could still be made based on a general theme that superseded three narrower themes (NEG, COG, and SOC) for both short-form versions. Without statistical adjustment for inconsistencies in response behaviour, DEMQOL-SF had a slightly more accurate estimate of HRQL differences between geographical regions than its parent version. DEMQOL-Proxy-SF, on the other hand, had a

more accurate estimate than its parent version for HRQL differences between dementia severity levels.

Taken together, HRQL assessments with short-form versions of DEMQOL measurement system have a similar basis and sensitivity as their parent versions for forming conclusions about individual differences in HRQL. There is also tentative evidence that DEMQOL-SF and DEMQOL-Proxy-SF are less prone than their parent versions to inconsistencies in response behaviour and hence hold wider feasibility for HRQL assessment across settings and populations.

5.1.3 Clinical and economic evaluation of longitudinal changes in HRQL

As people with dementia learn to maintain their HRQL in daily life, they may gradually change their definitions, priorities, and standards about HRQL. The nature of these subjective changes may help inform intervention strategies and influence conclusions about their effectiveness in improving HRQL. In this thesis, such potential changes were investigated with self- and informant report HRQL data that have been collected from a memory service clinic (baseline, 6-month, and 12-month) and a randomised clinical trial (baseline, 3-month, and 9-month).

Based on DEMQOL-SF and DEMQOL-Proxy-SF item data, there was plausible evidence to suggest that self-report responses carried a general theme of HRQL in which ‘negative’ emotion’ (NEG) and ‘worries about social relationship’ (SOC) held core relevance; whereas informant report responses carried a general theme of HRQL in which ‘worries about social relationship’ (SOC) held core relevance. These perceptions were found to be stable over time. The meaning (or themes) of

HRQL has not been re-conceptualised after initial assessments by the memory service clinic or in the randomised clinical trial. Furthermore, the study also found no evidence that re-prioritisation or re-calibration of internal standards has taken place when HRQL assessments were repeated within one year duration.

Given that differences that emerged over repeated HRQL assessments could not be attributed to re-conceptualisation, re-prioritisation, or re-calibration of internal standards, the observed gains in HRQL for both study samples likely reflected real changes in subjective HRQL over time. This strong form of measurement invariance provided further basis for examining the amount of measurement reliability in every assessment wave. Both DEMQOL-SF and DEMQOL-Proxy-SF showed similar levels of sensitivity to individual differences in HRQL across time. This suggests that longitudinal HRQL assessments can be based on changes in raw scores. Among the 17 items in DEMQOL-SF and DEMQOL-Proxy-SF respectively, preference-based algorithms had been developed for five items (DEMQOL-U) in the former and four items (DEMQOL-Proxy-U) in the latter for determining the perceived value of HRQL scenarios. Since the assignment of utility weights for preference-based items relies on raw score responses, the basis for economic evaluation of HRQL changes would be consistent with item response probabilities that reflect longitudinal changes in HRQL.

5.2 Limitations

When interpreting the findings set out in this thesis, a number of limitations should be taken into account. They have been reported in each empirical chapter and three broad issues are reiterated here.

First, a varying number of individuals in each study sample did not have HRQL data. These individuals in general had slightly more impaired health than those for whom self- and/or informant report HRQL was given. There is uncertainty in whether their HRQL perceptions and response behaviour are consistent with those reported in this research. This concern was initially addressed with multiple imputation to ameliorate the sample bias (Chapter 2). The major HRQL themes that emerged at this stage were employed for subsequent investigations. While the impact of missing data is likely to vary across these later stages, the same basis for forming conclusions about individual differences in HRQL was found to be tenable when examined cross-sectionally in two geographical regions (Chapter 3), and longitudinally in two clinical settings (Chapter 4).

The issue with missing data poses more threats on the longitudinal findings of this research. Given that missing data often reflects the challenging nature of the phenomenon under study, self- and informant report HRQL are likely to be unavailable for individuals who experienced more deterioration. A key concern as such is whether the same study conclusions would be reached if data had been available for individuals with more severe HRQL impairment at later time points. Impairment severity, in terms of cognition, neuropsychiatric symptoms and daily

functioning, did not show an apparent impact in the current research. The same conclusions about response shift were reached in two study samples, even though the clinical trial sample had more impaired health than the memory clinic cohort. This suggests that missing data bias due to severity of HRQL impairment may have had a limited impact on the research findings.

Second, with the methodological versatility afforded by latent variable modelling methods, a wide range of alternative modelling strategies and decision-making is possible in every stage of the current investigations. For instance, at the initial stage of examining the basis for forming conclusions with DEMQOL and DEMQOL-Proxy (Chapter 2), latent factors instead of testlets could have been employed to represent the 'excess' correlations between item pairs (e.g. DEMQOL-Proxy item 21 and 22 for APP domain). This decision would have allowed both bifactor EFA and CFA investigations to proceed with an identical set of HRQL themes even though some might hold less theoretical relevance to individual differences in HRQL. This in turn may provide a more consistent basis between bifactor EFA and CFA models for generating conclusions about factor collapse.

For the purpose of investigating inconsistencies in response behaviour (Chapter 3), the SEM framework of multi-group CFA instead of MIMIC model could have been employed for detecting DIF effects. As noted, MIMIC models only detect uniform DIF effects. If inconsistencies in response behaviour occurs only at high/low levels of HRQL (i.e. non-uniform DIF), this could be detected by multi-

group CFA. The use of MIMIC models on the other hand held a practical advantage of allowing for a concurrent investigation with geographical region, gender, and dementia severity. As such this strategy is potentially useful as a first-stage detection when the initial concern is presence of any DIF (Finch, 2005; Reininghaus et al., 2012; F. M. Yang et al., 2009). Given that the permutations of modelling strategies and decisions can generate different conclusions, replication of the current research findings is necessary.

Finally, themes that carry substantive relevance for HRQL are not limited to the ones based on DEMQOL measurement system. Given that other HRQL measures in dementia differ in content coverage, they may generate other findings about the basis for forming conclusions about HRQL, or question items that are prone to inconsistencies in response behaviour and/or response shift.

5.3 Study implications

5.3.1 HRQL assessment in clinical research and practice

In establishing an appropriate basis for making conclusions about HRQL with DEMQOL measurement system, this thesis supports the view that HRQL in dementia is most meaningfully interpreted as a general phenomenon in which the whole is greater than the sum of its parts. Responses on DEMQOL and DEMQOL-Proxy items were more sensitive to individual differences in HRQL when taken together for an overall conclusion than when they were considered as separate themes of individual differences that carry distinct implications for

clinical and policy decisions. This supports the view that HRQL conclusions should be based on overall total scores rather than subscale scores.

It has been argued that subscale scores should be calculated because HRQL by definition is a multidimensional concept and respective domain scores can help clarify treatment impact (Ettema et al., 2005; Perales et al., 2013). However, multidimensionality may be conceived as an unintended consequence of incorporating various ways in which HRQL has to be evaluated. On the other hand, the use of overall total scores does not detract attention from the ways in which treatment interventions have an impact on HRQL. As demonstrated, it is possible to illuminate the core themes of HRQL in self- and informant reports. Hence, the impact of treatment interventions may be clarified in terms of the themes that drive individual differences in overall HRQL.

5.3.2 Illness adaptation in dementia

Very little is known about how people adapt to the chronic and challenging circumstances living with dementia. Narrative reviews in this literature have highlighted that people with dementia may experience a shift in meanings, priorities, or expectations for subjective HRQL over the course of illness. An empirical inquiry was made with three waves of HRQL assessments that were conducted within a one-year window in two clinical contexts. This research found no evidence of response shift in self- and informant reports across repeated HRQL assessments in the context of a memory service clinic and a randomised clinical trial.

‘Negative emotion’ (NEG) and ’worries about social relationship’ (SOC) were consistently core themes in HRQL self-report, whereas the theme of SOC held core relevance in informant reports for the same period. As the clinical trial participants were recruited from old age psychiatry services and generally had more impaired health than those in the memory service sample, they were likely to be at later stages of diagnosis. This suggests that in the aftermath of a clinical diagnosis ‘worries about social relationship’ (SOC) is likely to persist and its core relevance for HRQL is unlikely to be re-conceptualised.

While a shift in meanings, priorities or expectations may occur as people with dementia cope with the chronic nature of their condition, these processes may unfold only gradually. Given that a key motivation is to determine whether response shift may obscure treatment impact in ways that imply an apparent lack of benefit, this research suggests that any response shift due to illness adaptation is not likely to have a major impact on group estimates of HRQL changes from observational or clinical trial research in dementia. Furthermore, the measurement of change can be based on the difference of overall total scores from repeated HRQL assessments with DEMQOL-SF and DEMQOL-Proxy-SF.

5.3.3 The value of life years in dementia

Treatment interventions in dementia employ disparate amount of resources for achieving their objectives due to a complex interplay of clinical and psychosocial outcomes. The extent in which they add value to life years has gained considerable policy interests with the advent of utility measurement in health

economics. Utility measurement estimates the perceived value of living in different states of health and calculates the gain or loss in value as health improves or deteriorates over time. This preference-based algorithm assumes that every health state carries a perceived value that does not change over time.

This thesis examined how response shift may alter the basis for assigning utility weights to calculate the perceived value of a health state in the preference-based algorithms of DEMQOL measurement system. Two key findings suggest that gains or losses in DEMQOL-U and DEMQOL-Proxy-U utility values are based on perceived values that do not change over time. First, the absence of response shift meant that there is no change in meanings, priorities, and expectations of HRQL. Second, DEMQOL-U and DEMQOL-Proxy-U items (in DEMQOL-SF and DEMQOL-Proxy-SF respectively) showed a consistent level of sensitivity to individual differences over repeated HRQL assessments. This meant that the measurement of change in HRQL can be based on a difference in raw scores over time. Since the assignment of utility weights for preference-based items relies on raw score responses, the measurement of change in utility values would be consistent with item response probabilities that reflect longitudinal changes in HRQL.

5.4 Future research directions

In addressing knowledge gaps about HRQL measurement in dementia, this thesis brought to light further research questions.

5.4.1 Importance of social network

Lawton (1994) suggested that social behaviour in people with dementia is 'a treatment goal that seems appropriate for an illness whose manifestations in general appear to represent estrangement from the external world'. This is consistent with a body of empirical literature demonstrating that social functioning plays a pivotal role in the illness experience (Frick et al., 2012; Hughes et al., 2013; Lou et al., 2013; MacRae, 2011) as well as healthy aging in general (Coyle & Dugan, 2012; Huxhold et al., 2013; Ichida et al., 2013; Rook et al., 2012). This research showed that 'worries about social relationship' (SOC) may be a core theme in HRQL self- and informant report. Demonstrating this with other generic and condition-specific HRQL measures would shed light on whether social functioning in dementia should be a key clinical and policy focus when evaluating treatment interventions.

5.4.2 Heterotypic continuity

HRQL in dementia may hold core themes that differ between self- and informant-report, community and residential home samples, as well as stages of illness and diagnosis. These complexities are compounded by potential shifts in priorities and expectations over the course of illness adaptation. Determining the impact and

value of treatment interventions require a coherent basis for the assessment and measurement of change in HRQL. This coherence may be found in the concept of 'heterotypic continuity' (Holmbeck et al., 2010) in which the same underlying phenomenon may be expressed differently at different stages of development. HRQL may be the same general phenomenon despite the underlying complexities. Support for this premise would allow research efforts to focus on what lies at the heart of HRQL in people with dementia at different times of need.

5.4.3 QALY estimates in the presence of response shift

Knowledge gaps remain in whether the perceived value of health states shows sufficient stability when accompanied by potential changes in the meanings, priorities, and expectations of HRQL in dementia. Understanding the interaction between response shift and utility measurement can strengthen the foundations for using QALY estimates in cost-effectiveness comparisons. Inter-disciplinary research is required to formulate preference-based algorithms that also capture the psychological processes of illness adaptation in dementia.

5.5 Concluding remarks

HRQL assessment is an undertaking that is fraught with practical challenges, as reflected in a systematic review that found only 10 out of 225 RCTs in dementia and mild cognitive impairment included HRQL as an endpoint (Scholzel-Dorenbos, van der Steen, Engels, & Olde Rikkert, 2007). To date, there are at least 15 condition-specific HRQL measures, with wide variation in conceptual coverage of what constitutes HRQL in dementia (Perales et al., 2013). There is a need to determine whether HRQL assessment has captured what is important to the target population (Halvorsrud & Kalfoss, 2007), so as to generate a coherent body of evidence to guide clinical and policy decisions (Bakas et al., 2012).

The basis for forming conclusions about individual differences in HRQL was illuminated with DEMQOL and DEMQOL-Proxy in this research. Items that were prone to inconsistencies in self- and informant report behaviour were identified. Based on this knowledge, DEMQOL-SF and DEMQOL-Proxy-SF were developed. These short-form versions, which included preference-based items, demonstrated strong psychometric properties for the measurement of change in HRQL. Taken together, this thesis strengthen the foundations for conducting clinical and economic evaluation of HRQL changes in treatment interventions for dementia across clinical and social care settings, and for the very old or very ill.

REFERENCES

Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life of people with stroke over time: true change or response shift? *Qual Life Res, 14*(3), 611-627.

Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, S. R. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then test and the individualized approaches. *J Clin Epidemiol, 58*(11), 1125-1133. doi: 10.1016/j.jclinepi.2005.03.003

Alexopoulos, G. S., Abrams, R. C., Young, R. C., & Shamoian, C. A. (1988). Cornell Scale for Depression in Dementia. *Biol Psychiatry, 23*(3), 271-284. doi: 10.1016/0006-3223(88)90038-8

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *J Abnorm Psychol, 112*(4), 545-557. doi: 10.1037/0021-843X.112.4.545

Alzheimer's Society. (2013). Statistics about dementia. Retrieved 15 May, 2014, from http://www.alzheimers.org.uk/statistics

Andrykowski, M. A., & Hunt, J. W. (1993). Positive psychosocial adjustment in potential bone marrow transplant recipients: Cancer as a psychosocial transition. *Psycho-Oncology, 2*(4), 261-276. doi: 10.1002/pon.2960020406

Armenakis, A. (1988). A review of research on the change typology. *Research in organizational change and development, 2*, 163-194.

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397 - 438.

Asparouhov, T., & Muthén, B. (2010a). Multiple imputation with Mplus *Mplus Technical Appendices.* from http://www.statmodel.com/download/Imputations7.pdf

Asparouhov, T., & Muthén, B. (2010b). Weighted Least Squares Estimation with Missing Data. *Mplus Technical Appendices.* from http://www.statmodel.com/techappen.shtml

Bakas, T., McLennon, S. M., Carpenter, J. S., Buelow, J. M., Otte, J. L., Hanna, K. M., . . . Welch, J. L. (2012). Systematic review of health-related quality of life models. *Health Qual Life Outcomes, 10*(1), 134. doi: 10.1186/1477-7525-10-134

Ballard, C., Hanney, M. L., Theodoulou, M., Douglas, S., McShane, R., Kossakowski, K., . . . investigators, D.-A. (2009). The dementia antipsychotic withdrawal trial (DART-AD): long-term follow-up of a randomised placebo-controlled trial. *Lancet Neurol, 8*(2), 151-157. doi: 10.1016/S1474-4422(08)70295-3

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-275). Mahwah, NJ: Lawrence Erlbaum.

Banerjee, S. (2007). Commentary on "Health economics and the value of therapy in Alzheimer's disease." Quality of life in dementia: development and use of a disease-specific measure of health-related quality of life in dementia. *Alzheimers Dement, 3*(3), 166-171. doi: 10.1016/j.jalz.2007.04.384

Banerjee, S. (2012). The macroeconomics of dementia--will the world economy get Alzheimer's disease? *Arch Med Res, 43*(8), 705-709. doi: 10.1016/j.arcmed.2012.10.006

Banerjee, S., Hellier, J., Dewey, M., Romeo, R., Ballard, C., Baldwin, R., . . . Burns, A. (2011). Sertraline or mirtazapine for depression in dementia (HTA-SADD): a randomised, multicentre, double-blind, placebo-controlled trial. *Lancet, 378*(9789), 403-411. doi: 10.1016/S0140-6736(11)60830-1

Banerjee, S., Smith, S. C., Lamping, D. L., Harwood, R. H., Foley, B., Smith, P., . . . Knapp, M. (2006). Quality of life in dementia: more than just cognition. An analysis of associations with quality of life in dementia. *J Neurol Neurosurg Psychiatry, 77*(2), 146-148. doi: 10.1136/jnnp.2005.072983

Banerjee, S., Willis, R., Matthews, D., Contell, F., Chan, J., & Murray, J. (2007). Improving the quality of care for mild to moderate dementia: an evaluation of the Croydon Memory Service Model. *Int J Geriatr Psychiatry, 22*(8), 782-788. doi: 10.1002/gps.1741

Barclay, R., & Tate, R. B. (2014). Response shift recalibration and reprioritization in health-related quality of life was identified prospectively in older men with and without stroke. *J Clin Epidemiol, 67*(5), 500-507. doi: 10.1016/j.jclinepi.2013.12.003

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol Bull, 107*(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Bernhard, J., Lowy, A., Maibach, R., Hurny, C., Swiss Group for Clinical Cancer Research, & Swiss Institute for Applied Cancer Research. (2001).

Response shift in the perception of health for utility evaluation. an explorative investigation. *Eur J Cancer, 37*(14), 1729-1735.

Binder, M., & Coad, A. (2013). "I'm afraid I have bad news for you..." Estimating the impact of different health impairments on subjective well-being. *Soc Sci Med, 87*, 155-167. doi: 10.1016/j.socscimed.2013.03.025

Black, B. S., Johnston, D., Morrison, A., Rabins, P. V., Lyketsos, C. G., & Samus, Q. M. (2012). Quality of life of community-residing persons with dementia based on self-rated and caregiver-rated measures. *Qual Life Res, 21*(8), 1379-1389. doi: 10.1007/s11136-011-0044-z

Brazier, J., & Fitzpatrick, R. (2002). Measures of health-related quality of life in an imperfect world: a comment on Dowie. *Health Econ, 11*(1), 17-19; discussion 21-12. doi: 10.1002/hec.669

Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol, 51*(11), 1115-1128.

Breetvelt, I. S., & Van Dam, F. S. (1991). Underreporting by cancer patients: the case of response-shift. *Soc Sci Med, 32*(9), 981-987. doi: 10.1016/0277-9536(91)90156-7

Brod, M., Stewart, A. L., Sands, L., & Walton, P. (1999). Conceptualization and measurement of quality of life in dementia: the dementia quality of life instrument (DQoL). *Gerontologist, 39*(1), 25-35. doi: 10.1093/geront/39.1.25

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychol Assess, 25*(1), 136-145. doi: 10.1037/a0029228

Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behav Res Ther, 41*(12), 1411-1426. doi: 10.1016/s0005-7967(03)00059-7

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *J Pers, 80*(4), 796-846. doi: 10.1111/j.1467-6494.2011.00749.x

Bucks, R. S., Ashworth, D. L., Wilcock, G. K., & Siegfried, K. (1996). Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. *Age Ageing, 25*(2), 113-120.

Carlson, M., Wilcox, R., Chou, C. P., Chang, M., Yang, F., Blanchard, J., . . . Clark, F. (2011). Psychometric properties of reverse-scored items on the

CES-D in a sample of ethnically diverse older adults. *Psychol Assess, 23*(2), 558-562. doi: 10.1037/a0022484

Cassileth, B. R., Lusk, E. J., Strouse, T. B., Miller, D. S., Brown, L. L., Cross, P. A., & Tenaglia, A. N. (1984). Psychosocial status in chronic illness. A comparative analysis of six diagnostic groups. *N Engl J Med, 311*(8), 506-511. doi: 10.1056/NEJM198408233110805

Cattell, R. B. (1996). Psychological theory and scientific method. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 1–18). Chicago: Rand McNally.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. *J Pers, 80*(1), 219-251. doi: 10.1111/j.1467-6494.2011.00739.x

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225. doi: DOI 10.1207/s15327906mbr4102_5

Cherepanov, D., Palta, M., Fryback, D. G., & Robert, S. A. (2010). Gender differences in health-related quality-of-life are partly explained by sociodemographic and socioeconomic variation between adult men and women in the US: evidence from four US nationally representative data sets. *Qual Life Res, 19*(8), 1115-1124. doi: 10.1007/s11136-010-9673-x

Clark, P. G. (1995). Quality of life, values, and teamwork in geriatric care: do we communicate what we mean? *Gerontologist, 35*(3), 402-411.

Cockerham, W. C., Hinote, B. P., & Abbott, P. (2006). Psychological distress, gender, and health lifestyles in Belarus, Kazakhstan, Russia, and Ukraine. *Soc Sci Med, 63*(9), 2381-2394. doi: 10.1016/j.socscimed.2006.06.001

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale. experience from the New Haven EPESE study. *J Clin Epidemiol, 53*(3), 285-289. doi: 10.1016/s0895-4356(99)00151-1

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods, 6*(4), 330-351. doi: 10.1037/1082-989x.6.4.330

Conway, M., & Ross, M. (1984). Getting What You Want by Revising What You Had. *Journal of Personality and Social Psychology, 47*(4), 738-748. doi: Doi 10.1037/0022-3514.47.4.738

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Qual Life Res, 18*(4), 447-460. doi: 10.1007/s11136-009-9464-4

Cooper, C., Mukadam, N., Katona, C., Lyketsos, C. G., Ames, D., Rabins, P., . . . World Federation of Biological Psychiatry - Old Age Taskforce. (2012). Systematic review of the effectiveness of non-pharmacological interventions to improve quality of life of people with dementia. *Int Psychogeriatr, 24*(6), 856-870. doi: 10.1017/S1041610211002614

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi: 10.1037/0021-9010.78.1.98

Coyle, C. E., & Dugan, E. (2012). Social isolation, loneliness and health among older adults. *J Aging Health, 24*(8), 1346-1363. doi: 10.1177/0898264312460275

Cummings, J. L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D. A., & Gornbein, J. (1994). The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. *Neurology, 44*(12), 2308-2314. doi: 10.1212/wnl.44.12.2308

Damschroder, L. J., Zikmund-Fisher, B. J., & Ubel, P. A. (2005). The impact of considering adaptation in health state valuation. *Soc Sci Med, 61*(2), 267-277. doi: 10.1016/j.socscimed.2004.11.060

Damschroder, L. J., Zikmund-Fisher, B. J., & Ubel, P. A. (2008). Considering adaptation in preference elicitations. *Health Psychol, 27*(3), 394-399. doi: 10.1037/0278-6133.27.3.394

Djernes, J. K. (2006). Prevalence and predictors of depression in populations of elderly: a review. *Acta Psychiatr Scand, 113*(5), 372-387. doi: 10.1111/j.1600-0447.2006.00770.x

Dolan, P. (1996). The effect of experience of illness on health state valuations. *J Clin Epidemiol, 49*(5), 551-564. doi: 10.1016/0895-4356(95)00532-3

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Med Care, 35*(11), 1095-1108.

Drapeau, A., Beaulieu-Prevost, D., Marchand, A., Boyer, R., Preville, M., & Kairouz, S. (2010). A life-course and time perspective on the construct validity of psychological distress in women and men. Measurement invariance of the K6 across gender. *BMC Med Res Methodol, 10*, 68. doi: 10.1186/1471-2288-10-68

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire-Short Version: an evaluation of reverse-worded and non-reverse-worded items via item response theory. *J Pers Assess, 94*(4), 427-437. doi: 10.1080/00223891.2012.662188

Ebesutani, C., Reise, S. P., Chorpita, B. F., Ale, C., Regan, J., Young, J., . . . Weisz, J. R. (2012). The Revised Child Anxiety and Depression Scale-Short Version: scale reduction via exploratory bifactor modeling of the broad anxiety factor. *Psychol Assess, 24*(4), 833-845. doi: 10.1037/a0027283

Ebesutani, C., Smith, A., Bernstein, A., Chorpita, B. F., Higa-McMillan, C., & Nakamura, B. (2011). A bifactor model of negative affectivity: fear and distress components among younger and older youth. *Psychol Assess, 23*(3), 679-691. doi: 10.1037/a0023234

Edelaar-Peeters, Y., Putter, H., Snoek, G. J., Sluis, T. A., Smit, C. A., Post, M. W., & Stiggelbout, A. M. (2012). The influence of time and adaptation on health state valuations in patients with spinal cord injury. *Med Decis Making, 32*(6), 805-814. doi: 10.1177/0272989X12447238

Ettema, T. P., Droes, R. M., de Lange, J., Mellenbergh, G. J., & Ribbe, M. W. (2005). A review of quality of life instruments used in dementia. *Qual Life Res, 14*(3), 675-686.

EuroQol Group. (1990). EuroQol--a new facility for the measurement of health-related quality of life. *Health policy, 16*(3), 199.

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295. doi: Doi 10.1177/0146621605275728

Finch, W. H., & French, B. F. (2008). Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*(5), 742-759. doi: Doi 10.1177/0013164407313370

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Med Care, 41*(7 Suppl), III75-III86. doi: 10.1097/01.MLR.0000076052.42628.CF

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9*(4), 466-491. doi: 10.1037/1082-989X.9.4.466

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state" : A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res, 12*(3), 189-198.

Frick, U., Irving, H., & Rehm, J. (2012). Social relationships as a major determinant in the valuation of health states. *Qual Life Res, 21*(2), 209-213. doi: 10.1007/s11136-011-9945-0

Fryback, D. G., Dunham, N. C., Palta, M., Hanmer, J., Buechner, J., Cherepanov, D., . . . Kind, P. (2007). US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. *Med Care, 45*(12), 1162-1170. doi: 10.1097/MLR.0b013e31814848f1

Gallo, J. J., Rabins, P. V., Lyketsos, C. G., Tien, A. Y., & Anthony, J. C. (1997). Depression without sadness: Functional outcomes of nondysphoric depression in later life. *Journal of the American Geriatrics Society, 45*(5), 570-578.

Goldberg, D., & Williams, P. (1988). *A user's guide to the General Health Questionnaire.* Basingstoke: NFER-Nelson.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12*(2), 133-157. doi: 10.1177/002188637601200201

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography (Reprinted from The Wall Street Journal, 1994). *Intelligence, 24*(1), 13-23. doi: Doi 10.1016/S0160-2896(97)90011-8

Green, C., Brazier, J., & Deverill, M. (2000). Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics, 17*(2), 151-165.

Groenvold, M., Fayers, P. M., Sprangers, M. A. G., Bjorner, J. B., Klee, M. C., Aaronson, N. K., . . . Mouridsen, H. T. (1999). Anxiety and depression in breast cancer patients at low risk of recurrence compared with the general population: A valid comparison? *J Clin Epidemiol, 52*(6), 523-530. doi: Doi 10.1016/S0895-4356(99)00022-0

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97-121). Washington, DC: American Psychological Association.

Guyatt, G. H. (2002). Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions". *Health Econ, 11*(1), 9-12; discussion 21-12. doi: 10.1002/hec.666

Guyatt, G. H., King, D. R., Feeny, D. H., Stubbing, D., & Goldstein, R. S. (1999). Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. *J Clin Epidemiol, 52*(3), 187-192. doi: 10.1016/s0895-4356(98)00157-7

Halvorsrud, L., & Kalfoss, M. (2007). The conceptualization and measurement of quality of life in older adults: a review of empirical studies published during 1994-2006. *European Journal of Ageing, 4*(4), 229-246. doi: 10.1007/s10433-007-0063-3

Hanmer, J., Lawrence, W. F., Anderson, J. P., Kaplan, R. M., & Fryback, D. G. (2006). Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores. *Med Decis Making, 26*(4), 391-400. doi: 10.1177/0272989X06290497

Hoe, J., Katona, C., Roch, B., & Livingston, G. (2005). Use of the QOL-AD for measuring quality of life in people with severe dementia--the LASER-AD study. *Age Ageing, 34*(2), 130-135. doi: 10.1093/ageing/afi030

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2008). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44*(1), 153-166. doi: 10.1007/s11135-008-9190-y

Holmbeck, G., Devine, K., & Bruno, E. (2010). Developmental issues and considerations in research and practice. In J. Weisz & A. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (2 ed., pp. 28-39). New York, NY: Guilford Press.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp Aging Res, 18*(3-4), 117-144. doi: 10.1080/03610739208253916

Horning, S. M., Melrose, R., & Sultzer, D. (2014). Insight in Alzheimer's disease and its relation to psychiatric and behavioral disturbances. *Int J Geriatr Psychiatry, 29*(1), 77-84. doi: 10.1002/gps.3972

Howard, G. S., Dailey, P. R., & Gulanick, N. A. (1979). The feasibility of informed pretests in attenuating response-shift bias. *Appl Psychol Meas, 3*(4), 481-494. doi: 10.1177/014662167900300406

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Appl Psychol Meas, 3*(1), 1-23. doi: 10.1177/014662167900300101

Hughes, T. F., Flatt, J. D., Fu, B., Chang, C. C., & Ganguli, M. (2013). Engagement in social activities and progression from mild to severe cognitive impairment: the MYHAT study. *Int Psychogeriatr, 25*(4), 587-595. doi: 10.1017/S1041610212002086

Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol, 53*(5), 459-468.

Huxhold, O., Fiori, K. L., & Windsor, T. D. (2013). The dynamic interplay of social network characteristics, subjective well-being, and health: the costs and benefits of socio-emotional selectivity. *Psychol Aging, 28*(1), 3-16. doi: 10.1037/a0030170

Ichida, Y., Hirai, H., Kondo, K., Kawachi, I., Takeda, T., & Endo, H. (2013). Does social participation improve self-rated health in the older population? A quasi-experimental intervention study. *Soc Sci Med, 94*, 83-90. doi: 10.1016/j.socscimed.2013.05.006

Inaba, A., Thoits, P. A., Ueno, K., Gove, W. R., Evenson, R. J., & Sloan, M. (2005). Depression in the United States and Japan: gender, marital status, and SES patterns. *Soc Sci Med, 61*(11), 2280-2292. doi: 10.1016/j.socscimed.2005.07.014

Jenkinson, C., Gray, A., Doll, H., Lawrence, K., Keoghane, S., & Layte, R. (1997). Evaluation of index and profile measures of health status in a randomized controlled trial. Comparison of the Medical Outcomes Study 36-Item Short Form Health Survey, EuroQol, and disease specific measures. *Med Care, 35*(11), 1109-1118.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory Bi-factor Analysis. *Psychometrika, 76*(4), 537-549. doi: 10.1007/s11336-011-9218-4

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The Role of Referent Indicators in Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(4), 642-657. doi: 10.1080/10705510903206014

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: effects of differential item functioning. *J Gerontol B Psychol Sci Soc Sci, 57*(6), P548-558. doi: 10.1093/geronb/57.6.P548

Kagawa-Singer, M. (1993). Redefining health: living with cancer. *Soc Sci Med, 37*(3), 295-304. doi: 10.1016/0277-9536(93)90261-2

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When More Pain Is Preferred to Less - Adding a Better End. *Psychological Science, 4*(6), 401-405. doi: DOI 10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., & Snell, J. (2000). Prediciting utility. In R. A. M. Hogarth (Ed.), *Insights in decision making* (pp. 673-693). Chicago: University of Chicago Press.

Kamata, A., & Bauer, D. J. (2008). A Note on the Relation Between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 136-153. doi: 10.1080/10705510701758406

Kifley, A., Heller, G. Z., Beath, K. J., Bulger, D., Ma, J., & Gebski, V. (2012). Multilevel latent variable models for global health-related quality of life assessment. *Stat Med, 31*(11-12), 1249-1264. doi: 10.1002/sim.4455

King-Kallimanis, B. L., Oort, F. J., Nolte, S., Schwartz, C. E., & Sprangers, M. A. (2011). Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Qual Life Res, 20*(10), 1527-1540. doi: 10.1007/s11136-010-9844-9

Kline, R. (2011). Identification *Principles and practice of structural equation modeling* (3 ed.). New York, NY: Guilford Press.

Knapp, M., Iemmi, V., & Romeo, R. (2013). Dementia care costs and outcomes: a systematic review. *Int J Geriatr Psychiatry, 28*(6), 551-561. doi: 10.1002/gps.3864

Knapp, M., Prince, M., Albanese, E., Banerjee, S., Dhanasiri, S., Fernandez, J.-L., . . . Stewart, R. (2007). *Dementia UK*. Alzheimer's Society Retrieved from http://www.alzheimers.org.uk/site/scripts/download_info.php?fileID=2.

Lawrence, V., Samsi, K., Banerjee, S., Morgan, C., & Murray, J. (2011). Threat to valued elements of life: the experience of dementia across three ethnic groups. *Gerontologist, 51*(1), 39-50. doi: 10.1093/geront/gnq073

Lawton, M. P. (1994). Quality of life in Alzheimer disease. *Alzheimer Dis Assoc Disord, 8 Suppl 3*(3), 138-150.

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., & Thogersen-Ntoumani, C. (2012). Method effects: the problem with

negatively versus positively keyed items. *J Pers Assess, 94*(2), 196-204. doi: 10.1080/00223891.2011.645936

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological methods, 4*(2), 192-211. doi: Doi 10.1037//1082-989x.4.2.192

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological methods, 18*(3), 285-300. doi: 10.1037/a0033266

Loewenstein, G., & Frederick, S. (1997). Predicting reactions to environmental change. In D. Bazerman, D. Messick, A. Tenbrunsel & K. Wade-Benzoni (Eds.), *Environment, ethics, and behaviour*. San Francisco: New Lexington Press.

Lou, V. W., Chi, I., Kwan, C. W., & Leung, A. Y. (2013). Trajectories of social engagement and depressive symptoms among long-term care facility residents in Hong Kong. *Age Ageing, 42*(2), 215-222. doi: 10.1093/ageing/afs159

Lucas-Carrasco, R., Lamping, D. L., Banerjee, S., Rejas, J., Smith, S. C., & Gomez-Benito, J. (2010). Validation of the Spanish version of the DEMQOL system. *Int Psychogeriatr, 22*(4), 589-597. doi: 10.1017/S1041610210000207

Lyketsos, C. G., Gonzales-Salvador, T., Chin, J. J., Baker, A., Black, B., & Rabins, P. (2003). A follow-up study of change in quality of life among persons with dementia residing in a long-term care facility. *Int J Geriatr Psychiatry, 18*(4), 275-281. doi: 10.1002/gps.796

MacRae, H. (2011). Self and other: The importance of social interaction and social relationships in shaping the experience of early-stage Alzheimer's disease. *Journal of Aging Studies, 25*(4), 445-456. doi: 10.1016/j.jaging.2011.06.001

Marsh, H. W. (1986). Negative Item Bias in Ratings Scales for Preadolescent Children - a Cognitive Developmental Phenomenon. *Developmental Psychology, 22*(1), 37-49. doi: Doi 10.1037/0012-1649.22.1.37

Marsh, H. W. (1989). Confirmatory Factor-Analyses of Multitrait-Multimethod Data - Many Problems and a Few Solutions. *Applied Psychological Measurement, 13*(4), 335-361. doi: Doi 10.1177/014662168901300402

Marsh, H. W. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*(4), 810-819.

Marsh, H. W. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling-a Multidisciplinary Journal, 11*(3), 320-341. doi: DOI 10.1207/s15328007sem1103_2

Marsh, H. W., & Bailey, M. (1991). Confirmatory Factor-Analyses of Multitrait-Multimethod Data - a Comparison of Alternative Models. *Applied Psychological Measurement, 15*(1), 47-70. doi: Doi 10.1177/014662169101500106

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is More Ever Too Much? The Number of Indicators per Factor in Confirmatory Factor Analysis. *Multivariate Behavioral Research, 33*(2), 181-220. doi: 10.1207/s15327906mbr3302_1

Marsh, H. W., Ludtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right--camouflaging misspecification with item parcels in CFA models. *Psychological methods, 18*(3), 257-284. doi: 10.1037/a0032773

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: traits, ephemeral artifacts, and stable response styles. *Psychol Assess, 22*(2), 366-381. doi: 10.1037/a0019225

Matud, M. P. (2004). Gender differences in stress and coping styles. *Personality and Individual Differences, 37*(7), 1401-1415. doi: DOI 10.1016/j.paid.2004.01.010

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357-367. doi: Doi 10.1177/014662169802200404

McDonald, R. P. (1999). *Test theory: A unified treatment*: Lawrence Erlbaum.

McPhail, S., & Haines, T. (2010). Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health Qual Life Outcomes, 8*, 65. doi: 10.1186/1477-7525-8-65

McTaggart-Cowan, H. M., Marra, C. A., Yang, Y., Brazier, J. E., Kopec, J. A., FitzGerald, J. M., . . . Lynd, L. D. (2008). The validity of generic and

condition-specific preference-based instruments: the ability to discriminate asthma control status. *Qual Life Res, 17*(3), 453-462. doi: 10.1007/s11136-008-9309-6

McTaggart-Cowan, H. M., Tsuchiya, A., O'Cathain, A., & Brazier, J. (2011). Understanding the effect of disease adaptation information on general population values for hypothetical health states. *Soc Sci Med, 72*(11), 1904-1912. doi: 10.1016/j.socscimed.2011.03.036

Menzel, P., Dolan, P., Richardson, J., & Olsen, J. A. (2002). The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. *Soc Sci Med, 55*(12), 2149-2158. doi: 10.1016/s0277-9536(01)00358-6

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi: 10.1007/bf02294825

Meszaros, V., Adam, S., Szabo, M., Szigeti, R., & Urban, R. (2014). The bifactor model of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS)--an alternative measurement model of burnout. *Stress Health, 30*(1), 82-88. doi: 10.1002/smi.2481

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42*(5), 869-874. doi: 10.1016/j.paid.2006.09.022

Millsap, R. E., & Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology, 73*(3), 574-584. doi: 10.1037/0021-9010.73.3.574

Mirowsky, J. (1996). Age and the gender gap in depression. *J Health Soc Behav, 37*(4), 362-380.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology, 43*(11), 2412-2414.

Moshagen, M., & Musch, J. (2014). Sample Size Requirements of the Robust Weighted Least Squares Estimator. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10*(2), 60-70. doi: Doi 10.1027/1614-2241/A000068

Moyle, W., Murfield, J. E., Griffiths, S. G., & Venturato, L. (2012). Assessing quality of life of older people with dementia: a comparison of quantitative self-report and proxy accounts. *J Adv Nurs, 68*(10), 2237-2246. doi: 10.1111/j.1365-2648.2011.05912.x

Mozley, C. G., Huxley, P., Sutcliffe, C., Bagley, H., Burns, A., Challis, D., & Cordingley, L. (1999). 'Not knowing where I am doesn't mean I don't

know what I like': cognitive impairment and quality of life responses in elderly people. *Int J Geriatr Psychiatry, 14*(9), 776-783.

Mulhern, B., Rowen, D., Brazier, J., Smith, S., Romeo, R., Tait, R., . . . Banerjee, S. (2013). Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. *Health Technol Assess, 17*(5), v-xv, 1-140. doi: 10.3310/hta17050

Muthén, B. (2013). Weighted least squares and the theta parameterization. *Mplus Language Addendum (Version 7.1).* from http://www.statmodel.com/download/Version7.1xLanguage.pdf

Muthén, B., & Asparouhov, T. (2002). Mplus Web Notes No. 4: Latent Variable Analysis With Categorical Outcomes: Multiple-Group And Growth Modeling In Mplus. from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B., du Toit, S., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. from http://www.statmodel.com/wlscv.shtml

Muthén, L., & Muthén, B. (1998-2012). *Mplus User's Guide. Seventh Edition.* Muthén & Muthén (Eds.), Retrieved from http://www.statmodel.com/ugexcerpts.shtml

NAO. (2007). *Improving services and support for people with dementia*. London, UK: National Audit Office Retrieved from http://www.nao.org.uk/report/improving-services-and-support-for-people-with-dementia/.

NICE SCIE. (2006). *Supporting people with dementia and their carers in health and social care*. London, UK: National Institute for Health and Clinical Excellence and the Social Care Institute for Excellence Retrieved from http://www.scie.org.uk/publications/misc/dementia/.

Norman, P., & Parker, S. (1996). The interpretation of change in verbal reports: Implications for health psychology. *Psychology & Health, 11*(2), 301-314. doi: Doi 10.1080/08870449608400259

Norton, S., Cosco, T., Doyle, F., Done, J., & Sacker, A. (2013). The Hospital Anxiety and Depression Scale: a meta confirmatory factor analysis. *J Psychosom Res, 74*(1), 74-81. doi: 10.1016/j.jpsychores.2012.10.010

Novella, J. L., Jochum, C., Jolly, D., Morrone, I., Ankri, J., Bureau, F., & Blanchard, F. (2001). Agreement between patients' and proxies' reports of quality of life in Alzheimer's disease. *Qual Life Res, 10*(5), 443-452.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Qual Life Res, 14*(3), 587-598. doi: DOI 10.1007/s11136-004-0830-y

Ozer, D., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. In S. Fiske, A. Kazdin & D. Schacter (Eds.), *Annual review of psychology* (Vol. 57, pp. 401-421). Palo Alto, CA: Annual Reviews.

Padilla, G. V., Mishel, M. H., & Grant, M. M. (1992). Uncertainty, appraisal and quality of life. *Qual Life Res, 1*(3), 155-165. doi: 10.1007/BF00635615

Perales, J., Cosco, T. D., Stephan, B. C., Haro, J. M., & Brayne, C. (2013). Health-related quality-of-life instruments for Alzheimer's disease and mixed dementia. *Int Psychogeriatr, 25*(5), 691-706. doi: 10.1017/S1041610212002293

Postulart, D., & Adang, E. M. (2000). Response shift and adaptation in chronically ill patients. *Med Decis Making, 20*(2), 186-193. doi: 10.1177/0272989x0002000204

Prince, M., Ferri, C. P., Acosta, D., Albanese, E., Arizaga, R., Dewey, M., . . . Uwakwe, R. (2007). The protocols for the 10/66 dementia research group population-based research programme. *BMC Public Health, 7*, 165. doi: 10.1186/1471-2458-7-165

Rabins, P. V., & Black, B. S. (2007). Measuring quality of life in dementia: purposes, goals, challenges and progress. *Int Psychogeriatr, 19*(3), 401-407. doi: 10.1017/S1041610207004863

Ready, R. E., Ott, B. R., & Grace, J. (2006). Insight and cognitive impairment: effects on quality-of-life reports from mild cognitive impairment and Alzheimer's disease patients. *Am J Alzheimers Dis Other Demen, 21*(4), 242-248. doi: 10.1177/1533317506290589

Reininghaus, U., McCabe, R., Burns, T., Croudace, T., & Priebe, S. (2012). The validity of subjective quality of life measures in psychotic patients with severe psychopathology and cognitive deficits: an item response model analysis. *Qual Life Res, 21*(2), 237-246. doi: 10.1007/s11136-011-9936-1

Reise, S. P. (2012). Invited Paper: The Rediscovery of Bifactor Measurement Models. *Multivariate Behav Res, 47*(5), 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess, 95*(2), 129-140. doi: 10.1080/00223891.2012.725437

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. *J Pers Assess, 92*(6), 544-559. doi: 10.1080/00223891.2010.496477

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res, 16 Suppl 1*(S1), 19-31. doi: 10.1007/s11136-007-9183-7

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annu Rev Clin Psychol, 5*, 27-48. doi: 10.1146/annurev.clinpsy.032408.153553

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull, 114*(3), 552-566.

Revell, A. J., Caskie, G. I., Willis, S. L., & Schaie, K. W. (2009). Factor structure and invariance of the Quality of Life in Alzheimer's Disease (QoL-AD) Scale. *Exp Aging Res, 35*(2), 250-267. doi: 10.1080/03610730902720521

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods, 17*(3), 354-373. doi: 10.1037/a0029315

Rook, K. S., Luong, G., Sorkin, D. H., Newsom, J. T., & Krause, N. (2012). Ambivalent versus problematic social ties: implications for psychological health, functional health, and interpersonal coping. *Psychol Aging, 27*(4), 912-923. doi: 10.1037/a0029246

Rosen, L. H., Beron, K. J., & Underwood, M. K. (2013). Assessing peer victimization across adolescence: measurement invariance and developmental change. *Psychol Assess, 25*(1), 1-11. doi: 10.1037/a0028985

Ross, M. (1989). Relation of Implicit Theories to the Construction of Personal Histories. *Psychological Review, 96*(2), 341-357. doi: Doi 10.1037/0033-295x.96.2.341

Sackett, D. L., & Torrance, G. W. (1978). The utility of different health states as perceived by the general public. *J Chronic Dis, 31*(11), 697-704. doi: 10.1016/0021-9681(78)90072-3

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *Br J Math Stat Psychol, 66*(2), 201-223. doi: 10.1111/j.2044-8317.2012.02049.x

Schmitt, N. (1982). The Use of Analysis of Covariance-Structures to Assess Beta-Change and Gamma-Change. *Multivariate Behavioral Research, 17*(3), 343-358. doi: DOI 10.1207/s15327906mbr1703_3

Scholzel-Dorenbos, C. J., van der Steen, M. J., Engels, L. K., & Olde Rikkert, M. G. (2007). Assessment of quality of life as outcome in dementia and MCI intervention trials: a systematic review. *Alzheimer Dis Assoc Disord, 21*(2), 172-178. doi: 10.1097/WAD.0b013e318047df4c

Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Qual Life Res, 15*(9), 1533-1550. doi: 10.1007/s11136-006-0025-9

Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med, 48*(11), 1531-1548.

Schwartz, C. E., & Sprangers, M. A. (2010). Guidelines for improving the stringency of response shift research using the thentest. *Qual Life Res, 19*(4), 455-464. doi: 10.1007/s11136-010-9585-9

Sijtsma, K. (2009a). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107-120. doi: 10.1007/s11336-008-9101-0

Sijtsma, K. (2009b). Reliability Beyond Theory and Into Practice. *Psychometrika, 74*(1), 169-173. doi: 10.1007/s11336-008-9103-y

Smith, D. M., Sherriff, R. L., Damschroder, L., Loewenstein, G., & Ubel, P. A. (2006). Misremembering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychol, 25*(6), 688-695. doi: 10.1037/0278-6133.25.6.688

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychol Assess, 12*(1), 102-111.

Smith, S. C., Lamping, D. L., Banerjee, S., Harwood, R., Foley, B., Smith, P., . . . Knapp, M. (2005). Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess, 9*(10), 1-93, iii-iv. doi: 10.3310/hta9100

Smith, S. C., Lamping, D. L., Banerjee, S., Harwood, R. H., Foley, B., Smith, P., . . . Knapp, M. (2007). Development of a new measure of health-related quality of life for people with dementia: DEMQOL. *Psychol Med, 37*(5), 737-746. doi: 10.1017/S0033291706009469

Smith, S. C., Murray, J., Banerjee, S., Foley, B., Cook, J. C., Lamping, D. L., . . . Mann, A. (2005). What constitutes health-related quality of life in dementia? Development of a conceptual framework for people with dementia and their carers. *Int J Geriatr Psychiatry, 20*(9), 889-895. doi: 10.1002/gps.1374

Sneeuw, K. C., Sprangers, M. A., & Aaronson, N. K. (2002). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease. *J Clin Epidemiol, 55*(11), 1130-1143.

Snow, A. L., Kunik, M. E., Molinari, V. A., Orengo, C. A., Doody, R., Graham, D. P., & Norris, M. P. (2005). Accuracy of self-reported depression in persons with dementia. *J Am Geriatr Soc, 53*(3), 389-396. doi: 10.1111/j.1532-5415.2005.53154.x

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47-103). Hillsdale, NJ: Erlbaum.

Spijker, A., Vernooij-Dassen, M., Vasse, E., Adang, E., Wollersheim, H., Grol, R., & Verhey, F. (2008). Effectiveness of nonpharmacological interventions in delaying the institutionalization of patients with dementia: a meta-analysis. *J Am Geriatr Soc, 56*(6), 1116-1128. doi: 10.1111/j.1532-5415.2008.01705.x

Sprangers, M. A. (1996). Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. *Cancer Treat Rev, 22 Suppl A*, 55-62. doi: 10.1016/S0305-7372(96)90064-X

Sprangers, M. A., & Aaronson, N. K. (1992). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol, 45*(7), 743-760.

Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med, 48*(11), 1507-1515. doi: 10.1016/S0277-9536(99)00045-3

Stansbury, J. P., Ried, L. D., & Velozo, C. A. (2006). Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) Scale. *J Pers Assess, 86*(1), 10-22. doi: 10.1207/s15327752jpa8601_03

Steeman, E., Tournoy, J., Grypdonck, M., Godderis, J., & De Casterle, B. D. (2011). Managing identity in early-stage dementia: maintaining a sense of being valued. *Ageing and Society, 33*(02), 216-242. doi: 10.1017/s0144686x11001115

Steiger, J. H. (1990). Structural Model Evaluation and Modification - an Interval Estimation Approach. *Multivariate Behavioral Research, 25*(2), 173-180. doi: DOI 10.1207/s15327906mbr2502_4

Steinberg, L., Sharp, C., Stanford, M. S., & Tharp, A. T. (2013). New tricks for an old measure: the development of the Barratt Impulsiveness Scale-Brief (BIS-Brief). *Psychol Assess, 25*(1), 216-226. doi: 10.1037/a0030550

Stiggelbout, A. M., & de Vogel-Voogt, E. (2008). Health state utilities: a framework for studying the gap between the imagined and the real. *Value Health, 11*(1), 76-87. doi: 10.1111/j.1524-4733.2007.00216.x

Takane, Y., & Deleeuw, J. (1987). On the Relationship between Item Response Theory and Factor-Analysis of Discretized Variables. *Psychometrika, 52*(3), 393-408. doi: Doi 10.1007/Bf02294363

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Med Care, 44*(11 Suppl 3), S152-170. doi: 10.1097/01.mlr.0000245142.74628.ab

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Qual Life Res, 16 Suppl 1*(S1), 33-42. doi: 10.1007/s11136-007-9184-6

Thies, W., Bleiler, L., & Alzheimer's Association. (2013). 2013 Alzheimer's disease facts and figures. *Alzheimers Dement, 9*(2), 208-245. doi: 10.1016/j.jalz.2013.02.003

Tomás, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining Method Effects Associated With Negatively Worded Items in Trait and State Global and Domain-Specific Self-Esteem Scales. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 299-313. doi: 10.1080/10705511.2013.769394

Trigg, R., Jones, R. W., & Skevington, S. M. (2007). Can people with mild to moderate dementia provide reliable answers about their quality of life? *Age Ageing, 36*(6), 663-669. doi: 10.1093/ageing/afm077

Trigg, R., Watts, S., Jones, R., & Tod, A. (2011). Predictors of quality of life ratings from persons with dementia: the role of insight. *Int J Geriatr Psychiatry, 26*(1), 83-91. doi: 10.1002/gps.2494

Tsevat, J., Cook, E. F., Green, M. L., Matchar, D. B., Dawson, N. V., Broste, S. K., . . . Goldman, L. (1995). Health values of the seriously ill. *Ann Intern Med, 122*(7), 514-520. doi: 10.7326/0003-4819-122-7-199504010-00007

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. doi: 10.1007/bf02291170

Ubel, P. A., Loewenstein, G., & Jepson, C. (2003). Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res, 12*(6), 599-607. doi: 10.1023/A:1025119931010

Ubel, P. A., Loewenstein, G., & Jepson, C. (2005). Disability and sunshine: can hedonic predictions be improved by drawing attention to focusing illusions or emotional adaptation? *J Exp Psychol Appl, 11*(2), 111-123. doi: 10.1037/1076-898X.11.2.111

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. *PLoS One, 8*(7), e68967. doi: 10.1371/journal.pone.0068967

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-70. doi: 10.1177/109442810031002

Vasse, E., Vernooij-Dassen, M., Cantegreil, I., Franco, M., Dorenlot, P., Woods, B., & Moniz-Cook, E. (2012). Guidelines for psychosocial interventions in dementia care: a European survey and comparison. *Int J Geriatr Psychiatry, 27*(1), 40-48. doi: 10.1002/gps.2687

Vogel, A., Mortensen, E. L., Hasselbalch, S. G., Andersen, B. B., & Waldemar, G. (2006). Patient versus informant reported quality of life in the earliest phases of Alzheimer's disease. *Int J Geriatr Psychiatry, 21*(12), 1132-1138. doi: 10.1002/gps.1619

Walters, S. J., & Brazier, J. E. (2005). Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res, 14*(6), 1523-1532.

Wang, W. C., & Shih, C. L. (2010). MIMIC Methods for Assessing Differential Item Functioning in Polytomous Items. *Applied Psychological Measurement, 34*(3), 166-180. doi: 10.1177/0146621609355279

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.),

Structural equation modeling: Concepts, issues, and applications (pp. 56-75). Thousand Oaks, CA: Sage Publications.

Willse, J. T., & Goodman, J. T. (2008). Comparison of multiple-indicators, multiple-causes- and item response theory-based analyses of subgroup differences. *Educational and Psychological Measurement, 68*(4), 587-602. doi: Doi 10.1177/0013164407312601

Wimo, A., & Prince, M. (2010). *World Alzheimer Report 2010: The global economic impact of dementia*. Alzheimer's Disease International Retrieved from http://www.alz.co.uk/research/worldreport/.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement, 73*(6), 913-934. doi: 10.1177/0013164413495237

Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behavioral Research, 44*(1), 1-27. doi: 10.1080/00273170802620121

World Health Organization, & Alzheimer's Disease International. (2012). Dementia: a public health priority. from http://www.who.int/mental_health/publications/dementia_report_2012/en/

Yang, F. M., & Jones, R. N. (2007). Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *J Clin Epidemiol, 60*(11), 1195-1200. doi: 10.1016/j.jclinepi.2007.02.008

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: an extension of the bi-factor item response theory model for use in differential item functioning. *J Psychiatr Res, 43*(12), 1025-1035. doi: 10.1016/j.jpsychires.2008.12.007

Yang, X., Li, J., & Shoptaw, S. (2008). Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Stat Med, 27*(15), 2826-2849. doi: 10.1002/sim.3111

Yang, Y., Sun, Y., Zhang, Y., Jiang, Y., Tang, J., Zhu, X., & Miao, D. (2013). Bifactor item response theory model of acute stress response. *PLoS One, 8*(6), e65291. doi: 10.1371/journal.pone.0065291

Yesavage, J., & Sheikh, J. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist, 5*(1), 165-173. doi: 10.1300/J018v05n01_09

Yoo, J. E. (2009). The Effect of Auxiliary Variables and Multiple Imputation on Parameter Estimation in Confirmatory Factor Analysis. *Educational and Psychological Measurement, 69*(6), 929-947. doi: Doi 10.1177/0013164409332225

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. *Doctoral dissertation, University of California, Los Angeles*. http://statmodel2.com/download/Yudissertation.pdf

Zarit, S. H., Reever, K. E., & Bach-Peterson, J. (1980). Relatives of the impaired elderly: correlates of feelings of burden. *Gerontologist, 20*(6), 649-655. doi: 10.1093/geront/20.6.649

Zhang, X. H., Li, S. C., Xie, F., Lo, N. N., Yang, K. Y., Yeo, S. J., . . . Thumboo, J. (2012). An exploratory study of response shift in health-related quality of life and utility assessment among patients with osteoarthritis undergoing total knee replacement surgery in a tertiary hospital in Singapore. *Value Health, 15*(1 Suppl), S72-78. doi: 10.1016/j.jval.2011.11.011

Zinbarg, R. E. (2006). Estimating Generalizability to a Latent Variable Common to All of a Scale's Indicators: A Comparison of Estimators for h. *Applied Psychological Measurement, 30*(2), 121-144. doi: 10.1177/0146621605278814

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. doi: 10.1007/s11336-003-0974-7

APPENDICES TO CHAPTER 2

Multiple imputation (DEMQOL) syntax

```
TITLE: Multiple imputation for baseline DEMQOL (Croydon)

DATA:
FILE = rawdem.dat ;

VARIABLE:
NAMES =
id
a1m0 a2m0 a3m0 a4m0 a5m0 a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0 a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0 a26m0 a27m0 a28m0
n12 age gender mxa0 kqa0 qa0 qc0 qd0 qe0 qf0 qg0 icd ;

!! NOTE: variable order will change after imputation

AUXILIARY =
id n12 age gender mxa0 icd kqa0 ;

USEVAR =
qa0 qc0 qd0 qe0 qf0 qg0          !! auxiliary variables for imputation
a1m0 - a28m0 ;

!! mmse (qa0), npi (qc0), gds (qd0), badl (qe0) zarit (qf0), ghq (qg0)

MISSING = All (-1234) ;

DATA IMPUTATION:
IMPUTE = a1m0 - a28m0 (c)  ;     !! impute as categorical variables

NDATASETS = 100 ;

SAVE = dem*.dat ;

ANALYSIS:
TYPE = Basic ;

OUTPUT:
Tech8 ;
```

Multiple imputation (DEMQOL-Proxy) syntax

```
TITLE: Multiple imputation for baseline DEMQOL-Proxy (Croydon)

DATA:
FILE = rawdemc.dat ;

VARIABLE:
NAMES =
id
b1m0 b2m0 b3m0 b4m0 b5m0 b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b31m0
n12 age gender mxa0 kqa0 qa0 qc0 qd0 qe0 qf0 qg0 icd ;

!! NOTE: variable order will change after imputation

AUXILIARY =
id n12 age gender mxa0 icd kqa0 ;

USEVAR =
qa0 qc0 qd0 qe0 qf0 qg0          !! auxiliary variables for imputation
a1m0 - a28m0 ;

MISSING = All (-1234) ;

DATA IMPUTATION:
IMPUTE = b1m0 - b31m0 (c)  ;     !! impute as categorical variables

NDATASETS = 100 ;

SAVE = demc*.dat ;

ANALYSIS:
TYPE = Basic ;

OUTPUT:
Tech8 ;
```

Exploratory bifactor analysis (DEMQOL) syntax

```
TITLE:
Hierarchical bifactor EFA for baseline DEMQOL (Croydon)

DATA:
FILE = demlist.dat ; !100 imputed data sets (n=1240)
TYPE = IMPUTATION ;

VARIABLE:
NAMES =
qa0 qc0 qd0 qe0 qf0 qg0
a1m0 a2m0 a3m0 a4m0 a5m0
a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0
a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0
a26m0 a27m0 a28m0
id n12 age gender
mxa0 icd kqa0 ;

!! use variable order in impute100dem.out (SAVEDATA INFORMATION)

MISSING = ALL (-1234) ;

USEVAR =
a1m0 a2m0 a3m0 a4m0 a5m0
a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0
a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0
a26m0 a27m0 a28m0 ;

CATEGORICAL = ALL ;

ANALYSIS:
ESTIMATOR = WLSMV ;
ROTATION = BI-GEOMIN (ORTHOGONAL) ;

MODEL:
fg f1 f2 f3 f4 f5 BY
a1m0 a2m0 a3m0 a4m0 a5m0
a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0
a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0
a26m0 a27m0 a28m0(*1) ;

OUTPUT: STDY ;
```

Exploratory bifactor analysis (DEMQOL-Proxy) syntax

```
TITLE:
Hierarchical bifactor EFA for baseline DEMQOL-Proxy (Croydon)

DATA:
FILE = demclist.dat ; !100 imputed data sets (n=1240)
TYPE = IMPUTATION ;

VARIABLE:
NAMES =
qa0 qc0 qd0 qe0 qf0 qg0
b1m0 b2m0 b3m0 b4m0 b5m0
b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0
b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0
b26m0 b27m0 b28m0 b29m0 b30m0
b31m0
id n12 age gender
mxa0 icd kqa0 ;

!! use variable order in impute100demc.out (SAVEDATA INFORMATION)

MISSING = ALL (-1234) ;

USEVAR =
b1m0 b2m0 b3m0 b4m0 b5m0
b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0
b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0
b26m0 b27m0 b28m0 b29m0 b30m0
b31m0;

CATEGORICAL = ALL ;

ANALYSIS:
ESTIMATOR = WLSMV ;
ROTATION = BI-GEOMIN (ORTHOGONAL) ;

MODEL:
fg f1 f2 f3 f4 f5 BY
b1m0 b2m0 b3m0 b4m0 b5m0
b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0
b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0
b26m0 b27m0 b28m0 b29m0 b30m0
b31m0 (*1) ;

OUTPUT: STDY ;
```

Bifactor confirmatory factor analysis (DEMQOL) syntax

```
TITLE: Bifactor CFA for baseline DEMQOL (CMS)
# Non-imputed data set (n=868)
# Testlet: 4

DATA:
FILE = rawdem.dat ;

DEFINE:
lon2 = a8m0 + a20m0 ;
soc2 = a21m0 + a22m0 ;
ovh2 = a27m0 + a28m0 ;
liv2 = a6m0 + a10m0 ;

VARIABLE:
NAMES =
id
A1M0 A2M0 A3M0 A4M0 A5M0
A6M0 A7M0 A8M0 A9M0 A10M0
A11M0 A12M0 A13M0 A14M0 A15M0
A16M0 A17M0 A18M0 A19M0 A20M0
A21M0 A22M0 A23M0 A24M0 A25M0
A26M0 A27M0 A28M0
n12 age gender
mxa0 kqa0
QA0 QC0 QD0 QE0 QF0 QG0
ICD ;

!! use variable order in rawdem.out

MISSING = ALL (-1234) ;

USEVAR =
a1m0 a2m0 a3m0 a4m0 a5m0
!!a6m0 a7m0 a8m0 a9m0 a10m0
a7m0 a9m0
a11m0 a12m0 a13m0 a14m0 a15m0
!!a16m0 a17m0 a18m0 a19m0 a20m0
a16m0 a17m0 a18m0 a19m0
!!a21m0 a22m0 a23m0 a24m0 a25m0
a23m0 a24m0 a25m0
!!a26m0 a27m0 a28m0
a26m0
lon2 soc2 ovh2 liv2 ;

CATEGORICAL = ALL lon2 soc2 ovh2 liv2 ;

ANALYSIS:
!ESTIMATOR = WLSMV ; !!Default
DIFFTEST = DFT_G4vG3a.dat ;
!DIFFTEST = DFT_G4vG3b.dat ;
!DIFFTEST = DFT_G4vG2.dat ;

MODEL:

qol BY
a1m0 a2m0 a3m0 a4m0 a5m0
a7m0 a9m0
a11m0 a12m0 a13m0 a14m0 a15m0
a16m0 a17m0 a18m0 a19m0
```

```
a23m0 a24m0 a25m0
a26m0
lon2 soc2 ovh2 liv2 ;

pos BY a3m0 a5m0 a1m0 liv2 ;
cog BY a19m0 a18m0 a17m0 a16m0 a15m0 a14m0 ;
neg BY a13m0 a12m0 a11m0 a4m0 ;
!soc BY a25m0 a24m0 a23m0 soc2 ;

!!G3S A
qol WITH pos@0 cog@0 neg@0 ;
pos WITH cog@0 neg@0 ;
cog WITH neg@0 ;

!!G3S B
!qol WITH pos@0 cog@0 soc@0 ;
!pos WITH cog@0 soc@0 ;
!cog WITH soc@0 ;

!!G4S
!qol WITH pos@0 cog@0 neg@0 soc@0 ;
!pos WITH cog@0 neg@0 soc@0 ;
!cog WITH neg@0 soc@0 ;
!neg WITH soc@0 ;

OUTPUT: STDYX Residual Modindices ;

SAVEDATA:
!DIFFTEST = DFT_G4vG3a.dat ;
!DIFFTEST = DFT_G4vG3b.dat ;
```

Bifactor confirmatory factor analysis (DEMQOL-Proxy) syntax

```
TITLE: Bifactor CFA for baseline DEMQOL-Proxy (CMS)
# Non-imputed data set (n=909)
# Testlet: 5

DATA:
FILE = rawdemc.dat ;

DEFINE:
app2 = b21m0 + b22m0 ;      !polyr=.817
mem2 = b12m0 + b14m0 ;      !polyr=.772
liv2 = b4m0 + b8m0 ;        !polyr=.771
fin2 = b24m0 + b25m0 ;      !polyr=.751
use2 = b29m0 + b30m0 ;      !polyr=.705

VARIABLE:
NAMES =
id
b1m0 b2m0 b3m0 b4m0 b5m0 b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b31m0
n12 age gender mxa0
kqa0 qa0 qc0 qd0 qe0 qf0 qg0 icd;

MISSING = ALL (-1234) ;

USEVAR =
!b1m0 b2m0 b3m0 b4m0 b5m0 b6m0 b7m0 b8m0 b9m0 b10m0
b1m0 b2m0 b3m0 b5m0 b6m0 b7m0 b9m0 b10m0
!b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b11m0 b13m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
!b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b23m0 b26m0 b27m0 b28m0
b31m0
app2 mem2 liv2 fin2 use2 ;

Categorical = ALL app2 mem2 liv2 fin2 use2 ;

ANALYSIS:
!ESTIMATOR = WLSMV ; !!Default
!DIFFTEST = DFT_G4vG3a.dat ;
DIFFTEST = DFT_G4vG3b.dat ;

MODEL:

qol BY
b1m0 b2m0 b3m0 b5m0 b6m0 b7m0 b9m0 b10m0
b11m0 b13m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b23m0 b26m0 b27m0 b28m0
b31m0
app2 mem2 liv2 fin2 use2 ;

neg BY b10m0 b9m0 b7m0 b5m0 b3m0 b2m0 ;
pos BY b11m0 b6m0 b1m0 liv2 ;
!soc BY b28m0 b27m0 use2 ;
cog BY b20m0 b19m0 b18m0 b17m0 b16m0 b15m0 b13m0 mem2 ;

!!G3A
!qol WITH neg@0 pos@0 soc@0 ;
```

```
!neg WITH pos@0 soc@0 ;
!pos WITH soc@0 ;

!!G3B
qol WITH neg@0 pos@0 cog@0 ;
neg WITH pos@0 cog@0 ;
pos WITH cog@0 ;

!!G4
!qol WITH neg@0 pos@0 soc@0 cog@0 ;
!neg WITH pos@0 soc@0 cog@0 ;
!pos WITH soc@0 cog@0 ;
!soc WITH cog@0 ;

OUTPUT: STDYX Residual Modindices ;

SAVEDATA:
!DIFFTEST = DFT_G4vG3a.dat ;
!DIFFTEST = DFT_G4vG3b.dat ;
```

Polychoric correlation matrix based on 100 imputed data sets of DEMQOL

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27
Q2	.34																										
Q3	.60	.29																									
Q4	.31	.48	.26																								
Q5	.45	.36	.49	.32																							
Q6	.48	.29	.52	.31	.51																						
Q7	.46	.52	.44	.43	.37	.30																					
Q8	.33	.33	.34	.31	.23	.16	.50																				
Q9	.38	.60	.32	.49	.38	.22	.63	.47																			
Q10	.43	.26	.52	.27	.52	.69	.35	.16	.18																		
Q11	.37	.42	.25	.56	.25	.18	.48	.30	.49	.13																	
Q12	.46	.48	.43	.63	.36	.39	.57	.50	.59	.36	.58																
Q13	.19	.35	.20	.52	.23	.36	.32	.27	.30	.31	.27	.41															
Q14	.22	.45	.17	.41	.18	.16	.37	.28	.52	.11	.36	.34	.31														
Q15	.19	.37	.06	.25	.15	.02	.32	.25	.45	.01	.37	.28	.25	.52													
Q16	.20	.36	.14	.29	.14	.08	.32	.23	.46	.07	.25	.32	.25	.56	.51												
Q17	.25	.44	.13	.42	.18	.10	.41	.31	.59	.04	.40	.41	.25	.69	.61	.61											
Q18	.27	.55	.20	.37	.33	.19	.46	.37	.60	.15	.44	.44	.34	.54	.58	.47	.61										
Q19	.24	.45	.19	.40	.25	.17	.42	.30	.48	.15	.41	.39	.27	.62	.55	.50	.67	.67									
Q20	.36	.28	.35	.29	.28	.20	.42	.80	.39	.17	.19	.48	.30	.25	.25	.25	.31	.38	.26								
Q21	.29	.43	.14	.35	.15	.15	.41	.32	.41	.01	.42	.42	.28	.38	.44	.34	.38	.50	.46	.38							
Q22	.27	.39	.19	.31	.11	.07	.40	.50	.43	-.03	.39	.37	.32	.34	.38	.32	.38	.46	.38	.50	.76						
Q23	.25	.35	.15	.30	.16	.08	.35	.37	.46	.01	.44	.33	.25	.42	.46	.38	.47	.54	.45	.34	.56	.68					
Q24	.22	.37	.07	.31	.21	.08	.33	.31	.43	.04	.39	.38	.25	.42	.47	.39	.53	.58	.48	.27	.46	.46	.66				
Q25	.28	.45	.17	.36	.18	.09	.35	.37	.57	.06	.33	.37	.32	.45	.43	.47	.52	.51	.55	.44	.53	.61	.65	.59			
Q26	.19	.26	.13	.25	.19	.12	.26	.18	.36	-.04	.25	.30	.31	.29	.44	.31	.39	.34	.33	.23	.47	.43	.47	.40	.57		
Q27	.37	.54	.31	.32	.31	.25	.41	.33	.52	.17	.39	.44	.27	.48	.36	.38	.51	.53	.58	.35	.49	.47	.51	.48	.61	.45	
Q28	.32	.43	.32	.32	.28	.36	.37	.21	.38	.26	.32	.39	.36	.40	.28	.28	.38	.46	.41	.20	.40	.38	.32	.31	.40	.37	.70

Polychoric correlation matrix based on 100 imputed data sets of DEMQOL-Proxy

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30
Q2	.33																													
Q3	.34	.50																												
Q4	.44	.13	.18																											
Q5	.47	.59	.52	.19																										
Q6	.56	.39	.39	.38	.50																									
Q7	.34	.64	.54	.14	.65	.38																								
Q8	.52	.18	.26	.77	.23	.41	.18																							
Q9	.32	.37	.51	.10	.42	.35	.44	.16																						
Q10	.42	.51	.58	.26	.62	.48	.58	.31	.55																					
Q11	.42	.19	.22	.47	.27	.44	.21	.46	.17	.33																				
Q12	.09	.39	.34	-.05	.30	.12	.35	-.02	.14	.31	-.06																			
Q13	.15	.21	.27	.02	.28	.20	.34	.00	.23	.26	.01	.42																		
Q14	.10	.37	.35	-.03	.37	.13	.41	-.02	.17	.37	.00	.77	.53																	
Q15	.07	.27	.27	-.04	.25	.09	.32	-.04	.16	.25	.02	.52	.49	.63																
Q16	.21	.31	.34	.01	.36	.26	.44	.06	.27	.35	.21	.30	.43	.42	.46															
Q17	.15	.38	.34	.09	.33	.17	.37	.08	.19	.35	.15	.47	.43	.62	.57	.56														
Q18	.21	.40	.42	.08	.43	.19	.46	.11	.22	.47	.18	.56	.48	.67	.53	.49	.66													
Q19	.15	.38	.36	.12	.35	.21	.42	.13	.25	.40	.12	.52	.47	.61	.54	.45	.60	.67												
Q20	.13	.32	.36	.10	.32	.12	.41	.09	.26	.36	.13	.39	.48	.48	.53	.43	.45	.65	.61											
Q21	.13	.26	.30	.20	.29	.18	.35	.17	.30	.36	.22	.25	.21	.30	.31	.33	.36	.37	.32	.34										
Q22	.17	.22	.29	.12	.28	.20	.35	.13	.22	.35	.14	.21	.20	.33	.35	.31	.36	.43	.37	.34	.82									
Q23	.14	.28	.26	.10	.26	.17	.32	.10	.20	.32	.10	.30	.35	.38	.34	.35	.36	.44	.46	.41	.52	.48								
Q24	.14	.34	.29	.07	.33	.16	.28	.14	.16	.29	.13	.28	.35	.39	.32	.42	.45	.45	.46	.40	.32	.36	.67							
Q25	.19	.34	.27	.07	.29	.25	.28	.07	.19	.32	.17	.36	.24	.38	.27	.39	.44	.41	.52	.36	.27	.31	.56	.75						
Q26	.21	.25	.33	.17	.28	.17	.32	.16	.23	.32	.07	.39	.35	.52	.40	.33	.41	.52	.54	.41	.28	.31	.45	.45	.45					
Q27	.11	.26	.24	.01	.30	.22	.30	.05	.17	.33	.10	.33	.39	.42	.42	.28	.38	.48	.48	.48	.40	.42	.56	.48	.44	.46				
Q28	.18	.31	.28	.14	.41	.34	.39	.14	.19	.50	.19	.29	.28	.39	.26	.30	.41	.46	.41	.32	.38	.41	.45	.38	.40	.35	.55			
Q29	.09	.22	.25	.13	.27	.19	.29	.08	.16	.29	.05	.27	.19	.35	.25	.23	.29	.31	.33	.32	.38	.39	.32	.23	.27	.49	.58	.45		
Q30	.16	.31	.33	.17	.35	.30	.33	.11	.25	.36	.11	.31	.29	.36	.34	.28	.35	.32	.36	.32	.34	.34	.38	.23	.28	.44	.51	.49	.71	
Q31	.14	.34	.22	.16	.30	.20	.32	.12	.20	.34	.06	.29	.11	.30	.31	.17	.27	.33	.37	.32	.28	.27	.28	.25	.28	.38	.40	.32	.32	.35

APPENDICES TO CHAPTER 3

Bifactor MIMIC model (DEMQOL) syntax

```
TITLE: CMS + 1066 DEMQOL (28 items)
! Model 1: CFA-28 1066UK
* Model 2: MIMIC

DATA:
FILE = a1066cms.dat ;

VARIABLE:
NAMES =
id nation age gender sev qol28s1
a1m0 a2m0 a3m0 a4m0 a5m0 a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0 a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0 a26m0 a27m0 a28m0 ;

MISSING = All (-1234) ;

USEVAR =
nation gender sev
a1m0 a2m0 a3m0 a4m0 a5m0 a6m0 a7m0 a8m0 a9m0 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0 a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0 a26m0 a27m0 a28m0 ;

CATEGORICAL = a1m0 - a28m0 ;

ANALYSIS:
!Estimator = WLSMV ;       !!default
!Estimator = MLR ;
!Integration = Montecarlo ;
!DIFFTEST = diff6x.dat ;

MODEL:
qol by
a1m0* a2m0 a3m0 a4m0 a5m0 a6m0 a7m0 a8m0 a9m0@1 a10m0
a11m0 a12m0 a13m0 a14m0 a15m0 a16m0 a17m0 a18m0 a19m0 a20m0
a21m0 a22m0 a23m0 a24m0 a25m0 a26m0 a27m0 a28m0;

pos by a10m0 a6m0 a5m0 a3m0 a1m0 ;
neg BY a13m0 a12m0 a11m0 a4m0 ;
cog BY a19m0 a18m0 a17m0 a16m0 a15m0 a14m0 ;
soc BY a25m0 a24m0 a23m0 a22m0 a21m0 ;

!! corr residual 'factors'
cr1 by a8m0* a20m0 (eq1) ;

!! unstandardised lambdas NOT available
cr1@1 ;

!! orthogonality
qol WITH pos@0 neg@0 cog@0 soc@0 cr1@0 ;
pos WITH neg@0 cog@0 soc@0 cr1@0 ;
neg WITH cog@0 soc@0 cr1@0 ;
cog WITH soc@0 cr1@0 ;
soc WITH cr1@0 ;
```

```
!MIMIC
qol on nation gender sev ;
pos on nation gender sev ;
cog on nation gender sev ;
neg on nation gender sev ;
soc on nation gender sev ;
cr1 on nation gender sev ;

!DIF items
a27m0 on nation ;
a2m0 on nation ;
a19m0 on nation ;
a3m0 on nation ;
a1m0 on nation ;
a13m0 on nation ;

SAVEDATA:
!DIFFTEST = diff6x.dat ;

OUTPUT: STDYX Residual Modindices(All) ;

!PLOT: TYPE = PLOT3;
```

Bifactor MIMIC model (DEMQOL-Proxy) syntax

```
TITLE: CMS + 1066 DEMQOL-Proxy (31 items)
! Model 1: CFA-31 1066UK
* Model 2: MIMIC

DATA:
FILE = b1066cms.dat ;

VARIABLE:
NAMES =
id nation age gender sev qol31s1
b1m0 b2m0 b3m0 b4m0 b5m0 b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b31m0 ;

MISSING = all (-1234) ;

USEVAR =
nation gender sev
b1m0 b2m0 b3m0 b4m0 b5m0 b6m0 b7m0 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b31m0 ;

CATEGORICAL = b1m0 - b31m0 ;

ANALYSIS:
!Estimator = WLSMV ;       !!default
!Estimator = MLR ;
!Integration = Montecarlo ;
!DIFFTEST = difftest.dat ;

MODEL:
qol by
b1m0* b2m0 b3m0 b4m0 b5m0 b6m0 b7m0@1 b8m0 b9m0 b10m0
b11m0 b12m0 b13m0 b14m0 b15m0 b16m0 b17m0 b18m0 b19m0 b20m0
b21m0 b22m0 b23m0 b24m0 b25m0 b26m0 b27m0 b28m0 b29m0 b30m0
b31m0 ;

neg BY b10m0 b9m0 b7m0 b5m0 b3m0 b2m0 ;
pos BY b11m0 b8m0 b6m0 b4m0 b1m0 ;
soc BY b30m0 b29m0 b28m0 b27m0 ;
cog BY b20m0 b19m0 b18m0 b17m0 b16m0 b15m0 b14m0 b13m0 b12m0 ;

!! corr residual 'factors'
cr1 by b21m0* b22m0 (eq1) ;
cr2 by b24m0* b25m0 (eq2) ;

!! unstandardised lambdas NOT available
cr1@1 ;
cr2@1 ;

!! orthogonality
qol WITH neg@0 pos@0 soc@0 cog@0 cr1@0 cr2@0 ;
neg WITH pos@0 soc@0 cog@0 cr1@0 cr2@0 ;
pos WITH soc@0 cog@0 cr1@0 cr2@0 ;
soc WITH cog@0 cr1@0 cr2@0 ;
cog WITH cr1@0 cr2@0 ;
cr1 with cr2@0 ;
```

```
!! MIMIC
qol on nation gender sev ;
neg on nation gender sev ;
pos on nation gender sev ;
cog on nation gender sev ;
soc on nation gender sev ;
cr1 on nation gender sev ;
cr2 on nation gender sev ;

!! DIF items
b11m0 on nation ;
b13m0 on nation ;
b3m0 on nation ;
b20m0 on nation ;
b2m0 on nation ;
b28m0 on gender ;
b4m0 on nation ;
b8m0 on nation ;
b16m0 on nation ;
b16m0 on sev ;
b20m0 on sev ;
b27m0 on nation ;
b9m0 on gender ;

SAVEDATA:
DIFFTEST = difftest.dat ;

OUTPUT: STDYX Residual Modindices(all) ;

!PLOT: TYPE = PLOT3;
```

DEMQOL-SF (item numbers from parent version in gray)

First I'm going to ask about your feelings. In the last week, have you felt…

			Not at all	A little	Quite a bit	A lot
1	1	cheerful?	☐	☐	☐	☐
2	2	worried or anxious?	☐	☐	☐	☐
3	4	frustrated?	☐	☐	☐	☐
4	7	sad?	☐	☐	☐	☐
5	8	lonely?	☐	☐	☐	☐
6	11	irritable?	☐	☐	☐	☐
7	12	fed-up?	☐	☐	☐	☐
8	13	that there are things that you wanted to do but couldn't?	☐	☐	☐	☐

Next, I'm going to ask you about your memory. In the last week, how worried have you been about…

			Not at all	A little	Quite a bit	A lot
9	14	forgetting things that happened recently?	☐	☐	☐	☐
10	17	your thoughts being muddled?	☐	☐	☐	☐
11	18	difficulty making decisions?	☐	☐	☐	☐
12	19	poor concentration?	☐	☐	☐	☐

Now, I'm going to ask you about your everyday life. In the last week, how worried have you been about…

			Not at all	A little	Quite a bit	A lot
13	21	how you get on with people close to you?	☐	☐	☐	☐
14	22	getting the affection that you want?	☐	☐	☐	☐
15	23	people not listening to you?	☐	☐	☐	☐
16	24	making yourself understood?	☐	☐	☐	☐
17	28	your health overall?	☐	☐	☐	☐

DEMQOL-Proxy-SF (item numbers from parent version in gray)

First I'm going to ask you about (your relative's) feelings. In the last week, would you say that (your relative) has felt…

			Not at all	A little	Quite a bit	A lot
1	3	frustrated?	☐	☐	☐	☐
2	5	sad?	☐	☐	☐	☐
3	7	distressed?	☐	☐	☐	☐
4	8	lively?	☐	☐	☐	☐

Next, I'm going to ask you about (your relative's) memory. In the last week, how worried would you say (your relative) has been about…

			Not at all	A little	Quite a bit	A lot
6	12	his/her memory in general?	☐	☐	☐	☐
7	14	forgetting things that happened recently?	☐	☐	☐	☐
8	17	forgetting what day it is?	☐	☐	☐	☐
9	18	his/her thoughts being muddled?	☐	☐	☐	☐

Now, I'm going to ask about (your relative's) everyday life. In the last week, how worried would you say (your relative) has been about…

			Not at all	A little	Quite a bit	A lot
10	22	keeping him/herself looking nice?	☐	☐	☐	☐
11	25	looking after his/her finances?	☐	☐	☐	☐
12	26	things taking longer than they used to?	☐	☐	☐	☐
13	27	getting in touch with people?	☐	☐	☐	☐
14	28	not having enough company?	☐	☐	☐	☐
15	29	not being able to help other people?	☐	☐	☐	☐
16	30	not playing a useful part in things?	☐	☐	☐	☐
17	31	his/her physical health?	☐	☐	☐	☐

APPENDICES TO CHAPTER 4

Configural invariance syntax for DEMQOL-SF bifactor longitudinal SEM model (syntax presented for Croydon Memory Service only, HTA-SADD analysis had only minor differences in terms of data set variables)

```
Title: Longitudinal bifactor CFA DEMQOL SF 17 (m 0 6 12)

Data:  File = dcroy.dat ;

Variable:
Names =
qa0 qa6 qa12 qc0 qc6 qc12 qd0 qd6 qd12 qe0 qe6 qe12 qf0 qf6 qf12
qg0 qg6 qg12 uv0 uv6 uv12
j1 j2 j3 j4 j5 j6 j7 j8 j9 j10 j11 j12 j13 j14 j15 j16 j17 j18 j19 j20
j21 j22 j23 j24 j25 j26 j27 j28
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 k11 k12 k13 k14 k15 k16 k17 k18 k19 k20
k21 k22 k23 k24 k25 k26 k27 k28
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18 r19 r20
r21 r22 r23 r24 r25 r26 r27 r28;

Missing = all (-1234) ;

Usevariables =
j1 j2 j4 j7 j8
j11 j12 j13 j14 j17 j18 j19
j21 j22 j23 j24 j28

k1 k2 k4 k7 k8
k11 k12 k13 k14 k17 k18 k19
k21 k22 k23 k24 k28

r1 r2 r4 r7 r8
r11 r12 r13 r14 r17 r18 r19
r21 r22 r23 r24 r28 ;


Categorical = ALL ;

Analysis:
Parameterization = Theta ;

Model:

jhrql by
j1* j2 j4 j7 j8
j11 j12 j13 j14 j17 j18 j19
j21 j22 j23 j24 j28 ;

jcog by j14* j17 j18 j19 ;

!! orthogonality constraints
jhrql with jcog@0 ;

!! fix variance (after check for factor collapse)
jhrql@1 ;
jcog@1 ;


khrql by
k1* k2 k4 k7 k8
```

```
k11 k12 k13 k14 k17 k18 k19
k21 k22 k23 k24 k28 ;

kcog by k14* k17 k18 k19 ;

!! orthogonality constraints
khrql with kcog@0 ;

!! fix variance (after check for factor collapse)
khrql@1 ;
kcog@1 ;


rhrql by
r1* r2 r4 r7 r8
r11 r12 r13 r14 r17 r18 r19
r21 r22 r23 r24 r28 ;

rcog by r14* r17 r18 r19 ;

!! orthogonality constraints
rhrql with rcog@0 ;

!! fix variance (after check for factor collapse)
rhrql@1 ;
rcog@1 ;


!! cross-occasion factor correlation constraints
jhrql with kcog@0 rcog@0 ;
jcog with khrql@0 rhrql@0 ;
khrql with rcog@0 ;
kcog with rhrql@0 ;

!! cross-occasion residual correlations
j1-j28 pwith k1-k28 ;
j1-j28 pwith r1-r28 ;
k1-k28 pwith r1-r28 ;

Savedata: DIFFTEST=inv.dat ;

Output: STDYX Modindices (All 3.84) ;
```

Configural invariance syntax for DEMQOL-SF bifactor longitudinal SEM model (syntax presented for HTA-SADD only, Croydon Memory Service analysis had only minor differences in terms of data set variables)

```
Title:
Longitudinal bifactor CFA DEMQOL PROXY SF 17 (wk 0 13 39)

Data:
File = hta31.dat ;

Variable:
Names =
male tx mmse0 mmse13 mmse39 npi0 npi13 npi39 badl0 badl13 badl39
csdd0 csdd13 csdd39 uvp0 uvp13 uvp39
j1 j2 j3 j4 j5 j6 j7 j8 j9 j10 j11 j12 j13 j14 j15 j16 j17 j18 j19 j20
j21 j22 j23 j24 j25 j26 j27 j28 j29 j30 j31
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 k11 k12 k13 k14 k15 k16 k17 k18 k19 k20
k21 k22 k23 k24 k25 k26 k27 k28 k29 k30 k31
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18 r19 r20
r21 r22 r23 r24 r25 r26 r27 r28 r29 r30 r31 ;

Missing = all (-1234) ;

Usevariables =

j3 j5 j7 j8 j10
j12 j14 j17 j18
j22 j25 j26 j27 j28 j29 j30 j31

k3 k5 k7 k8 k10
k12 k14 k17 k18
k22 k25 k26 k27 k28 k29 k30 k31

r3 r5 r7 r8 r10
r12 r14 r17 r18
r22 r25 r26 r27 r28 r29 r30 r31 ;

Categorical = ALL ;

Analysis: Parameterization = Theta ;

Model:
jhrql by
j3* j5 j7 j8 j10
j12 j14 j17 j18
j22 j25 j26 j27 j28 j29 j30 j31 ;

jneg by j3* j5 j7 j10 ;
jcog by j12* j14 j17 j18 ;

!! orthogonality constraints
jhrql with jneg@0 jcog@0 ;
jneg with jcog@0 ;

!! fix variance (after check for factor collapse)
jhrql@1 ;
jneg@1 ;
jcog@1 ;

khrql by
k3* k5 k7 k8 k10
k12 k14 k17 k18
k22 k25 k26 k27 k28 k29 k30 k31 ;
```

```
kneg by k3* k5 k7 k10 ;
kcog by k12* k14 k17 k18 ;

!! orthogonality constraints
khrql with kneg@0 kcog@0 ;
kneg with kcog@0 ;

!! fix variance (after check for factor collapse)
khrql@1 ;
kneg@1 ;
kcog@1 ;

rhrql by
r3* r5 r7 r8 r10
r12 r14 r17 r18
r22 r25 r26 r27 r28 r29 r30 r31 ;

rneg by r3* r5 r7 r10 ;
rcog by r12* r14 r17 r18 ;

!! orthogonality constraints
rhrql with rneg@0 rcog@0 ;
rneg with rcog@0 ;

!! fix variance (after checr for factor collapse)
rhrql@1 ;
rneg@1 ;
rcog@1 ;

!! cross-occasion factor correlation constraints
jhrql with kneg@0 kcog@0 rneg@0 rcog@0 ;
jneg with khrql@0 kcog@0 rhrql@0 rcog@0 ;
jcog with khrql@0 kneg@0 rhrql@0 rneg@0 ;
khrql with rneg@0 rcog@0 ;
kneg with rhrql@0 rcog@0 ;
kcog with rhrql@0 rneg@0 ;

!! cross-occasion residual correlations
j3-j31 pwith k3-k31 ;
j3-j31 pwith r3-r31 ;
k3-k31 pwith r3-r31 ;

Savedata: DIFFTEST=inv.dat ;

Output: STDYX Modindices (All 3.84) ;
```

Scalar invariance syntax for DEMQOL-SF bifactor longitudinal SEM model (syntax presented for Croydon Memory Service only, HTA-SADD analysis had only minor differences in terms of data set variables)

```
Title:
Longitudinal bifactor CFA DEMQOL SF 17 (m 0 6 12)

Data:  File = dcroy.dat ;

Variable:
Names =
qa0 qa6 qa12 qc0 qc6 qc12 qd0 qd6 qd12 qe0 qe6 qe12 qf0 qf6 qf12
qg0 qg6 qg12 uv0 uv6 uv12
j1 j2 j3 j4 j5 j6 j7 j8 j9 j10 j11 j12 j13 j14 j15 j16 j17 j18 j19 j20
j21 j22 j23 j24 j25 j26 j27 j28
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 k11 k12 k13 k14 k15 k16 k17 k18 k19 k20
k21 k22 k23 k24 k25 k26 k27 k28
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18 r19 r20
r21 r22 r23 r24 r25 r26 r27 r28;

Missing = all (-1234) ;

Usevariables =
j1 j2 j4 j7 j8
j11 j12 j13 j14 j17 j18 j19
j21 j22 j23 j24 j28

k1 k2 k4 k7 k8
k11 k12 k13 k14 k17 k18 k19
k21 k22 k23 k24 k28

r1 r2 r4 r7 r8
r11 r12 r13 r14 r17 r18 r19
r21 r22 r23 r24 r28 ;


Categorical = ALL ;

Analysis:
Parameterization = Theta ;
DIFFTEST = inv.dat ;

Model:

! cross-occasion factor loading equality
jhrql by j1*(g1) ;
jhrql by j2(g2) ;
jhrql by j4(g4) ;
jhrql by j7(g7) ;
jhrql by j8(g8) ;
jhrql by j11(g11) ;
jhrql by j12(g12) ;
jhrql by j13(g13) ;
jhrql by j14(g14) ;
jhrql by j17(g17) ;
jhrql by j18(g18) ;
jhrql by j19(g19);
jhrql by j21(g21) ;
jhrql by j22(g22) ;
jhrql by j23(g23) ;
jhrql by j24(g24) ;
jhrql by j28(g28) ;
```

```
jcog by j14*(c14) ;
jcog by j17(c17) ;
jcog by j18(c18) ;
jcog by j19(c19) ;

! orthogonality constraints
jhrql with jcog@0 ;

!! fix variance at time 1
jhrql@1 ;
jcog@1 ;


! cross-occasion factor loading equality
khrql by k1*(g1) ;
khrql by k2(g2) ;
khrql by k4(g4) ;
khrql by k7(g7) ;
khrql by k8(g8) ;
khrql by k11(g11) ;
khrql by k12(g12) ;
khrql by k13(g13) ;
khrql by k14(g14) ;
khrql by k17(g17) ;
khrql by k18(g18) ;
khrql by k19(g19);
khrql by k21(g21) ;
khrql by k22(g22) ;
khrql by k23(g23) ;
khrql by k24(g24) ;
khrql by k28(g28) ;

kcog by k14*(c14) ;
kcog by k17(c17) ;
kcog by k18(c18) ;
kcog by k19(c19) ;

! orthogonality constraints
khrql with kcog@0 ;

! cross-occasion factor loading equality
rhrql by r1*(g1) ;
rhrql by r2(g2) ;
rhrql by r4(g4) ;
rhrql by r7(g7) ;
rhrql by r8(g8) ;
rhrql by r11(g11) ;
rhrql by r12(g12) ;
rhrql by r13(g13) ;
rhrql by r14(g14) ;
rhrql by r17(g17) ;
rhrql by r18(g18) ;
rhrql by r19(g19);
rhrql by r21(g21) ;
rhrql by r22(g22) ;
rhrql by r23(g23) ;
rhrql by r24(g24) ;
rhrql by r28(g28) ;

rcog by r14*(c14) ;
rcog by r17(c17) ;
rcog by r18(c18) ;
rcog by r19(c19) ;


! orthogonality constraints
rhrql with rcog@0 ;
```

```
!! free variance at time 2
rhrql* ;
rcog* ;


!! cross-occasion factor correlation constraints
jhrql with kcog@0 rcog@0 ;
jcog with khrql@0 rhrql@0 ;
khrql with rcog@0 ;
kcog with rhrql@0 ;

!! cross-occasion residual correlations
j1-j28 pwith k1-k28 ;
j1-j28 pwith r1-r28 ;
k1-k28 pwith r1-r28 ;


! cross-occasion threshold equality
[j1$1 k1$1 r1$1] (11) ;
[j1$2 k1$2 r1$2] (12) ;
[j1$3 k1$3 r1$3] (13) ;

[j2$1 k2$1 r2$1] (21) ;
[j2$2 k2$2 r2$2] (22) ;
[j2$3 k2$3 r2$3] (23) ;

[j4$1 k4$1 r4$1] (41) ;
[j4$2 k4$2 r4$2] (42) ;
[j4$3 k4$3 r4$3] (43) ;

[j7$1 k7$1 r7$1] (71) ;
[j7$2 k7$2 r7$2] (72) ;
[j7$3 k7$3 r7$3] (73) ;

[j8$1 k8$1 r8$1] (81) ;
[j8$2 k8$2 r8$2] (82) ;
[j8$3 k8$3 r8$3] (83) ;

[j11$1 k11$1 r11$1] (111) ;
[j11$2 k11$2 r11$2] (112) ;
[j11$3 k11$3 r11$3] (113) ;

[j12$1 k12$1 r12$1] (121) ;
[j12$2 k12$2 r12$2] (122) ;
[j12$3 k12$3 r12$3] (123) ;

[j13$1 k13$1 r13$1] (131) ;
[j13$2 k13$2 r13$2] (132) ;
[j13$3 k13$3 r13$3] (133) ;

[j14$1 k14$1 r14$1] (141) ;
[j14$2 k14$2 r14$2] (142) ;
[j14$3 k14$3 r14$3] (143) ;

[j17$1 k17$1 r17$1] (171) ;
[j17$2 k17$2 r17$2] (172) ;
[j17$3 k17$3 r17$3] (173) ;

[j18$1 k18$1 r18$1] (181) ;
[j18$2 k18$2 r18$2] (182) ;
[j18$3 k18$3 r18$3] (183) ;

[j19$1 k19$1 r19$1] (191) ;
[j19$2 k19$2 r19$2] (192) ;
[j19$3 k19$3 r19$3] (193) ;

[j21$1 k21$1 r21$1] (211) ;
[j21$2 k21$2 r21$2] (212) ;
```

```
[j21$3 k21$3 r21$3] (213) ;

[j22$1 k22$1 r22$1] (221) ;
[j22$2 k22$2 r22$2] (222) ;
[j22$3 k22$3 r22$3] (223) ;

[j23$1 k23$1 r23$1] (231) ;
[j23$2 k23$2 r23$2] (232) ;
[j23$3 k23$3 r23$3] (233) ;

[j24$1 k24$1 r24$1] (241) ;
[j24$2 k24$2 r24$2] (242) ;
[j24$3 k24$3 r24$3] (243) ;

[j28$1 k28$1 r28$1] (281) ;
[j28$2 k28$2 r28$2] (282) ;
[j28$3 k28$3 r28$3] (283) ;


!! latent change over time
[khrql* ] ;
[kcog* ] ;
[rhrql* ] ;
[rcog* ] ;

!Savedata: DIFFTEST=inv.dat ;


Output: STDYX Modindices(ALL 3.84) ;
```

Scalar invariance syntax for DEMQOL-SF bifactor longitudinal SEM model (syntax presented for HTA-SADD only, Croydon Memory Service analysis had only minor differences in terms of data set variables)

```
Title:
Longitudinal bifactor CFA DEMQOL PROXY SF 17 (wk 0 13 39)

Data:
File = hta31.dat ;

Variable:
Names =
male tx mmse0 mmse13 mmse39 npi0 npi13 npi39 badl0 badl13 badl39
csdd0 csdd13 csdd39 uvp0 uvp13 uvp39
j1 j2 j3 j4 j5 j6 j7 j8 j9 j10 j11 j12 j13 j14 j15 j16 j17 j18 j19 j20
j21 j22 j23 j24 j25 j26 j27 j28 j29 j30 j31
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 k11 k12 k13 k14 k15 k16 k17 k18 k19 k20
k21 k22 k23 k24 k25 k26 k27 k28 k29 k30 k31
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18 r19 r20
r21 r22 r23 r24 r25 r26 r27 r28 r29 r30 r31 ;

Missing = all (-1234) ;

Usevariables =

j3 j5 j7 j8 j10
j12 j14 j17 j18
j22 j25 j26 j27 j28 j29 j30 j31

k3 k5 k7 k8 k10
k12 k14 k17 k18
k22 k25 k26 k27 k28 k29 k30 k31

r3 r5 r7 r8 r10
r12 r14 r17 r18
r22 r25 r26 r27 r28 r29 r30 r31 ;

Categorical = ALL ;

Analysis:
Parameterization = Theta ;
DIFFTEST = inv.dat ;


Model:
!! cross-occasion factor loading equality
jhrql by j3*(g3) ;
jhrql by j5(g5) ;
jhrql by j7 (g7) ;
jhrql by j8 (g8) ;
jhrql by j10 (g10) ;
jhrql by j12 (g12) ;
jhrql by j14 (g14) ;
jhrql by j17 (g17) ;
jhrql by j18 (g18) ;
jhrql by j22 (g22) ;
jhrql by j25 (g25) ;
jhrql by j26 (g26) ;
jhrql by j27 (g27) ;
jhrql by j28 (g28) ;
jhrql by j29 (g29) ;
jhrql by j30 (g30) ;
jhrql by j31 (g31) ;
```

```
jneg by j3* (n3) ;
jneg by j5 (n5) ;
jneg by j7 (n7) ;
jneg by j10 (n10) ;

jcog by j12* (c12) ;
jcog by j14 (c14) ;
jcog by j17 (c17) ;
jcog by j18 (c18) ;

!! orthogonality constraints
jhrql with jneg@0 jcog@0 ;
jneg with jcog@0 ;

!! fix variance at time 1
jhrql@1 ;
jneg@1 ;
jcog@1 ;

!! cross-occasion factor loading equality
khrql by k3*(g3) ;
khrql by k5(g5) ;
khrql by k7 (g7) ;
khrql by k8 (g8) ;
khrql by k10 (g10) ;
khrql by k12 (g12) ;
khrql by k14 (g14) ;
khrql by k17 (g17) ;
khrql by k18 (g18) ;
khrql by k22 (g22) ;
khrql by k25 (g25) ;
khrql by k26 (g26) ;
khrql by k27 (g27) ;
khrql by k28 (g28) ;
khrql by k29 (g29) ;
khrql by k30 (g30) ;
khrql by k31 (g31) ;

kneg by k3* (n3) ;
kneg by k5 (n5) ;
kneg by k7 (n7) ;
kneg by k10 (n10) ;

kcog by k12* (c12) ;
kcog by k14 (c14) ;
kcog by k17 (c17) ;
kcog by k18 (c18) ;

!! orthogonality constraints
khrql with kneg@0 kcog@0 ;
kneg with kcog@0 ;

!! free variance at time 2
khrql* ;
kneg* ;
kcog* ;


!! cross-occasion factor loading equality
rhrql by r3*(g3) ;
rhrql by r5(g5) ;
rhrql by r7 (g7) ;
rhrql by r8 (g8) ;
rhrql by r10 (g10) ;
rhrql by r12 (g12) ;
rhrql by r14 (g14) ;
rhrql by r17 (g17) ;
```

```
rhrql by r18 (g18) ;
rhrql by r22 (g22) ;
rhrql by r25 (g25) ;
rhrql by r26 (g26) ;
rhrql by r27 (g27) ;
rhrql by r28 (g28) ;
rhrql by r29 (g29) ;
rhrql by r30 (g30) ;
rhrql by r31 (g31) ;

rneg by r3* (n3) ;
rneg by r5 (n5) ;
rneg by r7 (n7) ;
rneg by r10 (n10) ;

rcog by r12* (c12) ;
rcog by r14 (c14) ;
rcog by r17 (c17) ;
rcog by r18 (c18) ;

!! orthogonality constraints
rhrql with rneg@0 rcog@0 ;
rneg with rcog@0 ;

!! free variance at time 3
rhrql* ;
rneg* ;
rcog* ;


!! cross-occasion threshold equality
[j3$1 k3$1 r3$1] (31) ;
[j3$2 k3$2 r3$2] (32) ;
[j3$3 k3$3 r3$3] (33) ;

[j5$1 k5$1 r5$1] (51) ;
[j5$2 k5$2 r5$2] (52) ;
[j5$3 k5$3 r5$3] (53) ;

[j7$1 k7$1 r7$1] (71) ;
[j7$2 k7$2 r7$2] (72) ;
[j7$3 k7$3 r7$3] (73) ;

[j8$1 k8$1 r8$1] (81) ;
[j8$2 k8$2 r8$2] (82) ;
[j8$3 k8$3 r8$3] (83) ;

[j10$1 k10$1 r10$1] (101) ;
[j10$2 k10$2 r10$2] (102) ;
[j10$3 k10$3 r10$3] (103) ;

[j12$1 k12$1 r12$1] (121) ;
[j12$2 k12$2 r12$2] (122) ;
[j12$3 k12$3 r12$3] (123) ;

[j14$1 k14$1 r14$1] (141) ;
[j14$2 k14$2 r14$2] (142) ;
[j14$3 k14$3 r14$3] (143) ;

[j17$1 k17$1 r17$1] (171) ;
[j17$2 k17$2 r17$2] (172) ;
[j17$3 k17$3 r17$3] (173) ;

[j18$1 k18$1 r18$1] (181) ;
[j18$2 k18$2 r18$2] (182) ;
[j18$3 k18$3 r18$3] (183) ;

[j22$1 k22$1 r22$1] (221) ;
```

```
[j22$2 k22$2 r22$2] (222) ;
[j22$3 k22$3 r22$3] (223) ;

[j25$1 k25$1 r25$1] (251) ;
[j25$2 k25$2 r25$2] (252) ;
[j25$3 k25$3 r25$3] (253) ;

[j26$1 k26$1 r26$1] (261) ;
[j26$2 k26$2 r26$2] (262) ;
[j26$3 k26$3 r26$3] (263) ;

[j27$1 k27$1 r27$1] (271) ;
[j27$2 k27$2 r27$2] (272) ;
[j27$3 k27$3 r27$3] (273) ;

[j28$1 k28$1 r28$1] (281) ;
[j28$2 k28$2 r28$2] (282) ;
[j28$3 k28$3 r28$3] (283) ;

[j29$1 k29$1 r29$1] (291) ;
[j29$2 k29$2 r29$2] (292) ;
[j29$3 k29$3 r29$3] (293) ;

[j30$1 k30$1 r30$1] (301) ;
[j30$2 k30$2 r30$2] (302) ;
[j30$3 k30$3 r30$3] (303) ;

[j31$1 k31$1 r31$1] (311) ;
[j31$2 k31$2 r31$2] (312) ;
[j31$3 k31$3 r31$3] (313) ;

!! cross-occasion factor correlation constraints
jhrql with kneg@0 kcog@0 rneg@0 rcog@0 ;
jneg with khrql@0 kcog@0 rhrql@0 rcog@0 ;
jcog with khrql@0 kneg@0 rhrql@0 rneg@0 ;
khrql with rneg@0 rcog@0 ;
kneg with rhrql@0 rcog@0 ;
kcog with rhrql@0 rneg@0 ;

!! cross-occasion residual correlations
j3-j31 pwith k3-k31 ;
j3-j31 pwith r3-r31 ;
k3-k31 pwith r3-r31 ;

!! latent change over time
[khrql* ] ;
[kneg* ] ;
[kcog* ] ;
[rhrql* ] ;
[rneg* ] ;
[rcog* ] ;

Savedata: DIFFTEST=inv.dat ;

Output: STDYX Modindices (All 3.84) ;
```