

**Approaches to the measurement of outcomes of  
chronic disease self-management interventions  
using a self-report inventory**

A thesis submitted in total fulfilment of the requirements  
for the degree of Doctor of Philosophy

**Sandra Nolte**

Bachelor of Business Administration (Hons)

School of Global Studies, Social Science & Planning (GSSSP)

Portfolio of Design and Social Context

**RMIT University, Australia**

March 2008

## Declaration

This is to certify that:

- (i) except where due acknowledgement has been made, the work is that of the candidate alone;
- (ii) the work has not been submitted previously, in whole or in part, to qualify for any other academic award;
- (iii) the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program;
- (iv) any editorial work, paid or unpaid, carried out by a third party is acknowledged;
- (v) ethics procedures and guidelines of the Human Research Ethics Committees of RMIT University and The University of Melbourne have been followed;
- (vi) the research contained in this thesis was supported by:
  - a research scholarship awarded by the School of Global Studies, Social Science & Planning (GSSSP), RMIT University;
  - a research scholarship awarded by the Centre for Rheumatic Diseases, Department of Medicine, The University of Melbourne;
  - an “International Mobility Activity” RMIT University travel grant;
  - the 2006 RMIT University Regional Award (onshore);
  - the 2006 Australian Disease Management Association Award; and
  - the 2006 International Society of Quality of Life Research (ISOQOL) Early Career Investigators' Award.

Sandra Nolte  
28 March 2008

## Acknowledgements

Many wonderful people have supported me since I commenced my studies in Australia four years ago. In particular, I wish to thank my supervisors Associate Professor Gerald Elsworth, Associate Professor Richard Osborne, and Professor Andrew Sinclair for guiding me through this research process. To Gerry, thank you for your ongoing support and for always having the time and patience to explain complex processes and sharing your vast knowledge with me. To Richard, thank you for your invaluable support and for giving me the opportunity to carry out this research at your centre and present my findings at several conferences. To Andy, thank you for taking me on board as a research student in 2004 and getting me in contact with the right people at the right time. Thank you all. I feel privileged that I had such outstanding support throughout my candidature.

I also wish to thank Professor Mirjam Sprangers and Associate Professor Frans Oort for their insightful comments during my visit to Amsterdam in September 2006 and over the last two years of my thesis. Further thanks go to Lucy Busija, Associate Professor Cliff da Costa, Dr John von Briesen for statistical advice and Professor Kate Lorig for her insightful comments on self-management interventions.

I thank the School of Global Studies, Social Science & Planning, RMIT University, and the Department of Medicine, The University of Melbourne, for partial funding of this research. I further wish to thank Dianne Ferguson, Mary Ljubanovic, Amanda Springer, Luke Tellefson, and Stella Vo for administrative support and the RMIT document delivery team for their never-ending help to retrieve publications from around the world.

Last but not least I wish to thank Duncan Graham for being there for me whenever I needed him and Joanne Jordan and Melissa King for their friendship and support during this time. I would further like to thank my friends and family in Germany who through phone calls, email and mail lightened up my days behind the computer.

Finally, to my mother Ursula Nolte who through her ongoing support helped me to get through this thesis. Thank you for your unconditional love. This thesis is dedicated to you.

# Table of contents

<b>Declaration</b> .....	<b>I</b>
<b>Acknowledgements</b> .....	<b>II</b>
<b>Table of contents</b> .....	<b>III</b>
<b>List of figures</b> .....	<b>VIII</b>
<b>List of tables</b> .....	<b>X</b>
<b>List of abbreviations</b> .....	<b>XIII</b>
<b>Abstract</b> .....	<b>1</b>
<b>Chapter 1 – Introduction</b> .....	<b>3</b>
1.1 Overview of the thesis .....	3
1.2 Background .....	6
1.2.1 The global burden of chronic disease .....	6
1.2.2 Chronic disease self-management interventions .....	8
1.2.3 The effectiveness of self-management interventions .....	13
1.2.3.1 Meta-analytic and other systematic reviews on self-management interventions ...	13
1.2.3.2 A systematic review of self-management interventions .....	16
1.2.3.3 Outcomes on which evaluations were based and how these were measured .....	26
1.2.3.4 Summary .....	32
1.2.4 The measurement of outcomes of self-management interventions .....	32
1.2.4.1 Research designs .....	34
1.2.4.2 The measurement of change .....	37
1.2.4.3 Confounding and bias in the measurement of outcomes of interventions .....	39
1.2.4.4 Response shift bias .....	43
1.2.4.5 Social desirability bias .....	48
1.3 Research questions .....	52
<b>Chapter 2 – Study design and data management</b> .....	<b>55</b>
2.1 Introduction .....	55
2.2 Study design .....	55
2.2.1 Setting .....	55
2.2.2 Ethics .....	55
2.2.3 Research design .....	55
2.3 Data collection, management and preparation .....	59
2.3.1 Recruitment of self-management courses .....	59
2.3.2 Data collection .....	60
2.3.3 Data screening prior to entry .....	64
2.3.4 Data entry and sample size .....	64
2.3.5 The distributional properties of the heiQ raw data .....	66
2.3.6 Treatment of missing data .....	68

## Chapter 2 (continued)

2.4	Summary.....	70
-----	--------------	----

## Chapter 3 – Approaches to the measurement of change with the heiQ..... 71

3.1	Introduction .....	71
3.2	Aims .....	71
3.3	Demographics.....	72
3.4	Pretest and posttest scores across heiQ-PP, heiQ-PPT, and heiQ-PPR .....	76
3.4.1	Specific methods .....	76
3.4.2	Results.....	77
3.4.3	Summary .....	81
3.5	Change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR.....	81
3.5.1	Specific methods .....	81
3.5.2	Results.....	83
3.5.3	Summary .....	86
3.6	Change scores across heiQ-PPR and heiQ-PPR Retro .....	86
3.6.1	Specific methods .....	86
3.6.2	Results.....	87
3.6.3	Summary .....	93
3.7	Discussion.....	93

## Chapter 4 – Statistical methods and the re-validation of the heiQ ..... 102

4.1	Introduction .....	102
4.2	Factor analysis.....	102
4.2.1	Introduction .....	102
4.2.2	Exploratory, confirmatory, unrestricted, and restricted factor analysis.....	103
4.2.3	Factorial simplicity, unidimensionality, homogeneity, and reliability.....	104
4.2.4	Summary .....	106
4.3	Structural equation modeling (SEM).....	106
4.3.1	Introduction .....	106
4.3.2	LISREL matrices and notation.....	107
4.3.3	Parameter estimation for non-normal ordinal data .....	112
4.3.4	Model evaluation.....	115
4.3.5	Summary .....	118
4.4	The factor structure of the Health Education Impact Questionnaire (heiQ) .....	118
4.4.1	Introduction .....	118
4.4.2	Specific methods .....	119
4.4.3	Results.....	119
4.4.4	Summary .....	129

## Chapter 5 – A model of measurement invariance to detect response shift..... 131

5.1	Introduction .....	131
-----	--------------------	-----

## Chapter 5 (continued)

5.2	Aims .....	133
5.3	The assessment of invariance of ordinal data in repeated measures.....	133
5.4	Response shift – heiQ-PP.....	142
5.4.1	Specific methods .....	142
5.4.2	Results.....	143
5.4.3	Summary .....	145
5.5	Response shift – heiQ-PPT .....	146
5.5.1	Specific methods .....	146
5.5.2	Results.....	146
5.5.3	Summary .....	148
5.6	Response shift – heiQ-PPR.....	148
5.6.1	Specific methods .....	148
5.6.2	Results.....	148
5.6.3	Summary .....	150
5.7	Measurement invariance – heiQ-PPR Retro.....	151
5.7.1	Specific methods .....	151
5.7.2	Results.....	151
5.7.3	Summary .....	154
5.8	Factor means of heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro.....	155
5.9	Discussion.....	156
<b>Chapter 6 – Change scores mediated by social desirability .....</b>		<b>162</b>
6.1	Introduction .....	162
6.2	Aims .....	162
6.3	The factor structure of the short form MC-C of the Marlowe-Crowne scale.....	162
6.3.1	Specific methods .....	162
6.3.2	Results.....	164
6.3.3	Summary .....	168
6.4	Social desirability – heiQ-PP.....	168
6.4.1	Specific methods .....	168
6.4.2	Results.....	172
6.4.3	Summary .....	174
6.5	Social desirability – heiQ-PPT .....	175
6.5.1	Specific methods .....	175
6.5.2	Results.....	175
6.5.3	Summary .....	178
6.6	Social desirability – heiQ-PPR.....	178
6.6.1	Specific methods .....	178
6.6.2	Results.....	179
6.6.3	Summary .....	180

## Chapter 6 (continued)

6.7	Social desirability – heiQ-PPR Retro .....	180
6.7.1	Specific methods .....	180
6.7.2	Results .....	181
6.7.3	Summary .....	182
6.8	Discussion.....	183
<b>Chapter 7 – Summary, conclusions, and future directions .....</b>		<b>186</b>
7.1	Introduction .....	186
7.2	Summary of the findings .....	186
7.3	Conclusions .....	189
7.4	Strengths.....	193
7.5	Limitations.....	194
7.6	Recommendations for future research.....	196
<b>References .....</b>		<b>199</b>
<b>Appendices .....</b>		<b>219</b>
Appendix 1	Systematic review of self-management interventions that were based on or similar to the Stanford programs.....	219
Appendix 2	Pretest heiQ including demographic variables.....	225
Appendix 3	Posttest heiQ-PP including the MC-C scale and demographic variables.....	229
Appendix 4	Posttest heiQ-PPT including the MC-C scale and demographic variables.....	234
Appendix 5	Posttest heiQ-PPR including the MC-C scale and demographic variables.....	239
Appendix 6	Course participation form .....	244
Appendix 7	Univariate and multivariate normality tests of the following heiQ data: pretests (n=666), retrospective pretests heiQ-PPR (n=189), posttests (n=603), posttests heiQ-PP (n=244), posttests heiQ-PPT (n=150), and posttests heiQ-PPR (n=209) .....	245
Appendix 8	Bivariate normality tests of the following heiQ data: pretests (n=666), retrospective pretests heiQ-PPR (n=189), posttests (n=603), posttests heiQ-PP (n=244), posttests heiQ-PPT (n=150), and posttests heiQ-PPR (n=209) .....	257
Appendix 9	Output homogeneity of variances and Brown-Forsythe ANOVA of pretests and posttests across heiQ-PP, heiQ-PPT, and heiQ-PPR.....	263
Appendix 10	Output homogeneity of variances and Brown-Forsythe ANOVA of change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR.....	264

Appendix 11	Chi-square significance tests ('decline', 'no change', 'improvement') across heiQ-PP, heiQ-PPT, and heiQ-PPR .....	265
Appendix 12	Paired t-tests of heiQ-PPR and heiQ-PPR Retro .....	266
Appendix 13	Chi-square significance tests ('decline', 'no change', 'improvement') across heiQ-PPR and heiQ-PPR Retro .....	267
Appendix 14	Histograms of actual (heiQ-PPR) and retrospective (heiQ-PPR Retro) change, illustrating proportions of participants in 'decline', 'no change', or 'improvement' .....	268
Appendix 15	Step 3 of Jöreskog's 3-step procedure; full model of the 42 heiQ items (n=949).....	276
Appendix 16	Formula of the Satorra-Bentler scaled difference chi-square test statistic (Satorra & Bentler, 2001) .....	278
Appendix 17	Exploratory Factor Analysis (CEFA) results of the MC-C scale .....	279



## List of figures

Figure		Page
1	An overview of the thesis structure .....	5
2	A chronic care framework; modified from the Chronic Care Model (Wagner et al., 1999) and the Chronic Conditions Framework (World Health Organization, 2002) .....	7
3	Flow chart of the content of the thesis, Part I.....	9
4	A program logic model of potential impacts of self-management interventions in terms of short-term, medium-term, and long-term outcomes © Copyright 2004. The Medical Journal of Australia – reproduced with permission.....	10
5	Examples of chronic disease self-management interventions © Copyright 2007. The Medical Journal of Australia – reproduced with permission.....	11
6	Flow chart of the search strategy .....	20
7	Degree of personal appraisal involved in the response process across outcomes that were frequently assessed in self-management interventions; modified from Schwartz & Rapkin (2004) .....	30
8	Degree of personal appraisal involved in the response process – illustrating the eight areas on which self-management programs are expected to impact (Osborne et al., 2007).....	31
9	Flow chart of the content of the thesis, Part II.....	33
10	True experimental designs (Campbell & Stanley, 1963).....	36
11	Response shift in the context of experimental designs.....	45
12	Flow chart of the content of the thesis, Part III.....	53
13	heiQ data collection .....	57
14	Item generation, heiQ construction, and heiQ validation phase .....	62
15	Study sample .....	65
16	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Positive and Active Engagement in Life .....	78
17	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Health-Directed Behaviour.....	78
18	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Skill and Technique Acquisition .....	79
19	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Constructive Attitudes and Approaches.....	79
20	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Self-Monitoring and Insight.....	79

<b>Figure</b>	<b>Page</b>
21	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Health Service Navigation ..... 80
22	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Social Integration and Support ..... 80
23	Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Emotional Well-Being ..... 80
24	Example of presenting proportions of people in categories of change ..... 83
25	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Positive and Active Engagement in Life ..... 88
26	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Health-Directed Behaviour ..... 88
27	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Skill and Technique Acquisition ..... 88
28	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Constructive Attitudes and Approaches ..... 89
29	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Self-Monitoring and Insight ..... 89
30	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Health Service Navigation ..... 89
31	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Social Integration and Support ..... 90
32	Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Emotional Well-Being ..... 90
33	Structural equation model – illustrating Positive and Active Engagement in Life ..... 107
34	Structural equation model – illustrating Positive and Active Engagement in Life, and Health-Directed Behaviour ..... 109
35	Structural equation model including ‘defensiveness’ as the mediating variable – illustrating Positive and Active Engagement in Life ..... 171

## List of tables

Table	Page
1	Effect sizes across systematic reviews on self-management interventions..... 15
2	Effect sizes of most frequently assessed outcomes in studies included in the systematic review of self-management interventions based on or similar to the Stanford curricula..... 22
3	Effect sizes of most frequently assessed outcomes in studies included in the systematic review of Section 1.2.3.2 grouped by performance-, perception-, and evaluation-based measures (Schwartz & Rapkin, 2004)..... 29
4	Demographic characteristics of participants who provided pretests (n=1,423) and comparison of study participants (n=949) versus those not included in the study (n=474)..... 73
5	Details on the courses across participants who provided pretests (n=1,423) and comparison of study participants (n=949) versus those not included in the study (n=474)..... 74
6	Demographic characteristics of respondents across the three randomised groups: heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314).....75
7	Details on the courses across respondents of the randomised groups: heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314)..... 76
8	Mean scores of pretests and actual posttests of heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314), respectively..... 77
9	Comparison of mean change scores derived from pretest-posttest data across heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314) ..... 84
10	Proportions of people in categories 'decline', 'no change', and 'improvement'; total sample (n=949), and comparison heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314) ..... 85
11	Actual and retrospective pretest data of group heiQ-PPR (n=314)..... 87
12	Comparison of mean change scores derived from actual pretest-posttest data (heiQ-PPR) and retrospective pretest-posttest data (heiQ-PPR Retro) of group heiQ-PPR (n=314) ..... 91
13	Proportions of people in categories 'decline', 'no change', and 'improvement' across heiQ-PPR and heiQ-PPR Retro (n=314)..... 92
14	Overview of the eight standard LISREL matrices, the Theta-delta-epsilon matrix, and the four vectors for the analysis of mean structures ..... 112
15	Step 1 of Jöreskog's 3-step procedure; eight one-factor models on heiQ pretests (n=949)..... 121
16	Summary of results of Step 2 of Jöreskog's 3-step procedure (n=949)..... 124
17	Step 3 of Jöreskog's 3-step procedure; full model of the 38 heiQ items (n=949) ..... 127

<b>Table</b>	<b>Page</b>
<b>18</b> Full model of the 38-item scale on heiQ pretest data of heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR / heiQ-PPR Retro (n=314).....	129
<b>19</b> 4-step procedure for the test of measurement invariance .....	141
<b>20</b> Fit indices of the response shift detection procedure, one-factor models of heiQ-PP .....	144
<b>21</b> Fit indices of the response shift detection procedure, four-factor models of heiQ-PP .....	145
<b>22</b> Fit indices of the response shift detection procedure, one-factor models of heiQ-PPT .....	147
<b>23</b> Fit indices of the response shift detection procedure, four-factor models of heiQ-PPT .....	148
<b>24</b> Fit indices of the response shift detection procedure, one-factor models of heiQ-PPR.....	149
<b>25</b> Fit indices of the response shift detection procedure, four-factor models of heiQ-PPR.....	150
<b>26</b> Fit indices of the response shift detection procedure, one-factor models of heiQ-PPR Retro.....	152
<b>27</b> Fit indices of the response shift detection procedure, four-factor models of heiQ-PPR Retro.....	154
<b>28</b> Change scores across heiQ-PP (n=331), heiQ-PPT (n=304), heiQ-PPR (n=314), and heiQ-PPR Retro (n=314) derived from a means model (SEM) based on only those items that met the condition of scalar and metric invariance (Step 4).....	155
<b>29</b> Confirmatory Factor Analysis of the MC-C, full sample (n=908).....	165
<b>30</b> Confirmatory Factor Analysis of the MC-C, heiQ-PP (n=318) .....	166
<b>31</b> Confirmatory Factor Analysis of the MC-C, heiQ-PPT (n=291).....	167
<b>32</b> Confirmatory Factor Analysis of the MC-C, heiQ-PPR / heiQ-PPR Retro (n=299) .....	167
<b>33</b> Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PP .....	172
<b>34</b> Regression of the posttests on the pretests (Gamma matrix), heiQ-PP .....	173
<b>35</b> Regression of 'defensiveness' (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PP .....	173
<b>36</b> Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PPT .....	175
<b>37</b> Regression of the posttests on the pretests (Gamma matrix), heiQ-PPT .....	176

<b>Table</b>	<b>Page</b>
<b>38</b> Regression of 'defensiveness' (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPT .....	176
<b>39</b> Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PPR .....	179
<b>40</b> Regression of the posttests on the pretests (Gamma matrix), heiQ-PPR .....	179
<b>41</b> Regression of 'defensiveness' (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPR.....	180
<b>42</b> Covariance between SD1, SD2, and the retrospective pretests (Phi matrix), heiQ-PPR Retro .....	181
<b>43</b> Regression of the posttests on the retrospective pretests (Gamma matrix), heiQ-PPR Retro .....	181
<b>44</b> Regression of 'defensiveness' (SD1) and the posttests on the retrospective pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPR Retro .....	182

## List of abbreviations

ACT	Australian Capital Territory
ANOVA	Analysis of variance
ASMP	Arthritis Self-Management Program
BDI	Beck Depression Index
BL	Baseline model
BUOA	Bone Up On Arthritis
CAA	Constructive Attitudes and Approaches
CDSMP	Chronic Disease Self-Management Program
CEFA	Comprehensive exploratory factor analysis
CFA	Confirmatory factor analysis
CG	Control group
CFI	Comparative fit index
CI	Confidence interval
ECVI	Expected cross-validation index
EFA	Exploratory factor analysis
ER	Emergency room
ES	Effect size
EWB	Emotional Well-Being
FQCI	Freiburg Questionnaire of Coping
FSR	Factor score regression coefficient
HDB	Health-Directed Behaviour
heiQ	Health Education Impact Questionnaire
heiQ-PP	Participants who received an actual pretest and an actual posttest
heiQ-PPT	Participants who received an actual pretest and an actual posttest plus transition questions at posttest
heiQ-PPR	Participants who received an actual pretest and an actual posttest plus retrospective pretest questions at posttest
HP	Health professional
HREC	Human Research Ethics Committee
HSN	Health Service Navigation
IG	Intervention group
LISREL	Linear Structural RELations
LX	Lambda X (factor loading)
NAMCIG	National Arthritis and Musculoskeletal Conditions Improvement Grant
MI	Modification index
NNT	Number needed to treat
OA	Osteoarthritis

PAE	Positive and Active Engagement in Life
PRELIS	PRE-processor for LISrel
QoL	Quality of life
RA	Rheumatoid arthritis
RCT	Randomised controlled trial
RML	Robust maximum likelihood
RMSEA	Root mean square error of approximation
RS	Response shift
SD	Standard deviation
SD1	Social desirability, factor 1 'defensiveness'
SD2	Social desirability, factor 2 'self-presentation'
SE	Standard error
SEM	Structural equation modeling
SIS	Social Integration and Support
SMI	Self-Monitoring and Insight
SPSS	Statistical Package for the Social Sciences
SRMR	Standardized root mean square residual
STA	Skill and Technique Acquisition
TAFE	Technical and further education
TD	Theta-delta (correlated error)
TX	Tau X (intercept of observed variable)
WLS	Weighted least squares
$\chi^2_{SB}$	Satorra-Bentler chi-square

## Abstract

The burden of chronic diseases is growing in both developed and developing countries and projections suggest that the global mortality from and prevalence of chronic diseases will further increase over the next two decades. The nature of chronic diseases is a challenge for current health systems; the management of chronic conditions is complex and it requires the contribution of a range of stakeholders. While healthcare providers, community partners, government and other organisations need to be involved in chronic disease care, individuals with a chronic condition, their family and/or carers are critical, albeit often neglected partners in the care process. Given that most day-to-day responsibilities fall on those affected, their active involvement is essential to ensure continuity of care. Hence, programs that are aimed at improving patients' ability to self-manage their condition are an important component of chronic disease management.

Several interventions are currently offered that are aimed at improving patients' skills to self-manage their chronic condition. Despite anecdotal support for these programs and individual studies suggesting that self-management programs are beneficial for a wide range of people, meta-analytic reviews suggest that these interventions are only marginally effective. Clinical benefits have been demonstrated for chronic conditions such as diabetes and hypertension, whereas interventions for musculoskeletal diseases have failed to show clear benefits.

A closer examination of self-management trials, however, suggests that the inconsistency in observed findings may not be related to specific disease groups but rather to the types of outcomes on which evaluations are based. While results that were derived from clinically assessed outcomes suggest positive impacts of self-management interventions, self-report outcomes such as disability and pain generally suggest small and inconsistent results. This observation therefore raises the question whether current studies – particularly those based on self-report outcomes – adequately reflect the effectiveness of self-management programs. Because of the complexity related to assessing and interpreting self-report outcomes, it is plausible that data derived from traditional pretest-posttest methods are confounded and are therefore poor indicators of program impacts.

As a result of the apparent uncertainty with regard to the effectiveness of self-management interventions, this thesis focused specifically on the validity of the traditional pretest-posttest method to measure program outcomes. The research design targeted the processes that people undergo when filling out questionnaires and whether this has an influence on self-report outcomes. This was achieved by developing a three-group research design. While actual pretest questionnaires were identical across groups, three questionnaire versions were distributed at posttest. One of the groups filled out a traditional posttest questionnaire,



whereas the other two groups were asked to provide data in addition to actual posttest questions, with one group of participants providing transition questions and one providing retrospective pretests. These questionnaires were then randomly distributed within self-management courses. Resulting datasets (three pretest-posttest samples; one retrospective pretest-posttest sample) were further examined for possible confounding effects of response shift and social desirability biases. Through the random allocation of the questionnaires it was ensured that data were not influenced by intra-group effects but rather that differences could exclusively be attributed to the design of the posttest questionnaires.

The thesis revealed that the design of the posttest questionnaire significantly influenced people's ratings of their posttest levels. In particular, when participants were asked to provide ratings of their retrospective pretest levels in addition to their actual posttest levels, their actual posttest levels were significantly higher than those of participants who did not perform this additional task at posttest.

Further analyses of the datasets used a factor-analytic approach of measurement invariance to explore whether response shift bias was a potential confounder of results. The analyses indicated that the observed differences between groups could not be explained by this bias. That is, at a group level, response shift bias did not seem to have confounded change scores based on actual pretest-posttest data. This finding was largely identical across datasets, i.e. only a relatively small number of items were found to be non-invariant across datasets. In contrast, when the factor-analytic model was applied to retrospective pretest-posttest data, more items were found to be non-invariant, indicating potential problems with this dataset.

Finally, the influence of socially desirable answers on obtained results is of concern in survey research. In view of the study sample and the research area, i.e. health behaviour research, a potential confounding effect through social desirability was explored. However, the model of partial mediation did not provide any appreciable bias through social desirability.

This research has provided important insight into the measurement of outcomes of chronic disease self-management interventions. While the threat to the validity of traditional pretest-posttest data due to confounding effects through response shift and social desirability biases could not be supported by the present data, the thesis has highlighted that the cognitive task that subjects are asked to perform when providing data at posttest significantly influenced self-reported outcomes at posttest. Given that previous research has predominantly focused on other aspects of validity – such as applying control group designs to circumvent common threats to internal and external validity – this study suggests that more attention must be paid to the design of questionnaires. The thesis concludes that further research, in particular into the influence of cognitive tasks on obtained scores, is important to improve the interpretation of outcomes data derived from participants of self-management courses.

# Chapter 1

## Introduction

# 1 Introduction

## 1.1 Overview of the thesis

The prevalence of chronic conditions is increasing worldwide. Chronic disease is a growing burden for healthcare systems in both developed and developing countries and projections show that the global mortality from chronic conditions will be exacerbated over the next two decades (Murray & Lopez, 1996; Yach *et al.*, 2004). Particularly in developed countries, where chronic diseases have replaced acute diseases as the primary cause of death, major health system reforms are critical to ensure adequate disease management now and in the future (World Health Organization, 2003). Despite some efforts to respond to this burden of chronic disease – for example, behavioural programs and pharmaceutical treatment – the increasing threat through chronic disease has been largely neglected; a neglect that has led to a gap between current reality of the disease burden and existing practice of chronic care (Beaglehole & Yach, 2003; Lawrence, 2005; Strong *et al.*, 2005; Yach *et al.*, 2004).

The nature of chronic diseases is not only a challenge for existing healthcare systems (World Health Organization, 2003) but the management of chronic disease is complex, requiring the contribution of a range of stakeholders. Apart from healthcare providers, community partners and governments need to be involved in chronic care which should eventually lead to better health outcomes (Wagner *et al.*, 1999; World Health Organization, 2002). Finally, individuals with a chronic condition and their families and/or carers are critical, albeit often underrated partners in the chronic care process (Von Korff *et al.*, 2002). Considering that individuals who have a long-term condition cannot be under permanent supervision of a health professional, most day-to-day challenges need to be managed by the patients themselves, i.e. the active involvement of individuals is required to ensure a continuous management of their chronic condition (Pittman *et al.*, 2005; Von Korff *et al.*, 2002). Hence, interventions that are aimed at improving patients' ability to manage their condition are an important component of overall chronic care (Wagner *et al.*, 1999; World Health Organization, 2002).

Several programs are currently available that are aimed at improving patients' skills to self-manage their chronic condition. Despite the apparent need for such interventions, current studies show inconsistent results regarding program impacts. Although many specific studies suggest that programs may be beneficial for a wide range of people (Barlow *et al.*, 2000; Fu *et al.*, 2003; Lorig *et al.*, 1993; Lorig, Sobel *et al.*, 1999), meta-analyses and other systematic reviews indicate that self-management interventions are only marginally effective for certain disease groups (Chodosh *et al.*, 2005; Newman *et al.*, 2004; Warsi *et al.*, 2003; Warsi *et al.*, 2004). In particular, interventions for arthritis have failed to show clear benefits (Chodosh *et al.*, 2005; Warsi *et al.*, 2003; Warsi *et al.*, 2004) causing some dispute within the research

community (Fries *et al.*, 2003; Solomon & Lee, 2003; Solomon *et al.*, 2002). In contrast, clear clinical benefits have been demonstrated for conditions such as hypertension and diabetes (Chodosh *et al.*, 2005; Warsi *et al.*, 2003; Warsi *et al.*, 2004).

A closer examination of these trials however suggests that the magnitude and inconsistency of these findings may be related to the types of outcomes that were assessed rather than the disease group. That is, where self-report measures such as symptoms and functioning were assessed – which are commonly measured in arthritis trials (Newman *et al.*, 2004) – results tended to be smaller and inconsistent. While it is plausible that self-management programs have less impact on these outcomes, this observation may alternatively suggest that studies do not adequately reflect program effects because of the complexity related to assessing these types of outcomes (Schwartz & Rapkin, 2004). Taking into consideration the many sources of potential bias in such measures (Cronbach, 1946; Paulhus, 1991; Podsakoff *et al.*, 2003; Webb *et al.*, 1966) it therefore seems plausible that current results – in particular those referring to self-report outcomes – are not trustworthy.

The aim of this thesis was to investigate issues pertaining to the measurement of self-report outcomes of chronic disease self-management programs. By applying a structured approach to the measurement of program outcomes, the research aimed to identify and quantify the potential influence of biases in self-report measures.

**Chapter 1** provides the background on chronic disease self-management interventions. An emphasis was placed on the types of outcomes that were assessed in published studies. The last section of the chapter introduces general concepts of the measurement of change. This includes an overview of biases that are frequently encountered in self-report outcomes with a detailed review of social desirability and response shift.

**Chapter 2** describes the methods used in this thesis. This incorporates the research design, recruitment of study participants, and data collection. Further, data screening and cleaning, treatment of missing data and data management are explained.

**Chapter 3** describes the demographic characteristics of the study sample and presents the first analyses of these data. Apart from reporting several basic results (participants' scores before the intervention (=pretest), after the intervention (=posttest), and computed change), the findings from an alternative way of presenting change are presented.

**Chapter 4** describes the main statistical techniques applied in the thesis. Given that these techniques required detailed knowledge of factor analysis and structural equation modeling (SEM), an introduction to these methods is provided. Chapter 4 also includes a re-validation

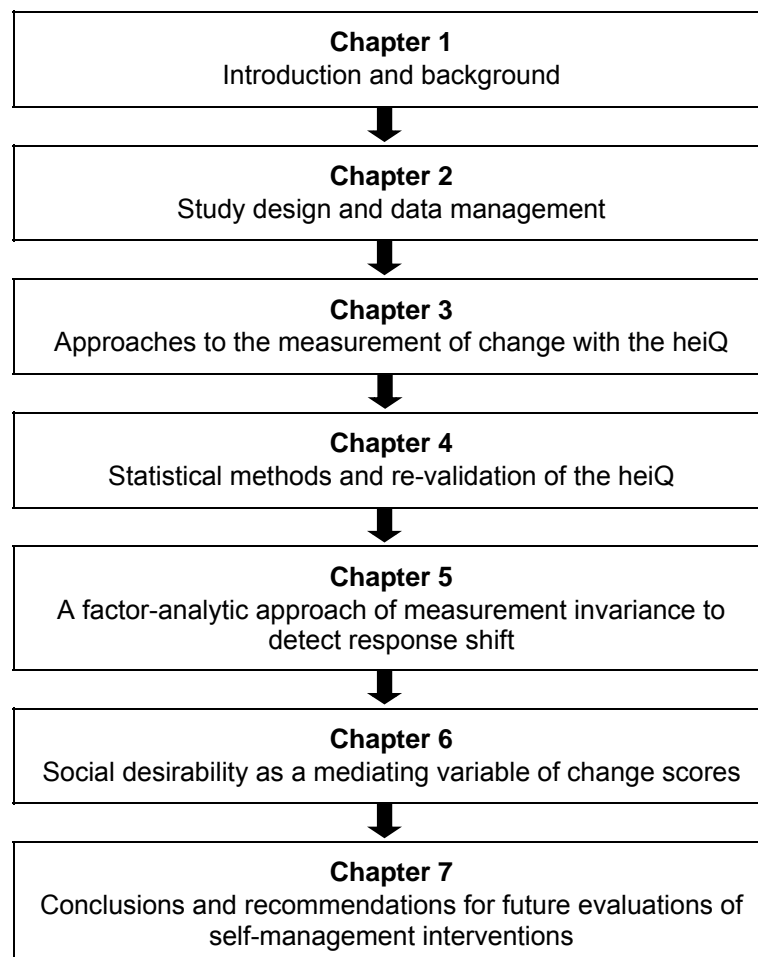
of the Health Education Impact Questionnaire (heiQ) given that the psychometric properties of this instrument were critical for subsequent analyses.

In **Chapter 5** a factor-analytic model of measurement invariance was applied to explore whether response shift bias could be detected in the study sample.

As a result of the findings of the analyses from Chapters 3 and 5, the potential influence of social desirability on the obtained scores was investigated in **Chapter 6**.

**Chapter 7** provides a discussion of the findings of the thesis. In particular, conclusions are drawn pertaining to the measurement of outcomes of self-management interventions and the interpretation of results when using different methods of determining program effects. The chapter concludes with recommendations for future research.

An overview of the thesis structure is provided in Figure 1.



**Figure 1** An overview of the thesis structure

## 1.2 Background

This section provides the theoretical base and rationale for the present research by reviewing the literature pertaining to: chronic disease, current evidence on the effectiveness of chronic disease self-management interventions, and issues related to the measurement of outcomes of these programs.

### 1.2.1 The global burden of chronic disease

The burden of chronic disease is increasing worldwide (Murray & Lopez, 1996). Recent estimates demonstrate that up to 59% of global deaths and 46% of disability are a result of chronic disease (World Health Organization, 2006). Risk factors that increase the probability of developing a chronic condition include physical inactivity, obesity, and tobacco and alcohol use (Australian Institute of Health and Welfare, 2002; Beaglehole & Yach, 2003; Yach *et al.*, 2004). Apart from behaviour-related causes, advances in the treatment of acute diseases and demographic trends such as ageing populations are reasons why chronic diseases have replaced acute diseases as the primary cause of death in developed countries (Australian Bureau of Statistics, 2004; Guterman, 2005; Lawrence, 2005; Statistisches Bundesamt, 2000). While this shift from acute to chronic diseases has occurred in these countries over the past 50 years, the prevalence of chronic conditions is a more recent phenomenon in developing countries where they are evolving at a fast pace (World Health Organization, 2003). Predictions suggest that the worldwide burden of chronic disease will be further exacerbated in the next two decades (Murray & Lopez, 1996).

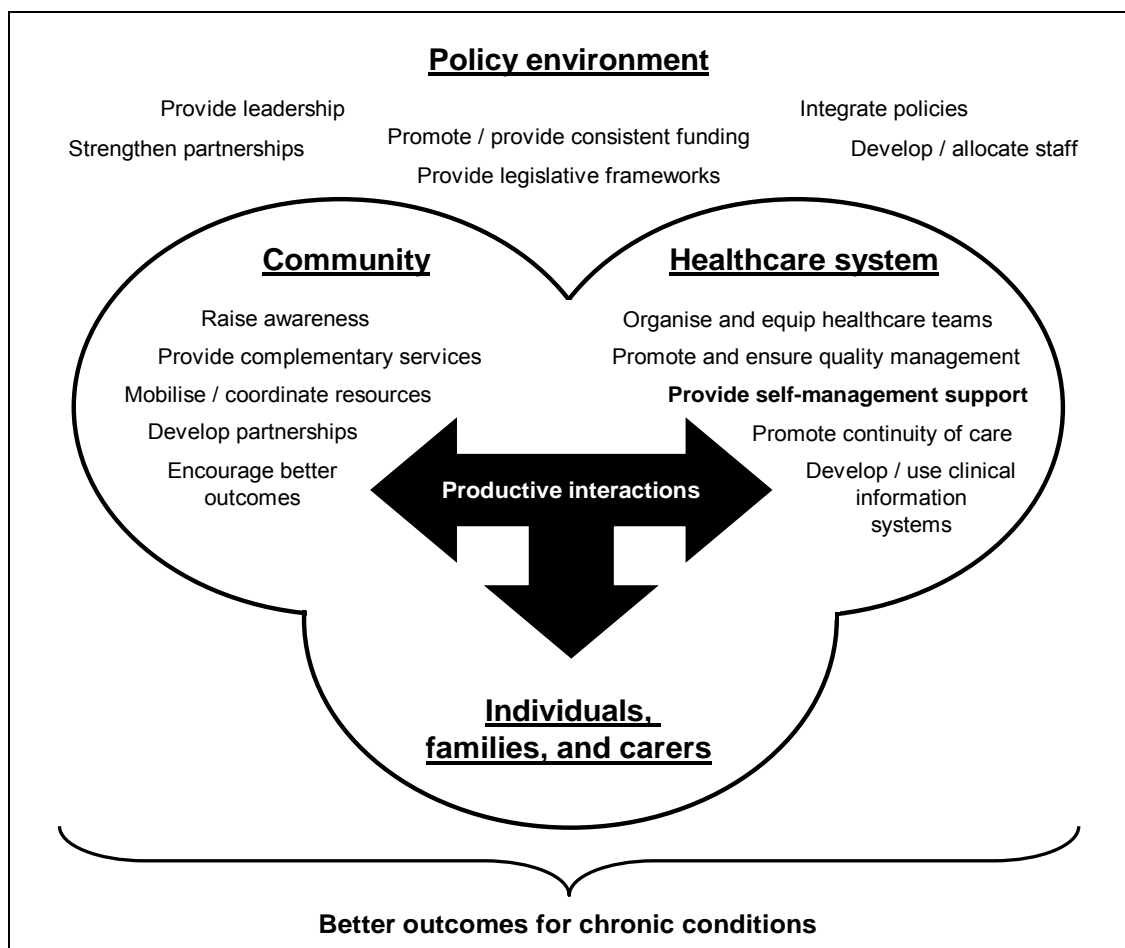
The shift from acute to chronic diseases in developed countries is a major challenge for their respective healthcare systems as well as the approach to the general treatment of disease. One of the main challenges is that the course of a chronic condition is different to that of an acute condition. While the latter is short-term and successful outcomes can be achieved through one healthcare provider (World Health Organization, 2002), chronic conditions are long-standing and more complex to manage (Yach *et al.*, 2004). Hence, no simple approach exists to the treatment of chronic diseases considering that programs need to be sustainable and geared towards the long-term management of disease (World Health Organization, 2003). Moreover, the involvement of a range of stakeholders such as healthcare teams, community partners, patients<sup>1</sup>, families and carers is required to optimise chronic disease management. However, holistic approaches that involve all of these stakeholders will not be

---

<sup>1</sup> Individuals who have a chronic condition are often referred to as “patients” in the literature. While it is acknowledged that they are not patients in a narrower sense but rather individuals who have one or more chronic conditions, both terms are used interchangeably in the present thesis.

successful until some systemic changes occur. Worldwide, health systems still follow an acute care model; they are not yet able to respond to the growing challenges of chronic disease. Hence, current health systems need to be re-oriented in the way they are structured, operated, and financed to meet the demand of chronic disease (World Health Organization, 2003; World Health Organization & Public Health Agency of Canada, 2005).

A systematic approach to chronic disease management has been described in the Chronic Conditions Framework (World Health Organization, 2002) which includes concepts from the Chronic Care Model (Wagner *et al.*, 1999). These models are representations of holistic approaches that define the process of managing chronic disease as a function of productive interactions between community partners, the healthcare system and individuals along with family members and/or carers. In addition to these stakeholders, governments and other organisations can support chronic disease management considerably by providing a positive policy environment (see Figure 2).



**Figure 2** A chronic care framework; modified from the Chronic Care Model (Wagner *et al.*, 1999) and the Chronic Conditions Framework (World Health Organization, 2002)

The model highlights that a positive policy environment can be achieved if governments integrate policies, provide leadership, and promote/provide funding. The responsibilities of the community partners include the mobilisation and coordination of resources, provision of complementary services, and the development of partnerships to support chronic disease management, while the necessary medical services are provided by the healthcare system. This includes organising and training healthcare teams, promoting continuity of care, and using clinical information systems to facilitate interaction and information flow between all stakeholders. Finally, the active engagement of individuals with chronic conditions and their families and/or carers is important to ensure productive interactions between these groups.

These interactions between the above groups can be further optimised through the provision of training that equip each group with specific knowledge and skills (Lawrence, 2005; World Health Organization, 2002). While education, feedback and reminders are typical programs for providers, reminders and education are offered for individuals (Weingarten *et al.*, 2002). Although all of the above programs are fundamental, patient education and self-management programs are of particular importance considering the nature of chronic diseases. Given that these conditions are long-term, it is unfeasible that individuals are in constant contact with healthcare providers. Consequently, regardless of the remaining stakeholders, most day-to-day challenges in the management of a chronic condition need to be faced by the individuals themselves, rendering self-management an inevitable component of chronic disease care (Pittman *et al.*, 2005; Von Korff *et al.*, 2002; Wagner *et al.*, 1999; World Health Organization, 2002).

### **1.2.2 Chronic disease self-management interventions**

In view of the increasing burden of chronic diseases introduced in the previous section and the need to provide self-management support, this section reviews programs that have been developed to improve individuals' capacity to self-manage (Wagner *et al.*, 1999) and their ability to live with their chronic condition (Lorig, 2003). Due to the large number of different self-management interventions that are currently available, the present section is limited to the definition of self-management as it applies to this thesis and a brief review of the different types of self-management programs with a focus on group-based interventions.

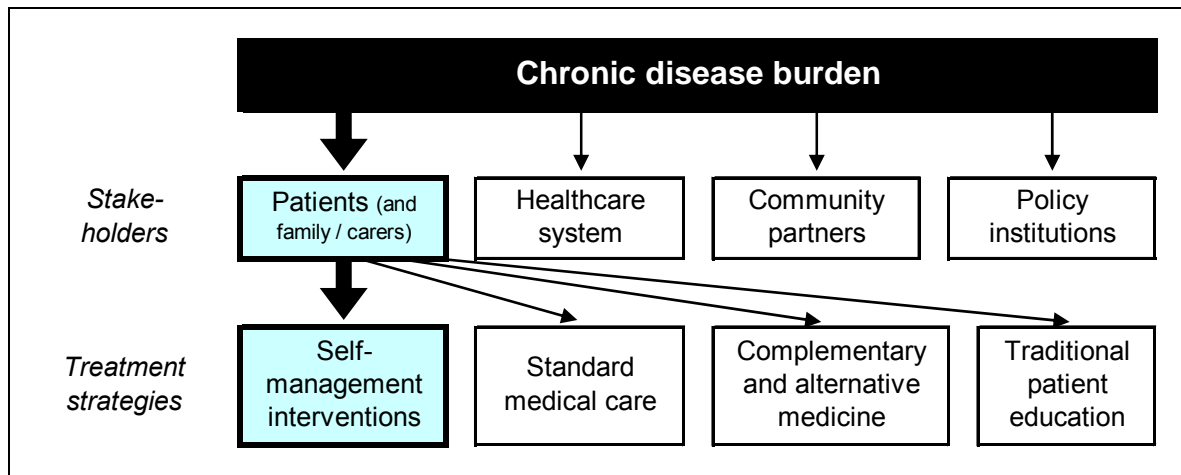
Interventions that are aimed at improving patients' ability to self-manage and live with their condition have long been recognised as an essential part of chronic disease management (Lorig & Holman, 1993; Lorig, 1982). As illustrated in Figure 3<sup>2</sup>, apart from standard medical

---

<sup>2</sup> This flow chart of the content of the thesis serves as a guide through the literature review and will be built up gradually through Chapter 1 (see also Figure 9 and Figure 12).



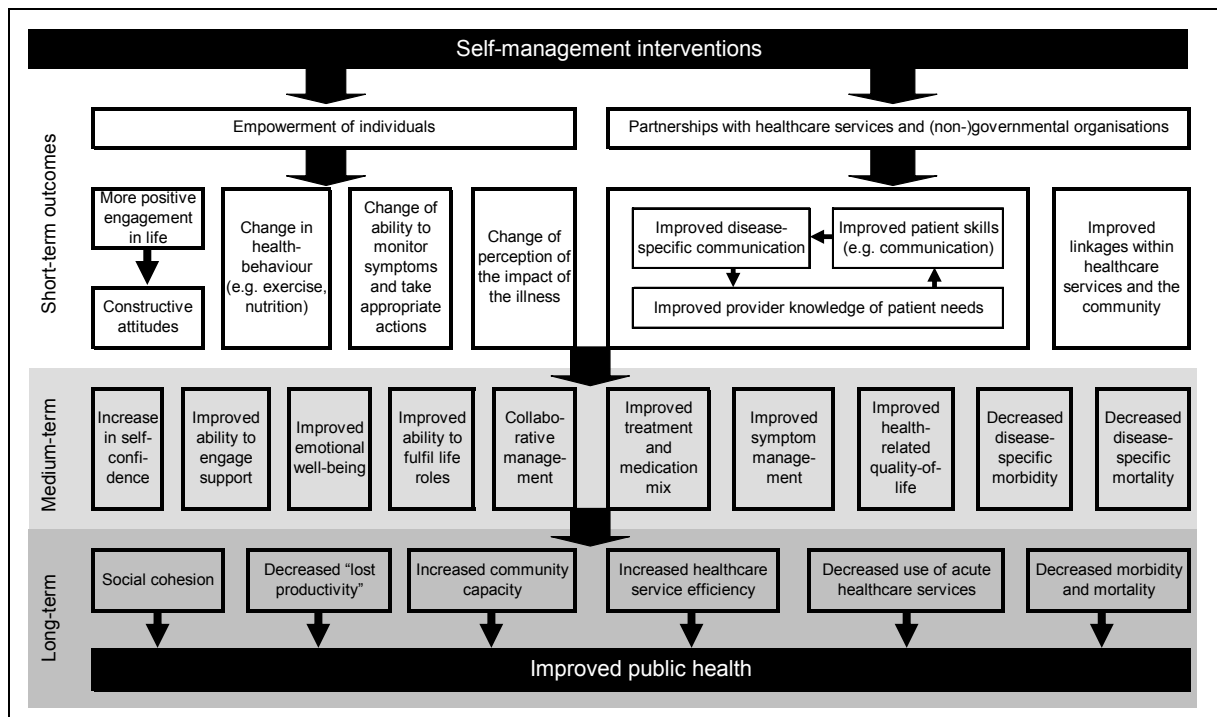
care, complementary and alternative medicine, and traditional patient education programs, self-management interventions (Von Korff *et al.*, 2002) are an important adjunct to the other types of treatments as they have the potential to fill the gap between patients' needs and the healthcare system as well as the community (Astin *et al.*, 2002; Barlow *et al.*, 2002).



**Figure 3** Flow chart of the content of the thesis, Part I

First, self-management shall be distinguished from traditional patient education. While the latter programs are aimed at improving patients' health behaviours and health status, self-management interventions focus on skills that help individuals live with their condition, solve problems, and strengthen partnerships with health professionals (Lorig, 2001). In the present thesis self-management is defined as the ability of individuals to monitor their condition and take appropriate actions to retain a satisfactory quality of life despite their disease (Lorig, 2003). This incorporates health behaviours such as exercise and proper nutrition, optimal use of medications, and an informed use of healthcare services. It also comprises skills that enable individuals to communicate effectively with health professionals, their family and/or carers as well as techniques to respond to physical and emotional issues related to their condition (Barlow *et al.*, 2002; Lorig, González, & Laurent, 1999).

To further elicit the main objectives of chronic disease self-management programs, Osborne *et al.* (2007) carried out several workshops in Australia in 2003 with representatives of a large range of stakeholders involved with chronic disease management. The first workshop included experts in self-management and was aimed at developing a program logic model for health education. The model, which was adapted to self-management interventions for the purpose of the present thesis, describes the different levels on which such programs are expected to have impacts (see Figure 4).



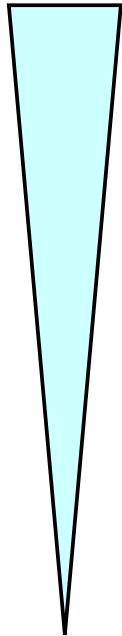
**Figure 4** A program logic model of potential impacts of self-management interventions in terms of short-term, medium-term, and long-term outcomes © Copyright 2004. The Medical Journal of Australia – reproduced with permission

Examples of short-term goals of self-management interventions are provision of information, empowering individuals and instigating change in participants' health behaviours, attitudes, and their perception of the impact of their illness. Furthermore, programs are aimed at strengthening partnerships with healthcare service providers and governmental as well as non-governmental organisations. Through improved communication skills (Lorig, González, & Laurent, 1999) disease-specific communication between patients and health professionals is expected to become more effective. At medium-term, individuals should have increased self-confidence, improved the management of their symptoms and experienced better health-related quality of life. Finally, in terms of long-term outcomes, such interventions have the potential to increase healthcare services efficiency, and decrease morbidity and mortality, and eventually they may improve public health outcomes if widely applied.

As part of the same project two further concept mapping workshops were carried out that were aimed at designing a questionnaire to assess self-management program effectiveness. Apart from the above stakeholders, individuals who had participated in a variety of self-management courses were also invited to these workshops. The workshops resulted in the definition of eight key areas on which self-management interventions are expected to have impacts (Osborne *et al.*, 2007): (1) Positive and Active Engagement in Life; (2) Health-Directed Behaviour; (3) Skill and Technique Acquisition; (4) Constructive Attitudes and

Approaches; (5) Self-Monitoring and Insight; (6) Health Service Navigation; (7) Social Integration and Support; and (8) Emotional Well-Being. The definition of self-management interventions as previously presented (page 9) is therefore extended to interventions that are aimed at improving individuals' ability to live with their chronic condition by instigating change in the above eight areas.

A wide range of self-management interventions is available in Australia. Programs vary from interventions at the population-level to programs for individuals (Jordan & Osborne, 2007). While interventions at the population-level are aimed at reaching a large number of people, group-based and one-on-one programs are more personalised and require more involvement from individuals. Furthermore, interventions differ in delivery mode. Group-based programs are currently delivered either in a class-setting or online, whereas programs for individual patients are generally delivered face-to-face, via telephone, or online. The different types of self-management interventions are illustrated in Figure 5 which includes practical examples of these programs.



Population	Type of intervention	Examples
	Television / multimedia campaigns	Quit campaign (anti-smoking)
	Written information	Publications from non-governmental organisations (e.g. Cancer Council, Arthritis Foundation, etc.)
	<b>Group-based program: formal / structured</b>	<b>Chronic Disease Self-Management Program; Arthritis Self-Management Course; "Bone Up On Arthritis" (BUOA)</b>
	Group-based program: ongoing cycle	Rehabilitation programs
	Group-based program (online)	UK National Health Service's Expert Patient Programme Online
	Online program (individual)	New South Wales Arthritis Foundation course
	Telephone coaching (individual)	Coaching patients On Achieving Cardiovascular Health (COACH) program
Individual	Face-to-face consultation (individual)	Flinders model of self-management care planning

**Figure 5** Examples of chronic disease self-management interventions © Copyright 2007. The Medical Journal of Australia – reproduced with permission

Given that group-based self-management interventions are the focus of the present thesis, the following overview is limited to this specific type of program. While again several group-based programs exist, the review provides an introduction to interventions based on curricula developed at Stanford University in the United States (Lorig *et al.*, 1985; Lorig, González, & Laurent, 1999), as these are the most common type of self-management interventions employed in Australia and the ones predominantly represented in the study sample. An overview of other disease-specific self-management interventions for chronic conditions such as asthma, cancer, cardiovascular diseases, diabetes, mental health, and pain has been provided by Redman (2001).

The Stanford programs are highly structured with licensed course leaders following a clearly defined protocol that is provided by the Stanford Patient Education Research Center (Lorig *et al.*, 1989; Lorig, González, & Laurent, 1999). The two main programs are the disease-specific Arthritis Self-Management Course (Lorig *et al.*, 1985) and the generic Chronic Disease Self-Management Program (Lorig, González, & Laurent, 1999). While the generic course was adapted from the former course and contains several elements of its curriculum, it is offered to a much broader audience as it is built on the assumption that people with any type of chronic disease face similar problems with regard to managing their condition (Lorig, Sobel *et al.*, 1999).

Both programs are run over a period of six to seven weeks with one two or two-and-a-half hour session each week. Groups consist of up to 15 participants and are generally led by two lay persons (Lorig, Ritter *et al.*, 2001). While these programs were initially designed to be peer-led only – and studies suggest that lay-led courses are equally effective to those run by health professionals (Cohen *et al.*, 1986; Lorig *et al.*, 1986) – self-management courses are now also run by either a combination of one lay leader and one health professional, or two health professionals (Lorig, Sobel *et al.*, 2001). Central components of the self-management sessions are action planning and problem-solving. Participants are also introduced to cognitive symptom management and techniques to deal with anger, fear, frustration, and depression. Further topics include muscle relaxation, fatigue management, exercise, and communication with healthcare teams and significant others. The sessions are interactive with the course leaders facilitating discussion between the participants (Lorig, González, & Laurent, 1999).

### 1.2.3 The effectiveness of self-management interventions

As outlined in Sections 1.2.1 and 1.2.2 self-management support is a central component of chronic disease management (Wagner *et al.*, 1999; World Health Organization, 2002). To ascertain whether group-based self-management interventions are effective, this section was aimed at providing a detailed review of current evidence. Published meta-analytic and other reviews are summarised in Section 1.2.3.1, while results of a systematic review of trials on self-management programs is presented in Section 1.2.3.2. In the final part of this section the reviewed studies were further examined with regard to the types of outcomes that were used to evaluate these interventions (see Section 1.2.3.3).

#### 1.2.3.1 Meta-analytic and other systematic reviews on self-management interventions

Self-management programs have been evaluated extensively. More than 2,000 publications were retrieved for the systematic review that was carried out in Section 1.2.3.2. Several recent meta-analytic and systematic reviews make the synthesis of outcomes of the studies manageable. Eight of these are presented in the following overview.

Apart from selecting reviews that explicitly used the term 'self-management', publications were also included that had reviewed 'psycho-educational' or 'psychological' interventions for chronic disease. In view of the distinction between 'patient education' and 'self-management' provided in Section 1.2.2, reviews of patient education were discarded. Further, the included reviews were selected on the basis of availability of summary effect sizes (ES), where ES is a standardised change score<sup>3</sup> (Cohen, 1988) that has been recommended to present health changes (Kazis *et al.*, 1989). As ES was applied throughout the thesis, further details on its calculation follow in Section 1.2.3.2. For this overview results are presented in summary ES, as reported in the selected publications, where a positive ES reflects improvement and a negative ES reflects deterioration on the target construct. Results were interpreted as small (ES~0.2), medium (ES~0.5), or large (ES~0.8) effects (Cohen, 1988). While the present overview relies on ES, one narrative review (Newman *et al.*, 2004) was included because of its comprehensiveness and qualitative overviews.

Most publications were based on disease-specific interventions. Two of these reviewed self-management programs for non-specific arthritis (Mullen *et al.*, 1987; Warsi *et al.*, 2003), one focused on osteoarthritis (Devos-Comby *et al.*, 2006), one reported outcomes for rheumatoid arthritis (Astin *et al.*, 2002), one focused on type 2 diabetes (Ismail *et al.*, 2004), and three

---

<sup>3</sup> In this thesis, unless stated otherwise, change scores refer to the comparison of participants' ratings of their respective present state on the target construct assessed at pretest (=before the intervention) with those assessed at posttest (=after the intervention). For simplification, these ratings are generally abbreviated to 'pretest scores' and 'posttest scores'.

reviewed several programs including arthritis, type 2 diabetes, and hypertension (Chodosh *et al.*, 2005; Newman *et al.*, 2004; Warsi *et al.*, 2004).<sup>4</sup>

Across reviews there was general consensus that self-management interventions can be beneficial for people with chronic conditions and programs were mostly regarded as an important adjunct to standard medical care (Astin *et al.*, 2002; Chodosh *et al.*, 2005; Devos-Comby *et al.*, 2006; Ismail *et al.*, 2004; Mullen *et al.*, 1987). While the reviews suggested somewhat larger effects for individuals with diabetes as well as hypertension compared with arthritis, two meta-analyses found clear evidence of publication bias in studies that targeted the former two disease groups (Chodosh *et al.*, 2005; Warsi *et al.*, 2004).

As shown in Table 1, reviews of *diabetes* studies suggested some statistically and clinically significant effects. While changes in fasting blood glucose levels were small, medium effects were found for glycosylated haemoglobin (HbA<sub>1c</sub>) and psychological variables (Chodosh *et al.*, 2005; Ismail *et al.*, 2004; Warsi *et al.*, 2004). Further, Newman *et al.* (2004) reported that 61% of studies showed reduced HbA<sub>1c</sub>, 54% demonstrated changes in self-management behaviours, and about one in three suggested improvements in psychological well-being. In contrast, programs had only little effect on individuals' quality of life (Newman *et al.*, 2004). However, two of the meta-analyses expressed concern regarding publication bias and so the results should be interpreted with caution (Chodosh *et al.*, 2005; Warsi *et al.*, 2004).

Reviews of interventions for people with *hypertension* showed some small to medium effects (see Table 1). While one meta-analysis found medium effects for diastolic and systolic blood pressure (Chodosh *et al.*, 2005), the other study reported negligible to small effects (Warsi *et al.*, 2004). Both reviews reported that there was evidence of publication bias and so again results should be interpreted with caution (Chodosh *et al.*, 2005; Warsi *et al.*, 2004).

Compared with the previous disease groups, *arthritis* programs suggested smaller effects. Despite these interventions receiving much attention in the literature, few studies reported medium or large effects. As presented in Table 1, negligible to small effects were found for disability, function, impairment, overall impact of osteoarthritis, pain, and physical outcomes. While the narrative review showed that 83% of studies indicated behaviour change and 60% found improved psychological well-being, considerably fewer studies reported positive effects for self-report outcomes such as disability, pain, painful and swollen joints, and symptoms (Newman *et al.*, 2004). In contrast to the reviews on diabetes and hypertension, no evidence of publication bias was reported for arthritis trials (Astin *et al.*, 2002; Chodosh *et al.*, 2005; Devos-Comby *et al.*, 2006; Warsi *et al.*, 2003; Warsi *et al.*, 2004).

---

<sup>4</sup> Data on asthma interventions had to be excluded because reviews included programs that provided only minimal education (Gibson *et al.*, 2002) or interventions included children (Warsi *et al.*, 2004).

**Table 1** Effect sizes across systematic reviews on self-management interventions<sup>1</sup>

Publication	Assessed outcome	ES <sup>2</sup>	95% CI
<u>Diabetes</u>			
Chodosh <i>et al.</i> , 2005	Fasting blood glucose	0.28	0.08-0.47
Ismail <i>et al.</i> , 2004	Fasting blood glucose	0.11	-0.42-0.65
Warsi <i>et al.</i> , 2004	Fasting blood glucose	0.11	-0.05-0.28
Chodosh <i>et al.</i> , 2005	Glycated haemoglobin	0.36	0.21-0.52
Ismail <i>et al.</i> , 2004	Glycated haemoglobin	0.32	0.07-0.57
Warsi <i>et al.</i> , 2004	Glycated haemoglobin	0.45	0.17-0.74
Chodosh <i>et al.</i> , 2005	Weight	-0.04	-0.07-0.16
Ismail <i>et al.</i> , 2004	Weight	0.00	-0.20-0.20
Ismail <i>et al.</i> , 2004	Psychological distress	0.58	0.20-0.95
<u>Hypertension</u>			
Chodosh <i>et al.</i> , 2005	Diastolic blood pressure	0.51	0.30-0.73
Warsi <i>et al.</i> , 2004	Diastolic blood pressure	0.10	-0.06-0.26
Chodosh <i>et al.</i> , 2005	Systolic blood pressure	0.39	0.28-0.51
Warsi <i>et al.</i> , 2004	Systolic blood pressure	0.20	0.01-0.39
<u>Osteoarthritis (OA), rheumatoid arthritis (RA), and other types of arthritis</u>			
Astin <i>et al.</i> , 2002 (RA)	Coping	0.46	0.09-0.83
Mullen <i>et al.</i> , 1987 (any arthritis)	Depression	0.28	0.15-0.42
Astin <i>et al.</i> , 2002 (RA)	Disability	0.27	0.12-0.42
Mullen <i>et al.</i> , 1987 (any arthritis)	Disability	0.09	-0.03-0.21
Warsi <i>et al.</i> , 2003 (any arthritis)	Disability	0.07	0.00-0.15
Chodosh <i>et al.</i> , 2005 (OA)	Function	0.06	0.02-0.10
Devos-Comby <i>et al.</i> , 2006 (OA)	Impairment	0.04	-0.25-0.34
Devos-Comby <i>et al.</i> , 2006 (OA)	Overall impact of OA	0.11	0.01-0.21
Astin <i>et al.</i> , 2002 (RA)	Pain	0.22	0.07-0.37
Chodosh <i>et al.</i> , 2005 (OA)	Pain	0.06	0.02-0.10
Mullen <i>et al.</i> , 1987 (any arthritis)	Pain	0.21	0.08-0.33
Warsi <i>et al.</i> , 2003 (any arthritis)	Pain	0.12	0.00-0.24
Devos-Comby <i>et al.</i> , 2006 (OA)	Physical outcomes	0.09	-0.01-0.19
Astin <i>et al.</i> , 2002 (RA)	Psychological outcomes	0.15	0.01-0.31
Devos-Comby <i>et al.</i> , 2006 (OA)	Psychological outcomes	0.20	0.08-0.33
Astin <i>et al.</i> , 2002 (RA)	Self-efficacy	0.35	0.11-0.59
Astin <i>et al.</i> , 2002 (RA)	Tender joints	0.15	0.09-0.39

<sup>1</sup> Effect sizes refer to comparisons of pretest with post intervention scores. They are presented in a way that positive signs mean improvement and negative signs mean deterioration on the target construct. ES were interpreted as small (ES~0.2), medium (ES~0.5), and large (ES~0.8) effects (Cohen, 1988).

<sup>2</sup> Effect sizes are reported as presented in the reviews. They most frequently refer to Cohen's d.

Legend

CI: confidence interval      ES: effect size      OA: osteoarthritis      RA: rheumatoid arthritis

## *Discussion*

Narrative and meta-analytic reviews suggest that self-management interventions can result in positive outcomes for some chronic conditions. While statistically and clinically significant effects were demonstrated for diabetes and hypertension, programs for arthritis appeared somewhat less effective. A closer examination of the included reviews however suggests that some meta-analyses were not only heterogeneous regarding the types of chronic diseases that were examined but evaluations also relied on different outcome measures. While studies on chronic diseases such as diabetes and hypertension regularly reported clinically assessed outcomes, trials on arthritis relied on self-report outcomes such as symptoms and functioning (Newman *et al.*, 2004) as there are no objective biological measures of disease severity. Considering that self-report outcomes are more complex to measure than clinical outcomes (Schwartz & Rapkin, 2004), it is plausible that current evaluations of arthritis trials are not an accurate reflection of program effects. Hence, current reviews may need to be interpreted more carefully in view of the types of outcomes on which they are based.

Further, the general usefulness of summary scores of such reviews needs to be considered. When summarising individual trials in meta-analytic reviews, the heterogeneity of programs, and the quality and design of the included studies constitute major drawbacks in view of the generalisability of findings (Eysenck, 1994; Knipschild, 1994). Because of this complexity, it is difficult to draw definite conclusions about the effectiveness of programs (Newman *et al.*, 2004). Although the present summary was restricted to specific program types, the reviews still included studies that did not match the types of self-management programs evaluated in the thesis. Some reviews contained studies with one-on-one sessions (Ismail *et al.*, 2004; Mullen *et al.*, 1987; Warsi *et al.*, 2003), other studies included exercise lessons, programs that consisted of phone calls or videotapes only (Chodosh *et al.*, 2005), or interventions without a formal curriculum (Warsi *et al.*, 2004). Considering these findings, the reviews can only provide a general overview of the effectiveness of a range of self-management courses. However, despite this limitation the reviews provide important insight into potential issues related to the comparison of studies that are based on different types of outcomes and the difficulties related to assessing self-report outcome measures.

### *1.2.3.2 A systematic review of self-management interventions*

The previous summary of reviews provided an overview of current evidence with regard to the effectiveness of self-management interventions. However, given that none of the reviews specifically summarised the types of interventions considered in this thesis, a systematic



review was performed on trials that were based on or similar to the Stanford protocol (Lorig *et al.*, 1985; Lorig, González, & Laurent, 1999).

### *Search strategy*

The criteria and rationale for selecting studies for the systematic review were as follows:

- (1) Inclusion of studies evaluating disease-specific or generic self-management interventions that were comparable to the programs assessed in the present thesis. If studies did not include a direct reference to Stanford (Lorig *et al.*, 1985; Lorig, González, & Laurent, 1999), studies were selected that evaluated interventions that included at least two of the three keywords 'problem-solving', 'action planning', and 'relaxation'. To be included in the review four characteristics had to be met by the self-management programs:
  - (a) Interventions were delivered in a group-setting;
  - (b) Interventions were based on a formal syllabus;
  - (c) Interventions ran between four and ten sessions within a period of three months;
  - (d) Interventions did not include any additional components such as exercise lessons, reinforcement techniques, individual consultations, and/or home visits.
- (2) Inclusion of studies between 1982 and 2006 because the first Stanford program (arthritis) was originally published in 1982 (Lorig, 1982).
- (3) Inclusion of randomised controlled trials (RCT) only. The search was limited to RCTs to keep the number of studies to a manageable size rather than follow the hierarchy of research designs where RCTs are considered the 'gold standard' (Sackett, 1994).<sup>5</sup> Further, the RCTs that were included compared an intervention group with a control group that did not receive any intervention but standard medical care.
- (4) Exclusion of studies that did not have sufficient power to detect a large-sized difference between intervention and control group means. That is, at  $\alpha=0.05$  a sample size = 26 is required to detect a large difference between the means of two independent samples (Cohen, 1992; Schwartz *et al.*, 2006).
- (5) Exclusion of studies that did not provide sufficient information on measured outcomes for the calculation of ES and the missing data could not be obtained from the authors.

---

<sup>5</sup> A more detailed discussion on research designs in the context of the present thesis, i.e. in the context of the measurement of outcomes of self-management interventions, follows in Section 1.2.4.

- (6) Exclusion of studies that did not assess any self-report outcome measures, i.e. studies purely assessing outcomes such as cost effectiveness or drug adherence were excluded.
- (7) Exclusion of studies on interventions for children or adolescents.

The systematic search was performed in December 2006 across the databases MEDLINE, EMBASE, CINAHL, and PsycINFO which are recommended for systematic reviews (Taal *et al.*, 2004). The search terms were 'self-management' or 'patient education', 'randomized' or 'randomised' or 'RCT', and 'chronic condition' or 'chronic disease' or 'arthritis' or 'asthma' or 'chronic obstructive pulmonary disease' or 'congestive heart failure' or 'COPD' or 'diabetes' or 'fibromyalgia' or 'hypertension' or 'musculoskeletal' or 'osteoarthritis' or 'osteoporosis' or 'pain' or 'rheumatoid' or 'stress'. These terms were derived from the types of programs that were included in the present thesis as well as from other reviews in this area (Chodosh *et al.*, 2005; Newman *et al.*, 2004). In PsycINFO, the search was further restricted to 'Journal articles only', languages 'English, German, Spanish' and 'Age groups 18 years and older'. In CINAHL the search was restricted to the above three languages and 'all adult'. Further studies were retrieved from The Cochrane Database of Systematic Reviews (The Cochrane Collaboration, 2007) and from reference lists of other systematic reviews and meta-analyses.

#### *Data analysis and presentation*

To make the results of the included trials comparable, ES were calculated for each outcome. Between-group treatment effects were calculated using Cohen's *d* (Cohen, 1988) which is one of the most commonly reported effect size indices (Rosnow & Rosenthal, 1996). It is calculated by subtracting the mean change score of the control group from the mean change score of the intervention group which is then divided by the pooled standard deviation (SD) of the two groups (Cohen, 1988; Rosnow & Rosenthal, 1996). The pooled SD is derived from the square root of the sum of the groups' variances of the baseline scores divided by two (Cohen, 1988). Further, the 95% confidence interval (CI) of each these ES estimates was calculated. It was derived from the estimated standard errors (SE) of the ES. After computing the SE for each of the outcome variables, the 95% CI of the ES was obtained by multiplying the square root of these SE by 1.96 (Hedges & Olkin, 1985).

Additionally, within-group ES were calculated for each treatment condition separately. This was useful to retain information about the source of between-group ES, i.e. whether changes in control group subjects (CG) were partially responsible for observed between-group ES. For example, it could be explored whether a large effect was caused by an improvement in

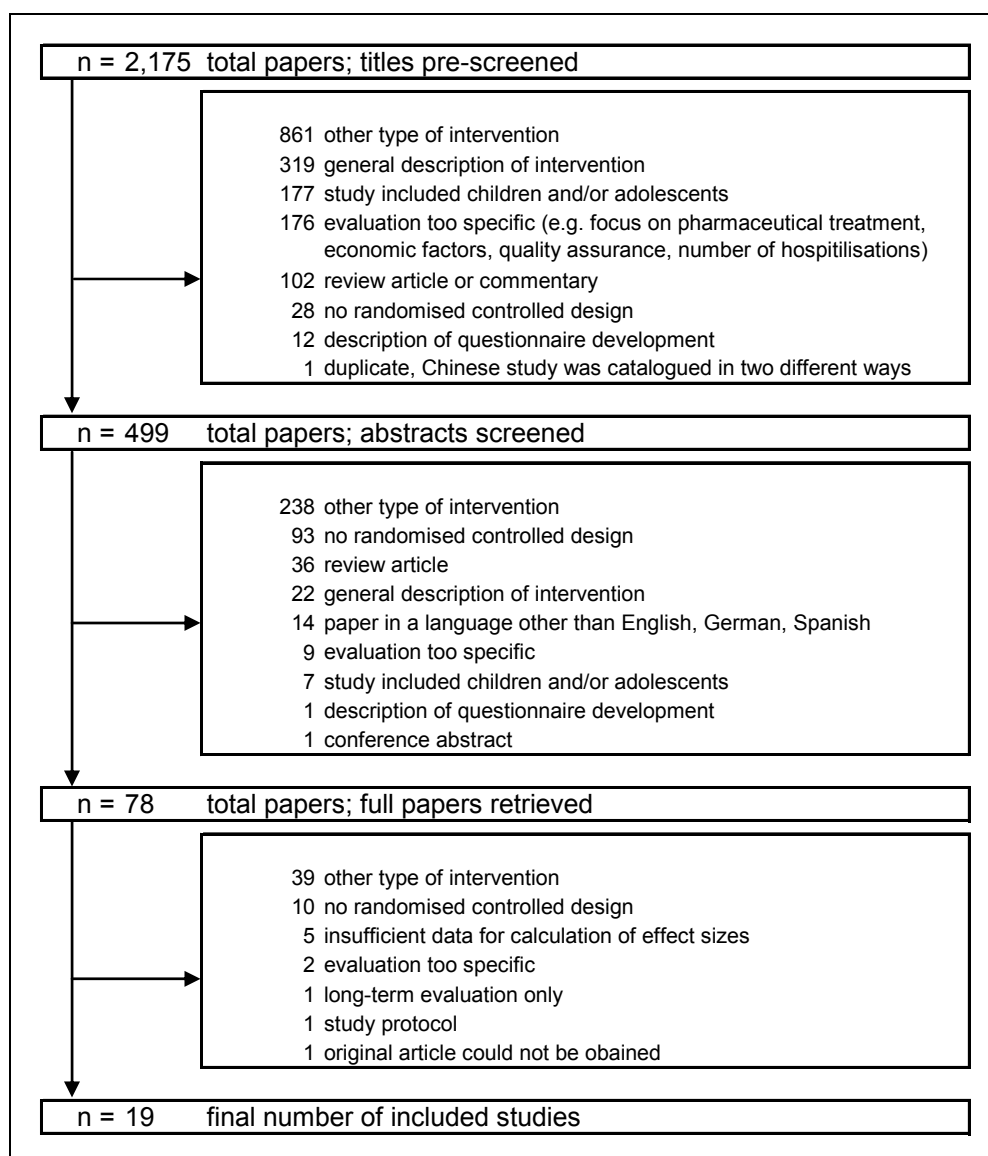
the intervention group (IG) or a decline in the control group, or whether a negligible between-group ES was caused by no change in either group or alternatively by a simultaneous change in both groups. In the same manner as in Section 1.2.3.1, reported effects are presented in a way that a positive ES reflects improvement and a negative ES reflects decline, and again obtained results were interpreted as small (ES~0.2), medium (ES~0.5), or large (ES~0.8) effects (Cohen, 1988).

While several studies included repeated measures with varying time periods, only the first post intervention assessment is reported in this review, i.e. no additional longitudinal data were included. Despite this restriction, follow-up scores relate to immediate post-assessment to assessment up to six months post intervention as this was the first post course data collection in some of the studies (Boesen *et al.*, 2005; Fu *et al.*, 2003; Haas *et al.*, 2005; Lorig, Sobel *et al.*, 1999).

Because of space constraints Table 2 provides a summary of the results and detailed results are provided in Appendix 1. In the summary table results are presented in a way that both the minimum and the maximum ES per outcome across studies is presented. Further, the median ES of the included studies was calculated. Hence, no summary scores such as those used in meta-analyses are reported as the main aim of the present review was to investigate the measures used in individual evaluations as well as effects across studies.

## *Results*

The result of the systematic search is presented in Figure 6. A total of 2,175 papers were identified in the systematic search. After pre-screening the titles, 1,676 publications were excluded as they failed to meet all inclusion criteria. The majority of studies were rejected because they evaluated other types of self-management interventions (n=861), were a general description of a program (n=319), or children and/or adolescents were included (n=177). Of the remaining 499 publications all abstracts were screened. The majority of these again did not meet all inclusion criteria with most exploring other types of interventions (n=238). This left 78 papers which were examined in full. Of these, 54 studies were excluded for similar reasons as above with the majority of trials evaluating other types of interventions (n=39). Eight of the remaining studies did not report sufficient data for the calculation of ES. After contacting the authors of each of these studies, three researchers were able to provide most missing data (Burckhardt *et al.*, 1994; Griffiths *et al.*, 2005; Taal *et al.*, 1993). No further information could be obtained from the remaining five authors (Clark *et al.*, 1992; Cohen *et al.*, 1986; Solomon *et al.*, 2002; Swerissen *et al.*, 2006; Worth, 2002). As a result, the review was based on a final number of 19 studies.



**Figure 6** Flow chart of the search strategy

The majority of the 19 trials investigated the effectiveness of arthritis-specific interventions. Seven studies evaluated the Arthritis Self-Management Course, with four targeting people with a range of musculoskeletal conditions (Barlow *et al.*, 2000; Lorig *et al.*, 1986; Lorig, González, & Ritter, 1999; Lorig *et al.*, 1989) and three focusing on osteoarthritis (Heuts *et al.*, 2005; Hopman-Rock & Westhoff, 2000; Keefe *et al.*, 1990). Three further trials evaluated alternative arthritis-specific interventions, with one being the ‘Bone Up On Arthritis’ program (Goepfing *et al.*, 1989) and two studies investigating group interventions for people with rheumatoid arthritis (Scholten *et al.*, 1999; Taal *et al.*, 1993). The remaining trials evaluated other disease-specific interventions targeted at people with back pain (Haas *et al.*, 2005; Von Korff *et al.*, 1998), chronic pain (LeFort *et al.*, 1998), fibromyalgia (Burckhardt *et al.*, 1994), and cancer (Boesen *et al.*, 2005), whereas four studies assessed the Chronic Disease Self-

Management Program (Fu *et al.*, 2003; Griffiths *et al.*, 2005; Lorig *et al.*, 2003; Lorig, Sobel *et al.*, 1999). One of the above trials (Lorig *et al.*, 1986) compared peer-led with professional-led courses. For the present study, given that the thesis did not differentiate between modes of instruction, the two groups were regarded as two separate trials.

The results of the systematic review are summarised in Table 2 and the detailed overview is provided in Appendix 1. Across trials more than 70 different variables were assessed with *depression*, *disability*, *pain* and *self-efficacy* being the most frequently assessed outcomes. Between one quarter and one third of studies collected data on outcomes such as *anxiety*, *communication with physician*, *fatigue*, *general health*, *knowledge*, *physical functioning* and *visits to physician*. In contrast, all remaining outcomes were assessed too infrequently to perform inter-study comparisons. Given that the majority of studies included participants with arthritis, the types of outcomes that were assessed were most similar to those reported in the systematic reviews on arthritis (see Section 1.2.3.1).

The impact of self-management interventions on levels of *anxiety* was assessed in five trials. Overall, negligible to small between- and within-group effects were found, with within-group ES being consistently larger than between-group ES (Barlow *et al.*, 2000; Boesen *et al.*, 2005; Burckhardt *et al.*, 1994; Griffiths *et al.*, 2005; Taal *et al.*, 1993).

*Communication with the physician* was also assessed in five trials. While observed between- and within-group effects were negligible to small (Barlow *et al.*, 2000; Fu *et al.*, 2003; Griffiths *et al.*, 2005; Lorig, Sobel *et al.*, 1999), one study reported medium-size within-group effects in a Spanish-speaking population (Lorig *et al.*, 2003).

A frequently assessed outcome was *depression* which was reported in 11 trials. Between- and within-group effects varied greatly across studies. While one trial (Scholten *et al.*, 1999) found medium and two trials observed small between- and within-group effects (Barlow *et al.*, 2000; LeFort *et al.*, 1998), all remaining trials showed negligible to small ES with maximum between-group effects of  $ES=0.13$  (Boesen *et al.*, 2005; Burckhardt *et al.*, 1994; Fu *et al.*, 2003; Goepfing *et al.*, 1989; Griffiths *et al.*, 2005; Lorig, González, & Ritter, 1999; Lorig *et al.*, 1989; Taal *et al.*, 1993). In three of these, however, the between-group effects were influenced by simultaneous improvements in intervention and control group subjects (Fu *et al.*, 2003; Lorig, González, & Ritter, 1999; Taal *et al.*, 1993).

Another outcome that was assessed in approximately two thirds of the included trials was *disability*. Reported effects again varied greatly across studies. Between- and within-group effects ranged from small negative (Keefe *et al.*, 1990) or negligible effects (Goepfing *et al.*, 1989; Lorig *et al.*, 1989; Lorig, Sobel *et al.*, 1999) to large positive effects (Scholten *et al.*, 1999).

**Table 2** Effect sizes of most frequently assessed outcomes in studies included in the systematic review of self-management interventions based on or similar to the Stanford curricula

Assessed outcome		n	Minimum	Median	Maximum	Range
Anxiety	d	5	0.02	0.10	0.20	0.18
	IG		0.11	0.21	0.31	0.20
	CG		0.00	0.08	0.25	0.25
Comm. phys.	d	5	-0.07	0.13	0.34	0.41
	IG		0.04	0.23	0.49	0.45
	CG		0.03	0.10	0.23	0.20
Depression <sup>1</sup>	d	11	0.00	0.12	0.64	0.64
	IG		0.07	0.21	0.57	0.50
	CG		-0.06	0.04	0.36	0.42
Disability	d	13	-0.18	0.14	1.42	1.60
	IG		-0.20	0.15	1.28	1.48
	CG		-0.43	-0.02	0.31	0.74
Fatigue	d	6	-0.01	0.19	0.29	0.30
	IG		0.12	0.18	0.40	0.28
	CG		-0.10	0.03	0.13	0.23
General health	d	7	-0.21	0.16	0.48	0.69
	IG		-0.11	0.17	0.52	0.63
	CG		-0.19	0.04	0.16	0.35
Knowledge	d	5	-0.05	0.78	1.11	1.16
	IG		0.37	0.95	1.28	0.91
	CG		0.04	0.17	0.42	0.38
Pain <sup>2</sup>	d	18	-0.28	0.10	0.43	0.71
	IG		-0.01	0.20	0.75	0.76
	CG		-0.34	0.11	0.80	1.14
Phys. funct.	d	4	-0.04	0.11	0.23	0.27
	IG		-0.06	0.06	0.26	0.32
	CG		-0.12	-0.01	0.03	0.15
Self-efficacy	d	10	0.05	0.30	0.72	0.67
	IG		0.02	0.40	0.64	0.62
	CG		-0.34	0.01	0.27	0.61
Visits phys.	d	8	-0.34	0.02	0.18	0.52
	IG		-0.11	0.13	0.18	0.29
	CG		0.02	0.12	0.21	0.19

<sup>1</sup> Scholten *et al.* (1999) assessed depression with the Freiburg Questionnaire of Coping with Illness (FQCI) and with the Beck Depression Index (BDI); these data were included as an average score of the two effect sizes (ES)

<sup>2</sup> Heuts *et al.* (2005) reported knee and hip pain separately; these data were included as an average score of the two ES

Legend

- CG: Effect size, control group
- Comm. phys.: Communication with physician
- d: Cohen's d, group difference (ES IG - ES CG)
- IG: Effect size, intervention group
- n: Number of studies included in the systematic review
- Phys. funct.: Physical functioning
- Visits phys.: Visits to physician

Six studies assessed effects on *fatigue*. The range of between-group ES was relatively small with between- and within-group ES showing a maximum of small effects (Barlow *et al.*, 2000; Boesen *et al.*, 2005; Burckhardt *et al.*, 1994; Fu *et al.*, 2003; Griffiths *et al.*, 2005; Lorig *et al.*, 2003). An exception was a medium within-group effect observed in Spanish-speaking course participants (Lorig *et al.*, 2003).

The impact of self-management interventions on *general health* was assessed in seven trials. Similar to most previous outcomes, effects varied greatly. While effects ranged from negative between-group ES (Haas *et al.*, 2005) to medium positive between- and within-group ES (Fu *et al.*, 2003; Lorig *et al.*, 2003), the majority of studies suggested negligible to small effects (Heuts *et al.*, 2005; LeFort *et al.*, 1998; Lorig, González, & Ritter, 1999; Lorig, Sobel *et al.*, 1999).

*Knowledge* was assessed in one quarter of the included studies all of which were arthritis-specific interventions (Goepfing *et al.*, 1989; Hopman-Rock & Westhoff, 2000; Lorig *et al.*, 1986; Lorig *et al.*, 1989). In contrast to previous results, between- as well as within-group effects were generally medium or large. The only exception was the study that compared lay-led courses with courses run by health professionals, with the former showing simultaneous medium-size improvements in subjects of both intervention and control group. In contrast, in self-management courses run by health professionals, improvements in intervention group subjects clearly exceeded those of the control group (Lorig *et al.*, 1986).

The impact of self-management programs on *pain* was assessed across all but two studies (Boesen *et al.*, 2005; Scholten *et al.*, 1999). Reported effects varied considerably and ranged from some trials observing small negative effects (Burckhardt *et al.*, 1994; Lorig *et al.*, 1986) to other studies showing medium positive between-group effects. The latter, however, were caused by increased pain in control subjects in some of the studies rather than improvement in experimental subjects (Hopman-Rock & Westhoff, 2000; LeFort *et al.*, 1998).

*Physical functioning* was assessed in four studies. In contrast to most previously presented outcomes, between- and within-group effects were largely consistent with all studies showing negligible to small effects (Barlow *et al.*, 2000; Burckhardt *et al.*, 1994; Heuts *et al.*, 2005; LeFort *et al.*, 1998).

*Self-efficacy* was assessed in ten trials. Again reported results varied greatly. Effects ranged from negligible (Heuts *et al.*, 2005) to above medium-size between- and within-group effects (LeFort *et al.*, 1998). Furthermore, small to medium between-group effects were reported in half of the studies (Barlow *et al.*, 2000; Burckhardt *et al.*, 1994; Fu *et al.*, 2003; Hopman-Rock & Westhoff, 2000; Lorig, González, & Ritter, 1999), whereas medium within-group ES were observed in four studies (Barlow *et al.*, 2000; Burckhardt *et al.*, 1994; Lorig, González,

& Ritter, 1999; Lorig *et al.*, 2003). It remains that between- as well as within-group effects varied considerably across studies.

Finally, the number of *visits to physician* was assessed in eight studies and again the effects were inconsistent across studies. Calculated ES ranged from some small decreases (Lorig *et al.*, 2003) to small increases in the number of visits (Lorig *et al.*, 1986). In contrast to other types of outcomes, the interpretability of this outcome is, however, difficult as it needs to be interpreted in relation to the quality of the disease-specific communication with the physician as well as subsequent emergency services use.

### *Discussion*

The systematic review of self-management interventions based on or similar to the Stanford curricula (Lorig *et al.*, 1985; Lorig, González, & Laurent, 1999) mainly included arthritis trials. Consequently, the outcomes reviewed in this section can be mostly compared with published meta-analyses on arthritis (see Section 1.2.3.1) that had generally shown negligible to small effects for *depression* (Mullen *et al.*, 1987), *disability* (Astin *et al.*, 2002; Mullen *et al.*, 1987; Warsi *et al.*, 2003) and *pain* (Astin *et al.*, 2002; Chodosh *et al.*, 2005; Mullen *et al.*, 1987; Warsi *et al.*, 2003), and somewhat larger effects for *self-efficacy* (Astin *et al.*, 2002). Although the present review did not calculate summary scores, results for the medians are similar to the summary scores of the meta-analyses with effects being generally small, while outcomes for *self-efficacy* were slightly larger. Therefore, this section essentially confirmed previously published reviews in that arthritis self-management interventions appear to have only small effects on participants. Given that variables such as *anxiety* and *effective communication with health professionals* are specifically targeted by the Stanford programs (Lorig, González, & Laurent, 1999), larger effects may have been anticipated. In contrast, the only outcome that showed clear benefits was *knowledge*. This outcome variable however is the only one that is not assessed by participant self-report but typically by knowledge tests.

Despite the above conclusions several aspects need to be considered when evaluating self-management programs. Firstly, this systematic review indicated inconsistent results across studies, rendering definite conclusions about program effectiveness on a range of outcomes such as depression or pain impossible. While summary scores of published meta-analyses obscure the results of individual studies, between-study comparisons in the present review indicated that single studies regularly differed by more than a medium-size ES (see Table 2). Taking into account that studies in this review predominantly assessed self-report outcomes, this range raises the question as to how reliably such outcomes can be measured (Schwartz & Rapkin, 2004). As discussed in Section 1.2.3.1, it is plausible that evaluations that rely on



participant self-report outcomes may not accurately reflect program impacts. Apart from potentially obscuring the magnitude of effects, observed inconsistencies across studies may be an additional indication that it is difficult to assess these outcomes reliably.

Secondly, further concerns arise in view of the types of outcomes that were assessed. As observed in a narrative review (Newman *et al.*, 2004), studies frequently assessed outcomes that are not particularly targeted by programs. While the Stanford curricula include topics on communication with the physician, emotions, and self-efficacy, it is questionable whether impacts on variables such as disability, fatigue, pain, or physical functioning can be expected as these outcomes are not specifically targeted by the Stanford protocols. In addition to the potential difficulty in measuring certain types of outcomes, it is essential that outcomes are assessed that match the objectives of the programs (see Section 1.2.2 on objectives of self-management programs as proposed by Osborne *et al.*, 2007). Further, while instruments to assess program outcomes were not specifically considered in this review, they differ in their relative sensitivity to measure change (Newman *et al.*, 2004). To make results comparable across studies, while also taking into account the objectives of specific self-management interventions, future research would be enhanced by the application of a mix of standard and program-specific outcome measures.

Thirdly, a further challenge of interpreting outcomes of self-management programs concerns the time frame in which outcomes can be expected to occur. The trials of the present review assessed post intervention outcomes ranging from direct post-assessment to several months after the intervention. It is beyond the scope of this thesis to consider this dimension further. Trials that are concerned with the effectiveness of self-management interventions, however, should take the dimension 'time' into account as different types of outcomes can be expected to occur at different time points. The program logic model of impacts of self-management programs presented in Section 1.2.2 can serve as a guide to arrange outcomes in terms of short-, medium-, and long-term effects (Osborne *et al.*, 2004).

Given that published reviews on self-management programs suggest generally small impacts on participants and the systematic review of Stanford courses show small, albeit inconsistent effects, several possible explanations were provided. While not all of the above aspects can be considered in this literature review, the inconsistency in outcomes across trials appeared to be most critical to understand the measurement of outcomes of self-management courses. As both Sections 1.2.3.1 and 1.2.3.2 suggested that a more detailed investigation of the types of reported outcomes may be useful, the following Section 1.2.3.3 is aimed at applying a systematic approach to categorise these outcomes.

### 1.2.3.3 *Outcomes on which evaluations were based and how these were measured*

Sections 1.2.3.1 and 1.2.3.2 provided reviews of current evidence about the effectiveness of self-management interventions. While the meta-analyses (see Section 1.2.3.1) showed some benefits for individuals with diabetes and hypertension, smaller and inconsistent effects were found for people with arthritis (see Sections 1.2.3.1 and 1.2.3.2). However, given that the latter trials typically rely on self-report outcomes and the former frequently measure clinically assessed outcomes, it is also possible that current evidence is not related to the disease group but rather to the types of outcomes that are evaluated. This section was therefore aimed at categorising the different types of outcomes that are typically assessed in self-management trials. To facilitate this exercise a quality of life appraisal model was used, where 'appraisal' denotes the cognitive processes carried out by the respondents when answering a question (Schwartz & Rapkin, 2004).

#### *Performance-, perception-, and evaluation-based measures*

In a recent article Schwartz and Rapkin (2004) suggested categorising outcomes that are commonly assessed in health research according to the level of cognitive appraisal involved in people's response processes. This model provides new insight into quality of life research as it questions whether current measurement models are applicable to those types of outcome measures that require cognitive appraisal (Krosnick, 1999; Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988). They proposed the following categorisation: *performance-, perception- and evaluation-based measures* (Schwartz & Rapkin, 2004).

Measurement models that are operationalised using statistical methods such as factor analysis or item response theory are built on the assumption that observed scores provide information about respondents' position on the scale of the underlying construct of interest. This understanding of the measurement process is assumed to be applicable to outcomes described as *performance-based measures*. While the assessment of people's performance may still be confounded by biases such as test anxiety or cheating, it is assumed that the interpretation of items of this type of measure is unequivocal, i.e. it is assumed that the interpretation of the items is stable across people and occasions. A timed walk of a specific distance is an example of these measures (Schwartz & Rapkin, 2004).

In a similar manner to the previous types of outcomes, the interpretation of an item that is considered a *perception-based measure* is assumed to be stable across people as well as occasions. However, in contrast to the previous type of outcome measures perception-based measures involve personal judgement in the response process. Hence, responses to such measures are dependent upon individuals who might consciously or unconsciously edit their

responses before providing an answer. For example, individuals who are asked to provide a judgement on how frequently they perform a certain exercise might be inclined to edit their final response in a socially desirable way. Hence, they distort their true levels by providing socially desirable answers. In spite of these potential biases, it remains that these types of measures are expected to converge across persons and occasions (Schwartz & Rapkin, 2004).

*Evaluation-based measures* are assumed to be intrinsic to the construct being measured. It is assumed that the interpretation of these items is unstable across people as well as across occasions. Persons, for example, who are asked to provide a judgment of their current level of pain, engage in cognitive appraisal processes when attending to such questions (Schwartz & Rapkin, 2004). These appraisal processes have been described to consist of the following four steps: 1) interpretation of the question, 2) retrieval of information relevant to answering the question, 3) processing of relevant information to make a judgment, and 4) formulation of the answer (Krosnick, 1999; Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988). Given that such cognitive appraisal processes are likely to differ across persons and/or occasions, it is evident that scores obtained from evaluation-based measures may be invalid and unreliable. Intra- as well as inter-person differences in any of the steps of this appraisal are conceivable, rendering comparisons over time and across people problematic (Schwartz & Rapkin, 2004).

#### *The application of performance-, perception-, and evaluation-based measures to outcomes of self-management interventions*

Through the use of the distinction between outcome measures (Schwartz & Rapkin, 2004), the frequently reported outcomes of self-management trials were categorised. The results of this exercise are presented in Table 3. This table has been arranged so that the results of intervention and control group from Table 2 have been classified by the category of outcome measure used (*performance-, perception-, and evaluation-based measures*). Given that this exercise was aimed at comparing respective magnitudes of effects between intervention and control group, Cohen's *d* is not repeated as it was not relevant for this exercise.

Of the outcomes that were presented in Sections 1.2.3.1 and 1.2.3.2, *clinical outcomes* and *knowledge* were most akin to the definition of *performance-based measures*. While it is assumed that no judgement is involved in measuring these outcomes, results may still be influenced by, for example, food or fluid intakes prior to blood tests or test anxiety. Given that clinical outcomes were assessed infrequently, only *knowledge* is discussed. When revisiting the review in Section 1.2.3.2, all but a lay-led self-management program (Lorig *et al.*, 1986)

suggested large effects in intervention subjects (Goepfinger *et al.*, 1989; Hopman-Rock & Westhoff, 2000; Lorig *et al.*, 1986; Lorig *et al.*, 1989). While there was some variability across trials, large differences between intervention and control subjects in their respective minima, medians, and maxima suggest that programs had an impact on this outcome and that it was possible to measure effects with the instruments that were used (Schwartz & Rapkin, 2004). For the purpose of the present exercise the results presented in Table 3 excluded the lay-led course as it was an outlier in the context of the other studies. Results indicate large program effects with differences in intervention group subjects across trials being relatively small (range=0.45) when considered in the context of the overall magnitude of effects.

Although *perception-based measures* involve some level of judgement, the interpretation of questionnaire items is assumed to be largely stable across persons and occasions (Schwartz & Rapkin, 2004). The variables *communication with the physician*, *physical functioning*, and *visits to the physician* were considered measures of this type. In view of the magnitude of effects for these outcomes (see Section 1.2.3.2), the results were inconsistent across trials. Effects ranged from negligible to small positive effects for intervention and control subjects in the variables *physical functioning* and *visits to the physician*, whereas up to medium effects were observed for *communication with the physician* in the intervention group. Hence, in contrast to *knowledge*, the range in scores of both intervention and control group subjects was relatively large in the context of the overall magnitude of effects. With ES ranging from 0.29 to 0.45 for intervention subjects and from 0.15 to 0.20 for control subjects, this suggests that perception-based measures may be less reliable and more difficult to assess compared with performance-based measures. As described previously, bias such as social desirability may have confounded results.

Finally, *evaluation-based measures* require the highest level of personal appraisal and as a consequence are assumed to have low stability across persons and occasions (Schwartz & Rapkin, 2004). Of the assessed outcomes presented in Sections 1.2.3.1 and 1.2.3.2, *anxiety*, *depression*, *disability*, *fatigue*, *health status*, *pain*, and *self-efficacy* were allocated to this type of measures. As presented in Table 3, intervention and control subjects showed inconsistent results across studies. This inconsistency was particularly striking across the variables *self-efficacy*, *disability*, *pain*, and *depression* with differences being of at least medium-size ES (range=0.50-1.48 for intervention subjects; range=0.42-1.14 for control subjects). Given that control subjects are expected to be largely stable across occasions as they did not receive an intervention, the variability in scores is substantial in these outcomes. In spite of these noticeable inconsistencies, effects on outcomes such as *anxiety*, *fatigue*, and *health status* were still comparatively inconsistent in view of overall small to medium effects across groups (range=0.20-0.63 for intervention subjects; range=0.23-0.35 for control subjects).

**Table 3** Effect sizes of most frequently assessed outcomes in studies included in the systematic review of Section 1.2.3.2 grouped by performance-, perception-, and evaluation-based measures (Schwartz & Rapkin, 2004)

Assessed outcome		n	Minimum	Median	Maximum	Range
Performance-based measures						
Knowledge	IG	4	0.83	0.98	1.28	0.45
	CG		0.04	0.14	0.42	0.38
Perception-based measures						
Visits phys.	IG	8	-0.11	0.13	0.18	0.29
	CG		0.02	0.12	0.21	0.19
Comm. phys.	IG	5	0.04	0.23	0.49	0.45
	CG		0.03	0.10	0.23	0.20
Phys. funct.	IG	4	-0.06	0.06	0.26	0.32
	CG		-0.12	-0.01	0.03	0.15
Evaluation-based measures						
General health	IG	7	-0.11	0.17	0.52	0.63
	CG		-0.19	0.04	0.16	0.35
Fatigue	IG	6	0.12	0.18	0.40	0.28
	CG		-0.10	0.03	0.13	0.23
Self-efficacy	IG	10	0.02	0.40	0.64	0.62
	CG		-0.34	0.01	0.27	0.61
Disability	IG	13	-0.20	0.15	1.28	1.48
	CG		-0.43	-0.02	0.31	0.74
Pain	IG	18	-0.01	0.20	0.75	0.76
	CG		-0.34	0.11	0.80	1.14
Anxiety	IG	5	0.11	0.21	0.31	0.20
	CG		0.00	0.08	0.25	0.25
Depression	IG	11	0.07	0.21	0.57	0.50
	CG		-0.06	0.04	0.36	0.42

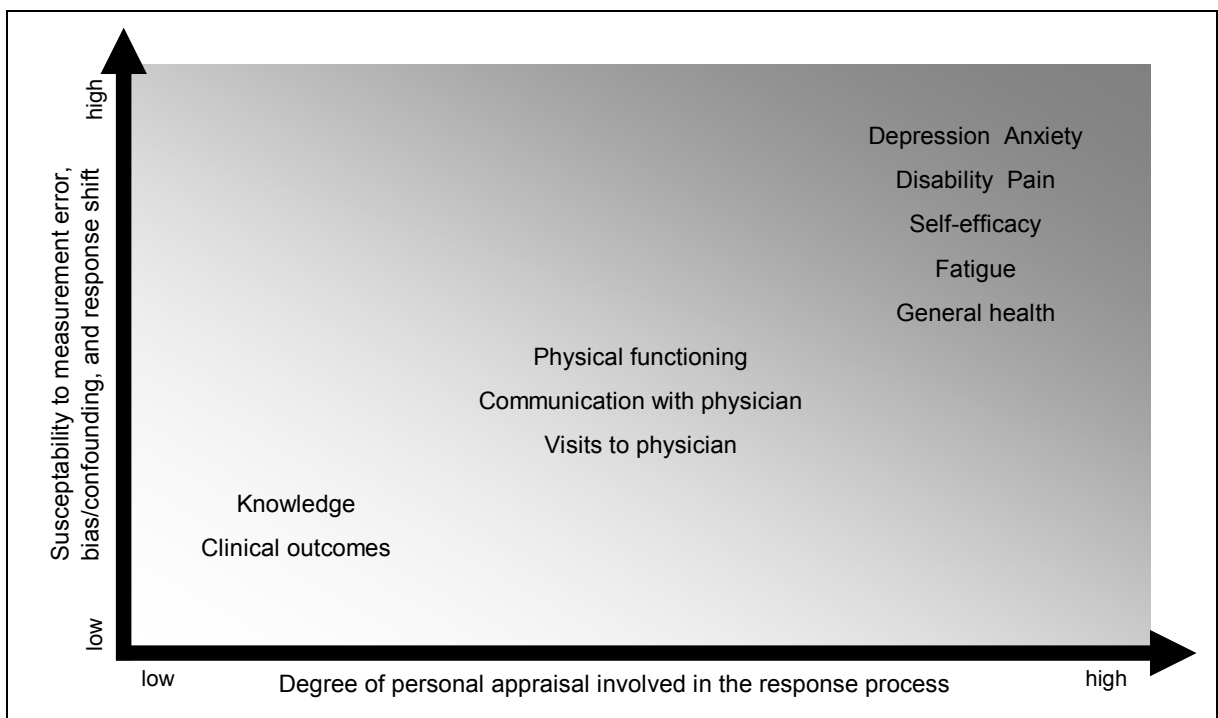
Legend

CG:	Effect size, control group
Comm. phys.:	Communication with physician
IG:	Effect size, intervention group
n:	Number of studies included in the systematic review
Phys. funct:	Physical functioning
Visits phys.:	Visits to physician

*Discussion*

In summary, *performance-based measures* showed some large effects for treatment subjects with relatively small inconsistencies in reported outcomes. In contrast, larger inconsistencies in results of intervention and control subjects were found for both *perception-* and *evaluation-based measures*. Although results of these measures tended to be in the expected direction, i.e. intervention subjects indicated larger effects than control subjects in respective minima,

medians and maxima, the inconsistency in results alluded to potential difficulties in assessing both types of outcomes. Hence, while the model of classifying outcomes as performance-, perception- and evaluation-based measures appears to be a useful way of approaching the measurement of self-report outcomes (Schwartz & Rapkin, 2004), grouping outcomes into three discrete categories was not possible in the current study. As a result, it is suggested that the different types of outcome measures be conceptualised on a continuum, described as the 'degree of appraisal involved in the response process'. As shown in Figure 7, with increasing degree of cognitive appraisal (x-axis) it is assumed that outcomes are susceptible to an increased risk of measurement error and/or bias (y-axis). The transition from the appraisal model to the refined model is facilitated by presenting the discussed outcomes in three groups, with those outcomes presented in the centre and to the right of the x-axis, i.e. the 'evaluative pole' of the x-axis, being similarly difficult to measure.

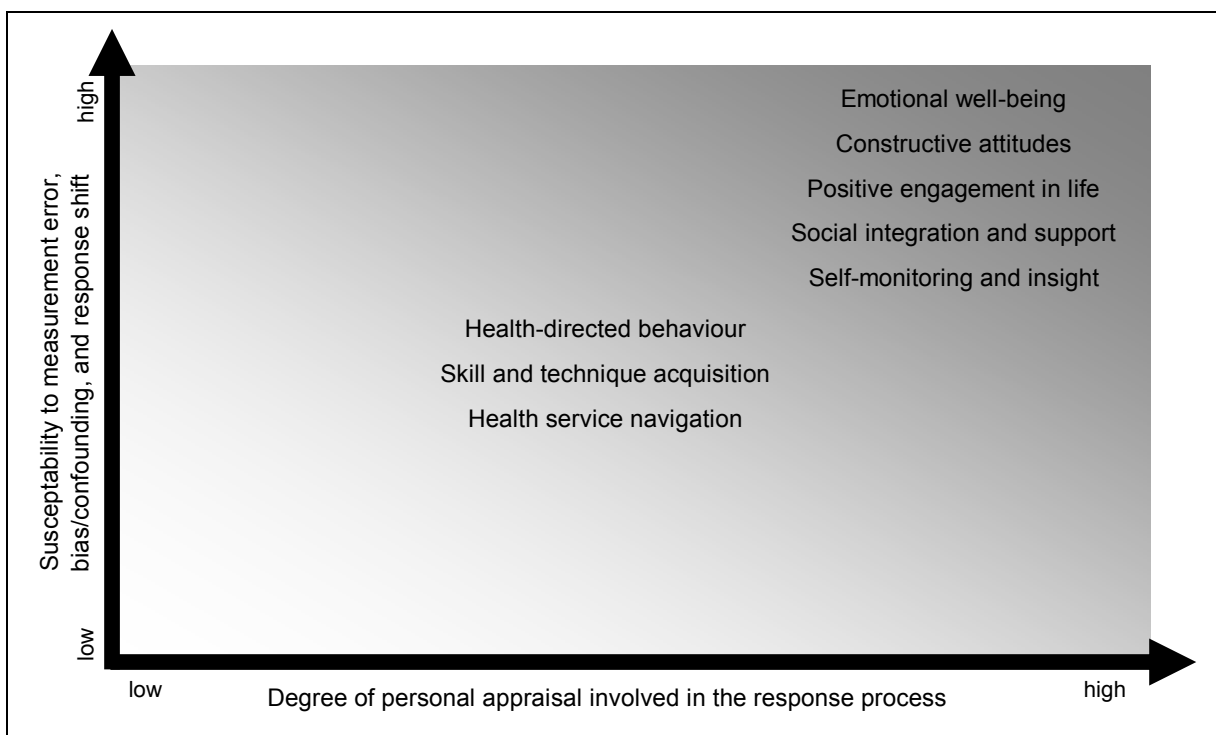


**Figure 7** Degree of personal appraisal involved in the response process across outcomes that were frequently assessed in self-management interventions; modified from Schwartz & Rapkin (2004)

Apart from differentiating between the different types of outcome measures, a further aspect in evaluations of self-management interventions relates to the types of outcomes used to assess the impact of these programs (see Section 1.2.3.2). Assuming that the measurement instruments that were applied validly capture the construct of interest (Brady, 1997), the selection of outcomes should be driven by the specific objectives of the intervention. While it is problematic if variables that are deemed important program outcomes are omitted in the

evaluation process, it is similarly problematic if variables are assessed that are not targeted by the program as was observed to be the case in some published self-management trials (Newman *et al.*, 2004).

As introduced in Section 1.2.2, Osborne and colleagues carried out various exercises to elicit the main objectives of self-management interventions. The project resulted in the definition of eight dimensions that are considered important areas in which programs are expected to impact (Osborne *et al.*, 2007). While some of these are similar to outcomes that have been frequently assessed, *clinical outcomes, disability, fatigue, general health, knowledge, pain, physical function, and physician visits* were not deemed key indicators of program impacts. Given that these new constructs play a central role in later parts of the present thesis, they are illustrated in Figure 8.



**Figure 8** Degree of personal appraisal involved in the response process – illustrating the eight areas on which self-management programs are expected to impact (Osborne *et al.*, 2007)

Given that these dimensions play a central role in the remainder of the thesis, this final exercise was aimed at exploring which of the outcomes that have been typically assessed in self-management trials correspond to the eight dimensions. *Health-Directed Behaviour* is a global concept of variables such as exercise, relaxation, and self-management behaviours. *Skill and Technique Acquisition* is in part a representation of cognitive symptom management and communication skills, with the latter outcome also being partially captured by *Health*

*Service Navigation* when related to communication with the physician. *Emotional Well-Being* subsumes anxiety and depression, whereas *Positive and Active Engagement in Life* and *Constructive Attitudes and Approaches* are not represented in outcomes that were frequently reported. At most, some aspects of self-efficacy such as participants developing more confidence in their capabilities to achieve a positive health outcome (Bandura, 1997) would be related to the latter dimension. *Self-Monitoring and Insight* may be matched with the ability to respond to changes in health through techniques such as cognitive symptom management and finally *Social Integration and Support* does not seem to relate to outcomes that have been assessed in previous self-management studies.

#### 1.2.3.4 Summary

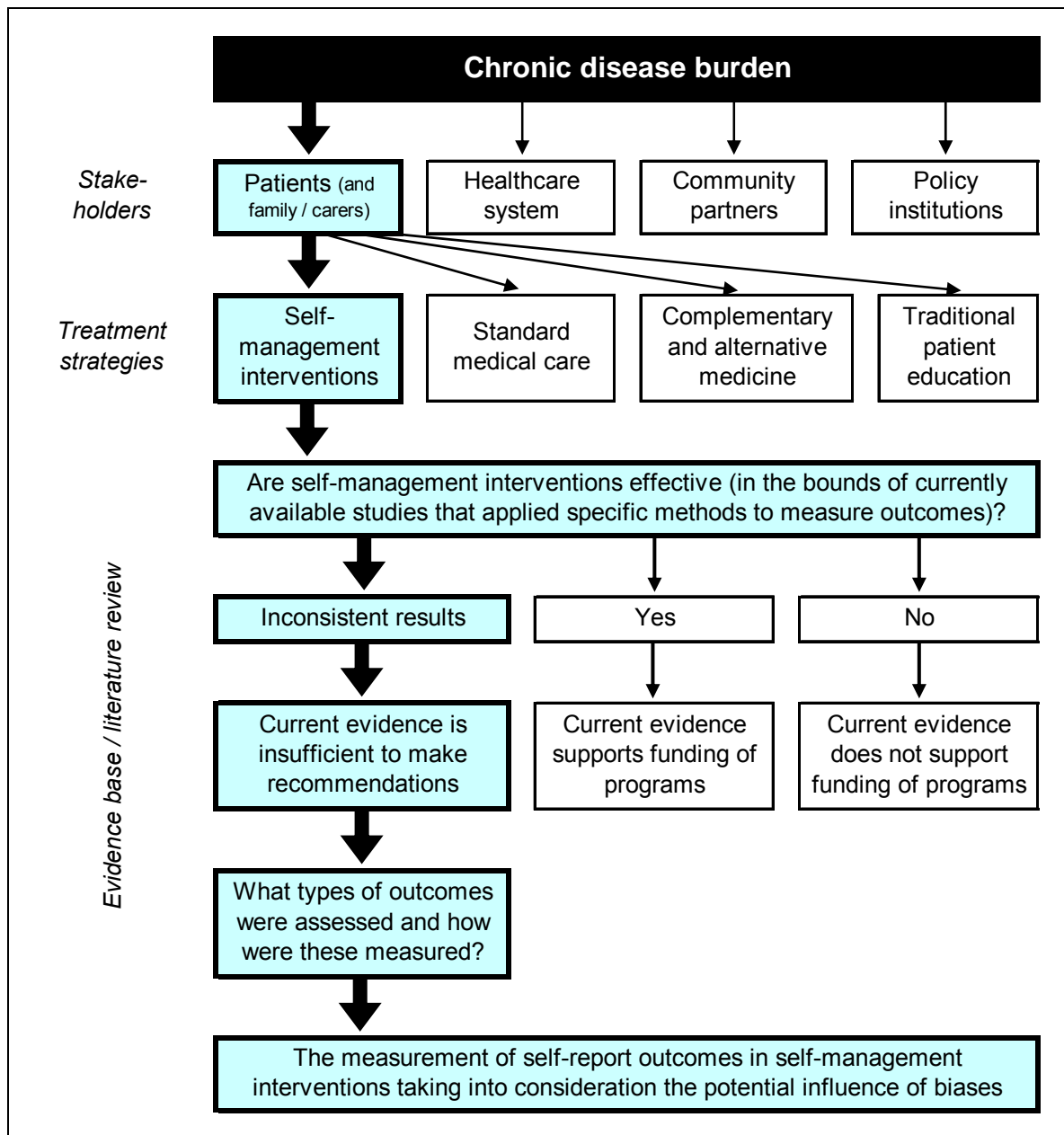
Sections 1.2.3.1 and 1.2.3.2 provided a review of the literature regarding current evidence about the effectiveness of self-management interventions. While some inconsistent results were observed, these were further examined in view of the characteristics of the types of outcomes that were assessed (see Section 1.2.3.3). It was concluded that with increasing degree of appraisal involved in the response process, outcomes are exposed to an increased risk of measurement error, confounding and bias. Apart from measurement aspects it was further observed that some frequently assessed outcomes did not match the objectives of self-management programs. While a new instrument has been developed to circumvent the latter issue (Osborne *et al.*, 2007), it remains that evaluations of self-management programs commonly rely on self-report outcomes that are complex to measure. The following Section 1.2.4 was therefore aimed at exploring issues related to the measurement of outcomes of self-management programs with particular focus on biases that are commonly encountered in self-report outcome measures.

### 1.2.4 The measurement of outcomes of self-management interventions

To facilitate the navigation through the literature review, the findings of the previous three sections and an overview of the final section of this chapter are illustrated in Figure 9. As introduced in Figure 3, the first two sections of this literature review provided an overview of the different stakeholders involved in chronic disease care as well as different treatment strategies to respond to the increasing burden of chronic disease. In view of the importance of involving individuals in chronic disease management, self-management was introduced as a key component of chronic care models. Given that such programs are the focus of this thesis, a comprehensive review of current evidence regarding program effectiveness was provided (see Sections 1.2.3.1 and 1.2.3.2). While some small positive effects were found,



self-report outcomes showed inconsistent results. Hence, when following the flow chart in Figure 9, the question about program effectiveness cannot be answered satisfactorily in view of current evidence. After investigating the different types of outcomes on which trials were based (see Section 1.2.3.3), it was concluded that the inconsistency in results may relate to the type of outcome, i.e. self-report outcomes, rather than specific disease groups.



**Figure 9** Flow chart of the content of the thesis, Part II

In view of these findings, the present section was aimed at reviewing issues pertaining to the measurement of self-report outcomes. This was approached by reviewing the different types of research designs that can be applied to assess program outcomes (Campbell & Stanley,

1963; Cook & Campbell, 1979). The remainder of the section focuses on the measurement of change. This includes general concepts as they apply to the present thesis and a review of biases that can be encountered in self-report outcomes with a specific focus on response shift and social desirability biases.

#### 1.2.4.1 *Research designs*

The measurement of outcomes of self-management programs is generally based on change scores derived from the comparison of participants' scores assessed before the intervention (=pretest) with scores assessed after the intervention (=posttest). To rule out the possibility that observed change scores in intervention subjects occurred for reasons other than the intervention, non-intervention control subjects are included in many studies. While this type of research design is frequently used in self-management trials, several alternative designs exist that shall be introduced hereafter.

One of the most comprehensive and frequently cited overviews of different research designs was provided by Campbell and Stanley (1963) who discussed these designs in the context of their internal and external validity. The main threats to internal validity of change scores can be summarised in the following way:

- (1) *History*, i.e. experiences that are gained between pretest and posttest measurement that are unrelated to the intervention.
- (2) *Maturation*, i.e. changes in treatment subjects that occurred between pretest and posttest as a result of passage of time; these are again unrelated to the intervention.
- (3) *Testing*, i.e. the influence of experiences that were gained in the pretest situation on the posttest results.
- (4) *Instrumentation*, i.e. a change in the observers, who alter the way they ask questions or extract answers through increased familiarity with the interview process, or a change in the calibration of a measurement instrument may cause changes in obtained scores.
- (5) *Statistical regression*, i.e. regression to the mean with extreme scores being closer to the mean at posttest than at pretest (Bland & Altman, 1994a, 1994b).
- (6) *Selection*, i.e. biases can occur as a result of non-random assignment to the comparison groups.
- (7) *Experimental mortality / attrition*, i.e. loss of intervention and control group subjects in a way that the final samples consist of groups of people that are not comparable anymore.
- (8) *Selection-maturation interaction*, i.e. if subjects were allocated in a non-random fashion to the comparison groups, the extent of maturation may differ across groups and may be mistaken for the influence of the treatment.

Further, the main threats to external validity are as follows (Campbell & Stanley, 1963):

- (1) *Interaction effect of pretesting*, i.e. the effects of pretesting described above may further result in study subjects being influenced in a way that resulting posttests would not be comparable with hypothetical posttest scores from the universe.
- (2) *Interaction effects of selection biases and the intervention*.
- (3) *Reactive effects of experimental arrangements* across comparison groups.
- (4) *Multiple-treatment interference*, i.e. if multiple interventions are carried out on the same subjects, earlier interventions are likely to have influences on later interventions.

While the above threats to the internal and external validity of measuring change cannot be ruled out as alternative explanations for observed variations in scores, different approaches to research designs are assumed to minimise some of these threats. The widely accepted distinction between these design approaches is the differentiation between pre-experimental, quasi-experimental and true experimental designs, with the latter being arguably superior to the other research designs as subjects are randomly allocated to the treatment or a control condition. Randomisation has been argued to be a viable way to rule out most threats to internal validity with the exception of *experimental mortality* (Campbell & Stanley, 1963).

Despite wide acceptance of this hierarchy of research designs, later publications appraise true experimental designs more critically with regard to their feasibility in field research. Apart from situations in which random allocation of subjects may not be feasible, desirable or ethical, it is recommended to design experiments in a way that these can still be used as quasi-experimental designs in case the randomisation breaks down (Cook & Campbell, 1979). It is also highlighted that even perfectly designed experiments are still not able to rule out all threats to internal validity, with the most important ones being: a) reactions of subjects dissatisfied with being allocated to a specific condition, b) Hawthorne/placebo effects in control group subjects (Sommer, 1968), and c) experimental mortality. Despite this critique the authors maintained that randomisation is still superior to other designs as it minimises the number of assumptions on which the findings need to be based (Cook & Campbell, 1979).

Given that true experimental designs are relevant in a later part of this section and because of their argued superiority to other research designs, these designs are introduced briefly. As shown in Figure 10, three true experimental designs exist: 1) *pretest-posttest control group design*, 2) *posttest-only control group design*, and 3) *Solomon four-group design*. In Figure 10 these designs are presented in a way that a ✓ at pretest and/or posttest means that data are collected at the respective occasion, and a ✓ at intervention means that subjects receive the treatment which consequently only applies to intervention groups (IG 1, 2, 3a, 3b).

The *pretest-posttest control group design* is a true experimental design that is frequently applied. In this type of research design subjects are randomly allocated to the intervention or a control group, with the latter being another type of intervention, a placebo-control condition, a waiting list control condition, or no intervention. While both groups provide data at pretest and at posttest (see Figure 10), it is assumed that most threats to internal validity can be accounted for by subtracting the computed change of the control group from the computed change of the intervention group. This type of design can also be extended to blinding the observers and subjects to the different treatment conditions (Campbell & Stanley, 1963).

	Research design	Group	Pretest	Intervention	Posttest
1.	Pretest-posttest control group	IG 1	✓	✓	✓
		CG 1	✓		✓
2.	Posttest-only control group	IG 2		✓	✓
		CG 2			✓
3.	Solomon four-group	IG 3a	✓	✓	✓
		CG 3a	✓		✓
		IG 3b		✓	✓
		CG 3b			✓

Legend

IG: Intervention group

CG: Control group

✓ Providing pretest and/or posttest data; receiving the intervention

**Figure 10** True experimental designs (Campbell & Stanley, 1963)

In contrast, the *posttest-only control group design* is based on the assumption that subjects that are randomly allocated to the comparison groups are identical at pretest. The collection of pretest data is not only assumed unnecessary as differences between the groups' posttest scores should reflect the impact of the intervention but this design further circumvents the threat to internal validity through a potential pretesting effect (Campbell & Stanley, 1963).

The *Solomon four-group design* is a combination of the two previous designs. Subjects are randomly allocated to one of four groups with one pair of experimental (IG 3a) and control group (CG 3a) providing pretest and posttest data, and the other pair providing posttest data only (IG 3b and CG 3b). Given that this type of design is able to quantify the potential pretesting effect, it is considered the strongest of true experimental research designs as it accounts for the potential threat that pretesting may have on internal and external validity (Campbell & Stanley, 1963).

### 1.2.4.2 The measurement of change

The measurement of change as well as the interpretation of change scores is an area that has been discussed extensively in the literature. A brief overview of this topic, which includes a definition of change as it pertains to the present thesis, is provided hereafter. More details on the specific methods that were used to measure change in this thesis follow in each of the analysis chapters (see Chapters 3 and 5 in particular).

The most common conceptualisation of change refers to a computed difference score that is derived from comparing scores that were assessed at two different points in time (Nunnally & Bernstein, 1994). Notwithstanding its frequent application, such a difference score however entails several sources of potential error apart from those that were introduced in the context of research designs (see Section 1.2.4.1). That is, the different types of measures that are used to compute change scores are susceptible to measurement error and confounding to varying degrees which may render an unambiguous interpretation of the results problematic. Hence, when measuring a construct, it is likely that a specific portion of its variance relates to something other than the content of the construct (Podsakoff *et al.*, 2003). While this portion of unexplained variance has been found to vary across measures and research areas, it can account for more than 40% of the total variance (Cote & Buckley, 1987). The larger the portion of unexplained variance the more problematic a) the interpretation of change and b) the interpretation of the relationship between different constructs, as there may be alternative explanations for why an association was found (Podsakoff *et al.*, 2003).

Observed change ( $\Delta_{\text{obs}}$ ) can be expressed as a function of variance due to true change ( $\Delta_{\text{true}}$ ) plus error variance ( $\sigma_{\text{error}}$ ):

$$\Delta_{\text{obs}} = \Delta_{\text{true}} + \sigma_{\text{error}}$$

Given that this equation is a simplistic representation of  $\Delta_{\text{obs}}$ , further details are discussed briefly. Firstly, the above equation is simplified in that it represents the difference score of a posttest minus a pretest score each of which is assumed to be measured with error. Given that change is not assessed directly but derived from comparing two measures, it would be more precise to illustrate  $\Delta_{\text{obs}}$  as a function of two observed scores plus error. While each score is measured with error, it can be assumed that the error of a variable correlates over time, i.e. in repeated measures it is expected that the errors of the pretest correlate with the respective errors of the posttest (Jöreskog & Sörbom, 1996-2001). Given that these concepts add more complexity to the presentation of change, it shall suffice to illustrate  $\Delta_{\text{obs}}$  as above for the purpose of this overview.

Secondly, following from the previous Section 1.2.4.1, the component  $\Delta_{\text{true}}$  of  $\Delta_{\text{obs}}$  can further be understood as a combination of true change due to an intervention plus true change due

to, for example, history, maturation, and other events that are unrelated to the intervention (Campbell & Stanley, 1963). While change unrelated to the intervention poses a threat to the validity of assessing the true effect of treatments – which can partially be prevented through the application of appropriate research designs (see Section 1.2.4.1) – it is assumed that change that is not attributable to the intervention is still related to the construct that is being measured. Given that this thesis was only concerned with issues related to components of the variance that do not refer to the content of the construct, this distinction was not of further concern in this research.

Thirdly, it is generally accepted that  $\sigma_{\text{error}}$  consists of two components: random and specific error variance (Child, 1990; Nunnally & Bernstein, 1994). While both are problematic in the assessment of scores, the specific component can have more serious consequences as it may introduce systematic errors in observed scores that are unrelated to the construct being measured (Podsakoff *et al.*, 2003). In the context of the present thesis, this specific error is defined as variance due to confounding or bias that is neither due to random error nor due to true change on the construct of interest. Following from this distinction between random and specific error, the definition of change as presented on page 37 is extended to equation:

$$\Delta_{\text{obs}} = \Delta_{\text{true}} + (\sigma_{\text{random measurement error}} + \sigma_{\text{specific error due to bias}})$$

In the context of the validity of deriving a measure of change from scores assessed over time, it remains to be acknowledged that – as described before – error components are expected to correlate in repeated measures (Jöreskog & Sörbom, 1996-2001). Hence, while specific error is problematic for the validity of single scores, it is less problematic in the assessment of change, provided that the specific error component of the above equation is constant. The problem arises when the specific error is not constant across measurement occasions. That is, specific error then becomes a threat to the validity of change scores as the bias contained in single scores does not cancel out over time.

In sum, there are several potential problems with regard to the measurement of change and its unambiguous interpretation. While random error is problematic, specific error that is not constant over time poses an even more serious threat to the validity of change scores as it provides a rival explanation for why an association between different constructs was found (Podsakoff *et al.*, 2003). As introduced in Section 1.2.3.3, in particular outcomes that require appraisal are susceptible to such bias which may be an explanation for why inconsistencies in the results in Sections 1.2.3.1 and 1.2.3.2 were found. Depending on the nature and the magnitude of specific error, this component of  $\Delta_{\text{obs}}$  needs to be considered when measuring outcomes of self-management interventions. The remainder of this section is concerned with a review of biases that are common in self-report outcomes.

### 1.2.4.3 *Confounding and bias in the measurement of outcomes of interventions*

The present section was aimed at providing an overview of biases encountered in self-report outcomes. Given that a wide range of such biases has been identified in the literature, this section is limited to a brief description of the most prominent of these specific errors. At first, some general definitions of terms that are used throughout the remainder of the thesis are provided.

In the context of the thesis ‘confounders’ and ‘biases’ are defined as a portion of the variance of a score that is attributable to specific errors. Formally the two terms can be distinguished. ‘Confounding’ is the distortion of effects resulting from an insufficient control of measurement error, whereas ‘bias’ is a systematic deviation of results from ‘true’ scores because of errors in planning and execution of trials (Sachs, 2002). Notwithstanding this distinction researchers use the term ‘bias’ more frequently, even in situations where ‘confounding’ would be more appropriate. To avoid any ambiguity in the present research only the term ‘bias’ is used, with ‘bias’ representing the part of the variance that is independent of the content of the measured construct and that differs from random measurement error.

Further, ‘response bias’, ‘response style’ and ‘response set’ are generally used in the context of specific measurement errors. While ‘response bias’ is used as a generic term to describe respondents’ tendency to attend to items in a systematic way that is unrelated to the content of the items (Cronbach, 1946; Paulhus, 1991), the distinction between ‘response style’ and ‘response set’ is less clear. While both are specific forms of ‘response bias’, they have been used in different contexts (Paulhus, 1991; Rorer, 1965). Again to avoid ambiguity ‘response style’ is used throughout the remainder of the present thesis to describe individuals’ tendency to exhibit a certain style when responding to an item of a questionnaire that is unrelated to the item content (Baumgartner & Steenkamp, 2001).

A large amount of research into distortion of scores due to biases exists and a selection of biases that were deemed important in the context of this thesis is presented hereafter.<sup>6</sup> While all of these may contaminate data in a way that a given item not only assesses the item content but additionally something that is independent of the item of interest (Cronbach, 1946), the list consists of a range of biases that are different in nature. While some refer to intentional or unintentional response styles such as acquiescence or social desirability bias, other forms of bias may be a result of conscious or unconscious psychological processes such as Hawthorne effect or response shift.

---

<sup>6</sup> For a comprehensive review of common method biases and recommendations on how to deal with them statistically and/or through procedural remedies the reader may refer to Podsakoff *et al.* (2003).

The following list provides a brief introduction to these biases including further explanations where necessary:

- *Acquiescence* describes the tendency to provide confirming responses and to agree with items irrespective of their content (Cronbach, 1946). Some authors differentiate between 'agreement acquiescence' which describes the tendency to agree to an item regardless of its wording and 'acceptance acquiescence' which is the tendency to endorse an item regardless of its content (Morf & Jackson, 1972; Paulhus, 1991). Several methods exist to control for this prominent bias (Billiet & McClendon, 2000; Hofstee *et al.*, 1998; Lentz, 1938; Paulhus, 1991). Despite being recognised for its importance (Jackson & Messick, 1961), there is little consensus regarding the severity of the influence of acquiescence bias (Ferrando *et al.*, 2003). Some suggest that questionnaires need to be corrected for acquiescence variance (Hofstee *et al.*, 1998), while others deem its influence negligible (Rorer, 1965; Rorer & Goldberg, 1965);
- *Consistency motif* is the tendency of respondents to try to be consistent across a set of similar items (Podsakoff *et al.*, 2003);
- *Effort justification* describes a situation in which respondents may feel that they did not receive positive effects from participating in an intervention. In view of the time and effort they invested into their participation (Howard, Ralph *et al.*, 1979), this perception however may result in cognitive dissonance (Aronson & Mills, 1959; Festinger, 1957; Hill & Betz, 2005). Consequently, they adjust their scores to avoid this cognitive conflict;
- *End-aversion / central tendency bias* describes the propensity to avoid the extremes of a response scale (Choi & Pak, 2005; Podsakoff *et al.*, 2003);
- *Extremity bias* is the opposite of the previous bias, i.e. it describes the propensity to make strong and determined rather than weak or indecisive statements (Paulhus, 1991);
- *Faking bad* is the tendency of respondents to try to appear worse than they are, i.e. they may try to appear sicker to be chosen to receive a treatment (Choi & Pak, 2005);
- *Gambling* is the propensity to always provide a response, i.e. if in doubt people choose to give any answer as opposed to not providing a response (Cronbach, 1946);
- *Halo effect* describes a situation in which an impression of a person/situation carries over to the rating of other areas, i.e. later ratings would either be confounded by the first rating or by an overall evaluation of a person/situation (Nunnally & Bernstein, 1994);



- *Hawthorne / placebo effect* describes a situation in which control subjects of a trial show effects that are due to factors other than the treatment that was intended to be evaluated (Franke, 1979; Franke & Kaul, 1978; Sommer, 1968);
- When individuals engage in an *implicit theory of stability or change* when answering questions, they infer their response from a comparison of their current state with a past state. For example, if they think that they received benefits from a treatment, they provide their answer in a way that a positive difference is shown (Ross, 1989; Schwarz *et al.*, 1998). This theory only applies when people are asked to provide retrospective ratings in addition to their 'now'-ratings, i.e. people use their posttests as a benchmark to construct their levels at pretest. In the context of this theory it is highlighted that respondents rarely remember their past states but infer from the present (Norman, 2003; Ross, 1989).
- *Positive / negative affectivity* refers to the respective trait of a respondent, i.e. people who are high on positive/negative affectivity tend to generally provide more positive/negative answers (Podsakoff *et al.*, 2003);
- As previously introduced in the context of internal threats to validity (Campbell & Stanley, 1963), the *pretesting effect* describes the influence of a pretest which may itself induce change rather than the treatment/intervention (Webb *et al.*, 1966);
- *Recall bias / memory effect* applies to research situations in which there is either a long time gap between repeated measures or in retrospective assessments. While distorted memories may be interpreted as an important adjustment to certain situations or states, it poses a threat to the accuracy of scores on past states and/or inferring change or stability (Loftus *et al.*, 1991; Pearson *et al.*, 1992; Ross, 1989);
- *Response shift* is a situation in which respondents change their perspective as a result of a treatment or an intervention (Howard & Dailey, 1979; Sprangers, 1989). This type of bias is of particular importance in the evaluation of self-management courses as it has been found to occur in up to 70% of course participants (Osborne *et al.*, 2006). Given its importance in the context of this thesis, response shift bias is described in more detail in Section 1.2.4.4;
- *Role selection* results from subjects who take on a certain role as a result of being part of an experiment, i.e. their perceived role expectations may not or only partially reflect who they are or how they would behave outside the research setting (Webb *et al.*, 1966);
- As introduced in Section 1.2.3.3, it can be assumed that with increasing difficulty of a question, respondents need to engage in a cognitive response process (Krosnick, 1999; Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988). If people fully engage in such

process this is commonly referred to as optimising (Krosnick & Alwin, 1987). In situations such as questionnaires being too long, questions too difficult, and/or response options being too extensive, people might become fatigued or de-motivated. Consequently, they might engage in *satisficing*, i.e. they either do not engage in all four steps of the response process or they do not execute any of the steps but search for a cue in the question that helps them choose an answer (Holbrook *et al.*, 2003; Krosnick, 1999);

- *Social desirability* is another prominent bias which has received frequent attention in the literature (Loevinger, 1959; Paulhus, 1991). This response style reflects the propensity of people to endorse items that reflect socially desirable traits and/or behaviours (Nunnally & Bernstein, 1994). Apart from its general popularity across research areas, this bias may play a particularly important role in the measurement of outcomes of self-management interventions. Firstly, in health-related outcomes it is likely that individuals are inclined to present themselves or certain health-behaviours in a more positive light. Secondly, in evaluations of self-management programs, participants often fill out the questionnaires in the presence of the course leaders and this may trigger socially desirable responses as the respondents may be aware that they are indirectly evaluating the performance of the leader. As a result, some respondents might provide answers with the intention to please the course leader. In view of the prominence of this bias and its likely importance in the context of this thesis, social desirability is described in more detail in Section 1.2.4.5;
- The *tendency to deny* such as denying symptoms has been described as the opposite to acquiescence (Loevinger, 1959). In later publications this bias appears to be used more frequently in the context of social desirability, i.e. 'denial' of socially undesirable traits and behaviours (Millham & Kellogg, 1980; Ramanaiah *et al.*, 1977);
- *Transient mood state*, i.e. participants' self-report outcomes may be biased by the mood they are in when filling out a questionnaire (Podsakoff *et al.*, 2003);

The above list provides an overview of the different types of biases encountered in self-report outcomes. The remainder of the literature review focuses on those biases that were deemed most relevant for the thesis. While *response shift* seems to have been frequently overlooked in overviews of biases (Paulhus, 1991; Podsakoff *et al.*, 2003), there is increasing evidence that individuals' change in perspective between two measurement occasions may pose a serious, albeit underestimated threat to the validity of change scores that are derived from pretest-posttest comparisons (Howard & Dailey, 1979; Sprangers, 1989). Particularly in the context of self-management interventions, response shift bias may need to be considered to be able to interpret program outcomes (Osborne *et al.*, 2006). As a consequence, this bias is discussed in detail in the following Section 1.2.4.4. Furthermore, *social desirability* has been proposed as a particular threat to the validity of change scores derived from retrospective

pretest data (Hill & Betz, 2005; van de Vliert *et al.*, 1985), a form of questionnaire design that is investigated in this thesis. This bias is therefore discussed in detail in Section 1.2.4.5.

#### 1.2.4.4 *Response shift bias*

“Some women in the course were very depressed; they isolate themselves and I am a go-go person. It made me realize that I wasn't depressed at all, just getting older and can't do things as quick as I used to.” (Osborne *et al.*, 2006)

The quote above highlights the potential change in perspective that may occur as a result of participating in a self-management course. Although this change may be a wanted outcome of an intervention (Golembiewsky *et al.*, 1976; Howard, Ralph *et al.*, 1979; Sprangers, 1989), it is a potential bias in survey research (Howard & Dailey, 1979; Howard, Ralph *et al.*, 1979). Taking into account that such *response shifts* may affect as many as 70% of participants of self-management courses (Osborne *et al.*, 2006), this bias poses a threat to the validity of self-report outcomes such as those assessed in self-management trials (see Section 1.2.3). Response shift is therefore introduced in more detail in this section of the literature review. The following review is divided into two parts. While the first part provides a general overview of the concept and historical development of this phenomenon, the second part introduces methods that have been proposed for the detection of this bias. Apart from an exploratory study (Osborne *et al.*, 2006), the present thesis is the first study to systematically explore this bias in the context of self-management programs.

#### *The conceptualisation of response shift*

The notion of response shift was first described in the area of organisational behaviour. In the mid 1970s Golembiewsky *et al.* (1976) identified three types of change: 1) *alpha*, 2) *beta*, and 3) *gamma* change. *Alpha* change was defined as a change from one state (pretest) to another state (posttest) assuming that both assessments are based on a stable calibration of the measurement instrument, while *beta* change was described as a change that resulted from a new calibration of this scale at posttest. Finally, *gamma* change was defined as a redefinition of the scale's content (Golembiewsky *et al.*, 1976). While these early publications highlighted that a distinction between these levels of change is essential to interpret change scores adequately, it was also emphasised that it facilitates the formulation of program goals such as a change at the gamma level being a potential wanted outcome of some programs (Golembiewsky *et al.*, 1976; Zmud & Armenakis, 1978). While response shift as an outcome is widely accepted (Howard, Ralph *et al.*, 1979; Osborne *et al.*, 2006; Sprangers, 1989), most

studies however do not explore this bias as an outcome in its own right but focus on potential confounding effects in the measurement of program outcomes.

A few years later Howard *et al.* (1979b) conducted a series of studies on educational training courses. Without reference to the distinction between the different levels of change, they introduced the term *response shift* as a potential source of bias in the comparison of pretest-posttest data. Analogous to Golembiewsky *et al.* (1976), they stressed that a change in a subject's perception, i.e. a change in the metric of pretest and posttest measures, needs to be distinguished from actual changes (Howard, Ralph *et al.*, 1979). While their definition of response shift contained elements of beta and gamma change, they used a somewhat more general description of this phenomenon with response shift being defined as a change in subjects' basis for rating their level on a given construct and a change in their understanding of this construct (Howard & Dailey, 1979; Howard, Ralph *et al.*, 1979). Further, they classified response shift as an *instrumentation effect*, i.e. as introduced in Section 1.2.4.1, a change in the measurement instrument is a threat to the internal validity of change scores (Campbell & Stanley, 1963; Howard, Ralph *et al.*, 1979).

The synthesis of the work of Campbell and Stanley (1963) and Golembiewsky *et al.* (1976), which was also described by other authors during this period (Armenakis & Zmud, 1979), has the advantage of understanding the practical relevance of response shift in the context of research designs. Although earlier research did not refer to the actual concept of response shift, an *instrumentation effect* was ascribed to changes in the calibration of the measuring instrument (Campbell & Stanley, 1963) – where a recalibration of the measure implies that pretest and posttest scores are not based on a common metric (Cronbach & Furby, 1970). Given that it is this common metric that is critical for an unbiased comparison of pretests and posttests – with response shift threatening such unbiased comparison – response shift could logically be defined in the context of threats to internal validity (Howard, Ralph *et al.*, 1979).

While response shift has obvious implications for the validity of within-subject change scores, it also challenges between-subject and between-group comparisons (Howard, Ralph *et al.*, 1979; Sprangers, 1989). If treatment subjects have a response shift it is assumed that their posttest scores compared with their pretest scores not only show program effects and effects due to other reasons such as history or maturation but also response shift effects (Howard, Ralph *et al.*, 1979). As a consequence, *within-subject* change scores are confounded by recalibrations and/or redefinitions of the target construct. Given that such response shift may vary across subjects, *between-subject* comparisons become similarly confounded as each subject may experience different types and magnitudes of response shift. Finally, a comparison between intervention and control groups is also contaminated if response shifts occur. Given that response shift is assumed to be treatment-induced, control subjects should

be “free” of this bias, rendering *between-group* comparisons invalid (Howard, Ralph *et al.*, 1979; Sprangers, 1989; Sprangers *et al.*, 1999).

To facilitate the conceptualisation of response shift, Figure 11 is an extension of Figure 10 as it illustrates response shift in the context of true experimental designs. Again, the presence of a ✓ at pretest and/or posttest means that data are collected, and a ✓ at intervention means that subjects of that group receive the treatment.

In Figure 11 response shift is presented in a way that the grey-shaded areas are assumed to be potentially confounded by this bias which would then have implications on several levels. In designs 1 and 3a, within-subject comparisons in the intervention group are confounded because pretest data are based on a different metric compared with posttest data. Moreover, across research designs between-group comparisons are confounded as the posttests of the intervention group are assumed to be influenced by response shift bias, whereas the control group would not have experienced a treatment-induced response shift. Finally, intervention group subjects within the grey-shaded boxes are not comparable at posttest anymore as it may be assumed that each individual was affected by response shift in a different way and/or to a different extent (between-subject comparisons).

	Research design	Group	Pretest	Intervention	Posttest
1.	Pretest-posttest control group	IG 1	✓	✓	✓
		CG 1	✓		✓
2.	Posttest-only control group	IG 2		✓	✓
		CG 2			✓
3.	Solomon four-group	IG 3a	✓	✓	✓
		CG 3a	✓		✓
		IG 3b		✓	✓
		CG 3b			✓

Legend

IG: Intervention group

CG: Control group

✓ Providing pretest and/or posttest data; receiving the intervention

Grey shaded box Potential response shift in intervention subjects that may confound within-subject, within-group and between-group comparisons

**Figure 11** Response shift in the context of experimental designs

Since the early 1990s response shift has also received increased attention in health research with one of the first studies exploring this concept in psychiatric and psychotherapy patients (Stieglitz, 1990). The reason for the response shift phenomenon being increasingly applied in

this area is because it provides a plausible explanation for paradoxes observed in trials on severely ill patients. For example, it has been regularly found that patients indicated quality of life levels that were comparable to those of their healthy counterparts (Ahmed *et al.*, 2005; Albrecht & Devlieger, 1999; Breetvelt & Van Dam, 1991; Rees *et al.*, 2002; Schwartz & Rapkin, 2004; Sprangers & Schwartz, 1999). Discrepancies have also been found between patients' self-evaluations and clinical measures, i.e. despite poor clinical outcomes, patients reported comparatively high quality of life scores (Daltroy *et al.*, 1999; Kagawa-Singer, 1993; Sprangers & Schwartz, 1999; Wilson & Cleary, 1995). Hence, the response shift hypothesis is an explanation of these seemingly illogical findings in health research.

The most substantial work on response shift in health research was published in the late 1990s with conceptualisations that are now widely applied (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999). Building on work from research areas such as organisational behaviour (Golembiewsky *et al.*, 1976) and education (Howard, Dailey *et al.*, 1979; Howard *et al.*, 1981; Howard, Ralph *et al.*, 1979; Howard, Schmeck *et al.*, 1979), response shift was defined as a change in subjects' self-evaluation of a given target construct. Such change was described to result from either 1) a *recalibration* of the scale that underlies the measure of this construct, 2) a *reprioritisation* of components of the construct, or 3) a *redefinition* of the construct (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999). While the definitions were inherent in *beta* and *gamma* change (Golembiewsky *et al.*, 1976), it was stressed that dividing gamma change into the components 2) and 3) was necessary as a change in values (reprioritisation) was an equally important component of the response shift theory. While these hypothesised types of response shift may occur simultaneously, any combination of the three is possible adding yet another layer of complexity to this phenomenon (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999).

#### *Techniques to detect response shift*

As the notion of the response shift phenomenon was introduced more than three decades ago, a large range of response shift detection methods has been developed. Considering the range of existing methods the present review needed to be limited to a brief introduction to those detection methods that were applied in this thesis. Two comprehensive overviews of different approaches to detect response shift have been provided by Schwartz & Sprangers (1999) and Thompson & Hunt (1996).

A common method to detect response shift is a design approach (Schwartz & Sprangers, 1999) that consists of the collection of *retrospective pretest* data. These retrospective pretest data are generally collected either simultaneous to or in close proximity to posttest data after

the conclusion of a program. This approach is based on the assumption that subjects provide their retrospective pretests from the same perspective as their posttests, i.e. the comparison of the two scores is assumed to be free of response shift. While several researchers have demonstrated the superiority of this method over pretest-posttest comparisons, they also stressed that these retrospective pretests should not replace actual pretests given that the comparison of the two pretests provides critical information on the direction and magnitude of response shift (Howard & Dailey, 1979; Schwartz & Sprangers, 1999). Further support for using retrospective pretests for the computation of change scores was provided by other researchers, who showed that change scores derived from retrospective pretests correlated higher with objective criteria than change scores based on actual pretests (Howard *et al.*, 1981; Howard, Ralph *et al.*, 1979; Skeff *et al.*, 1992; Stieglitz, 1990). Therefore, the former change measure may be a more accurate reflection of respondents' perceived change (Schwartz & Sprangers, 1999).

Despite apparent advantages of retrospective pretesting, this method is not without criticism. As introduced in Section 1.2.4.3, the validity of retrospective pretests has been questioned for reasons of *recall bias* and *implicit theory of stability or change*. While the former refers to memory distortion (Campbell & Stanley, 1963; Loftus *et al.*, 1991; Pearson *et al.*, 1992; Ross & MacDonald, 1997), the latter refers to the possibility that individuals infer from their 'now'-ratings (posttests) to their pretest state, i.e. they 'construct' their pretests rather than trying to recall (Norman, 2003; Ross, 1989; Schwarz *et al.*, 1998). There has also been concern that retrospective pretests are more prone to social desirability than actual pretests (Hill & Betz, 2005; van de Vliert *et al.*, 1985). Several researchers however refuted this criticism (Howard *et al.*, 1981; Sprangers, 1989; Terborg *et al.*, 1980). And finally the simultaneous assessment of posttest and retrospective pretest data has caused some concern with regard to possible confounding of both data (Randolph & Elloy, 1989; van de Vliert *et al.*, 1985). While this potential influence has rarely been explored, three studies were found, none of which could find evidence of such dependency (Howard, Ralph *et al.*, 1979; Sprangers & Hoogstraten, 1989; Terborg & Davis, 1982). In sum, collecting retrospective pretests may not always be preferable but it has been strongly recommended for situations in which response shift is likely to occur (Howard, Schmeck *et al.*, 1979).

One of the most frequently used statistical methods – a *factor-analytic approach* – has also been developed since the mid 1970s (Golembiewsky *et al.*, 1976; Schwartz & Sprangers, 1999). While it was initially designed to detect gamma change (Golembiewsky *et al.*, 1976), this method was soon extended to the assessment of beta change (Schmitt, 1982). The most sophisticated version of the approach was developed in more recent years, incorporating current terminology of response shift (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999) into the factor model (Oort, 2005b; Oort *et al.*, 2005). Given that this model assumes

familiarity with factor analysis and SEM, it is not presented here but in later chapters of the thesis. While a more detailed introduction to these specific statistical techniques is provided in Chapter 4, the above model is fully introduced in Chapter 5 where it was applied.

In conclusion, both the retrospective and the factor-analytic approach have been shown to result in similar findings (Schmitt *et al.*, 1984; Visser *et al.*, 2005). To date they seem to be the most promising methods to detect response shift, in particular at a group level.

#### 1.2.4.5 *Social desirability bias*

Social desirability is one of the major sources of response bias in survey research (DeMaio, 1984). As briefly discussed in Section 1.2.4.3, this bias may be particularly important in the measurement of outcomes of self-management interventions. One of the reasons why these interventions may be prone to social desirability bias may be that participants and the course leaders build rapport over the course of the intervention. As a result, participants may feel the need to please the course leaders when filling out the questionnaire, particularly at the end of interventions. In view of this potential threat to the validity of scores derived from the course participants, it is surprising that this bias has rarely been explored in past studies. Only two studies out of more than 100 controlled trials of self-management programs considered this bias as a potential covariate (Glasgow *et al.*, 1992; Vlaeyen *et al.*, 1996). Because of this lack of research into social desirability bias in the context of self-management, this section provides an overview of the bias across research areas. The review is divided into two parts. The first part provides a review of the conceptualisation of social desirability and the second part introduces the Marlowe-Crowne Social Desirability scale (Crowne & Marlowe, 1960).

#### *The conceptualisation of social desirability*

Although there are several elements to its conceptualisation, social desirability can generally be described as a response style that is exhibited by respondents who endorse items that represent socially desirable traits and/or behaviours (Nunnally & Bernstein, 1994). Hence, individuals who seek to present themselves in a favourable light would be considered as providing socially desirable answers (Crowne & Marlowe, 1964; DeMaio, 1984). Moreover, social desirability has frequently been described in terms of two dimensions: 1) the need for *social approval*, i.e. individuals try to create a positive impression of themselves to receive approval from others (*impression management*), and 2) *self-deception* or *defensiveness*, i.e. individuals try to avoid disapproval by denying socially undesirable traits and/or behaviours (Crowne & Marlowe, 1964; Millham, 1974; Moorman & Podsakoff, 1992; Paulhus, 1984).



While social desirability is often referred to as a response style (Nunnally & Bernstein, 1994; Paulhus, 1991), others discussed alternative ways to conceptualise social desirability. In the context of the most frequently used social desirability scales, McCrae and Costa (1983) discussed the necessity to differentiate between response styles and personality traits. In their study they considered social desirability scales to be measures of personality traits rather than a response style; a view, that has been supported by others (Kozma & Stones, 1987). If defined as a trait, however, the application of scales to measure social desirability is problematic because respondents who are truly conscientious, trustworthy, and honest might not only be accused of faking or lying but it would also be difficult to discriminate between the honest respondent and those respondents who provide answers in a socially desirable way (McCrae & Costa, 1983).

Moreover, it has been debated whether social desirability represents a personality construct or whether it may be thought of as a response tendency or a characteristic of certain items, i.e. the way items are written may trigger socially desirable responses from respondents who are prone to respond in this way (DeMaio, 1984; Nederhof, 1985). Despite the potential importance of regarding social desirability bias as an item characteristic (DeMaio, 1984), the exploration of social desirability on the item and/or scale level is deemed more complex than its assessment on the respondent level (Moorman & Podsakoff, 1992). This might explain why researchers typically explore social desirability as a response tendency as opposed to a characteristic of an item (Moorman & Podsakoff, 1992; Paulhus, 1991).

The occurrence of social desirability may further vary according to the survey method. It has been found that social desirability is less pronounced in questionnaire-based surveys, while it is more likely that responses are biased by social desirability in personal interviews (DeMaio, 1984; Schwarz & Oyserman, 2001). Other studies found evidence that phone interviews are again more prone to social desirability than personal interviews (Holbrook *et al.*, 2003).

Social desirability has also been found to be related to demographic variables. Research suggested that this bias is more pronounced in older women (Ray, 1988; Visser *et al.*, 1989) and women who have a lower socio-economic background (Kalliopuska, 1992; Visser *et al.*, 1989). Socially desirable responding has also been found to correlate negatively with years of formal schooling, i.e. the more years of schooling the smaller the likelihood of providing socially desirable answers (Deshields *et al.*, 1995). Finally, social desirability has been found to correlate positively with age, i.e. the older the respondent the more likely a bias through social desirability (Deshields *et al.*, 1995; Komarahadi *et al.*, 2004).

Finally, *perception-* as well as *evaluation-based measures* (Schwartz & Rapkin, 2004) tend to be more sensitive to social desirability than other types of questions. As discussed in Section 1.2.3.3, respondents engage in cognitive appraisal when attending to self-report outcomes

that require such process (Schwartz & Rapkin, 2004). While the presented four-step process is the common conceptualisation of this cognitive appraisal, the last step 'formulation of the answer' (Krosnick, 1999; Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988) can be further divided into an answer editing process and the final response (Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988). In this editing process respondents may adjust their response in a socially desirable way for reasons such as self-presentation. As discussed in Section 1.2.4.3, evaluations of self-management courses may be particularly prone to social desirability as participants may want to please the course leader.

### *Measures of social desirability*

A range of social desirability scales has been developed that researchers can choose from according to their specific research interests (Crowne & Marlowe, 1960; Edwards, 1957; Edwards & Walsh, 1964; Hays *et al.*, 1989; Paulhus, 1984). As the 33-item Marlowe-Crowne (MC) Social Desirability scale (Crowne & Marlowe, 1960) was used in the present research, only this social desirability measure is described in detail. A comprehensive overview of other widely used measures of social desirability has been provided by Paulhus (1991).

The MC scale (Crowne & Marlowe, 1960) is one of the most popular and widely used indices of social desirability (Barger, 2002). It was designed as an alternative to the Edwards Social Desirability scale (Edwards, 1957) which had been criticised for its significant correlations with many personality inventories, rendering the interpretation of scores difficult (Crowne & Marlowe, 1964). The MC scale has been commonly described as a measure of a person's need for approval. While the authors defined social desirability in terms of two dimensions, i.e. the *need for approval* and the *avoidance of disapproval* (Crowne & Marlowe, 1964; Paulhus, 1991), the scale has been considered as a measure of a single dimension (Crowne & Marlowe, 1964; Leite & Beretvas, 2005).

Despite the theoretical conceptualisation of the MC scale (Crowne & Marlowe, 1964), studies that applied factor-analytic techniques to validate this measure found inconsistent results. Support for a two-factor hypothesis was provided by two independent researchers with each proposing a model consisting of two correlated factors named *impression management* and *self-deceptive enhancement* (Paulhus, 1984), or *attribution* and *denial* (Ramanaiah *et al.*, 1977). In contrast, other studies found that the MC scale is multi-dimensional, i.e. neither a one-factor nor a two-factor solution showed satisfactory model fit (Ballard, 1992; Barger, 2002; Crino *et al.*, 1983; Leite & Beretvas, 2005; Loo & Loewen, 2004; Loo & Thorpe, 2000). While some researchers expressed concern with regard to the interpretation of the MC scale scores (Ballard, 1992; Barger, 2002), the results of most studies should be treated with

caution. Only two of the above studies applied rigorous statistical techniques to explore the psychometric properties of the MC scale (Barger, 2002; Leite & Beretvas, 2005). Furthermore, the generalisability of studies may be questionable given that almost all used samples of students (Ballard, 1992; Ballard *et al.*, 1988; Barger, 2002; Fischer & Fick, 1993; Fraboni & Cooper, 1989; Loo & Thorpe, 2000).

The length of the MC scale (33 items) may be a burden for some respondents, particularly if the scale is used in conjunction with other instruments. As a consequence, a range of short forms has been developed. Reynolds' (1982) short forms MC-A (11 items), MC-B (12 items), and MC-C (13 items) alongside the Strahan and Gerbasi (1972) short forms X1 and X2 (each 10 items) and XX (20 items) are the most frequently applied ones (Barger, 2002; Leite & Beretvas, 2005). These short forms are each based on a subset of items of the original scale that are measured on a two-point 'true-false' response scale in the same manner as the full MC scale (Crowne & Marlowe, 1960).

Similar to the original scale, opinions differ regarding the usefulness of the MC short forms. While some researchers suggested that all scales are unsatisfactory (Ballard, 1992; Barger, 2002), others demonstrated that they are even improvements over the original scale (Fischer & Fick, 1993; Loo & Loewen, 2004; Loo & Thorpe, 2000). In a similar manner to the original scale, most studies on the short forms should be treated with caution. Apart from one study (Barger, 2002) none of the studies applied rigorous statistical methods. Further, studies only explored model fit of a one-factor solution, whereas no study could be found that tested any of the MC short forms for potential two- or multi-factor solutions such as those applied to the full MC scale (Paulhus, 1984; Ramanaiah *et al.*, 1977). Of all short forms, Reynolds' MC-C (Reynolds, 1982) has been explored most extensively (Zook & Sipps, 1985) and it is now one of the short forms that is most frequently used (Andrews & Meyer, 2003; Frasure-Smith *et al.*, 1999; Leake *et al.*, 1999). This 13-item short form has been described as a reliable alternative to the full MC scale (Reynolds, 1982; Robinette, 1991; Zook & Sipps, 1985) with acceptable internal consistency (Andrews & Meyer, 2003; Ballard, 1992; Loo & Thorpe, 2000; Reynolds, 1982; Zook & Sipps, 1985).

In summary, the large body of literature on social desirability suggests that this bias can be a potentially serious threat to the validity of scores in survey research. While the severity of this bias may be dependent on the survey method as well as the demographic characteristics of the respondents, social desirability may contaminate self-report outcomes. Although current evidence is somewhat inconclusive regarding the application of the MC scale, the full 33-item scale and Reynolds' short form MC-C remain measures of social desirability that are most frequently applied to assess this bias.

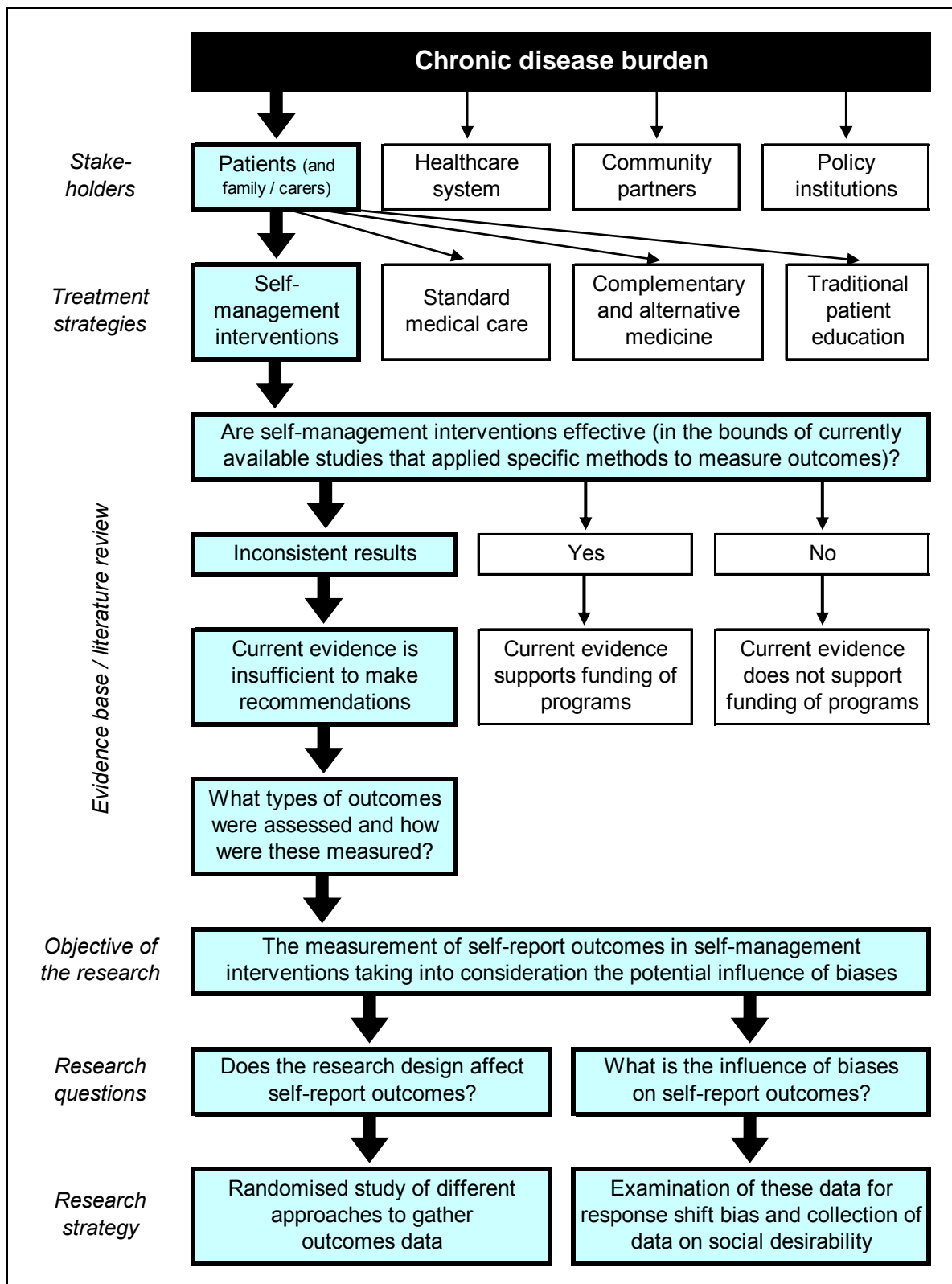
### 1.3 Research questions

Based on current evidence it is difficult to draw conclusions with regard to the effectiveness of chronic disease self-management programs. While there was a trend towards knowledge and clinically assessed outcomes suggesting larger effects than self-report outcomes such as symptoms or functioning, results for outcomes derived from participant self-report were generally inconsistent across trials, rendering overall interpretation of results problematic. An explanation for observed inconsistencies may be confounding of self-report outcomes as a result of biases (see Section 1.2.4). To date self-management studies have largely neglected the influence of bias. Consequently, current and alternative methods of measuring change need to be explored to assess whether confounding exists and the potential extent to which outcomes of self-management interventions are affected.

The overall objective of this study was to explore the validity of the traditional pretest-posttest design and compare outcomes with two alternative designs. Apart from comparing derived measures of change across methods, the influence of confounding through *response shift* and *social desirability* biases was explored. The research questions of the present thesis are as follows:

- I. Does the application of differently designed questionnaires at posttest alter conclusions about the value of programs when effectiveness is assessed from change scores derived from pretest and posttest measures? (Chapter 3);
- II. Are conclusions about program effectiveness different when deriving change scores from retrospective in place of actual pretest data? (Chapter 3);
- III. Can response shift be detected in actual pretest-posttest data when applying a model of measurement invariance? (Chapter 5);
- IV. Are the model parameters invariant across retrospective pretests and actual posttests? (Chapter 5);
- V. Can bias through social desirability be detected in change scores derived from actual pretest-posttest data? (Chapter 6);
- VI. Can bias through socially desirable responses be detected in change scores derived from retrospective pretest-posttest data? (Chapter 6).

To visualise the objectives and research questions of this thesis, which resulted from the apparent lack of evidence as well as the dearth in research on biases in the measurement of outcomes of self-management interventions, Figure 9 was further extended to Figure 12.



**Figure 12** Flow chart of the content of the thesis, Part III

As illustrated in Figure 12, the provided literature review revealed inconsistent findings about the effectiveness of self-management interventions, particularly where self-report measures

were used (see Section 1.2.3). Given that no definite conclusions could be drawn with regard to the value of these programs, the objective of this thesis was to explore the measurement of outcomes of self-management courses with a focus on self-report measures. As described in Section 1.1, the research questions were approached by investigating a range of methods to measure outcomes of self-management interventions (see Section 2.2.3 for a detailed description of the research design). These contrasting approaches to collecting outcomes data were systematically examined for the presence of response shift bias. Further data were gathered on participants' tendency to provide socially desirable responses.

Apart from the investigation of several methods of measuring outcomes of self-management programs, it was also aimed to explore the psychometric properties of the Health Education Impact Questionnaire (heiQ). This instrument was developed in Australia in 2003 (Osborne *et al.*, 2007) and is used in Australia and other countries. Given that all analyses of this thesis were based on scores derived from this instrument, the heiQ was re-validated in Section 4.4.

# Chapter 2

## Study design and data management

## **2 Study design and data management**

### **2.1 Introduction**

The present chapter provides details about the overall study design of the thesis including a description of the research setting, ethics, research design, participant recruitment, and data collection. The latter topic includes a description of the questionnaires that were used to gather these data. Additionally, the processes of data screening, handling, management, and preparation of the datasets for subsequent analyses are explained.

### **2.2 Study design**

#### **2.2.1 Setting**

The research was conducted at the Arthritis Foundation of Victoria Centre for Rheumatic Diseases, Royal Melbourne Hospital, in Parkville, Australia. Data were obtained from community health centres, hospitals, and Non-Governmental Organisations across the following Australian states and territories: the Australian Capital Territory (ACT), New South Wales, Queensland, South Australia, and Victoria.

#### **2.2.2 Ethics**

The University of Melbourne and the RMIT University Human Research Ethics Committees (HREC) approved the study in 2004 (HREC reference numbers 030305; SETNBAPP 5004).

#### **2.2.3 Research design**

The present research was a systematic approach to the measurement of outcomes of self-management interventions. All data concerning program outcomes were assessed through self-report derived from course participants who filled out the Health Education Impact Questionnaire (heiQ). Although the heiQ is a relatively new instrument, it was chosen for the following reasons: a) it has been shown to have strong psychometric properties (Osborne *et al.*, 2007), b) it is currently the only instrument specifically designed to measure impacts of self-management courses (further details on the heiQ are provided in Section 2.3.2), and c) in view of the rather complex study design presented hereafter and the inclusion of additional measures to assess potential bias in scores (see Section 2.3.2), using the heiQ as the sole



assessment instrument of program outcomes was aimed at minimising the burden on respondents. It is acknowledged that this, however, may limit the findings of the thesis to be specific to this particular measure.

As shown in Figure 13, after the recruitment of course participants, standardised heiQ data were collected at the beginning of each self-management intervention (=pretest heiQ, see Appendix 2 for an example of the questionnaire). In contrast, three differently designed heiQs were distributed at the end of each course (=posttest heiQ). To avoid that the latter data were influenced by potential intra-group effects, all posttest questionnaires were randomised within courses following a specific randomisation procedure. These posttest heiQs were designed as follows:

- *heiQ-PP*

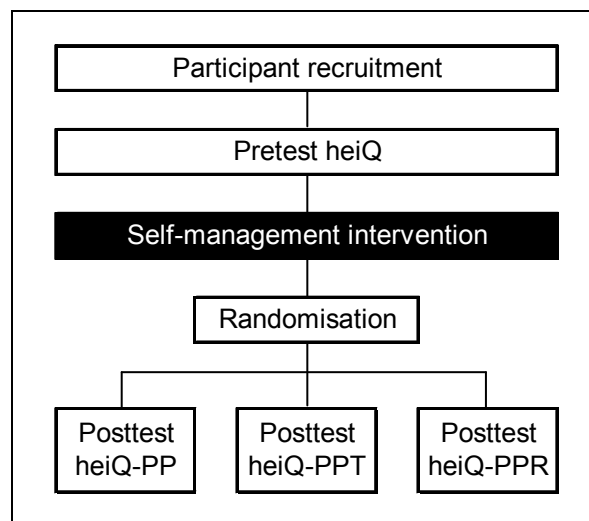
At the end of self-management courses, participants of this group were asked to provide ratings of their current feelings about the respective content of the heiQ items (=actual posttest). In the same manner as at pretest, ratings were provided on a six-point Likert scale (Likert, 1932) ranging from 'strongly disagree' to 'strongly agree'. This posttest heiQ was therefore identical to the pretest heiQ in that it collected data on how people were feeling at the time of filling out the questionnaire. This group was labelled 'heiQ-PP' with the first 'P' representing the pretest heiQ and the second 'P' representing the posttest questionnaire that, in this case, consisted of posttest questions only. As a consequence, change scores derived from posttest minus pretest data could be calculated for this group (see Appendix 3 for an example of heiQ-PP).

- *heiQ-PPT*

This group was termed 'heiQ-PPT' as course participants were asked to provide a direct assessment of their perceived change in addition to ratings of their actual posttest levels. After rating their current levels at posttest ('P'), respondents then provided answers to the same set of items in the form of transition questions ('T'). While answers to the first set of questions were again provided on a six-point ordinal scale, responses to the transition questions were provided on a five-point ordinal scale with response options ranging from 'much worse' to 'much better' with 'the same' as the midpoint. Consequently, apart from computed change scores based on pretest-posttest data, a direct assessment of people's perceived change was available for this group. However, the latter change scores were not used in later analyses as only the potential influence of the transition questions on the ratings of the actual posttests was of interest in the present thesis (see Appendix 4 for an example of heiQ-PPT).

- *heiQ-PPR / heiQ-PPR Retro*

The remaining third of participants were asked to provide an indirect assessment of their perceived change in addition to ratings of their actual posttest levels. Apart from actual posttests ('P'), they then answered the same set of questions with reference to their perceived levels at pretest, i.e. their level on respective heiQ items at the start of the self-management course (=retrospective pretest, 'R'). Both sets of questions were provided on the same six-point Likert scale as used in the previously described heiQs, with the exception of the transition questions. As a result of this design, a second set of change scores was available for this group that could be derived from the comparison of posttest minus retrospective pretest data. Given that the two change scores were treated as separate datasets in later analyses, they were named 'heiQ-PPR' (=pretest-posttest data) and 'heiQ-PPR Retro' (=retrospective pretest-posttest data) to differentiate between the two change measures (see Appendix 5 for an example of heiQ-PPR).



**Figure 13** heiQ data collection

Considering that this study design was specifically developed for this thesis, the rationale for choosing the design is discussed hereafter:

Following from the findings with regard to current evidence about self-management program effectiveness (see Section 1.2.3) and subsequent examination of the literature pertaining to biases in self-report outcomes (see Section 1.2.4), there was sufficient reason to question the appropriateness of comparing self-report ratings provided at pretest with those provided at posttest. As a consequence, this research was aimed at investigating whether change

scores derived from comparing pretest with posttest data are valid to measure outcomes of self-management interventions.

To explore whether the validity concerns about the traditional pretest-posttest design were justified – in particular in connection with response shift bias – retrospective pretests were collected in addition to actual pretest and actual posttest data (*heiQ-PPR*). The collection of retrospective pretest data has been recommended for the following two reasons (see Section 1.2.4.4): 1) retrospective pretests are based on the assumption that they are provided from the same perspective as actual posttest data, i.e. the comparison of the scores is assumed to be free of response shift bias; 2) the comparison of actual pretest data with retrospective pretest data is assumed to provide information on the magnitude as well as the direction of response shift (Howard & Dailey, 1979; Schwartz & Sprangers, 1999). To ensure that the ratings of the retrospective pretest items were provided from the same perspective as the ratings of the posttest items, posttest data were collected first as recommended by several authors (Howard, Ralph *et al.*, 1979; Sprangers *et al.*, 1999).

While it was expected that the data obtained from *heiQ-PPR* would provide important insight into the validity of a retrospective way of measuring outcomes of self-management courses, two potential problems were identified if these data were to be interpreted in isolation:

Firstly, retrospective pretest data are generally provided in close proximity to the posttest data. While this simultaneous assessment – as discussed before – is necessary to ensure that retrospective pretests are provided in relation to the posttests (Howard & Dailey, 1979), it bears the risk of resulting in a potential interdependence of the two scores. That is, once respondents are familiar with the task of providing answers from two different perspectives, i.e. first from the posttest and then from the retrospective pretest perspective, they may alternatively start rating their posttests relative to their perceived retrospective pretest levels. As a result of this concern, it was decided to randomly distribute questionnaire *heiQ-PP* to investigate whether the ratings of the actual posttests were dependent upon the presence of retrospective pretest questions. In view of the random allocation, it was assumed that the comparison of the respective ratings of the actual posttest levels of *heiQ-PP* and *heiQ-PPR* would indicate whether the additional task of providing retrospective pretest data influenced the ratings of the actual posttests of *heiQ-PPR*.

The second issue pertaining to *heiQ-PPR* concerned the task difficulty relating to providing retrospective pretests in close proximity to posttests. It can be assumed that the cognitive task of differentiating between current (=posttest) and past (=retrospective) states may be demanding for some subjects. An overly demanding task has the potential to introduce new biases such as *satisficing*, i.e. participants do not engage in all steps of the response process when providing their answers (see Section 1.2.4). As a result, an alternative retrospective

assessment was developed (*heiQ-PPT*), founded on the assumption that the cognitive task of providing a direct estimation of one's perceived change in the form of transition questions was less difficult than providing retrospective pretests in addition to posttests. The rationale for randomly distributing questionnaire *heiQ-PPT* in addition to the other posttest *heiQs* was to assess whether a second cognitive task alone would influence ratings of the actual posttests or whether a potential influence was related to the type of cognitive task people were asked to perform.

Finally, it shall be discussed why the study design was developed rather than making use of other research designs such as those including control subjects. One of the main reasons for developing this design was related to the objective of this thesis. The thesis was aimed at investigating methods of measuring change and exploring potential confounding effects in change scores. Hence, the question of validity was explored from a different angle compared with previous approaches to internal and external validity (Campbell & Stanley, 1963; Cook & Campbell, 1979). That is, this research was not aimed at determining the effectiveness of programs per se – for which control subjects would have provided important information (see Section 1.2.4.1) – but the thesis was concerned with a) the influence of the questionnaire design on people's ratings of their posttest levels and b) the possible confounding effects of response shift and social desirability bias on the results of the evaluation of self-management programs. Thus, instead of manipulating the variable 'intervention', the variable 'instrument' was manipulated, with the randomisation being based on the assumption that randomised subjects were comparable at the beginning and at the end of courses. Consequently, any observed differences in scores could be attributed to the design of the respective posttest *heiQ* rather than competing interventions. Moreover, issues pertaining to the use of control groups in the context of response shift bias have been discussed in Section 1.2.4.4.

## **2.3 Data collection, management and preparation**

### **2.3.1 Recruitment of self-management courses**

In 2003 the Centre for Rheumatic Diseases received a National Arthritis and Musculoskeletal Conditions Improvement Grant (NAMCIG) from the Commonwealth Department of Health and Ageing to develop a National Quality and Monitoring System for self-management programs in Australia (Osborne & Whitfield, 2004). An extensive database of Australian self-management course leaders had been created as part of the project which was made available for the recruitment of courses for this study. For this, all course leaders from the database received a letter inviting them to take part in the research project. Throughout the data collection period further courses were recruited through the promotion of the study at

several national conferences, strong rapport with the Australian arthritis foundations – who are major providers of self-management courses in Australia – and word-of-mouth of course leaders who already took part in the research. The final database comprised n=625 leaders who were offered involvement in the study and who received regular information and updates about the research.

### **2.3.2 Data collection**

Once course leaders were part of the study, they were asked to forward information on their course schedules to the researcher. Approximately two weeks before the start of a course, leaders were then sent a heiQ package that consisted of a letter with basic information about the study, a leaflet explaining the data collection procedure, pretest and posttest heiQs, a course participation form, and envelopes for the posttest heiQs. The additional envelopes were distributed to enable participants to make their evaluation of the course confidentially. Once all data were gathered, the course leaders returned all forms to the researcher. Given that no personal information was recorded by the respondents, the researcher was blinded to the participants' identity. To ensure a correct matching between pre- and posttest heiQs all questionnaires were labelled with unique identification numbers. Given that the initial matching of the heiQs was carried out by the course leaders, some safety measures were included to ensure that the matching was done correctly. These safety measures are described in more detail in the context of *Demographic data* and *Course participation form*.

Data were collected from February 2005 to December 2006. The minimum requirement for the sample size of each of the datasets was mainly derived from recommendations by Carroll (1978) who developed a calculation for the minimum sample size for factor analyses. In order to establish  $m$  factors he proposed a sample size of  $2m + 2^m$  (Carroll, 1978). As described hereafter, the heiQ consists of eight factors (Osborne *et al.*, 2007); hence, n=272 was considered the minimum sample size for the analyses. It was aimed to reach a sample size of n=300 per group. These sample size requirements were derived from factor analysis as there are too few recommendations for SEM (Hair *et al.*, 2006). No definite suggestions exist for this type of analysis because the required sample size is dependent on the size of the model, the strength of the relationship of the items, the size of the loadings, the number of indicators per factor, the distribution of the data, and the parameter estimation procedure (Bentler & Dudgeon, 1996; Boomsma & Hoogland, 2001; Muthén & Muthén, 2002; Tanaka, 1987). It has only been proposed that the sample size should not drop below n=200, as this can lead to improper solutions (Boomsma & Hoogland, 2001).

### *The Health Education Impact Questionnaire (heiQ)*

For the development of the Health Education Impact Questionnaire (heiQ) state-of-the-art techniques were applied. As introduced in Section 1.2.2 and further presented in Figure 4, this included the generation of a program logic model for health education interventions (Osborne *et al.*, 2004), and concept mapping workshops (Batterham *et al.*, 2002; Trochim & Linton, 1986; Trochim *et al.*, 2004) which were attended by a wide range of stakeholders (patients, course leaders, health professionals, health managers, program funders and policy makers). Furthermore, a range of statistical techniques such as exploratory factor analysis (EFA) using the computer program CEFA<sup>7</sup> (Browne *et al.*, 2004), and confirmatory factor analysis (CFA) using LISREL<sup>8</sup> Version 8.5 (Jöreskog & Sörbom, 1996-2001) were applied. Additional details on the development of the heiQ can be found in Osborne *et al.* (2007).

After the first phase of item and construct development, 69 candidate statements were tested on a construction sample (n=591). The statistical analyses led to the selection of 42 items that were again tested in another sample (n=598). The final validation confirmed this set of 42 items that constitute the eight dimensions that were previously introduced in Section 1.2.2 in the context of the objectives of self-management programs (Osborne *et al.*, 2007):

- (1) Positive and Active Engagement in Life (five items);
- (2) Health-Directed Behaviour (four items);
- (3) Skill and Technique Acquisition (five items);
- (4) Constructive Attitudes and Approaches (five items);
- (5) Self-Monitoring and Insight (seven items);
- (6) Health Service Navigation (five items);
- (7) Social Integration and Support (five items);
- (8) Emotional Well-Being (six items).

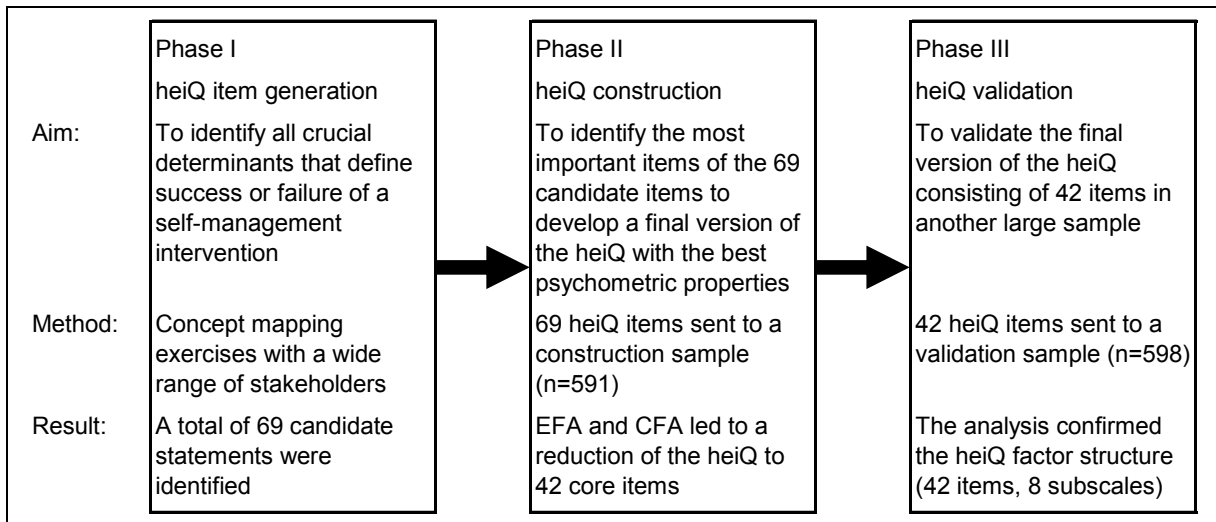
One subscale (Emotional Well-Being) is composed of items scored in the reverse direction to the other seven subscales.

The development and validation of the heiQ is again summarised in Figure 14. The different phases are separated into item generation, heiQ construction, and heiQ validation.

---

<sup>7</sup> CEFA (Comprehensive Exploratory Factor Analysis) is a computer program designed to estimate factor solutions in an exploratory way (Browne *et al.*, 2004). It is one of the recommended programs to carry out this type of factor analysis (McDonald, 2005).

<sup>8</sup> LISREL (LInear Structural RELations) is a program that combines CFA and linear regression. A more detailed introduction to LISREL follows in Section 4.3.2.



**Figure 14** Item generation, heiQ construction, and heiQ validation phase

### *Reynolds' short form MC-C of the Marlowe-Crowne Social Desirability scale*

In addition to the core heiQ items, all posttest questionnaires contained a measure of social desirability to assess people's tendency to provide socially desirable answers. As introduced in Section 1.2.4.5, the MC scale (Crowne & Marlowe, 1960) is one of the most widely used scales of social desirability (Barger, 2002). Because of its length, however, it was decided that the original scale was too long to be included in the current study as the posttest heiQs already consisted of several pages. To reduce the burden on the respondents, Reynolds' 13-item short form MC-C (Reynolds, 1982) was included instead (see page 5 of each respective posttest heiQ provided in Appendix 3, Appendix 4, and Appendix 5). As mentioned in Section 1.2.4.5, the response options of the MC-C are dichotomised with eight items being keyed 'false', i.e. 'false' is interpreted as the socially desirable response, while the remaining five items are keyed 'true' such that the answer 'true' reflects social desirability (Reynolds, 1982). Further details on the factor structure of the MC-C follow in Section 6.3.

### *Demographic data*

A range of demographic variables was collected at the end of each pre- and posttest. Most of these variables were used for the description of the obtained sample and the comparison of the randomised groups to ensure that there were no systematic differences between the respondents of the three posttest questionnaires. Given that the distribution of the heiQ was part of an ongoing quality assurance project, some further demographic data were collected. While these were not used in the present thesis, they are included in the following list for

completeness as they were part of the heiQ packages that were sent out (see Appendix 3 through to Appendix 5).

At pretest, the following demographic data were collected from the course participants:

- Age
- Gender
- Postcode
- Number of people living in the participant's household
- Aboriginal or Torres Strait Islander background
- Birth place
- Primary language spoken at home
- Education
- Chronic conditions that currently trouble or have troubled the participant
- Main health problem
- Previous participation in any self-management course
- Smoking status
- Employment status
- Data on healthcare concession cards
- Private health insurance status
- Plans to lose weight
- Height
- Weight

At posttest, the following demographic data were collected:

- Age
- Gender
- Postcode
- Smoking status
- Height
- Weight

As can be seen, some demographic data were assessed twice, i.e. at pretest and at posttest. The main reason for repeating the collection of these data was to facilitate the linkage between each pretest and posttest. Given that the matching of these questionnaires was carried out by the course leaders by means of the unique identification numbers, this initial matching was outside the control of the researcher. Therefore, age, gender, and postcode were used to ensure that the questionnaires had been matched correctly, i.e. the researcher cross-checked these variables between pretest and posttest before data entry.



### *Course participation form*

Apart from collecting data from the participants, course leaders were also asked to provide some information. This course participation form (see Appendix 6) assessed data on course type, date of the first session, course duration, details on the organisation that was running the course, and course venue as well as data on the course leaders such as contact details, course leader status (peer leader or health professional), details about their training, and number of courses they had conducted over the previous 12 months. Finally, the course leaders were asked to take note of each participant's course attendance. They were asked to provide information on each participant's gender and identification numbers, with the latter being necessary for the match between questionnaires and course attendance. This form was another safety measure to ensure the correct linkage of pretest and posttest heiQs. On a few occasions course leaders had to be contacted to verify the data.

### **2.3.3 Data screening prior to entry**

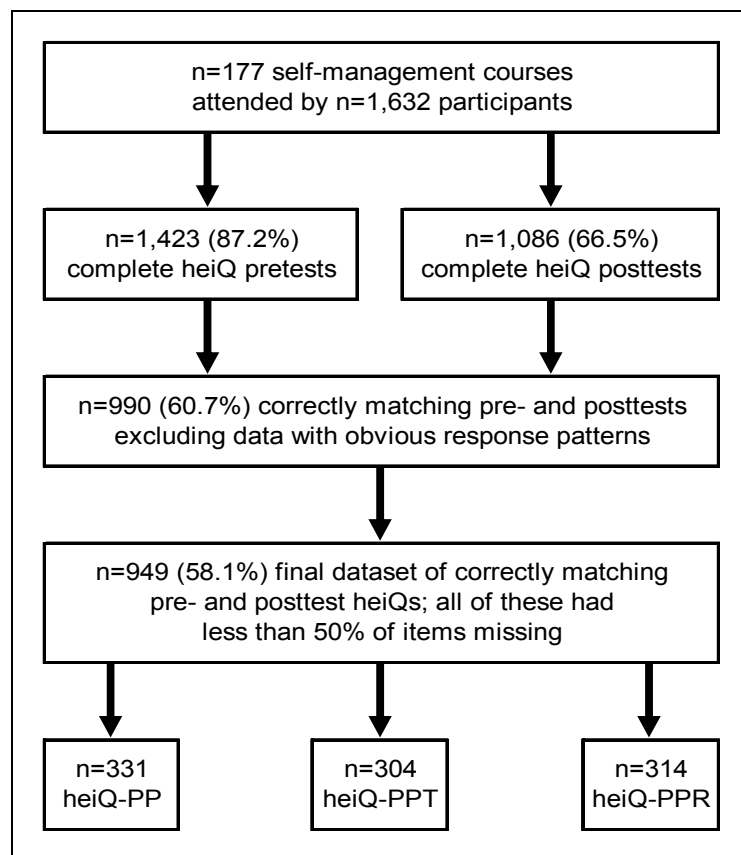
Once the heiQ packages comprising pre- and posttest heiQs and the course participation form were received at the Centre for Rheumatic Diseases, some steps were undertaken before data entry. As described in Section 2.3.2, it was ensured that the pre- and posttests were matched correctly. While an incorrect match was observed in approximately one tenth of all cases, almost all of these were corrected by comparing the demographic variables *age*, *gender*, and *postcode*. The data of six subjects had to be discarded as no matching pair was found. In addition to this exercise, all heiQs were screened for data integrity such as obvious response patterns. Given that this instrument was designed in a way that subscale Emotional Well-Being is reversed (see Section 2.3.2), i.e. answers to these items mean the opposite to answers to items of the remaining seven heiQ subscales, obvious response patterns could be detected easily. It was therefore assumed that respondents who had consistently chosen the same response option had filled out the heiQ regardless of the respective item content. As a result, these data were discarded which led to the deletion of an additional six cases. After these pre-screening exercises, all data were entered into the heiQ database.

### **2.3.4 Data entry and sample size**

The heiQ database was built in Microsoft Access. It was designed in a user-friendly way that suited the purpose of the thesis. Separate tables were created for general information about the course, participant attendance rates, course leader information, pretests, posttests, and the MC-C scale. Once entered, all questionnaires were filed and locked in a secure place at

the Centre for Rheumatic Diseases. These will be stored until 2013 following the conditions of the University of Melbourne Human Research Ethics Committee. The filing cabinets can only be accessed by authorised staff of the centre.

As illustrated in Figure 15, data from n=177 self-management courses had been obtained by the end of the data collection period in December 2006. These involved a total of n=1,632 participants, of whom n=1,423 (87.2%) had provided heiQ pretests and n=1,086 (66.5%) had provided heiQ posttests. These data were then exported into SPSS for further preparation. Given that only matching pretests and posttests were extracted from the database, a total of n=990 (60.7%) questionnaires remained. In SPSS the variable 'nmiss' was created that provided information on the number of items that each participant had missed. This variable was created for each set of 42 heiQ items separately, i.e. pretests, posttests of heiQ-PP, heiQ-PPT, and heiQ-PPR, transition questions of heiQ-PPT, and retrospective pretests of heiQ-PPR. As will be explained in Section 2.3.6, it was decided that at least 50% of items (>21 items) in any one of these questionnaires needed to have been provided by participants to be acceptable. Therefore, all cases with 'nmiss'>21 were discarded from later analyses. This led to a final sample size of n=949 (58.1%) pretests and posttests which was made up of n=331 for group heiQ-PP, n=304 for group heiQ-PPT, and n=314 for group heiQ-PPR.



**Figure 15** Study sample

### 2.3.5 The distributional properties of the heiQ raw data

Before handling the remaining missing data, the raw data were examined for their univariate, multivariate, and bivariate distributional properties. This was critical for the analyses because specific measures need to be undertaken to ensure that the appropriate statistical methods are applied given the characteristics of the data. In particular, if data are non-normally distributed, the application of inappropriate techniques can have large influences on results (West *et al.*, 1995). To accommodate non-normal data distributions, data transformations or specific parameter estimation methods are available (Bollen, 1989; Hair *et al.*, 2006; Satorra & Bentler, 1994). Further details on the parameter estimation method that was applied in this thesis will follow in Section 4.3.3.

#### *Univariate and multivariate normality*

The univariate and multivariate normality checks were undertaken in LISREL<sup>9</sup> version 8.72 (Jöreskog & Sörbom, 1996-2001). This program, which was used for most analyses in the thesis, offers a convenient data screening facility through its data pre-processor program PRELIS 2.72 (Jöreskog & Sörbom, 1996-2002; Jöreskog & Yang, 1996). Given that these tests only run for continuous data, all items were temporarily declared continuous instead of ordinal to be able to run the tests. At first, all tests were carried out on the full samples of the heiQ pretests and posttests, i.e. the data of the three randomised groups were collapsed. In a second step, however, the tests were carried out on the randomised samples separately to ensure that the distributional properties were similar across groups. This step further included normality tests on the retrospective pretests of group heiQ-PPR. Tests on the transition questions were omitted as these data were not used in later analyses.

The results of the heiQ pretests indicated that all but five items were negatively skewed. In contrast, the kurtosis of the data was less consistent. One fourth of items (n=10) showed normal kurtosis, 19 items were leptokurtic, i.e. the distribution was 'pointy' with 'thin' tails, and the remaining 13 items were platykurtic, i.e. they had a 'flat' distribution with comparatively 'fat' tails. Similar to the pretests the majority of items (n=38) of the posttests were negatively skewed. The pattern of kurtosis of the posttests, however, differed from the pretests in that all but two items showed a significant non-normal distribution with the majority of items (n=32) being leptokurtic. It was further observed that most items with a platykurtic distribution

---

<sup>9</sup> Although commonly used when referring to the computer program, LISREL is also known as a synonym for analysis of covariance structures (Nunnally & Bernstein, 1994). In the context of this thesis, LISREL was exclusively used in reference to the software.

were items of the Emotional Well-Being subscale, i.e. items that had reversed scoring (see Section 2.3.2). Details of the results of the tests are provided in Appendix 7.

When dividing the posttest sample into the three groups, the results for univariate normality were largely the same compared with the results of the collapsed posttest sample. Across samples all but five items had negative skewness. With regard to kurtosis, results were also similar to the full posttest sample, i.e. about 75% of items had a leptokurtic distribution and all items of Emotional Well-Being were platykurtic. The only obvious difference between the three samples was that all items of the Social Integration and Support subscale had normal kurtosis in heiQ-PPR, while most of these were non-normal in the other two samples. For the retrospective pretests (heiQ-PPR Retro) results were generally similar to the results of the previous samples with the majority of items (n=35) having negative skewness. In contrast to the other samples, however, half of these items had normal kurtosis. Of the remaining items, 15 were moderately platykurtic and six were leptokurtic (see Appendix 7).

Given that the majority of the items departed from univariate normality, the data could not be multinormal (West *et al.*, 1995). Largest departures from multivariate normality were found in the collapsed sample of actual posttests, followed by the pretests, the posttests of heiQ-PPT, heiQ-PP, and heiQ-PPR. Least departure from multivariate normality was observed in the sample of retrospective pretests of group heiQ-PPR (see Appendix 7).

### *Bivariate normality*

The previous tests indicated that most heiQ pretest, posttest, and retrospective pretest items departed from normality. While non-normality can be caused by limited sample sizes, it can also be a result of ordinal scaling of items (Schumacker & Lomax, 2004), as was the case in the present samples. One way of dealing with non-normality is the transformation of data (Hair *et al.*, 2006; Schumacker & Lomax, 2004). Alternatively it has been recommended to use the moment matrix in combination with an asymptotic covariance matrix to handle non-normal data (Jöreskog, 2002-2005; Schumacker & Lomax, 2004). As explained in more detail in Section 4.3.3, the latter method was applied in the present study, i.e. matrices based on polychoric correlations<sup>10</sup> and asymptotic covariances were used to estimate the model parameters. A necessary condition for using polychoric correlations, however, is that the data are bivariate normal (Jöreskog, 2002-2005). While these correlations have been found to be robust against small departure from bivariate normality, it has been recommended to apply the bivariate test of close fit (Jöreskog, 2002-2005) which is similar to the root mean square

---

<sup>10</sup> Further details on the computation of polychoric correlations follow in Section 4.3.3.

error of approximation (RMSEA) measure (Steiger, 1990). If this index is significant, then the application of polychoric correlations is problematic (Jöreskog, 2002-2005).

After applying the bivariate test of close fit on each heiQ dataset, it was found that only one pair out of 861 possible combinations of item pairs ( $(42 \cdot 41)/2$ ) did not meet the requirement of close fit in each of the following datasets: the collapsed sample of the actual posttests and the sample of posttests of group heiQ-PP. Across the remaining samples all items met the requirement of close fit. It could therefore be assumed that the method applied in the present analyses was appropriate (see Appendix 8 for the results of the bivariate test of close fit).

### 2.3.6 Treatment of missing data

Finally, each sample had to be prepared in view of missing data, i.e. some of the included cases had missed up to 21 items (see Section 2.3.4). When handling samples with missing data, several issues have to be taken into account that shall be discussed briefly. Firstly, the need for complete data depends on the statistical technique used for the analyses. In the case of SEM, as applied in the present thesis, it is assumed that the variables of the moment matrix follow a Wishart distribution (Jöreskog, 1979). Hence, complete data are required for the probability density function (Brown, 1994; Marcoulides & Schumacker, 1996b). Secondly, it is necessary to investigate the reasons for data missingness, i.e. some missingness mechanisms render certain techniques to handle missing data inappropriate.

A classification of data missingness mechanisms was introduced by Rubin (Rubin, 1976). After substantial extension of his original suggestions these mechanisms are now commonly referred to as patterns in which items are missing completely at random (*MCAR*), missing at random (*MAR*), or not missing at random (*NMAR*). In *MCAR*, cases with missing values are a random subsample of the full dataset, i.e. the missingness mechanism is independent of any data values. In contrast, the assumption in *MAR* is less restrictive. The missingness depends on the observed variables; however, it does not depend on missing values. Finally, in *NMAR*, the missingness depends on both missing and observed values (Little & Rubin, 1989, 2002). In the data of this study no obvious missingness patterns existed, with the exception of a pattern that could be attributed to a mistake in the printing of the first wave of heiQ-PPT, i.e. the last item of the first page had been accidentally omitted.<sup>11</sup> It was further observed that only up to 5% of items in any sample were missing and that 90% of subjects had generally missed no more than two items. Therefore, given that no missingness pattern

---

<sup>11</sup> This also explains the comparatively small sample size of the raw data of heiQ-PPT in the normality tests performed in the previous Section 2.3.5 (see Appendix 7 and Appendix 8).

was observed and only few items were missing, it was assumed that the data missingness of the present data was ignorable, i.e. it was assumed to be MCAR.

While the method of handling missing data should be selected on the basis of the amount of missing values as well as these missingness mechanisms (Arbuckle, 1996), the strengths and weaknesses of each method need to be considered as well. The range of techniques for handling missing data and each method's strengths and weaknesses has been discussed extensively (Allison, 2003; Little & Rubin, 1989; Schafer & Graham, 2002). Therefore, this final description of the preparation of the present samples is limited to a brief description of the techniques that were applied. After careful examination of the techniques, a combination of the following two methods for handling the missing heiQ data was chosen:

Firstly, a derivative of the frequently used *listwise deletion* was applied. In listwise deletion all cases with missing data are deleted. While this method generally leads to the deletion of too many cases and biases are introduced if data are not MCAR (Little & Rubin, 2002; Marcoulides & Schumacker, 1996a), it was decided to restrict the deletion of cases to those who had provided less than 50% of items in any one questionnaire (see Section 2.3.4). The main reason for the deletion of these cases was to reduce any potential bias introduced by the second technique applied to the missing data in which all remaining missing values were replaced. Given that the deletion of these cases reduced the proportion of missing items to a maximum of 3.3% missing values across datasets, it was assumed that the following missing data replacement technique would not change the original properties of the data.

Secondly, all remaining missing data were replaced by the *expectation-maximisation (EM) algorithm* (Dempster *et al.*, 1977). Despite criticism that it may lead to biased estimates and biased standard errors (Allison, 2003; Enders, 2001), other research has suggested that this method leads to least biased estimates regardless of the missingness mechanisms (Gold *et al.*, 2003), and the EM algorithm was also found to be one of the few recommended methods for missing data replacement (Schafer & Graham, 2002). Before applying this method, the following aspects were considered: a) although one of the assumptions of such algorithms is that data are multivariate normal (Enders, 2001), research has suggested that this method is justifiable in the non-normal case (Yuan & Bentler, 2000). And b) while most research on such algorithms used continuous data (Allison, 2003; Gold & Bentler, 2000; Gold *et al.*, 2003; Muthén *et al.*, 1987), a practical application on quality of life data suggests that this algorithm is appropriate for the ordinal case (Lee *et al.*, 2005). Moreover, given that the heiQ items are measured on a six-point scale, it can be assumed that the scales approximate an underlying continuity (von Briesen, personal communication, March 16, 2007).

PRELIS was used for the missing value imputation. Because of the ordinal nature of the heiQ data, the EM algorithm imputed discrete values. Given that no further cases were lost in this

process, the final samples consisted of n=949 complete pre- and posttest heiQs, with n=331 for heiQ-PP, n=304 for heiQ-PPT, and n=314 for heiQ-PPR.

## **2.4 Summary**

The present chapter provided the background on the research design of the thesis including data collection and data management. After careful screening, preparation, and replacement of all missing data, the final datasets were then ready for analyses.

# Chapter 3

Approaches to the  
measurement of  
change with the  
heiQ



### **3 Approaches to the measurement of change with the heiQ**

#### **3.1 Introduction**

In view of the rather inconsistent results with regard to the effectiveness of self-management programs (see Section 1.2.3), this research set out to investigate several designs to measure outcomes of such interventions. For this, a research design was developed that compared the traditional pretest-posttest design with two pretest-posttest designs that asked course participants to additionally provide retrospective data at posttest (see Section 2.2.3). Given that participants who had been randomly allocated to one of the alternative designs had to perform a second cognitive task at posttest, the first analysis chapter of the thesis was aimed at investigating potential effects of the research design on obtained scores.

This chapter is divided into four parts. After a description of the demographic characteristics of the study participants, the first part of the analyses attends to the comparison of mean scores of actual pretest and posttest data across the three randomised groups (heiQ-PP; heiQ-PPT; heiQ-PPR). In the second part of the analyses, change scores based on pretest-posttest data are described. This incorporates a comparison of a) mean change scores and b) proportions of participants who were classified as 'decline', 'no change', or 'improvement'. In the final part of the analyses, change scores derived from retrospective pretest data were compared with change scores derived from actual pretest data (heiQ-PPR; heiQ-PPR Retro). The last two sets of analyses were aimed at exploring whether conclusions about program effectiveness differed across datasets (heiQ-PP; heiQ-PPT; heiQ-PPR; heiQ-PPR Retro) and whether conclusions were dependent on the method of presenting change, i.e. mean change scores or proportions of participants in pre-defined categories of change (see Section 1.3).

#### **3.2 Aims**

The aims of the chapter were:

- 3.a To describe the demographic characteristics of the study sample including comparisons across samples with complete and incomplete data, and across randomised groups;
- 3.b To assess whether the participants of the three randomised groups differed at pretest;
- 3.c To investigate whether the participants of the randomised groups had responded differently to the questions that constituted the actual posttest, i.e. whether the different tasks at posttest had influenced people's ratings of their actual posttest levels;

- 3.d To test whether conclusions about program effectiveness differed across groups when assessing a) each group's mean change scores and b) proportions of participants in pre-defined categories of change (heiQ-PP; heiQ-PPT; heiQ-PPR);
- 3.e To test whether conclusions about program effectiveness differed across actual and retrospective pretest-posttest data when assessing a) each group's mean change scores and b) proportions of subjects in pre-defined categories of change (heiQ-PPR; heiQ-PPR Retro).

### 3.3 Demographics

As described in Section 2.3.4 and illustrated in Figure 15, a total of 1,423 course participants provided pretest data. The demographic characteristics of these participants are described in this section. Given that some of these, however, could not be included in the analyses as no matching posttest data were available, this section also explored whether differences existed between study participants (n=949) and those participants who had to be excluded (n=474). Finally, it was investigated whether there were differences in the demographic characteristics of participants across heiQ-PP, heiQ-PPT, and heiQ-PPR.

Participants' characteristics are presented in Table 4. Of the 1,423 course participants, 76% were female. Participants' age ranged from 19 to 98 years with a mean age of 62 years. Twelve percent reported their formal education to be up to primary school, 30% up to year 8, 24% up to year 12, 18% had a TAFE (Australian Technical and Further Education) diploma, and 16% had a university degree. At the time of data collection, most course participants (66%) were retired, while only 5% were working full-time. The majority of participants (91%) reported English as their primary language and 73% reported having been born in Australia. Finally, participants were asked to indicate their chronic condition(s). As multiple responses were possible, the list of conditions in Table 4 adds up to more than 100%. Most frequently reported chronic diseases were: osteoarthritis (47%), depression (30%), asthma (21%), and diabetes (21%). Given that participants had the opportunity to indicate conditions in addition to those on the provided list (see page 4 of the pretest heiQ, Appendix 2), an additional 41% of 'other conditions' were reported.

When comparing the group of individuals whose data could not be included in the analyses (n=474) with the study participants (n=949), only few differences were observed. Significant differences were found for birth place and primary language with a smaller number of study participants having been born outside Australia (24.9% versus 30.2%) and slightly more participants (92.0% versus 87.6%) reporting English to be their primary language. A further

difference was found for diabetes with 22.1% of study participants reporting this condition compared with 17.2% of subjects who were excluded from the analyses.

**Table 4** Demographic characteristics of participants who provided pretests (n=1,423) and comparison of study participants (n=949) versus those not included in the study (n=474)\*

	Total n=1,423		Study participants n=949		Individuals who could not be included n=474	
	n	%	n	%	n	%
Gender						
Female	1,065	75.9	716	76.3	349	74.9
Male	339	24.2	222	23.7	117	25.1
Age						
Mean (standard deviation)	62.0 (13.3)		62.3 (13.0)		61.4 (14.0)	
Range	19-98		19-90		22-98	
Education						
Primary education	155	11.8	97	11.0	58	13.2
Up to year 8	395	30.0	265	30.2	130	29.5
Year 9 to 12	321	24.4	218	24.8	103	23.4
TAFE	241	18.3	167	19.0	74	16.8
University	207	15.8	131	14.9	76	17.2
Employment status						
Full-time	58	4.9	28	3.1	30	6.7
Part-time	122	9.0	81	9.0	41	9.1
Unemployed	97	7.2	67	7.4	30	6.7
Home duties	168	12.5	117	12.9	51	11.3
Retired	892	65.9	600	66.4	292	64.9
Other	17	1.3	11	1.2	6	1.3
Birth place*						
Australia	1,024	73.4	702	75.1	322	69.8
Born elsewhere	372	26.9	233	24.9	139	30.2
Main language*						
English	1,263	90.6	859	92.0	404	87.6
Other	132	9.9	75	8.0	57	12.4
Chronic condition (more than one could be selected)						
Asthma	288	21.0	199	21.6	89	19.6
Cancer	76	5.5	52	5.7	24	5.3
Coronary heart disease	181	13.2	115	12.5	66	14.5
Depression	406	29.6	267	29.0	139	30.6
Diabetes*	281	20.7	203	22.1	78	17.2
Fibromyalgia	153	11.1	102	11.1	51	11.2
Osteoarthritis	644	46.9	435	47.3	209	46.0
Osteoporosis	206	15.0	136	14.8	70	15.4
Rheumatoid arthritis	238	17.5	149	16.2	89	19.6
Other	566	41.4	391	42.5	175	38.9

\* Significant differences at  $p < 0.05$  level (Chi-square and t-test, respectively)

As shown in Table 5, further data were collected on the course type, course duration, and venue where the course was held (see Section 2.3.2). The majority of subjects attended the Chronic Disease Self-Management Program (Lorig, González, & Laurent, 1999), while 16% attended the Arthritis Self-Management Course (Lorig *et al.*, 1985), 4% attended an osteoporosis course (Osteoporosis Victoria, 2001), and 5% attended other self-management courses. Given that most self-management courses followed the Stanford curriculum (Lorig *et al.*, 1985; Lorig, González, & Laurent, 1999), the majority of interventions (87%) were run over six or seven weeks. Of the remaining courses 5% were run over four and 8% were run over a period of up to 12 weeks. Most courses were held at community health centres (54%), while 13% were located at hospitals, 10% at arthritis foundations, and 24% at other venues. When comparing the group of individuals whose data could not be included in the analyses (n=474) with the study participants (n=949), it was observed that slightly more study participants had attended a self-management course that lasted up to 12 weeks.

**Table 5** Details on the courses across participants who provided pretests (n=1,423) and comparison of study participants (n=949) versus those not included in the study (n=474)\*

	Total n=1,423		Study participants n=949		Individuals who could not be included n=474	
	n	%	n	%	n	%
Course type						
Arthritis	216	15.5	139	15.0	77	16.5
Generic chronic disease	1,054	75.5	696	75.0	358	76.5
Osteoporosis	62	4.4	41	4.4	21	4.5
Other	64	5.0	52	5.6	12	2.6
Course duration*						
4 weeks	71	5.0	47	5.0	24	5.1
6-7 weeks	1,241	87.2	817	86.1	424	89.5
Up to 12 weeks	111	8.1	85	9.0	26	5.5
Venues where courses were held						
Arthritis foundation	127	9.5	80	8.8	47	10.6
Community health centre	732	54.3	511	56.3	221	49.8
Hospital	175	13.0	114	12.6	61	13.7
Other	318	23.6	203	22.4	115	25.9

\* Significant differences at  $p < 0.05$  level (Chi-square)

Finally, participants' characteristics across the three groups (heiQ-PP; heiQ-PPT; heiQ-PPR) were compared (see Table 6). Results suggested that the subjects across the randomised samples were largely identical. One of the few significant differences was observed for birth place with a higher percentage of people of group heiQ-PPR having been born in Australia

(73.5% of heiQ-PP; 70.8% of heiQ-PPT; 80.9% of heiQ-PPR). Another significant difference was observed for people who reported having rheumatoid arthritis. Significantly more people of group heiQ-PPR (21.6%) reported having this condition compared with 15.9% of group heiQ-PP and 10.9% of group heiQ-PPT.

**Table 6** Demographic characteristics of respondents across the three randomised groups: heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314)

	heiQ-PP n=331		heiQ-PPT n=304		heiQ-PPR n=314	
	n	%	n	%	n	%
<b>Gender</b>						
Female	244	74.2	231	77.3	241	77.7
Male	85	25.8	68	22.7	69	22.3
<b>Age</b>						
Mean (standard deviation)	62.2 (13.2)		62.0 (13.4)		62.6 (12.4)	
Range	19-90		19-88		20-86	
<b>Education</b>						
Primary education	31	9.8	35	12.8	31	10.7
Up to year 8	100	31.7	79	28.9	86	29.7
Year 9 to 12	82	26.0	68	24.9	68	23.4
TAFE	59	18.7	54	19.8	54	18.6
University	43	13.7	37	13.6	51	17.6
<b>Employment status</b>						
Full-time	13	4.2	6	2.1	9	3.0
Part-time	21	6.7	25	8.7	35	11.6
Unemployed	28	8.9	21	7.3	18	5.9
Home duties	42	13.4	36	12.5	39	12.9
Retired	204	65.2	195	67.7	201	66.3
Other	5	1.6	5	1.7	1	0.3
<b>Birth place*</b>						
Australia	241	73.5	211	70.8	250	80.9
Born elsewhere	87	26.5	87	29.2	59	19.1
<b>Main language</b>						
English	301	92.0	270	90.6	288	93.2
Other	26	8.0	28	9.4	21	6.8
<b>Chronic condition (more than one could be selected)</b>						
Asthma	69	21.5	64	21.8	66	21.6
Cancer	17	5.3	17	5.8	18	5.9
Coronary heart disease	42	13.1	30	10.2	43	14.1
Depression	96	29.9	79	27.0	92	30.1
Diabetes	71	22.1	56	19.1	76	24.8
Fibromyalgia	37	11.5	34	11.6	31	10.1
Osteoarthritis	146	45.5	134	45.7	155	50.7
Osteoporosis	47	14.6	41	14.0	48	15.7
Rheumatoid arthritis*	51	15.9	32	10.9	66	21.6
Other	140	43.8	121	41.3	130	42.5

\* Significant differences at  $p < 0.05$  level (Chi-square and ANOVA, respectively)

No significant differences were found between participants of the three randomised groups with regard to course details (see Table 7).

**Table 7** Details on the courses across respondents of the randomised groups: heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314)

	heiQ-PP n=331		heiQ-PPT n=304		heiQ-PPR n=314	
	n	%	n	%	n	%
Course type						
Arthritis	57	17.6	42	14.1	40	13.0
Generic chronic disease	230	71.2	223	75.1	243	78.9
Osteoporosis	18	5.6	13	4.4	10	3.2
Other	18	5.6	19	6.4	15	4.9
Course duration						
4 weeks	20	6.0	17	5.6	10	3.2
6 weeks	279	84.3	259	85.2	279	88.9
Up to 12 weeks	32	9.7	28	9.2	25	8.0
Venues where courses were held						
Arthritis foundation	31	9.8	22	7.5	27	9.0
Community health centre	176	55.7	163	55.8	172	57.3
Hospital	39	12.3	37	12.7	38	12.7
Other	70	22.2	70	24.0	63	21.0

\* Significant differences at  $p < 0.05$  level (Chi-square)

### 3.4 Pretest and posttest scores across heiQ-PP, heiQ-PPT, and heiQ-PPR

#### 3.4.1 Specific methods

As described in the introduction of this chapter, this section attends to each group's raw data that were provided at the beginning (=pretest) and at the end of self-management courses (=posttest). While it was important for later comparisons to ensure that participants of the randomised groups had not differed at pretest, in the second part of the analyses it was explored whether the cognitive tasks that respondents had to perform at posttest (heiQ-PP; heiQ-PPT; heiQ-PPR) influenced ratings of their actual posttest levels.

Given that the heiQ data violated some assumptions of parametric tests (see Section 2.3.5), Brown-Forsythe analysis of variance (ANOVA) was applied. This method has the advantage of being robust against departure from normality as it decreases the probability of falsely rejecting the null hypothesis (Brown & Forsythe, 1974). Reported posthoc procedures were based on Tukey which is recommended when sample sizes and variances are similar (Field, 2005), as was the case in the present samples.

### 3.4.2 Results

#### *Pretests across heiQ-PP, heiQ-PPT, and heiQ-PPR*

The variances of the pretests were homogeneous across groups in all subscales. Further, the statistical analyses suggested that the pretest scores of the participants across heiQ-PP, heiQ-PPT, and heiQ-PPR did not differ. As shown in Table 8, mean pretest scores differed by no more than 0.15 on a six-point scale (see Appendix 9 for the significance tests).

**Table 8** Mean scores of pretests and actual posttests of heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314), respectively<sup>1</sup>

		heiQ-PP		heiQ-PPT		heiQ-PPR	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
1. Positive and Active Engagement in Life	Pretest	4.47	(0.92)	4.51	(1.02)	4.42	(0.98)
	Posttest*	<u>4.78</u>	<u>(0.78)</u>	<u>4.87</u>	<u>(0.71)</u>	<u>5.00</u>	<u>(0.74)</u>
2. Health-Directed Behaviour	Pretest	4.31	(1.18)	4.42	(1.12)	4.30	(1.16)
	Posttest*	<u>4.65</u>	<u>(0.98)</u>	<u>4.85</u>	<u>(0.85)</u>	<u>4.83</u>	<u>(1.00)</u>
3. Skill and Technique Acquisition	Pretest	4.08	(0.92)	4.17	(0.95)	4.10	(0.96)
	Posttest*	<u>4.64</u>	<u>(0.72)</u>	<u>4.79</u>	<u>(0.67)</u>	<u>4.90</u>	<u>(0.69)</u>
4. Constructive Attitudes and Approaches	Pretest	4.51	(0.93)	4.57	(0.96)	4.42	(1.02)
	Posttest*	<u>4.72</u>	<u>(0.85)</u>	<u>4.82</u>	<u>(0.86)</u>	<u>4.90</u>	<u>(0.86)</u>
5. Self-Monitoring and Insight	Pretest	4.73	(0.65)	4.79	(0.67)	4.74	(0.68)
	Posttest*	<u>4.96</u>	<u>(0.55)</u>	<u>5.03</u>	<u>(0.50)</u>	<u>5.16</u>	<u>(0.52)</u>
6. Health Service Navigation	Pretest	4.62	(0.90)	4.65	(0.92)	4.64	(0.95)
	Posttest*	<u>4.84</u>	<u>(0.81)</u>	<u>4.83</u>	<u>(0.86)</u>	<u>5.00</u>	<u>(0.79)</u>
7. Social Integration and Support	Pretest	4.26	(1.13)	4.27	(1.17)	4.16	(1.21)
	Posttest	4.43	(1.06)	4.53	(1.03)	4.50	(1.13)
8. Emotional Well-Being	Pretest	3.29	(1.23)	3.28	(1.26)	3.29	(1.21)
	Posttest	3.57	(1.16)	3.55	(1.22)	3.54	(1.18)

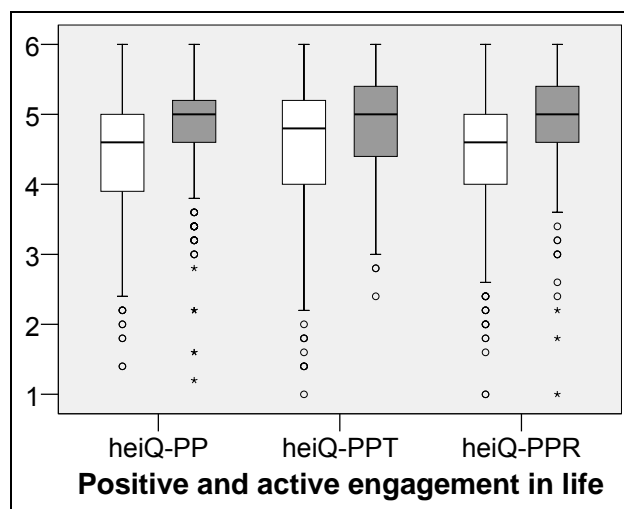
\* Significant differences between heiQ-PP, heiQ-PPT, and heiQ-PPR; robust ANOVA (Brown-Forsythe),  $p < 0.05$

<sup>1</sup> Scores of a group are underlined if they differed from one of the other groups, with lines on the same height indicating a significant difference between these groups

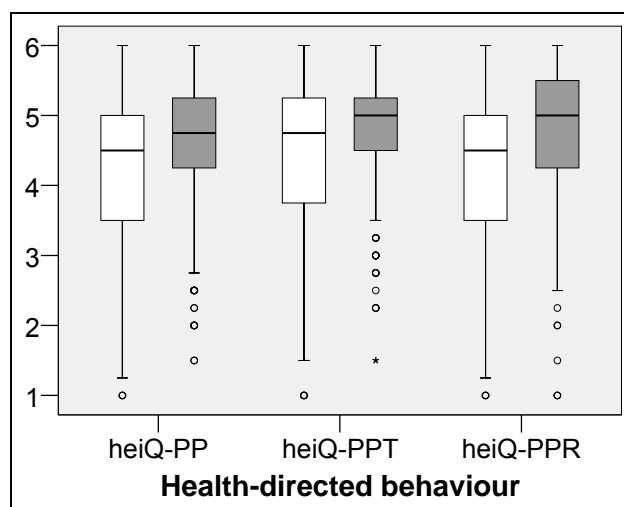
#### *Posttests across heiQ-PP, heiQ-PPT, and heiQ-PPR*

With the exception of subscale Health-Directed Behaviour, variances of actual posttests were homogeneous across groups. Overall, mean differences were found in six subscales (see Table 8). The results of these analyses are also visualised in the form of boxplots (Figure 16 to Figure 23). Although the present comparisons focused on mean scores, boxplots provide useful information on the distribution of the pre- and posttest data (Norman & Streiner, 2000).

Differences across groups in the first six subscales varied greatly. When comparing mean posttest scores of heiQ-PP with those of heiQ-PPT, the former group showed lower posttest levels than the latter group in two subscales (Health-Directed Behaviour; Skill and Technique Acquisition). In contrast, the posttest scores of heiQ-PP were significantly lower than those of heiQ-PPR in six heiQ subscales, the exceptions being Social Integration and Support, and Emotional Well-Being. Finally, when comparing heiQ-PPT with heiQ-PPR posttest scores of heiQ-PPT were found to be lower than those of heiQ-PPR in two subscales (Self-Monitoring and Insight; Health Service Navigation). To facilitate the differentiation between groups Table 8 is prepared in a way that scores of a group are underlined if they differed from one of the other groups, with lines on the same height indicating a significant difference between two groups (see Appendix 9 for the significance tests).

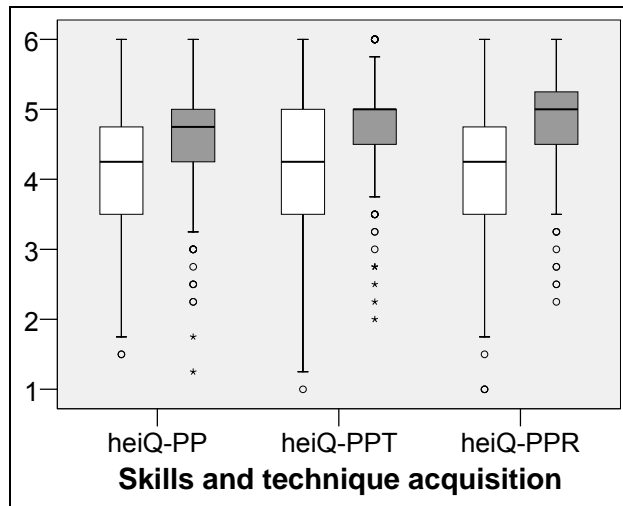


**Figure 16** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Positive and Active Engagement in Life

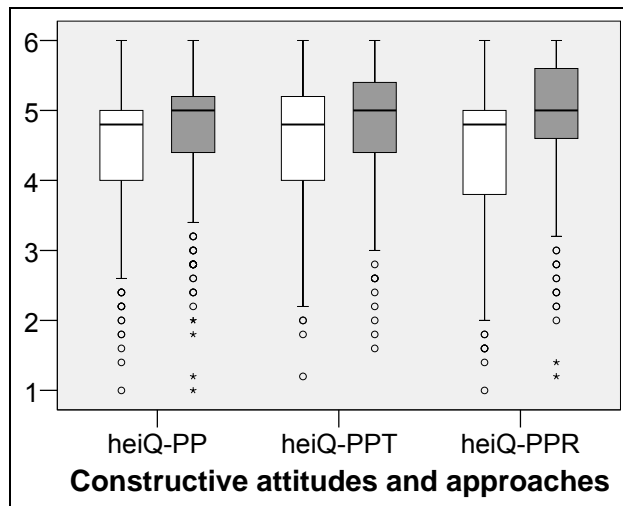


**Figure 17** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Health-Directed Behaviour

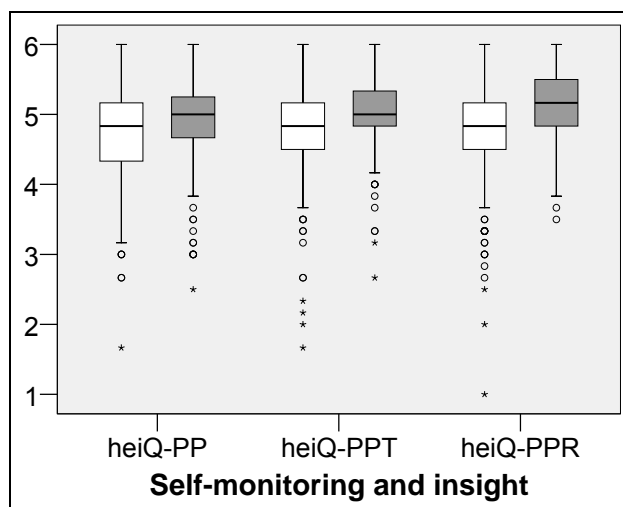




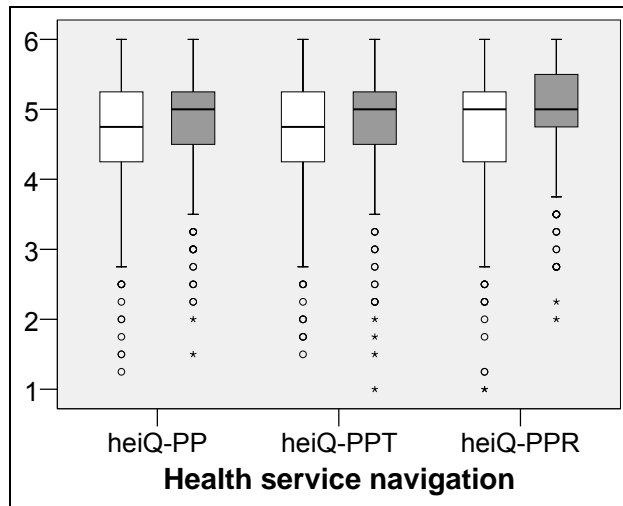
**Figure 18** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Skill and Technique Acquisition



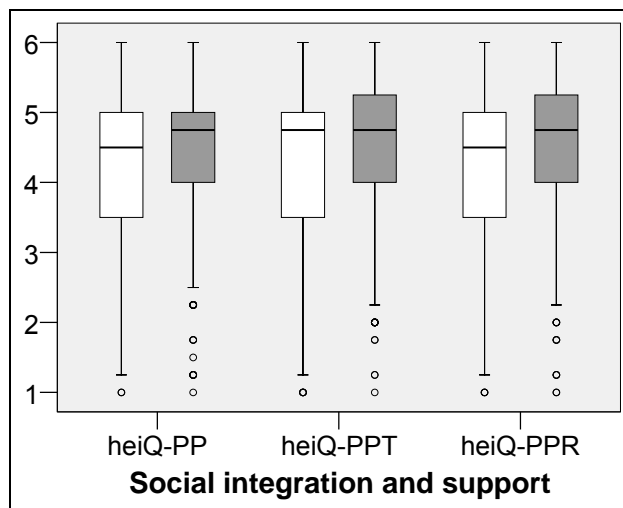
**Figure 19** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Constructive Attitudes and Approaches



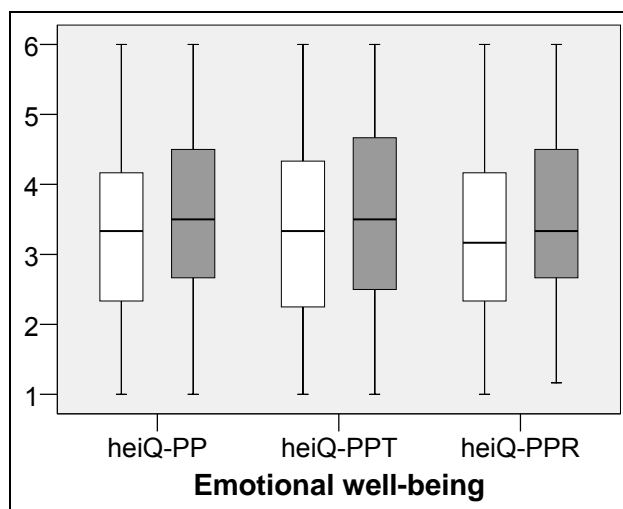
**Figure 20** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Self-Monitoring and Insight



**Figure 21** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Health Service Navigation



**Figure 22** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Social Integration and Support



**Figure 23** Pretest (white) versus posttest (grey) data across heiQ-PP, heiQ-PPT, and heiQ-PPR; Emotional Well-Being

### **3.4.3 Summary**

The previous analyses indicated some significant mean differences between actual posttest scores of heiQ-PP, heiQ-PPT, and heiQ-PPR. Given that the participants of the three groups had not differed significantly at pretest and the three posttest questionnaires were randomly distributed within self-management courses, it was assumed that the observed differences in actual posttests could be attributed to features of the posttest questionnaires. Hence, asking participants to perform a second task in addition to answering actual posttests influenced ratings of actual posttest levels.

This effect was particularly pronounced when comparing the actual posttest scores of group heiQ-PP with those of group heiQ-PPR. That is, providing retrospective pretests in addition to actual posttests influenced ratings of actual posttests in six of eight heiQ subscales, with heiQ-PPR providing significantly higher ratings of their actual posttests than heiQ-PP. This effect was less pronounced in participants who were asked to provide an assessment of their perceived change in the form of transition questions in addition to actual posttests. Group heiQ-PPT provided significantly higher ratings of their actual posttests than heiQ-PP in only two heiQ subscales. Hence, only the inclusion of a second cognitive task at posttest in the form of retrospective pretests had a substantial influence on ratings of actual posttest levels compared with people who filled out the traditional posttest questionnaire (heiQ-PP). Finally, it was observed that the actual posttest scores of heiQ-PPT – although not significantly higher than those of heiQ-PP – were elevated in a way that they were only significantly lower than those of heiQ-PPR in two heiQ subscales.

## **3.5 Change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR**

### **3.5.1 Specific methods**

In the previous Section 3.4.2 it was found that participants' ratings of their actual posttest levels were influenced by the additional tasks they were asked to perform at posttest. Given that the randomised groups had not differed at pretest, resulting mean change scores were also affected by these differences in actual posttest scores. The purpose of this section was to investigate whether these differences in computed change scores had an influence on overall conclusions about program effectiveness across groups.

This section is structured in a way that at first actual mean change scores across groups are computed. By applying robust ANOVA it was aimed to explore whether resulting differences in mean change scores would lead to a different set of conclusions about the value of self-management interventions across heiQ-PP, heiQ-PPT, and heiQ-PPR. In the second part of

this section an alternative method of presenting change was applied (see below). Given that the alternative method was based on effect sizes (ES), the concept of ES – which was briefly introduced in Section 1.2.3 – is described first.

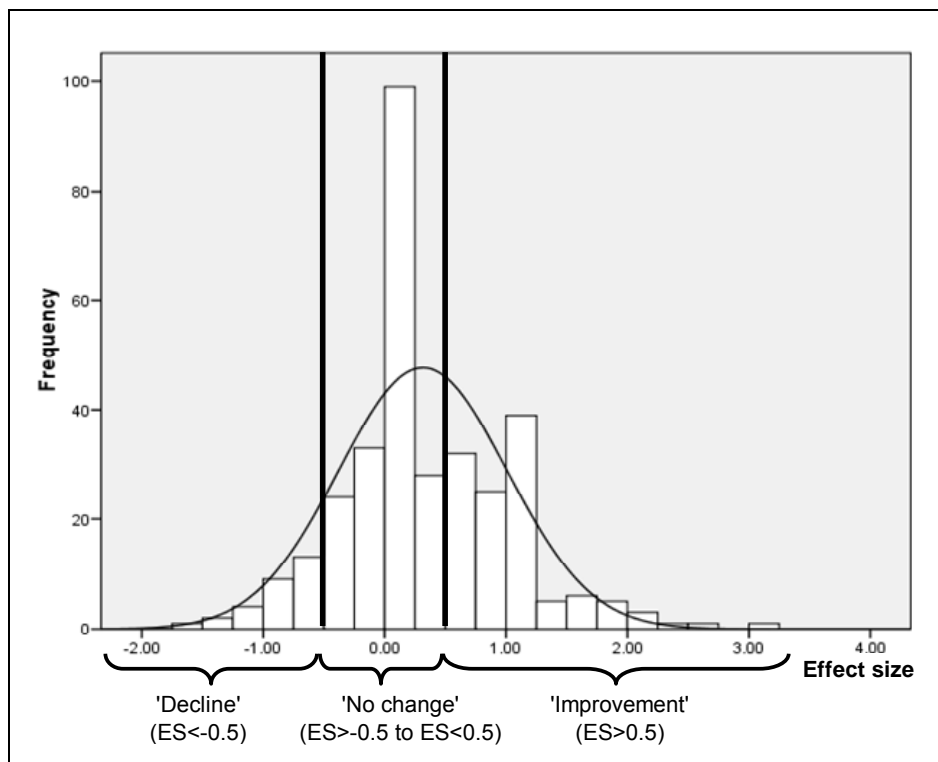
*Effect sizes* (ES) are standardised scores that belong to the family of distribution-based measures of change (Lydick & Epstein, 1993). Several ways of calculating ES exist, with Cohen's *d* (Cohen, 1988) and Hedges' *g* (Hedges & Olkin, 1985) being the most popular indices (Rosnow & Rosenthal, 1996). ES is calculated by dividing the difference between two group means either by the pooled standard deviation of the two groups at baseline (Cohen, 1988) or by the standard deviation of the control group at baseline (Guyatt *et al.*, 2002). This standardisation of change scores has several advantages: a) it allows for an easier interpretation of results, b) it facilitates the comparison of scores across different measures and/or studies, and c) ES can be used as a comparison standard to aid the interpretation of results (Kazis *et al.*, 1989). One such benchmark was introduced and applied in Section 1.2.3, with ES~0.2 interpreted as small, ES~0.5 as medium, and ES~0.8 as large effects (Cohen, 1988). This interpretation of ES is frequently applied and generally accepted, and has been supported by other authors (Samsa *et al.*, 1999). It was used throughout this study.

While Cohen's definition of ES is considered a starting point for interpreting the results of evaluations (Samsa *et al.*, 1999), its major drawback is that it obscures information on the distribution of scores. Given that a group-based measure does not provide any information about individual subjects, it has been suggested that results should be presented in terms of proportions of people reaching/exceeding a pre-defined threshold (Guyatt *et al.*, 2002). This approach is similar to *number needed to treat* (NNT) analysis (Walter, 2001), i.e. the inverse of this proportion can be translated to the number of participants needed to attend a self-management course to achieve substantial improvement in one participant (Norman, 2005; Wyrwich *et al.*, 2005). This method of presenting program outcomes is increasingly popular in areas such as clinical trials (Wyrwich *et al.*, 2005) as results are easy to interpret (Guyatt *et al.*, 2002) and also easy to communicate to stakeholders.

In this thesis, each participant's effect size – the intra-individual effect size – was calculated by using the difference in each participant's score before and after a self-management intervention divided by the standard deviation of the group's baseline scores (Guyatt *et al.*, 2002; Wyrwich *et al.*, 2005). These intra-individual effect sizes were then matched against a threshold which was set at ES=0.5. Hence, individuals who showed at least a medium effect size were considered as having benefited from attending a self-management intervention.

Given that the health of people with chronic conditions fluctuates, it was also important to record those individuals who had experienced a substantial decline during the self-management course to interpret the number of people who benefited relative to those who

declined. Hence, scores are presented in a way that  $ES < -0.5$  denotes 'decline',  $ES = -0.5$  to  $0.5$  indicates 'no change', and  $ES > 0.5$  represents 'improvement'. This method of presenting change is also illustrated in Figure 24. The area to the left of the first vertical bar shows the number of people who experienced a decline, the area between the two bars represents people who had only minimal changes or no change, while the area to the right of the second vertical bar shows the number of participants who received benefits from attending a self-management course. Further, chi-square tests were applied to determine whether the differences between the groups were significant. Because of some very low frequencies in the 'decline' category, significance tests were based on the comparison of 'improvement' with 'no improvement', i.e. the categories 'decline' and 'no change' were combined.



**Figure 24** Example of presenting proportions of people in categories of change

### 3.5.2 Results

#### *Mean change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR*

The variances of the mean change scores across the three groups were homogeneous in six of the eight subscales. As expected, some large differences in mean change scores between groups were observed with seven subscales indicating significant differences (see Table 9). To facilitate the differentiation between the three groups, Table 9 is prepared in a similar way to Table 8 in that the scores of a group are underlined if they differed from one of the other

groups with lines on the same height indicating a significant difference between two groups (see Appendix 10 for the significance tests).

Differences were again largest between heiQ-PP and heiQ-PPR with the latter showing significantly larger mean change scores compared with the former in seven heiQ subscales. In particular, Positive and Active Engagement in Life, Skill and Technique Acquisition, and Constructive Attitudes and Approaches showed some large discrepancies between scores. It was additionally observed that mean change scores of heiQ-PPT did not differ from heiQ-PP in any subscale, while mean change scores of heiQ-PPT differed significantly from heiQ-PPR in four heiQ subscales. That is, people who had filled out transition questions in addition to questions about their actual posttest levels (heiQ-PPT) showed significantly smaller effects in four subscales than participants who had filled out retrospective pretests (heiQ-PPR).

**Table 9** Comparison of mean change scores derived from pretest-posttest data across heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314)<sup>1</sup>

		heiQ-PP		heiQ-PPT		heiQ-PPR	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
1. Positive and Active Engagement	Change*	0.31	(0.67)	0.36	(0.82)	0.58	(0.78)
2. Health-Directed Behaviour	Change*	0.34	(0.89)	0.43	(1.00)	0.53	(1.00)
3. Skill and Technique Acquisition	Change*	0.56	(0.84)	0.62	(1.01)	0.80	(0.99)
4. Constructive Attitudes	Change*	0.20	(0.74)	0.25	(0.76)	0.48	(0.88)
5. Self-Monitoring and Insight	Change*	0.23	(0.60)	0.24	(0.63)	0.42	(0.69)
6. Health Service Navigation	Change*	0.22	(0.70)	0.18	(0.76)	0.36	(0.83)
7. Social Integration and Support	Change*	0.17	(0.80)	0.26	(0.85)	0.34	(0.90)
8. Emotional Well-Being	Change	0.28	(0.91)	0.27	(0.96)	0.25	(0.91)

\* Significant differences between heiQ-PPT, heiQ-PPT, and heiQ-PPR; robust ANOVA (Brown-Forsythe),  $p < 0.05$

<sup>1</sup> Scores of a group are underlined if they differed from one of the other groups with lines on the same height indicating a significant difference between these groups

#### *Decline, no change, and improvement across heiQ-PP, heiQ-PPT, and heiQ-PPR*

Again several differences between the groups were observed when subjects were grouped into the categories 'decline', 'no change', or 'improvement'. As shown in Table 10, significant differences were found in seven heiQ subscales with larger proportions of people showing substantial benefits in heiQ-PPR compared with the other two groups. In contrast, only few differences were found between heiQ-PP and heiQ-PPT.

**Table 10** Proportions of people in categories ‘decline’, ‘no change’, and ‘improvement’; total sample (n=949), and comparison heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR (n=314)

		Total	heiQ-PP	heiQ version heiQ-PPT	heiQ-PPR
<b>1. Positive and Active Engagement in Life*</b>					
Decline	n (%)	62 (6.5%)	29 (8.8%)	23 (7.6%)	10 (3.2%)
No change	n (%)	516 (54.4%)	184 (55.6%)	178 (58.6%)	154 (49.0%)
Improvement	n (%)	371 (39.1%)	118 (35.6%)	103 (33.9%)	150 (47.8%)
<b>2. Health-Directed Behaviour*</b>					
Decline	n (%)	87 (9.2%)	34 (10.3%)	31 (10.2%)	22 (7.0%)
No change	n (%)	534 (56.3%)	198 (59.8%)	171 (56.3%)	165 (52.5%)
Improvement	n (%)	328 (34.6%)	99 (29.9%)	102 (33.6%)	127 (40.4%)
<b>3. Skill and Technique Acquisition*</b>					
Decline	n (%)	93 (9.8%)	32 (9.7%)	38 (12.5%)	23 (7.3%)
No change	n (%)	300 (31.6%)	116 (35.0%)	97 (31.9%)	87 (27.7%)
Improvement	n (%)	556 (58.6%)	183 (55.3%)	169 (55.6%)	204 (65.0%)
<b>4. Constructive Attitudes and Approaches*</b>					
Decline	n (%)	93 (9.8%)	36 (10.9%)	36 (11.8%)	21 (6.7%)
No change	n (%)	535 (56.4%)	203 (61.3%)	174 (57.2%)	158 (50.3%)
Improvement	n (%)	321 (33.8%)	92 (27.8%)	94 (30.9%)	135 (43.0%)
<b>5. Self-Monitoring and Insight*</b>					
Decline	n (%)	83 (8.7%)	31 (9.4%)	31 (10.2%)	21 (6.7%)
No change	n (%)	519 (54.7%)	195 (58.9%)	174 (57.2%)	150 (47.8%)
Improvement	n (%)	347 (36.6%)	105 (31.7%)	99 (32.6%)	143 (45.5%)
<b>6. Health Service Navigation*</b>					
Decline	n (%)	140 (14.8%)	45 (13.6%)	54 (17.8%)	41 (13.1%)
No change	n (%)	453 (47.7%)	165 (49.8%)	150 (49.3%)	138 (43.9%)
Improvement	n (%)	356 (37.5%)	121 (36.6%)	100 (32.9%)	135 (43.0%)
<b>7. Social Integration and Support*</b>					
Decline	n (%)	106 (11.2%)	38 (11.5%)	40 (13.2%)	28 (8.9%)
No change	n (%)	580 (61.1%)	220 (66.5%)	177 (58.2%)	183 (58.3%)
Improvement	n (%)	263 (27.7%)	73 (22.1%)	87 (28.6%)	103 (32.8%)
<b>8. Emotional Well-Being</b>					
Decline	n (%)	144 (15.2%)	44 (13.3%)	51 (16.8%)	49 (15.6%)
No change	n (%)	499 (52.6%)	181 (54.7%)	153 (50.3%)	165 (52.5%)
Improvement	n (%)	306 (32.2%)	106 (32.0%)	100 (32.9%)	100 (31.8%)

\* Significant differences between heiQ-PP, heiQ-PPT, and heiQ-PPR; chi-squares differences based on the comparison of ‘improvement’ versus ‘no improvement’ ( $p < 0.05$ )

In a similar manner to previous findings differences were most pronounced when comparing heiQ-PP with heiQ-PPR, with six subscales showing pronounced differences in proportions of participants indicating ‘improvement’ (see Appendix 11 for the significance tests).

### **3.5.3 Summary**

The analyses of the present section largely confirmed the findings of the preceding Section 3.4 in that a second cognitive task at posttest not only influenced participants’ ratings of their actual posttest levels but also resulting mean change scores. In particular, asking subjects to provide ratings of their pretest levels in retrospect in addition to ratings of their actual posttest levels, led to significantly larger mean change scores in seven subscales when compared with scores of subjects who did not have to perform an additional task at posttest (heiQ-PP). In contrast, when participants were asked to provide a direct assessment of their perceived change in addition to ratings of their actual posttests (heiQ-PPT), mean change scores were slightly larger; however, these mean differences were not statistically significant when compared with change scores of heiQ-PP. Further, while the comparison of posttest levels of heiQ-PPT and heiQ-PPR had indicated only two significant differences, mean change scores of heiQ-PPR were significantly larger than those of heiQ-PPT in half of the heiQ subscales.

Differences across groups were not only observed in computed mean change scores but also when people were classified as ‘decline’, ‘no change’, or ‘improvement’. In most heiQ subscales the proportions of participants in the ‘improvement’ category were substantially higher in heiQ-PPR compared with either heiQ-PP or heiQ-PPT.

## **3.6 Change scores across heiQ-PPR and heiQ-PPR Retro**

### **3.6.1 Specific methods**

In this section mean change scores based on retrospective pretest and actual posttest data are reported. As these data were only available from heiQ-PPR (see Section 2.2.3), the analyses were based on the comparison of dataset heiQ-PPR Retro with this group’s dataset of actual pre- and posttests (heiQ-PPR). The analyses were aimed at investigating whether respondents had provided congruent change scores across heiQ-PPR and heiQ-PPR Retro. It was explored whether the collection of pretest data at two different points in time, i.e. at the beginning (=actual) and at the end (=retrospective) of courses, resulted in different change scores. Given that scores were based on identical posttest data, they were equally affected by the comparatively high levels of actual posttest scores of heiQ-PPR (see Section 3.4.2).



In a similar manner to the comparison of the randomised groups, the first part of the section compares mean scores of the raw data of heiQ-PPR, i.e. actual pretest data were compared with retrospective pretest data (paired t-tests). In the next step, mean change scores were computed. The comparisons of change scores across heiQ-PPR and heiQ-PPR Retro were again based on a) mean change scores (paired t-tests) and b) proportions of participants in the three categories 'decline', 'no change', and 'improvement'. Because of small frequencies in the 'decline' category, 'decline' and 'no change' were again combined for the significance tests (see Section 3.5).

### 3.6.2 Results

#### *Actual versus retrospective pretests across heiQ-PPR and heiQ-PPR Retro*

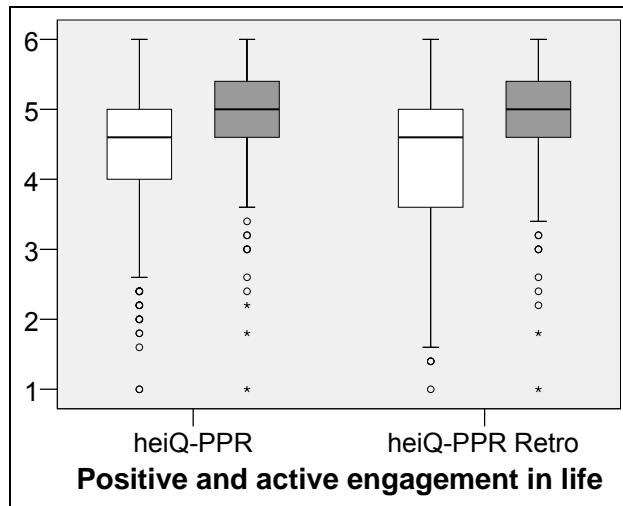
Paired t-tests showed some significant differences in scores with retrospective pretests being significantly lower than actual pretests in three heiQ subscales. As presented in Table 11, Health-Directed Behaviour however was the only heiQ subscale with pronounced differences between the two scores. All remaining subscales showed differences no larger than 0.12.

**Table 11** Actual and retrospective pretest data of group heiQ-PPR (n=314)

	Actual pretest		Retrospective pretest	
	Mean	(SD)	Mean	(SD)
1. Positive and Active Engagement in Life*	4.42	(0.98)	4.32	(1.08)
2. Health-Directed Behaviour*	4.30	(1.16)	4.01	(1.21)
3. Skill and Technique Acquisition	4.10	(0.96)	4.13	(0.94)
4. Constructive Attitudes and Approaches	4.42	(1.02)	4.38	(1.09)
5. Self-Monitoring and Insight*	4.74	(0.68)	4.62	(0.78)
6. Health Service Navigation	4.64	(0.95)	4.66	(1.01)
7. Social Integration and Support	4.16	(1.21)	4.18	(1.27)
8. Emotional Well-Being	3.29	(1.21)	3.20	(1.24)

\* Significant differences between actual and retrospective pretest; Paired t-test,  $p < 0.05$

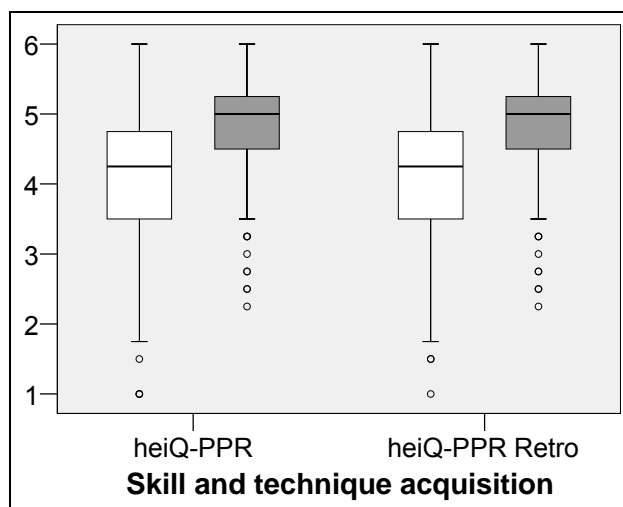
The results of these comparisons of actual and retrospective pretests of group heiQ-PPR are also visualised in Figure 25 to Figure 32. Again boxplots were chosen to present the data (see Section 3.4.2). For completeness, actual and retrospective pretest scores are presented along with actual posttest scores as the latter were used to compute the respective change scores. As all data were provided from the same subjects, these posttests, however, are identical across heiQ-PPR and heiQ-PPR Retro (see grey-shaded boxplots).



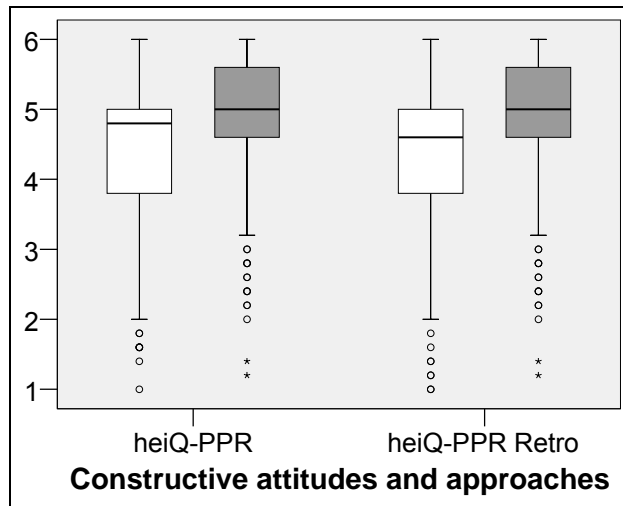
**Figure 25** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Positive and Active Engagement in Life



**Figure 26** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Health-Directed Behaviour



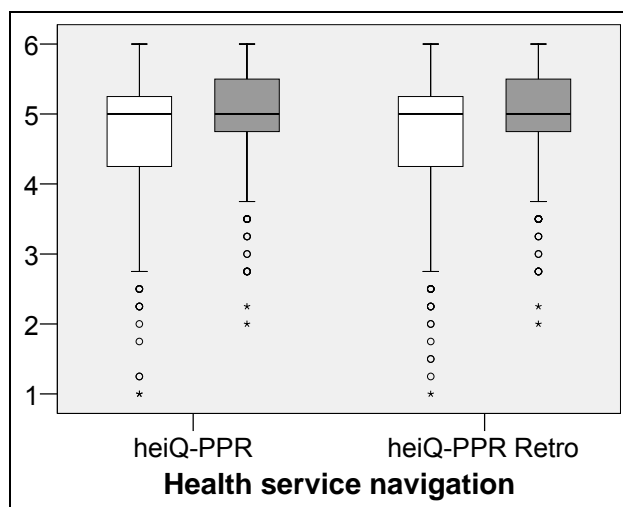
**Figure 27** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Skill and Technique Acquisition



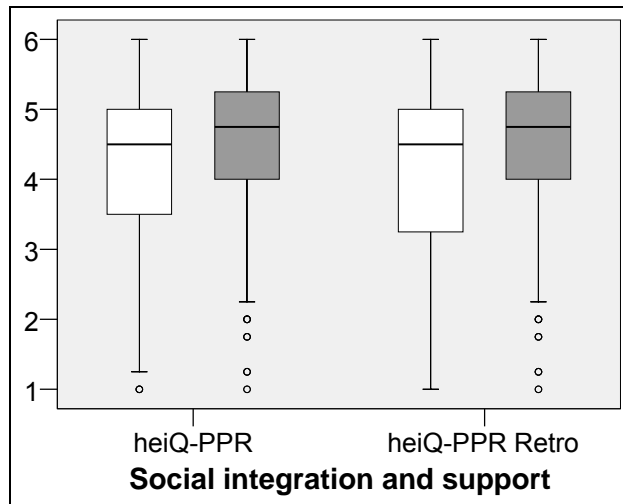
**Figure 28** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Constructive Attitudes and Approaches



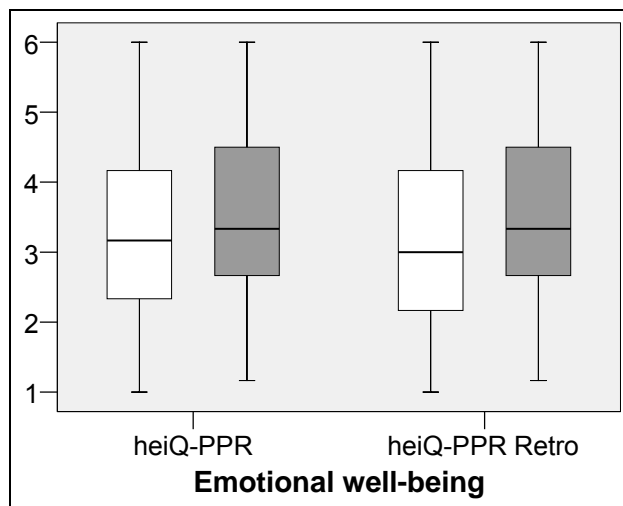
**Figure 29** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Self-Monitoring and Insight



**Figure 30** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Health Service Navigation



**Figure 31** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Social Integration and Support



**Figure 32** Actual/retrospective pretest (white) versus posttest (grey) data of group heiQ-PPR; Emotional Well-Being

*Mean change scores across heiQ-PPR and heiQ-PPR Retro*

As a consequence of the observed differences in actual and retrospective pretest scores, paired t-tests showed that change scores differed significantly in three of the eight subscales. Change scores derived from retrospective pretests were significantly larger in Positive and Active Engagement in Life, Health-Directed Behaviour, and Self-Monitoring and Insight (see Table 12). As observed previously, only heiQ subscale Health-Directed Behaviour however indicated large discrepancies between the two datasets with a mean change of 0.53 for heiQ-PPR, compared with 0.82 for heiQ-PPR Retro. The remaining seven subscales showed differences no larger than 0.11 (see Appendix 12 for the significance tests).

**Table 12** Comparison of mean change scores derived from actual pretest-posttest data (heiQ-PPR) and retrospective pretest-posttest data (heiQ-PPR Retro) of group heiQ-PPR (n=314)

		heiQ-PPR		heiQ-PPR Retro	
		Mean	(SD)	Mean	(SD)
1. Positive and Active Engagement in Life	Change*	0.58	(0.78)	0.68	(0.86)
2. Health-Directed Behaviour	Change*	0.53	(1.00)	0.82	(0.90)
3. Skill and Technique Acquisition	Change	0.80	(0.99)	0.77	(0.87)
4. Constructive Attitudes and Approaches	Change	0.48	(0.88)	0.52	(0.75)
5. Self-Monitoring and Insight	Change*	0.42	(0.69)	0.53	(0.69)
6. Health Service Navigation	Change	0.36	(0.83)	0.34	(0.68)
7. Social Integration and Support	Change	0.34	(0.90)	0.32	(0.62)
8. Emotional Well-Being	Change	0.25	(0.91)	0.34	(0.74)

\* Significant differences between heiQ-PPR and heiQ-PPR Retro; Paired t-test,  $p < 0.05$

#### *Decline, no change, and improvement across heiQ-PPR and heiQ-PPR Retro*

When grouping respondents into ‘decline’, ‘no change’, or ‘improvement’, three subscales suggested differences between heiQ-PPR and heiQ-PPR Retro. However, in contrast to the previous presentation of mean change scores, only one of these indicated a larger proportion of subjects in the ‘improvement’ category in heiQ-PPR Retro, while the other two subscales showed a larger proportion of participants in the ‘improvement’ category in heiQ-PPR. As presented in Table 13, the largest discrepancy in scores was observed in Health Service Navigation with 43.0% of participants indicating ‘improvement’ in heiQ-PPR compared with 19.1% of subjects in heiQ-PPR Retro. Despite this somewhat inconsistent pattern in results across the two datasets, none of the remaining five heiQ subscales suggested differences between heiQ-PPR and heiQ-PPR Retro (see Appendix 13 for the significance tests).

In contrast to the distribution of proportions in the ‘improvement’ category, the distribution of scores in the ‘decline’ category showed a more consistent pattern between the two datasets. Across subscales on average 8.6% of participants were classified as ‘decline’ in heiQ-PPR, whereas substantially fewer subjects (1.7% across subscales) were classified as ‘decline’ in heiQ-PPR Retro. Hence, in retrospect fewer participants indicated ‘decline’ compared with mean change scores derived from actual pretests that participants had provided at the start of a self-management intervention. The smaller proportion of participants indicating ‘decline’ in retrospect, however, resulted in a larger proportion of people in the ‘no change’ category rather than observing more participants in the ‘improvement’ category in heiQ-PPR Retro compared with heiQ-PPR. A more detailed illustration of the distribution of the scores of the two datasets is presented in Appendix 14 in the form of histograms.

**Table 13** Proportions of people in categories 'decline', 'no change', and 'improvement' across heiQ-PPR and heiQ-PPR Retro (n=314)

		heiQ-PPR		heiQ-PPR Retro	
<b>1. Positive and Active Engagement in Life</b>					
Decline	n (%)	10	(3.2%)	5	(1.6%)
No change	n (%)	154	(49.0%)	162	(51.6%)
Improvement	n (%)	150	(47.8%)	147	(46.8%)
<b>2. Health-Directed Behaviour*</b>					
Decline	n (%)	22	(7.0%)	3	(1.0%)
No change	n (%)	165	(52.5%)	157	(50.0%)
Improvement	n (%)	127	(40.4%)	154	(49.0%)
<b>3. Skill and Technique Acquisition</b>					
Decline	n (%)	23	(7.3%)	9	(2.9%)
No change	n (%)	87	(27.7%)	119	(37.9%)
Improvement	n (%)	204	(65.0%)	186	(59.2%)
<b>4. Constructive Attitudes and Approaches</b>					
Decline	n (%)	21	(6.7%)	4	(1.3%)
No change	n (%)	158	(50.3%)	191	(60.8%)
Improvement	n (%)	135	(43.0%)	119	(37.9%)
<b>5. Self-Monitoring and Insight</b>					
Decline	n (%)	21	(6.7%)	5	(1.6%)
No change	n (%)	150	(47.8%)	167	(53.2%)
Improvement	n (%)	143	(45.5%)	142	(45.2%)
<b>6. Health Service Navigation*</b>					
Decline	n (%)	41	(13.1%)	4	(1.3%)
No change	n (%)	138	(43.9%)	250	(79.6%)
Improvement	n (%)	135	(43.0%)	60	(19.1%)
<b>7. Social Integration and Support*</b>					
Decline	n (%)	28	(8.9%)	3	(1.0%)
No change	n (%)	183	(58.3%)	251	(79.9%)
Improvement	n (%)	103	(32.8%)	60	(19.1%)
<b>8. Emotional Well-Being</b>					
Decline	n (%)	49	(15.6%)	10	(3.2%)
No change	n (%)	165	(52.5%)	225	(71.7%)
Improvement	n (%)	100	(31.8%)	79	(25.2%)

\* Significant differences between heiQ-PPR and heiQ-PPR Retro; chi-square tests based on comparison of 'improvement' versus 'no improvement' ( $p < 0.05$ )

### 3.6.3 Summary

The analyses of this section suggested that mean pretest scores derived from participants in retrospect, i.e. at the end of self-management courses, were significantly lower in three of the eight heiQ subscales than pretest scores assessed at the start of self-management courses. Consequently, computed change scores based on retrospective pretests (heiQ-PPR Retro) were significantly larger in these three subscales compared with those derived from actual pretests (heiQ-PPR). Compared with the analyses in Section 3.5, however, fewer differences between the datasets were observed which were also smaller in magnitude than differences between heiQ-PP, heiQ-PPT, and heiQ-PPR.

When choosing an alternative method of presenting change by way of computing proportions of people in different categories of change, a somewhat inconsistent pattern was observed. In two heiQ subscales heiQ-PPR showed significantly larger proportions of participants in the 'improvement' category, whereas in one heiQ subscale heiQ-PPR Retro showed significantly larger proportions in this category. In a similar manner to the computed mean change scores, five heiQ subscales did not indicate any differences between heiQ-PPR and heiQ-PPR Retro and again observed differences were less pronounced than those observed in Section 3.5.

## 3.7 Discussion

### *Differences in mean change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR*

In the first two sections of the present chapter it was found that mean levels of actual posttest scores and resulting mean change scores were consistently largest for heiQ-PPR. Given that subjects of the randomised groups had not differed at pretest, observed differences between groups could therefore be attributed to the specific questionnaire design of heiQ-PPR. Hence, asking participants to provide ratings of their pretest levels in retrospect in addition to ratings of their actual posttest levels led to significantly higher posttest scores. This resulted in mean change scores being significantly larger in seven of eight heiQ subscales compared with those of subjects who did not have to perform an additional task at posttest (heiQ-PP). In contrast, when participants were asked to provide a direct assessment of their perceived change in addition to actual posttests (heiQ-PPT), their mean change scores were slightly but not significantly larger than those of heiQ-PP. Finally, mean change scores of heiQ-PPT differed significantly from those of heiQ-PPR in four of eight subscales (see Section 3.5).

This discussion focuses on explanations why larger mean change scores were regularly observed in heiQ-PPR. Given that effects were substantially less pronounced in heiQ-PPT, the observed findings must be related to the nature of the second cognitive task rather than

the presence of a second cognitive task. As described in Section 2.2.3, subjects of heiQ-PPR filled out each questionnaire item first with reference to their current state (=actual posttest) and then with reference to their past state (=retrospective pretest). Although there had not been any reference to 'change' in the instruction to this posttest heiQ (see Appendix 5), it can be assumed that most respondents were soon aware that they were providing an indirect assessment of change. Hence, they may have provided respective self-ratings relative to each other, i.e. retrospective pretest levels relative to posttest levels and/or vice versa.

The following explanations for the consistently larger mean change scores observed in group heiQ-PPR across seven heiQ subscales are proposed:

- a) It is plausible that the cognitive task related to questionnaire heiQ-PPR, i.e. differentiating between current (=posttest) and past (=retrospective) states, was more challenging than the tasks of the other two groups. Research has suggested that greater task difficulties are related to increased *satisficing* (Krosnick, 1999; Krosnick & Alwin, 1987; Lam & Bengo, 2003). That is, where a cognitive task is too demanding, respondents provide an answer that they believe is 'satisfactory' instead of *optimising* their answer (Krosnick, 1991, 1999; Krosnick & Alwin, 1987). Hence, they do not engage in all four steps of the response process as defined in Section 1.2.3.3 (Krosnick, 1999; Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988). In the present study it is plausible that at least some of the respondents had difficulties when filling out heiQ-PPR. In the context of 'satisfactory answer' it could then be assumed that these subjects would have inadvertently provided ratings of posttest levels that were higher than their 'true' levels (see Section 2.2.3).
- b) It is plausible that the higher posttest levels of group heiQ-PPR may have been caused by *social desirability* (Crowne & Marlowe, 1964; Paulhus, 1991). Knowing that they were providing an indirect assessment of their change and therefore a rating of the quality of the course and the course leader, participants may have felt inclined to attest improvement. To please the course leaders they may have increased their posttest levels beyond their 'true' levels. Social desirability bias has also been found to be associated with the previous strategy to *satisfice* (Lam & Bengo, 2003).

In contrast to group heiQ-PPR, participants who provided answers to transition questions (heiQ-PPT) may have responded to the transition questions in a *socially desirable* way, whereas their actual posttest scores may have been unaffected by this bias. As actual posttest scores were not provided in direct relation to another score such as retrospective pretest scores in heiQ-PPR, participants' responses to the posttest questions may have been independent of their responses to the transition questions.



- c) When people participate in an intervention, they invest at least some energy, time, or money to be able to attend (Howard, Ralph *et al.*, 1979). As a consequence, it can be assumed that these participants would expect to gain at least some benefits from their attendance. When filling out a questionnaire such as heiQ-PPR, they may then realise that their perceived posttest level is not much higher than their perceived retrospective pretest level. Hence, they may feel that there is a gap between their personal investment and the benefits they thought they should have received. If this gap is large it may lead to conflicting cognitions between perceived benefits and expected benefits that can cause a *cognitive dissonance* (Aronson & Mills, 1959; Festinger, 1957; Hill & Betz, 2005). As a result of trying to avoid conflicting cognitions (Hill & Betz, 2005), some respondents may have increased their ratings of their posttest levels relative to their retrospective pretest levels to make themselves feel comfortable that the self-management course indeed had been valuable and useful. This effect has also been referred to as *effort justification bias* (Aronson & Mills, 1959; Hill & Betz, 2005; Sprangers & Hoogstraten, 1988).
- d) Finally, it is possible that *response shift bias* may have influenced people's ratings. With reference to previous research on a sample that was comparable to the study sample, it can be assumed that at least some of the subjects of the present sample experienced a response shift between pretest and posttest (Osborne *et al.*, 2006). In view of the random allocation of the posttest questionnaires in this research, it can be assumed that similar proportions of participants across groups had such a response shift. As a result of this change in subjects' internal scale between actual pretest and posttest, a comparison of the two scores may be confounded (Howard & Dailey, 1979). That is, if the heiQ items at the end of the self-management course were filled out from a different perspective than the heiQ items at the beginning of the course, resulting change scores would be invalid (see Section 1.2.4 for a review of response shift bias).

When interpreting the present findings in the context of response shift theory, it could be inferred that all mean change scores derived from actual pretest-posttest data were affected by subjects who had a response shift. However, considering that participants were confronted with different cognitive tasks when filling out the posttest heiQ, it is possible that confounding through response shift differed across groups. For example, it is possible that the task of providing ratings of retrospective pretest levels simultaneous to ratings of actual posttest levels (heiQ-PPR) made participants provide the ratings of their posttest levels relative to their perceived retrospective pretest levels. Hence, providing ratings of retrospective pretest levels at posttest may have alleviated a response shift effect on ratings of actual posttest levels. Given that subjects of the other groups were not asked to provide a retrospective rating of their pretest levels at posttest, it is assumed that they did not interpret posttest questions in terms of their hypothetical pretest levels. If

these assumptions hold, then the findings suggest that response shift bias did not affect posttest scores of heiQ-PPR as severely as those of heiQ-PP and heiQ-PPT.

In summary, the design of the posttest heiQs influenced ratings of actual posttest levels in a way that mean change scores were consistently larger in heiQ-PPR compared with the other two groups. The observed influence of the design of the posttest questionnaire was so strong that different conclusions about program effectiveness would be derived. When interpreting change scores of both heiQ-PP and heiQ-PPT, it might be concluded that self-management courses have small impacts, whereas change scores of heiQ-PPR suggest medium effects.

However, it remains to be explored whether the higher posttest levels of heiQ-PPR are a more accurate reflection of subjects' levels at posttest relative to their actual pretest levels or whether the simultaneous assessment of posttest and retrospective pretest data may have confounded the self-ratings (Randolph & Elloy, 1989; van de Vliert *et al.*, 1985). As described in the literature review (see Section 1.2.4.4), the few studies that investigated this potential dependency of the scores did not find evidence of such dependency (Howard, Ralph *et al.*, 1979; Sprangers & Hoogstraten, 1989; Terborg & Davis, 1982). Nevertheless, in this study an interdependence of the scores cannot be ruled out considering the observed differences in actual posttest levels, in particular between heiQ-PP and heiQ-PPR.

#### *Decline, no change, and improvement across heiQ-PP, heiQ-PPT, and heiQ-PPR*

Following the observed differences in mean change scores across groups, the comparison of proportions of participants in categories 'decline', 'no change', and 'improvement' suggested similar results. With the exception of heiQ subscale Emotional Well-Being, the proportions of participants showing 'improvement' were consistently larger in heiQ-PPR compared with either heiQ-PP or heiQ-PPT (see Section 3.5).

Because of the different way of analysing, presenting, and interpreting the data, the overall conclusions about program effectiveness are not as strongly affected as the comparisons of mean change scores. Despite significant chi-square tests, results across groups showed a largely similar finding with about one third to one half of subjects indicating 'improvement'. In the context of *NNT*, these results can be translated to the number of people needed to attend a self-management course to achieve one participant reaching or exceeding the pre-defined threshold of  $ES=0.5$  (Norman, 2005; Wyrwich *et al.*, 2005). In the broader context of program effectiveness, all datasets indicated that self-management interventions were effective with two to three people needed to attend a course to achieve a substantial improvement in one

participant<sup>12</sup>. Hence, when reporting proportions of participants in pre-defined categories of change, conclusions about program effectiveness seem to be less affected by the design of the posttest heiQ than was observed for mean change scores.

#### *Differences in mean change scores across heiQ-PPR and heiQ-PPR Retro*

In addition to comparing mean change scores across the randomised groups, actual mean change scores of group heiQ-PPR were compared with this group's change scores based on retrospective pretest and posttest data (heiQ-PPR Retro). It was found that in three subscales change scores derived from retrospective pretest data were significantly larger than those derived from actual pretest data. Considering that the same posttest scores were applied to compute the two change scores, these differences were caused by differences in the pretest scores, i.e. in retrospect subjects reported having been worse compared with their ratings of the same items at actual pretest in three of eight heiQ subscales.

Similar explanations to those proposed for the higher posttest levels of group heiQ-PPR may be relevant in the present case. The following explanations for the few lower retrospective pretest scores compared with actual pretest scores are proposed:

- a) It is possible that the presumably greater task difficulty for subjects of group heiQ-PPR led to increased *satisficing* (Krosnick, 1999; Krosnick & Alwin, 1987; Lam & Bengo, 2003). In addition to or instead of providing higher posttest levels, participants may have provided their retrospective pretest levels in a way that they believed was 'satisfactory' (in terms of 'sufficient') for the researcher (Krosnick, 1991, 1999; Krosnick & Alwin, 1987). In the present case, it can be assumed that 'satisfactory' is equivalent to providing lower ratings of pretest levels in retrospect.
- b) By providing ratings of retrospective pretest levels in close proximity to the actual posttest questions, answers to the retrospective pretests may have also been affected by *socially desirable* responses (Crowne & Marlowe, 1964; Paulhus, 1991). Hence, relative to their posttest levels, respondents may have felt that it is socially desirable if their retrospective pretest levels were the same or below their posttest levels to avoid reporting decline. While previous research has suggested that retrospective pretests may be even less vulnerable to social desirability (Howard *et al.*, 1981), this confounding cannot be ruled out in this type of questionnaire design.

---

<sup>12</sup> Osoba (personal communication, August 9, 2006) suggested that cancer treatments are considered effective if as many as ten patients are treated to achieve benefits in one patient. In the context of self-management programs he deemed programs that reach a ratio of 3:1 as beneficial.

- c) It is also conceivable that lower retrospective compared with actual pretest scores are a result of *effort justification bias* (Aronson & Mills, 1959; Hill & Betz, 2005). By providing low retrospective pretest levels relative to their actual posttest levels, respondents may have wanted to make themselves feel comfortable that the intervention was beneficial.
- d) Somewhat related to effort justification bias is the possibility that some of the subjects based their responses on an *implicit theory of stability or change* (see Section 1.2.4.3) By comparing their current state with their past state they may have assessed if a difference was apparent (Ross, 1989; Schwarz *et al.*, 1998). By using their actual posttest levels as a benchmark, they then constructed their retrospective pretest levels in a way that the difference resulted in stability or change. This theory also highlights that people generally infer from present states rather than recalling past states (Norman, 2003; Ross, 1989).
- e) Finally, the appropriateness of using retrospective pretest scores shall be discussed. As described in Section 1.2.4.4, these scores are frequently criticised for being vulnerable to *recall bias*, i.e. scores are considered unreliable which is founded on the assumption that people are unable to accurately recall their past states (Loftus *et al.*, 1991; Pearson *et al.*, 1992; Ross, 1989). In contrast, given that the comparison of actual pretests and posttests may be biased through a change in participants' perspective between the two ratings (Howard & Dailey, 1979), retrospective pretest data are used as a *remedy to circumvent response shift bias*. Pretest data that participants provide in retrospect are built on the assumption that ratings of these questions are provided from the same perspective as the one underlying the ratings of the actual posttest items (Howard, Ralph *et al.*, 1979; Sprangers, 1989). Although neither of these theories explains the magnitude of the respective ratings of actual and retrospective pretests, both must be taken into account when considering retrospective pretest data as a substitute for actual pretest data.

In summary, when comparing mean change scores derived from actual pretests (heiQ-PPR) with those derived from retrospective pretests (heiQ-PPR Retro), only few differences were observed. While three subscales showed statistically significant differences between the two measures of change – with heiQ-PPR Retro indicating larger effects – only subscale Health-Directed Behaviour suggested substantially different results. That is, in retrospect people felt that their health behaviours were much worse before the course than they had indicated at actual pretest. With regard to overall conclusions about program effectiveness, differences between heiQ-PPR and heiQ-PPR Retro were relatively small across all eight subscales. That is, regardless of the method used, results suggest medium effects of self-management interventions. However, for further interpretation of the results it has to be taken into account that change scores of both heiQ-PPR and heiQ-PPR Retro were derived from the same high posttest scores.

This finding is partially in line with previous research in this area. It has been reported that about half of those studies that compared change scores based on retrospective as opposed to actual pretest scores showed similar results across methods. The other half of the studies, however, observed significantly larger change scores for the retrospective method (Terborg *et al.*, 1980). While the effect of the posttest design on actual posttest levels was discussed in the context of heiQ-PP, heiQ-PPT, and heiQ-PPR, previous research has also investigated the reverse, i.e. the influence of the posttest levels on retrospective pretest levels. Although these studies did not observe such influence (Sprangers & Hoogstraten, 1989; Terborg & Davis, 1982), it is possible that this was the case in the present study. That is, in heiQ-PPR, posttest levels may have been influenced by people's ratings of their retrospective pretest levels, retrospective pretest levels may have been influenced by the posttest levels, or both ratings may have influenced each other.

#### *Decline, no change, and improvement across heiQ-PPR and heiQ-PPR Retro*

Similar to the observed differences between the two computed mean change scores of group heiQ-PPR, the majority of subscales did not suggest any difference between heiQ-PPR and heiQ-PPR Retro when looking at proportions of participants in different categories of change. The three subscales that indicated statistically significant differences, however, suggested a somewhat inconsistent pattern. When scores were derived from retrospective pretest data (heiQ-PPR Retro), the proportions of people in the 'improvement' category were significantly larger in Health-Directed Behaviour. In contrast, in subscales Health Service Navigation and Social Integration and Support the proportions of participants in the 'improvement' category were significantly smaller when scores were derived from heiQ-PPR Retro.

While these results appear to be in contrast to observed differences in mean change scores, they need to be explored more closely. Firstly, across both methods of reporting change, i.e. mean change scores and proportions of people in categories of change, the majority of heiQ subscales did not indicate a significant difference between heiQ-PPR and heiQ-PPR Retro. Hence, overall conclusions about program effectiveness are largely the same across the two methods of presenting program outcomes in that the application of either method (heiQ-PPR or heiQ-PPR Retro) would lead to similar conclusions. Secondly, of those subscales that had shown significant differences between mean change scores, only one showed a substantial difference of 0.29 (Health-Directed Behaviour), whereas the remaining two indicated a rather small difference (Positive and Active Engagement in Life; Self-Monitoring and Insight) of 0.10 and 0.11, respectively. The alternative method of presenting change also found significant differences between heiQ-PPR and heiQ-PPR Retro in subscale Health-Directed Behaviour, i.e. again the two methods of presenting results are not as dissimilar as it initially appears.

Despite this closer examination of results it however needs to be discussed why significantly larger proportions of participants in the 'improvement' category were observed in heiQ-PPR as opposed to heiQ-PPR Retro. That is, it was somewhat surprising to observe significant differences in the proportion of participants in the 'improvement' category in Health Service Navigation, and Social Integration and Support given that the groups' mean change scores had been almost identical (see Table 12). The explanation can be found in the distribution of the scores across heiQ-PPR and heiQ-PPR Retro (see Table 13). While heiQ-PPR showed on average 8.6% of participants in the 'decline' category, heiQ-PPR Retro indicated substantially fewer subjects in this category (1.7% on average). Therefore, the respective distribution of scores of heiQ-PPR and heiQ-PPR Retro explains why different conclusions can be obtained when applying two methods of presenting change (see also the histograms illustrated in Appendix 14). The observation that people hardly report decline in retrospect has also been found in previous studies (Howard & Dailey, 1979).

In summary, given that five subscales had not indicated any significant difference between effects as shown by heiQ-PPR and heiQ-PPR Retro respectively, overall conclusions about program effectiveness are similar, i.e. they are largely independent of the method used. In terms of NNT it can again be concluded that self-management interventions are generally effective across most heiQ subscales. Without affecting overall conclusions about program effectiveness, it was however also found that the retrospective method suggested that up to five people are needed to attend a self-management course to achieve substantial improvement in one person in the subscales Health Service Navigation, and Social Integration and Support, i.e. heiQ-PPR Retro suggested somewhat less positive effects in these areas compared with heiQ-PPR.

### *Concluding remarks*

The analyses of this chapter led to the following results with regard to the research questions posed in Section 1.3:

- I. Does the application of differently designed questionnaires at posttest alter conclusions about the value of programs when effectiveness is assessed from change scores derived from pretest and posttest measures?
  - The inclusion of retrospective pretest questions at posttest (heiQ-PPR) influenced results in a way that conclusions about program effectiveness are different to those that would be drawn from traditional pretest-posttest data (heiQ-PP). Effects are less

pronounced when comparing heiQ-PP with heiQ-PPT, i.e. change scores of subjects who provided transition questions in addition to actual posttest levels (heiQ-PPT).

II. Are conclusions about program effectiveness different when deriving change scores from retrospective in place of actual pretest data (heiQ-PPR; heiQ-PPR Retro)?

- The conclusions about program effectiveness are largely independent of the method used, i.e. independent of using retrospective pretest data as opposed to actual pretest data to calculate change scores. For further interpretation of these results, however, it has to be taken into account that both change measures are derived from relatively high posttest scores of group heiQ-PPR.

While the analyses of this chapter provided answers to the above research questions, the findings raised further questions that need to be addressed in subsequent analyses. In particular, the observed differences in actual posttest scores across heiQ-PP, heiQ-PPT, and heiQ-PPR are an important finding as – to the author’s knowledge – the present study is the first in this area to investigate effects of the questionnaire design on reported levels. The following list provides a summary of possible explanations for the obtained results:

- *Satisficing*;
- *Social desirability*;
- *Effort justification bias*;
- *Implicit theory of stability or change*;
- A diminished occurrence of *response shift bias* in scores of heiQ-PPR.

It is only feasible to explore some of these in this thesis. Given that research into *satisficing*, *effort justification*, and *implicit theory of stability or change* require data not collected in the present study – in particular qualitative data – the exploration of these explanations must be left for future research. The available data and statistical methods, however, enable the exploration of the role of *response shift* and *social desirability biases*, with the former being investigated in Chapter 5, while the latter is examined in Chapter 6. Given that the statistical techniques applied in both Chapters 5 and 6 are more sophisticated than those applied in the present chapter and require a measurement instrument with good psychometric properties, Chapter 4 provides an introduction to these techniques as well as a re-validation of the heiQ before carrying out further analyses.

# Chapter 4

## Statistical methods and the re-validation of the heiQ



## 4 Statistical methods and the re-validation of the heiQ

### 4.1 Introduction

This chapter introduces a family of procedures based on factor analysis which were used for the re-validation of the heiQ and formed the basis of the statistical analyses applied in Chapters 5 and 6. Given that the analyses are based on data that were measured on a six-point ordinal scale (see Section 2.2.3), the specific issues related to the analysis of this type of data are also discussed. The chapter concludes with the re-validation of the heiQ.

### 4.2 Factor analysis

#### 4.2.1 Introduction

Factor analysis is a multivariate analysis technique designed to solve statistical problems at a group level. Its purpose is to reduce a large set of observed variables to a small number of latent constructs (Bollen & Arminger, 1991; Cattell, 1988; Jöreskog, 1979) where a latent construct is defined as an underlying concept or factor that is common to some or all items of a given dataset (Child, 1990; Harman, 1976). The relationship between an observed variable and a factor is expressed as a *factor loading*, with the square of the factor loading providing information on how much variance in each observed variable can be explained by the factor (Hair *et al.*, 2006). Factor loadings of  $>0.30$  are considered statistically significant.<sup>13</sup> Despite statistical significance, in practice loadings of  $>0.50$  are generally considered the minimum and loadings in excess of 0.70 are desirable to obtain a good set of indicators of a construct (Hair *et al.*, 2006). While the factor loading provides information on how much variance the item shares with its factor, the *factor score regression coefficient* indicates the relative importance of that item in the context of the factor (Nunnally & Bernstein, 1994).

While it is the aim of factor analysis to obtain a set of items where each share a large amount of variance with the latent variable, this factor generally does not account for all variance of an item. That is, a certain amount of an item's variance is unique to that variable (Jöreskog, 1979). Mathematically this can be expressed as:  $(1 - \text{squared loading}) = \text{unique variance}$ . For example, an item with a loading of 0.70 has a unique variance of 0.51, i.e. it shares only half of its variance with the factor (Nunnally & Bernstein, 1994). Unique variance can again be subdivided into error variance and specific variance. *Error variance* represents imperfections in the measurement, while *specific variance* describes the variance component that an item might share with items that are not included in the analysis. Because of these specific

---

<sup>13</sup> The statistical significance of factor loadings depends on the sample size. A statistically significant loading of 0.30 is based on a sample size of  $n=350$ . Details on statistical significance of factor loadings are provided by Hair *et al.* (2006).

variances, factor loadings or even the whole factor structure may change when altering the number of included variables (Child, 1990; Harman, 1976). While this distinction between error and specific variance is plausible in theory, it is not possible to distinguish between the two when applying factor analysis (Child, 1990). As a result, they are subsumed under the umbrella term *residual variance* (Jöreskog, 1979).

In the case of a factor model consisting of more than one factor, i.e. included items load on more than one latent variable, this factor model can be described in terms of factor pattern that is defined by the loadings of each of the items on the respective factors. While the factor pattern is simple in a one-factor model as it only consists of loadings of all items on one factor, in a multiple-factor solution this factor pattern can be described by the patterns of loadings of the individual items on different factors. Optimally the factor model consists of items that have large loadings on only one factor, while having minor or zero loadings on the remaining factors. Mathematically the overall factor solution can be expressed in terms of the vector of observed variables being equal to the product of factor pattern and matrix of factor loadings plus the matrix of unique variances (Jöreskog, 1969; Mulaik, 1972).

Since the development of factor analysis more than a century ago (Spearman, 1904), ample research has been undertaken that has led to various refinements of the original technique. Given that the remainder of the thesis makes frequent use of the specific terminology of factor analysis, the most important concepts related to this technique are introduced briefly in the following sections.

#### **4.2.2 Exploratory, confirmatory, unrestricted, and restricted factor analysis**

A common way of describing factor analysis is to distinguish between exploratory (EFA) and confirmatory (CFA) approaches (Bollen, 1989). *EFA* is an approach in which the researcher explores a sample without prior knowledge about the data, i.e. the researcher has no *a priori* hypotheses about the factor solution (Jöreskog, 1969). Hence, an exploratory approach is generally aimed at obtaining further information about the properties of the data such as the correlational relationship between observed variables, the number of underlying factors, or the overall factor structure (Child, 1990). In simple terms, the concept of EFA consists of grouping all highly correlating variables together to determine the underlying factor structure of a questionnaire (Fayers & Hand, 1997). This is generally achieved by factor rotation, i.e. the reference axes are manipulated until a factor solution is found. When rotating the factors, it is further possible to determine whether they are allowed to correlate (oblique rotation) or whether they are strictly uncorrelated, i.e. orthogonal to each other (Child, 1990).

In contrast, *CFA* is generally aimed at confirming the factor structure of a predefined model (Child, 1990; Fayers & Hand, 1997). Hence, the researcher has prior knowledge about the data and specifies *a priori* hypotheses about the factor solution (Jöreskog, 1969). In spite of testing hypotheses, however, *CFA* is not always strictly confirmatory, i.e. by either modifying a predefined model until it is acceptable or specifying several competing models (Jöreskog, 1969, 1993), it can be used in an exploratory way to find an optimal factor solution when initial model fit is not satisfactory. Consequently, the differentiation between *EFA* and *CFA* is often not clear (Bollen, 1989) and it has been suggested illustrating them on a continuum moving from *EFA* to *CFA* (Mulaik, 1972) rather than describing them in discrete categories.

As a result of the difficulty in differentiating between *EFA* and *CFA*, researchers frequently use the terms *unrestricted* and *restricted* approaches instead (McDonald, 2005). These two approaches can be clearly separated from each other in terms of the restrictions imposed on the factor model. In *unrestricted factor analysis* there are no constraints on the factor model, i.e. no constraints on factor patterns, loadings or residual variances are imposed. Solutions are generally obtained through factor rotation but these are often non-unique because of the missing constraints in the model (Jöreskog, 1969; Mulaik, 1972). One of the recommended programs that allows for this type of factor analysis (McDonald, 2005) is *CEFA* (Browne *et al.*, 2004) which was used in the development of the *heiQ* as well as in a later part of this thesis (see Sections 2.3.2 and 6.3).

In contrast, *restricted factor analysis* imposes restrictions on the factor model. In particular, the solution of the product of factor patterns and factor loadings is restricted. While the factor solution can again be non-unique, a unique solution can generally be obtained by imposing several independent restrictions on the above parameters (Mulaik, 1972). Also, few sufficient conditions have been defined that make a restricted factor model identified. Firstly, per factor at least three observed variables are needed with significant loadings on this factor while having zero loadings on the other factors. Secondly, factors with only two items with non-zero loadings that again have zero loadings on the other factors are correlated with these other latent variables (McDonald, 1999, 2005). A range of computer programs is available for restricted analysis (McDonald, 2005) including *LISREL* (Jöreskog & Sörbom, 1996-2001) which was used in all restricted factor analyses of the thesis.

### **4.2.3 Factorial simplicity, unidimensionality, homogeneity, and reliability**

Apart from distinguishing between the different applications of factor analysis, it is also useful to discuss the types of factor solutions that can be obtained. When designing a questionnaire the aim is to identify a solution that consists of items that have a significant loading on one of

the factors while having negligible loadings on the remaining factors. Items that possess this property are referred to as being *unifactorial* or *factorially simple* (Bentler, 1977; Kaiser, 1974). Information on the property of each item is generally obtained in unrestricted models which can then be used in the specification of restricted models. Given that the unrestricted solution often indicates that items have high loadings on one factor while also having minor loadings on others, it has become common practice to set these minor loadings to zero in the restricted factor model (Ferrando & Lorenzo-Seva, 2000). In the case of all items in a given factor solution being unifactorial, the solution is said to be composed of *independent clusters* (McDonald, 2005). There are no clear guidelines as to how many items have to be factorially simple for a factor solution to be acceptable but recommendations range from two (Cattell, 1988) to three to four unifactorial items per factor (Carroll, 1978).

It may be sufficient to have a minimum number of unifactorial items per factor for a factor solution to be acceptable. However, a scale that is not *unidimensional*, i.e. at least some of the items of the construct are factorially complex, means that subjects differ along this scale in more than one way (Nunnally & Bernstein, 1994). This, in turn, has some implications for the interpretability of the results: a) the interpretation of a change in an observed variable is ambiguous as it is unknown which underlying concept caused the observed change, and b) if a considerable relationship exists between two items of different factors, this may allude to problems with the construct and discriminant validity of the questionnaire (Hair *et al.*, 2006).

*Homogeneity* is yet another term closely related to factorial simplicity and unidimensionality. A scale is not considered homogeneous if the content of the scale is diverse. In contrast to the other terms, heterogeneity can also be caused by too much random error. It remains that items of each construct must be unidimensional to be interpretable unambiguously (Nunnally & Bernstein, 1994).

While the previous three terms factorial simplicity, unidimensionality, and homogeneity are conceptually related, *reliability* describes another property of a scale. It is an index that shows the extent to which a questionnaire is consistent across multiple measurements (Hair *et al.*, 2006). It is therefore concerned with the replicability of results. One of the most widely used indices of reliability is coefficient alpha (Cronbach, 1951; Guttman, 1945).<sup>14</sup> While its lower limit is generally 0.70, this limit may have to be set higher with an increasing number of items as coefficient alpha is influenced by the number of items in the scale. In contrast, if the research is of exploratory nature the cut-off value may need to be lower (Hair *et al.*, 2006). Finally, while high coefficient alpha is a necessary condition for a scale to be unidimensional,

---

<sup>14</sup> Coefficient alpha is also referred to as Guttman-Cronbach alpha. According to McDonald (1999) the coefficient is often incorrectly attributed to Cronbach (1951), while Guttman (1945) was the first to publish coefficient alpha. In the present thesis this index will be referred to as coefficient alpha.

a high value is still not a sufficient condition for the scale to have unidimensional properties (Nunnally & Bernstein, 1994).

After introducing the types of factor solutions that can be obtained in factor analysis, these shall be further discussed in the context of this thesis. Firstly, unidimensionality of each heiQ subscale was considered an important condition for the measurement of outcomes of self-management interventions. If any one subscale was composed of factorially complex items, it would not be possible to ascertain in which area the self-management intervention had had an impact. The factor patterns were therefore specified in a way that each item loaded on one factor only. Secondly, no correlated errors of items across factors were allowed. For the same reason as before, an association between items of different factors suggests problems with the validity of an instrument (Hair *et al.*, 2006). Hence, in the heiQ re-validation it was aimed to avoid inter-factor correlated errors. Thirdly, correlated errors within the same factor (intra-factor) had to be minimal. When items of the same factor share excess unique variance, this again alludes to problems with the validity and/or dimensionality of a scale (Hair *et al.*, 2006). Finally, it was aimed to demonstrate coefficient alpha of 0.70 or higher for each heiQ subscale to be reasonably reliable.

#### **4.2.4 Summary**

The previous sections provided an overview of the terminology used in the context of factor analysis. In particular, the distinction between the different applications of factor analysis was important because these are frequently used throughout the remainder of the thesis. In view of the different types of solutions that can be obtained in factor analyses, these concepts were further discussed with reference to their application in this thesis. Therefore, the basic conditions for the planned re-validation of the heiQ were established including a discussion of topics pertaining to unidimensionality, correlated errors, and coefficient alpha.

### **4.3 Structural equation modeling (SEM)**

#### **4.3.1 Introduction**

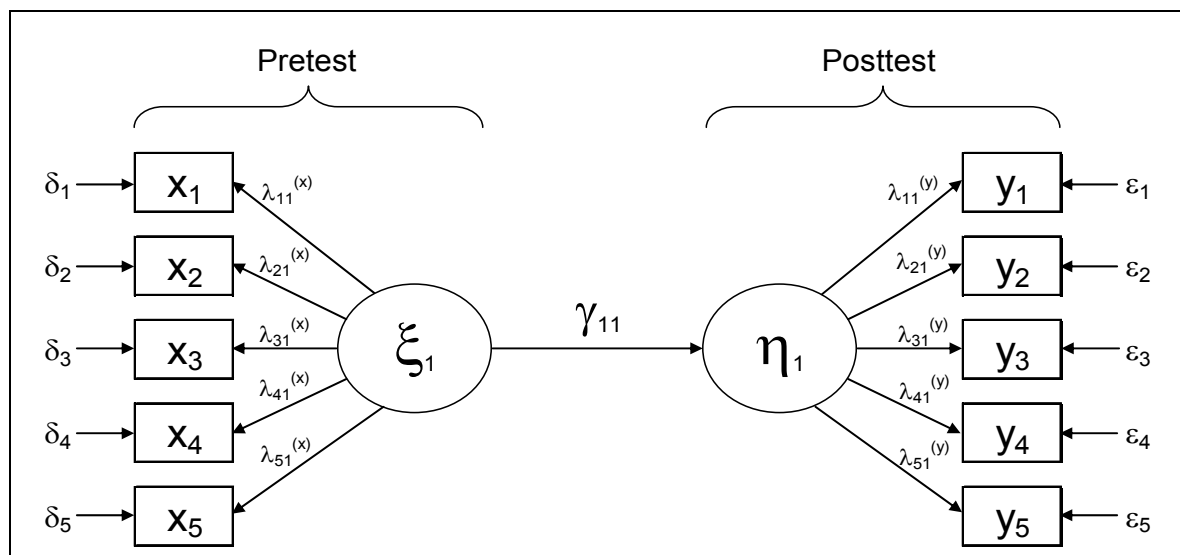
Factor analysis forms a cornerstone of structural equation modeling (SEM) which was the main statistical method applied in Chapters 5 and 6. SEM is a technique that combines factor analysis and regression analysis to solve multivariate research questions at a group level (Bollen & Arminger, 1991). The technique is also often referred to as the analysis of the structure of correlation/covariance matrices (Bollen, 1989; Rust & Golombok, 1999). The

generally accepted procedure of SEM involves the following five steps (Bollen & Long, 1993; Jöreskog, 1993):

- (1) model specification;
- (2) model identification;
- (3) model estimation;
- (4) model evaluation / testing model fit;
- (5) model modification / respecification.

### 4.3.2 LISREL matrices and notation

The generic SEM model is composed of two measurement models and a structural model. The *measurement models* define the relations between observed and latent variables and can therefore be compared with the factor model described in Section 4.2. In contrast, the *structural model* represents the regression of one latent variable on another latent variable (Bollen, 1989; Byrne, 1998). To make further descriptions of this method easier, a full SEM model is depicted in Figure 33. It illustrates heiQ subscale Positive and Active Engagement in Life and incorporates the measurement model of the pretest (X-model), the structural path from the latent variable of the pretest to the latent variable of the posttest, and the Y-model, i.e. the measurement model of the posttest.



**Figure 33** Structural equation model – illustrating Positive and Active Engagement in Life (see Table 14 for a legend of the LISREL notation)

In illustrations of SEM models, observed variables are generally represented by squares and latent variables are represented by circles. Curves with arrows pointing in both directions represent *correlations* (not shown in Figure 33, see Figure 34) and a straight arrow with a single head represents a causal path from the base of the arrow to the point of the arrow (Bollen, 1989). The base is generally referred to as the *exogenous variable*, whereas the variable that the arrow is pointed to is referred to as the *endogenous variable*.

As mentioned in Section 2.3.5, the SEM computer program LISREL version 8.72 (Jöreskog & Sörbom, 1996-2001) was used for the analyses in Chapters 5 and 6. Given that the notation is specific to this software, an overview of the terminology is provided hereafter.

In LISREL the full SEM model is composed of thirteen matrices/vectors. Eight of these are part of the standard LISREL output with four relating to the measurement models. The first of these matrices contains the factor loadings of the observed variables on the latent variables. In LISREL notation this matrix is referred to as the *Lambda-X matrix* ( $\Lambda_x$ ) with the latent variable being called Ksi ( $\xi$ ) and the regression coefficients from  $\xi$  to the observed variables being  $\lambda_x$  (see Figure 33). A second matrix of the X-model contains the error variances of the observed variables and is referred to as the *Theta-delta matrix* ( $\Theta_\delta$ ). Both these matrices of parameters of the X-model are mirrored in the Y-model with the *Lambda-Y matrix* ( $\Lambda_y$ ) relating the latent variable Eta ( $\eta$ ) to the observed y-variables, while the *Theta-epsilon matrix*  $\Theta_\epsilon$  contains error variances of the y-variables. Mathematically the relationship between the different variables, the factor loadings, and the error variances can be expressed as follows:

$$x = \Lambda_x \xi + \theta_\delta$$

$$y = \Lambda_y \eta + \theta_\epsilon$$

Hence, each observed variable can be expressed as a function of their loading on the latent variable plus error variance (Bollen, 1989).

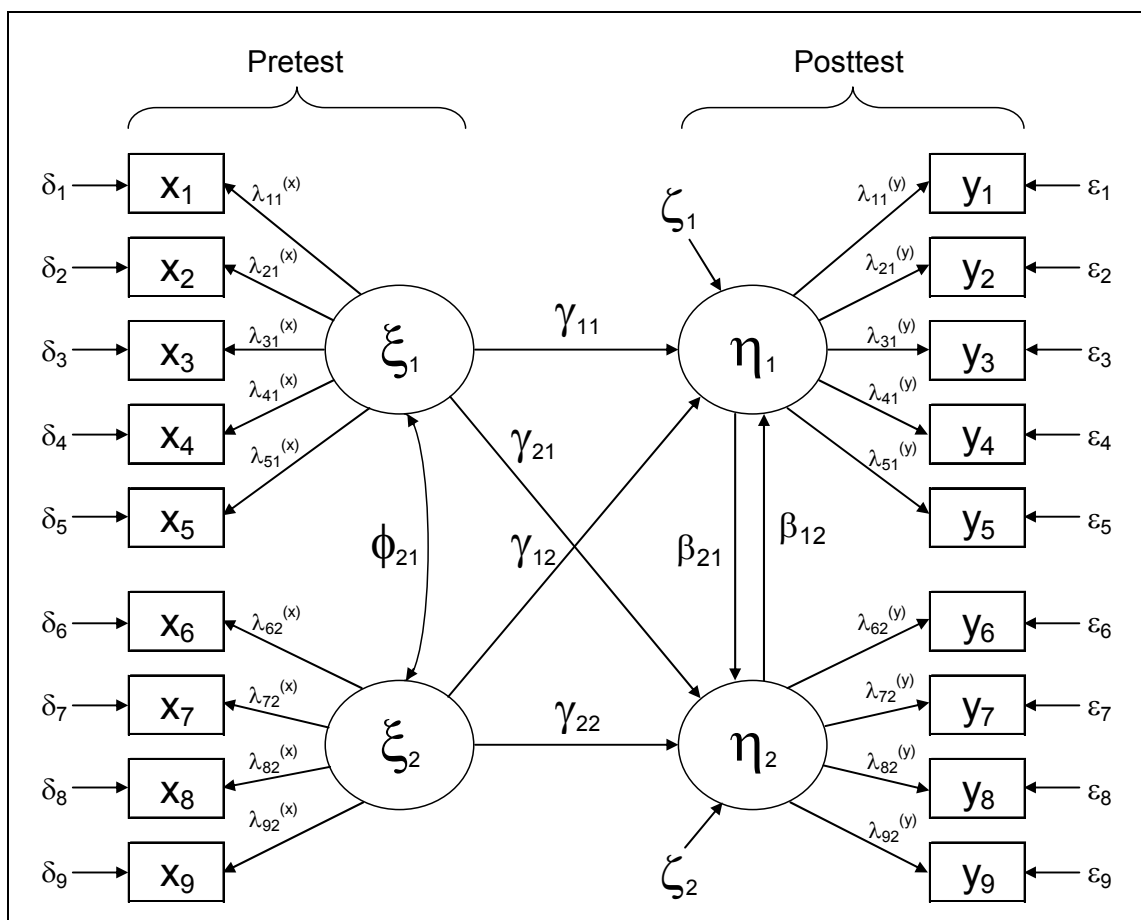
While the first four matrices of the standard LISREL output describe the two measurement models, the remaining four matrices are part of the structural model. Of these, three matrices contain information that is rather important for models that contain at least two latent exogenous and two latent endogenous variables. Figure 34 is therefore an extension of Figure 33 to better represent these matrices. It illustrates the heiQ subscales Positive and Active Engagement in Life, and Health-Directed Behaviour.

The first matrix of the structural model is the *Phi matrix* ( $\Phi$ ). It is a variance-covariance matrix containing the variances of each latent variable of the X-model and the covariances of these  $\xi$ 's (see Figure 34). The second matrix is the *Gamma matrix* ( $\Gamma$ ) which contains regression

coefficients that relate the latent exogenous  $\xi$  to the latent endogenous variable  $\eta$ . These coefficients are depicted as  $\gamma_{11}$ ,  $\gamma_{21}$ ,  $\gamma_{12}$ , and  $\gamma_{22}$  in Figure 34. To the right of Figure 34 two further matrices can be defined. The *Beta matrix* (B) is a regression matrix that describes the association between the  $\eta$ 's, with models being referred to as nonrecursive models in case the paths between these  $\eta$ 's go in both directions (Bollen, 1989). Finally, the *Psi matrix* ( $\Psi$ ) contains the error variances and covariances Zeta ( $\zeta$ ) (Byrne, 1998; Jöreskog & Sörbom, 1996-2001). These result from the assumption that the endogenous latent variables are measured with error that is not explained by the model (Schumacker & Lomax, 2004). In SEM it is further assumed that these error terms correlate, i.e.  $\zeta_1$  and  $\zeta_2$  are correlated (not shown in Figure 34). The relationships between these matrices can be expressed as:

$$\eta = B\eta + \Gamma\xi + \zeta$$

This equation shows that  $\eta$  can be expressed as a function of its regression on other  $\eta$ 's plus the regression of  $\eta$  on  $\xi$  plus the error term  $\zeta$  (Bollen, 1989). Applied to the model in Figure 34, this equation is obtained by adding all arrows that are pointed to  $\eta$ , i.e.  $\eta_1$  would be expressed as:  $\eta_1 = \beta_{12}\eta_2 + \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1$ .



**Figure 34** Structural equation model – illustrating Positive and Active Engagement in Life, and Health-Directed Behaviour (see Table 14 for a legend of the LISREL notation)



Finally, the standard LISREL model is based on the following assumptions (Bollen, 1989; Jöreskog & Sörbom, 1996-2001; Nunnally & Bernstein, 1994):

- (1)  $\delta$  and  $\xi$  are uncorrelated;
- (2)  $\varepsilon$  and  $\eta$  are uncorrelated;
- (3)  $\zeta$  and  $\xi$  are uncorrelated;
- (4)  $\delta$ ,  $\varepsilon$  and  $\zeta$  are mutually uncorrelated;
- (5) the diagonal elements of B are zero.

As mentioned before, five further LISREL matrices exist that are not part of the standard LISREL output. Unless explicitly specified, these additional matrices default to zero matrices (Jöreskog & Sörbom, 1996-2001). One of the additional matrices is the *Theta-delta-epsilon matrix*  $\Theta_{\delta\varepsilon}$ , a matrix that contains the correlated errors between the x- and y-variables. This matrix is a covariance matrix (Jöreskog & Sörbom, 1993) and is typically used in situations with repeated measures. Given that in repeated measures the same variable is measured at least twice over a period of time, it is assumed that a portion of the error variance correlates, i.e. the portion of the error that is specific to the variable and not random measurement error is expected to correlate over time (Jöreskog & Sörbom, 1996-2001). Most analyses of this thesis involved repeated measures. Hence,  $\Theta_{\delta\varepsilon}$  was frequently included in the models. As a consequence of defining  $\Theta_{\delta\varepsilon}$ , the 4<sup>th</sup> assumption of the standard LISREL model as outlined above was relaxed in a way that  $\theta_{\delta}$  and  $\theta_{\varepsilon}$  of the same item were allowed to correlate.

The remaining four matrices – or more specifically vectors (Bollen, 1989) – that are not part of the standard LISREL model are used in analyses of mean structures. They are convenient to be included in certain circumstances such as those when researchers want to compare the latent means of multiple groups. In general, the parameters of these matrices are not identified unless several constraints are imposed on the model. In multigroup models only few conditions need to be imposed, whereas in single groups multiple conditions are necessary (Jöreskog & Sörbom, 1996-2001). To model these mean structures in LISREL four intercept terms are added to the model. These are the mean parameters of the observed variables and the mean parameters of the latent variables (Jöreskog & Sörbom, 1996-2001; Schumacker & Lomax, 2004). The intercepts of the observed variables are called *Tau X* ( $\tau_x$ ) for the x-variables and *Tau Y* ( $\tau_y$ ) for the y-variables. Mathematically they can be expressed as follows (Jöreskog & Sörbom, 1996-2001; Sörbom, 1974):

$$x = \tau_x + \Lambda_x \xi + \theta_{\delta}$$

$$y = \tau_y + \Lambda_y \eta + \theta_{\varepsilon}$$

These equations are extensions of the previous equations with the addition of the intercepts. The vector of these intercepts contains the expected value of the observed variable when its latent variable is zero and is interpreted in the same way as the constant term in a regression equation (Bollen, 1989), i.e. the point of intersection with the y-axis. While intercept terms are used to define the mean of an observed variable, it is generally not equal to the mean. Instead, the mean of an observed variable can be understood as a construction of the means of the latent variables and the structural coefficients (Byrne, 1998), i.e. the observed mean is determined by the intercept and the latent mean multiplied by the factor loading (Oort, 2005b) and it is only equal to the intercept when the latent mean is zero (Bollen, 1989).

While the previous vectors referred to the means of the observed variables, the remaining two vectors contain the parameters of the factor means. These are referred to as *Kappa* ( $\kappa$ ) and *Alpha* ( $\alpha$ ) with the former representing the mean of  $\xi$  and the latter referring to the mean of  $\eta$ . The determination of the mean of  $\xi$  is straightforward, i.e. it can be derived from the equation for the expected value of  $x$  (Bollen, 1989):

$$E(x) = \tau_x + \Lambda_x \kappa$$

In contrast, the estimation of the mean of  $\eta$  is more complex. Mathematically  $\eta$  can be expressed by extending the equation of the structural model (Bollen, 1989):

$$\eta = \alpha + B\eta + \Gamma\xi + \zeta$$

Hence,  $\eta$  is now not only a function of the structural coefficients but it additionally contains the intercept parameter  $\alpha$ . The expected value of  $\eta$  can then be expressed as (Bollen, 1989):

$$\begin{aligned} E(\eta) &= (I - B)^{-1}(\alpha + \Gamma\xi + \zeta) \\ &= (I - B)^{-1}(\alpha + \Gamma\kappa) \end{aligned}$$

As expressed in the last equation, the mean of  $\eta$  is determined by the structural parameters in the Beta and the Gamma matrix and the intercept parameters  $\alpha$  and  $\kappa$ . Following from the above equations, the vectors of the means of the  $y$ -variables can be expressed as follows (Bollen, 1989; Jöreskog & Sörbom, 1996-2001):

$$E(y) = \tau_y + \Lambda_y(I - B)^{-1}(\alpha + \Gamma\kappa)$$

While the first part of the equation resembles the equation for the expected value of  $x$ , the second part includes the rather complex expression for the expected values of  $\eta$  (Jöreskog & Sörbom, 1996-2001).

For a final overview of the LISREL matrices and vectors, these are summarised in Table 14. Given that the Theta-delta-epsilon matrix and the four vectors of the means model are not part of the standard LISREL model, these are shaded in grey. Apart from the symbols of the parameters, the name of each of the matrices and vectors and their respective symbols, a short description is provided.

**Table 14** Overview of the eight standard LISREL matrices, the Theta-delta-epsilon matrix, and the four vectors for the analysis of mean structures

Matrix / vector number	Parameter symbol	Name	Matrix / vector symbol	Description
1	$\lambda_x$	Lambda-X	$\Lambda_x$	Factor loadings in the X-model
2	$\theta_\delta$	Theta-delta	$\Theta_\delta$	Residuals of the x-variables
3	$\lambda_y$	Lambda-Y	$\Lambda_y$	Factor loadings in the Y-model
4	$\theta_\epsilon$	Theta-epsilon	$\Theta_\epsilon$	Residuals of the y-variables
5	$\phi$	Phi	$\Phi$	Variances and covariances of the latent exogenous variables
6	$\gamma$	Gamma	$\Gamma$	Causal path between latent endogenous and latent exogenous variable
7	$\beta$	Beta	B	Causal path between latent endogenous variables
8	$\psi$	Psi	$\Psi$	Variances and covariances of the endogenous error terms
9	$\theta_{\delta\epsilon}$	Theta-delta-epsilon	$\Theta_{\delta\epsilon}$	Correlated errors between the x- and y-variables
10	$\tau_x$	Tau-X	$T_x$	Intercepts of the x-variables
11	$\tau_y$	Tau-Y	$T_y$	Intercepts of the y-variables
12	$\kappa$	Kappa	K	Intercepts of the latent exogenous variables
13	$\alpha$	Alpha	A	Intercepts of the latent endogenous variables

### 4.3.3 Parameter estimation for non-normal ordinal data

Several methods exist for the parameter estimation of the LISREL matrices. Given that the method is crucial for later data interpretations, the selection of the method should be based on the distributional properties and the scaling of the data. As mentioned in Section 2.3.5, because of the *non-normal* distribution of the heiQ data, all present analyses were based on the moment matrix and its asymptotic covariance matrix (Jöreskog, 2002-2005; Schumacker & Lomax, 2004). While the distributional properties of the data were one reason for the selection of the parameter estimation method that was used in this thesis, the selection of

this technique was also based on the ordinal properties of the heiQ scale (see Section 2.2.3). Given that the latter reason has not yet been discussed, all issues related to this particular scaling of questionnaire items are introduced hereafter.

The main challenge of *ordinal data* is the interpretation of obtained scores as these data have no units of measurement. Inter-person and intra-person across-occasion comparisons are problematic if data are ordinal because no statement can be made about the magnitude of the difference between two scores (Jöreskog & Sörbom, 1996-2002; Stucki *et al.*, 1996). Consequently, the input matrices as well as the parameter estimation need to be based on techniques that are appropriate for ordinal data (Jöreskog & Sörbom, 1996-2002; Muthén, 1984; Olsson, 1979). As a result of developments over the past decades, such techniques are now readily available in SEM computer programs such as the recent LISREL versions (Jöreskog, 2002-2005; Jöreskog & Sörbom, 1996-2002). Based on first publications in the late 1970s and mid 1980s (Muthén, 1984; Olsson, 1979) the LISREL user can now request polychoric correlations with the asymptotic covariance matrix, which are the input matrices recommended to be used with ordinal data (Hipp & Bollen, 2003; Jöreskog, 1994, 2002-2005).

It is helpful to provide some information on the theory behind *polychoric correlations*. When calculating these correlations it is assumed that a normally distributed continuous variable underlies each ordinal variable (Jöreskog, 1994; Olsson, 1979). This approach is in line with assertions that ordinal data are intrinsically quantitative (Agresti, 1984). That is, although ordinal variables should not be treated as if they were continuous (Jöreskog & Sörbom, 1996-2002), they should neither be treated as if they were categorical, as it is often the case when researchers apply nominal methods (Agresti, 1984). This assumption of an underlying continuity can be visualised by imagining the normal curve being cut into sections according to the frequency distributions in each response category. The points where the normal curve is being cut are commonly referred to as *thresholds* (Olsson, 1979) and are determined from the univariate marginal distribution function. These thresholds are then used to estimate the polychoric correlations which are derived from the bivariate marginal likelihoods based on the thresholds (Jöreskog, 1994; Olsson, 1979). The computation of polychoric correlations is the first step that is necessary for the parameter estimation of ordinal data. These correlations can then be used as the basis for calculating the matrix of asymptotic variances and covariances of these correlations, commonly named the asymptotic covariance matrix. This matrix forms the second essential input matrix (Jöreskog, 2002-2005).

For the parameter estimation of ordinal data, the weighted least squares (WLS) method has been recommended (Jöreskog & Sörbom, 1996-2001). The main drawback of this method, however, is that it requires a large sample size because the asymptotic covariance matrix

needs to be inverted (Jöreskog, 2002-2005). Given that these sample requirements are often not met, *robust maximum likelihood* (RML) has been described to be an alternative method to analyse ordinal data (Jöreskog, 2002-2005). RML has the further advantage of providing the Satorra-Bentler chi-square ( $\chi^2_{SB}$ ) which corrects for non-normality (Satorra & Bentler, 1988, 1994). This chi-square takes the multivariate kurtosis of the data into account (Curran *et al.*, 1996; Hu & Bentler, 1995) and it has been shown to work well regardless of data distribution (Bollen, 1989; Hu & Bentler, 1995; Jöreskog, 2002-2005), ordinal scaling (DiStefano, 2002; Jöreskog, 2002-2005), and with samples as small as  $n=200$  (Bentler & Yuan, 1999; Jöreskog, 2002-2005). For these reasons, RML was considered appropriate for the analysis of the heiQ data. Moreover, RML has been applied in previous published work on the heiQ (Osborne *et al.*, 2007), i.e. it was useful to employ the same method to make results comparable.

The application of these ordinal techniques in LISREL generally involves the following three steps (Jöreskog, 1990, 1994; Jöreskog & Moustaki, 2001):

- (1) Estimation of the thresholds in PRELIS, LISREL's data pre-processor program that was mentioned in Section 2.3.5;
- (2) Estimation of the input matrices based on polychoric correlations and asymptotic covariances in PRELIS;
- (3) Estimation of the model parameters in LISREL.

Unless it is necessary to impose further conditions on the thresholds, the first two steps are generally carried out in one PRELIS step. That is, when requesting the matrix of polychoric correlations, PRELIS automatically estimates all thresholds and bases the requested matrix on these thresholds. The computation of correlations, however, also implies that the means and standard deviations of the underlying continuous variables are standardised to zero and one respectively. This standardisation, however, may not be optimal because changes in the frequency distributions of variables may indicate that means and/or standard deviations of the underlying continuous variables changed. Jöreskog (2002-2005) therefore developed an *alternative parameterisation*. By computing covariances instead of polychoric correlations, it is achieved that the first two thresholds are held constant so that the means and the standard deviations can be estimated (Jöreskog, 2002-2005). This alternative parameterisation is also particularly important in models that involve observed and latent means. Given that some models in the present research included the analysis of mean structures, the alternative parameterisation was applied.

#### 4.3.4 Model evaluation

Following the five steps introduced in Section 4.3.1, it is necessary to evaluate whether the model fits the data once the model has been estimated. Several criteria can be used to judge whether or not the model is acceptable. To assess how well the data are represented by the model, it has been recommended to examine the fit statistics as well as the residual matrix (Browne *et al.*, 2002). However, the main problem pertaining to the interpretation of the fit statistics is that neither the selection of the indices nor the cut-off values has been universally agreed upon. In particular, the cut-off points that are currently applied are somewhat arbitrary as they evolved from experience rather than theory. Hence, they must be treated cautiously. It is useful to interpret them relative to other models and relative to the content area (Marsh *et al.*, 2004) as high values may be crucial in developed areas, whereas low values may be suitable in less developed areas (Bollen, 1989). Finally, fit indices only provide information on whether or not a model fits the data but they cannot prove causality or plausibility (Bollen, 1989), i.e. even in the case of good fit statistics, subjective judgement is still necessary for the final decision regarding model fit (Browne & Cudeck, 1989; Marsh *et al.*, 2004).

In this thesis the model evaluation was guided by three criteria. Firstly, a combination of indices was chosen for a comprehensive assessment of model fit, i.e. a range of qualitatively different fit statistics was applied (Bollen & Long, 1993; Marsh *et al.*, 1996; Tanaka, 1993). To maximise the range of indices, at least one of each of the following three categories was selected: absolute fit, absolute misfit, and incremental fit indices (Bollen, 1989; Browne *et al.*, 2002). The cut-off points were based on suggestions in the literature (Browne & Cudeck, 1989; Hair *et al.*, 2006; Hu & Bentler, 1999). Secondly, while these indices were used as a means to support the evaluation of the model fit, the interpretation and subsequent model re-specifications were also guided by the research questions (Browne & Cudeck, 1989; Marsh *et al.*, 1988; Marsh *et al.*, 2004). Thirdly, modification indices (MIs), standardised residuals, and other components of the model as provided in the LISREL output were closely examined as they provide crucial information on the model (Bollen & Long, 1993). MIs are particularly helpful as they can lead to significant improvements of the model fit.

The following paragraphs provide an introduction to the different categories of fit indices, with those that were selected for the evaluation of the present models described in more detail.

##### *Absolute fit indices*

These types of fit indices are directly derived from the fit of the model, i.e. they do not rely on the comparison to a baseline model (Browne *et al.*, 2002). The most prominent index of this

category is the  $\chi^2$  statistic (Gerbing & Anderson, 1993) which is based on the comparison of the model covariance matrix with the sample covariance matrix. Hence, if a non-significant  $\chi^2$  is obtained, this indicates that the two matrices do not differ significantly, i.e. it indicates that the model fits well (Bollen, 1989). However, the application of  $\chi^2$  has been criticised for being inflated by sample size<sup>15</sup>, size of the model, and data non-normality, while being understated as more parameters are added to the model (Bollen, 1989; Browne & Cudeck, 1989; Hu & Bentler, 1995). To respond to some of these issues, derivatives of the  $\chi^2$  statistic have been developed (Bollen, 1989; Jöreskog, 1993; Kline, 2005; Tanaka, 1993). The one applied in the thesis is the  $\chi^2_{SB}$  which corrects for data non-normality (Satorra & Bentler, 1988, 1994).

For the present thesis some limitations are attached to the interpretation of  $\chi^2_{SB}$ . In view of the relatively large sample size in the re-validation ( $n=949$ ) and size of the model consisting of 42 variables and eight factors,  $\chi^2_{SB}$  was interpreted relative to the sample size as well as complexity of the models. Given that the  $\chi^2_{SB}$  was not taken as the sole indicator of model fit (Bollen & Long, 1993), the notion of fit was relaxed in a way that a significant  $\chi^2_{SB}$  was considered acceptable if the remaining fit indices indicated satisfactory fit. In contrast, those fit indices that are less directly influenced by sample and model size were expected to meet at least the requirements of acceptable fit (Browne & Cudeck, 1989).

#### *Absolute misfit indices*

In the same manner as the former family of fit statistics, absolute misfit indices do not rely on the comparison to a baseline model. A main characteristic of misfit indices is that a small value is a sign of good model fit, with zero indicating perfect fit (Browne *et al.*, 2002). Two indices that are often applied – and which were used in this study – are SRMR (standardised root mean square residuals) and RMSEA (root mean square error of approximation).

While most goodness-of-fit indices are somewhat related, SRMR has been found to perform uniquely different to most fit indices (Hu & Bentler, 1999). Despite its relative independence, a further reason explaining why SRMR is a crucial component of model evaluation is that it helps detect potential sources of model misfit (Browne *et al.*, 2002). The closer the SRMR statistic is to zero, the smaller the remaining residuals in the model and the better model fit. It is calculated by taking the square root of the mean squared residuals, i.e. the average deviation of observed and predicted variances and covariances (Hair *et al.*, 2006). Although no clear cut-off value exists, SRMR of up to 0.08 is generally considered acceptable (Byrne, 1998; Hair *et al.*, 2006).

---

<sup>15</sup> The sample size effect has been found to be negligible when target models were true, but substantial when models were false (Marsh *et al.*, 1988).

The *RMSEA* is an absolute misfit index that attempts to correct for sample size and model complexity (Hair *et al.*, 2006; Steiger, 1990). By incorporating both sample size and degrees of freedom and by meeting the requirement of model parsimony, *RMSEA* has become one of the most popular indices (Browne & Cudeck, 1989; Browne & Du Toit, 1991; Fan *et al.*, 1999; Jöreskog & Sörbom, 1993). Although it appears to overrate model fit when samples get small, *RMSEA* has been described to be largely unaffected by sample size and shown to work well across a number of conditions (Fan *et al.*, 1999). A further advantage is that its distributional properties are known which allows for the calculation of confidence intervals (Hair *et al.*, 2006; MacCallum *et al.*, 1996). It has been suggested that a value of  $<0.05$  indicates close fit, while a value of  $<0.08$  is considered acceptable fit (Browne & Cudeck, 1989). Hu and Bentler (1999) undertook simulation studies on combinations of fit indices. They found that a combination of a cut-off value of  $<0.06$  for *RMSEA* with *SRMR*  $<0.09$  (or  $<0.10$ ) yielded best results (Hu & Bentler, 1999).

In this thesis the guidelines for the interpretation of *SRMR* and *RMSEA* were set as follows: provided all remaining fit indices were satisfactory, models were immediately accepted with *SRMR* of 0.05 or less (Byrne, 1998). Further, models with *SRMR* between 0.05 and 0.10 were generally inspected more closely before making a decision on acceptance or rejection of the model. However, in case all remaining fit indices were acceptable, the absolute cut-off for *SRMR* was set at 0.10. For *RMSEA*, both its actual value and the according confidence interval (CI) were taken into consideration as has been suggested by several researchers (Jöreskog, 2002-2005; MacCallum *et al.*, 1996). The reason for employing the CI of *RMSEA* is to be able to report the results with 90% confidence (Browne & Cudeck, 1989). The cut-off value for *RMSEA* was set at 0.06 (Hu & Bentler, 1999) with the upper bound of its CI being no larger than 0.10 (Hair *et al.*, 2006).

#### *Incremental / comparative fit indices*

In contrast to the absolute fit indices, incremental fit indices rely on the comparison of the target model to an alternative or a baseline model (Marsh *et al.*, 1988). They generally fall in the range between 0 and 1 (Marsh *et al.*, 1996) with numbers closer to 1 indicating better model fit (Hair *et al.*, 2006). While a range of incremental fit indices exists, each has benefits and shortcomings. Of these, the comparative fit index (*CFI*) (Bentler, 1990) appears to be one of the most suitable indices. It has been shown to work well across a range of conditions (Fan *et al.*, 1999; Gerbing & Anderson, 1993; Marsh *et al.*, 1996) such as being largely unaffected by sample size (Fan *et al.*, 1999) and being normed between 0 and 1 (Bentler, 1990; Hair *et al.*, 2006). While the latter property comes at the expense of a small downward



bias with decreasing sample size (Bentler, 1990; Gerbing & Anderson, 1993), this is outweighed by its otherwise strong performance (Bentler, 1990).

Incremental fit indices are generally considered to indicate unsatisfactory model fit when they drop below 0.90 (Hair *et al.*, 2006). Given that it has been recommended to report a range of statistics (Bollen & Long, 1993; Marsh *et al.*, 1996; Tanaka, 1993), it has become common practice to report cut-off values in the context of a second fit index as previously shown in the context of absolute misfit indices. For CFI, different combinations of cut-off values have been suggested. For example, in combination with SRMR, cut-off values of 0.92 (Hair *et al.*, 2006) or 0.96 (Hu & Bentler, 1999) have been recommended. Hence, the minimum value for CFI was set at 0.92, while models with CFI between 0.92 and 0.96 were inspected more closely.

#### **4.3.5 Summary**

This section provided an introduction to SEM which included a discussion of the parameter estimation technique for ordinal data and issues related to the evaluation of model fit. While the latter process is relatively subjective (Marsh *et al.*, 2004) and cut-off values are not to be applied in a definitive way (Bollen & Long, 1993; Gerbing & Anderson, 1993; Tanaka, 1993), it was necessary to set some guidelines to make subsequent analyses comprehensible. The fit indices that were selected to judge upon model fit ( $\chi^2_{SB}$ , SRMR, RMSEA, and CFI) were therefore assumed to provide a basis for the evaluation of the model, whereas each cut-off value was interpreted relative to the remaining fit indices to ensure that the fit of each model was judged in a broader context. Further aspects of this broader context were other model components such as the size of respective factor loadings. Finally, subjective judgement of the researcher was considered essential in this process of evaluating model fit.

### **4.4 The factor structure of the Health Education Impact Questionnaire (heiQ)**

#### **4.4.1 Introduction**

In preparation for subsequent analyses, the final section of this chapter describes the factor structure of the heiQ. As stated in Section 3.7, the analyses of Chapters 5 and 6 required the application of a questionnaire with good psychometric properties. Given that these analyses investigated response shift and social desirability bias, the quality of the data was important as it was critical that the degree of model fit could exclusively be traced back to the research questions rather than weaknesses in the questionnaire. Given that the heiQ is a relatively new instrument and its validation was based on subjects drawn from a different population

(Osborne *et al.*, 2007) than the subjects of the current sample, it was considered necessary to re-validate the heiQ. While the initial validation sample had been recruited from a broader population with almost half of the subjects being outpatients of the Royal Melbourne Hospital (Osborne *et al.*, 2007), the sample of the present research solely consisted of participants of self-management courses (see Section 2.2.3).

#### 4.4.2 Specific methods

Given that the criteria for the execution of factor analysis were discussed in Section 4.2, the description of the specific methods of the heiQ re-validation is kept concise. The re-validation was carried out on pretest heiQs of all course participants who were included in the research (n=949). Following from Section 4.3.3, RML was used for the parameter estimation hence the input matrices were polychoric covariances and the asymptotic covariance matrix. The re-validation followed Jöreskog's 3-step approach (Jöreskog, 1993). In this approach the hypothesised factor structure is gradually built up with each step providing additional information about the property of each item in the context of its factor and in the context of the whole questionnaire. The three steps can be summarised as follows:

- (1) In the first step, one-factor models are defined, i.e. in the present case eight one-factor models were specified. These models are then examined for potential correlated errors ( $\theta_{ij}$ ) between items that are hypothesised to belong to the same factor;
- (2) This step is followed by a specification of two-factor models. As a result of applying every combination of the eight subscales, 28 models were specified. In this step each item is inspected for a) additional correlated errors with items of the same factor (intra-factor), b) correlated errors with items of different factors (inter-factor), and c) significant loadings on factors other than their hypothesised factor (cross-loadings);
- (3) Finally, the full factor model combining all eight subscales was specified. In this last step, suggested MIs were inspected and information that had been obtained in the previous two steps was used as these steps had already alluded to potential problem items.

#### 4.4.3 Results

##### *Step 1 – one-factor models*

The eight one-factor models largely resulted in good fit of each respective model. As shown in Table 15, the first two heiQ subscales showed excellent fit with both suggesting a non-

significant  $\chi^2_{SB}$ . The remaining indices were RMSEA=0.0 (90% CI, 0.0-0.022), CFI=1.0, SRMR=0.007, and coefficient alpha=0.85 for Positive and Active Engagement in Life, and RMSEA=0.0 (90% CI, 0.0-0.058), CFI=1.0, SRMR=0.007, and coefficient alpha=0.82 for Health-Directed Behaviour.

The third subscale Skill and Technique Acquisition indicated some problems. While model fit was acceptable, item “I am very good at using aids and devices to make my life easier” (q3\_2) had a substantially lower loading (0.45) and small factor score regression coefficient compared with the other items of the subscale. It was therefore decided to exclude this item to ensure that this heiQ subscale consisted of strong indicators only. The fit statistics of the reduced four-item factor were:  $\chi^2_{SB}(2)=8.2$  ( $p=0.016$ ), RMSEA=0.057 (90% CI, 0.021-0.100), CFI=1.0, SRMR=0.023, and coefficient alpha=0.79 with all factor loadings >0.60.

Constructive Attitudes and Approaches showed good model fit with  $\chi^2_{SB}(5)=14.4$  ( $p=0.013$ ), RMSEA=0.045 (90% CI, 0.019-0.072), CFI=1.0, SRMR=0.021, and coefficient alpha=0.85.

Self-Monitoring and Insight consisting of seven heiQ items indicated acceptable model fit. In view of a large correlated error between “I know what things can trigger my health problems and make them worse” (q5\_3) and “I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me” (q5\_7), model fit was compared with a) allowing for this correlated error or b) excluding one of the items. Given that the content of the two items was somewhat similar and both a) and b) showed a similar improvement to model fit, it was decided to exclude one of the items to avoid correlated errors in this subscale. This led to an exclusion of item q5\_7 as it had a slightly smaller factor loading and smaller factor score regression coefficient than item q5\_3, and model fit was substantially better when q5\_7 was excluded. The fit statistics of the reduced six-item subscale were:  $\chi^2_{SB}(9)=32.4$  ( $p<0.001$ ), RMSEA=0.052 (90% CI, 0.034-0.072), CFI=0.99, SRMR=0.036, and coefficient alpha=0.72.

The fit statistics of Health Service Navigation suggested somewhat suboptimal fit. With high standardised residuals for item “I confidently give healthcare professionals the information they need to help me” (q6\_3), and later models regularly indicating problems with this item, it was decided to exclude it. The fit indices of the four-item model were:  $\chi^2_{SB}(2)=7.7$  ( $p=0.022$ ), RMSEA=0.055 (90% CI, 0.018-0.098), CFI=1.0, SRMR=0.019, and coefficient alpha=0.82.

The initial model fit of subscale Social Integration and Support was acceptable. Given that item “When I feel ill, my family and carers really understand what I am going through” (q7\_5) indicated some high standardised residuals, the model was again tested excluding this item. Given that model fit improved significantly, it was decided to discard this item. The fit indices of the reduced four-item subscale were:  $\chi^2_{SB}(2)=3.0$  ( $p=0.225$ ), RMSEA=0.023 (90% CI, 0.0-0.072), CFI=1.0, SRMR=0.011, and coefficient alpha=0.86.

Finally, Emotional Well-Being indicated acceptable model fit with  $\chi^2_{SB}(9)=34.5$  ( $p<0.001$ ), RMSEA=0.055 (90% CI, 0.036-0.075), CFI=1.0, SRMR=0.026, and coefficient alpha=0.89.

**Table 15** Step 1 of Jöreskog's 3-step procedure; eight one-factor models on heiQ pretests (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>1. Positive and Active Engagement in Life</b>				
q1_1 / Q11	I am doing interesting things in my life	0.86	0.27	0.308
q1_2 / Q3	Most days I am doing some of the things I really enjoy	0.72	0.48	0.173
q1_3 / Q7	I try to make the most of my life	0.75	0.44	0.166
q1_4 / Q16	I have plans to do enjoyable things for myself during the next few days	0.74	0.46	0.178
q1_5 / Q30	I feel like I am actively involved in life	0.79	0.38	0.198
Fit statistics: $\chi^2_{SB}(5)=1.8$ , $p=0.881$ ; RMSEA=0.0 (90% CI, 0.0; 0.022); CFI=1.0; SRMR=0.007. Coefficient alpha: 0.85				
<b>2. Health-Directed Behaviour</b>				
q2_1 / Q40	I walk for exercise, for at least 15 minutes per day, most days of the week	0.81	0.35	0.286
q2_2 / Q5	I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	0.77	0.42	0.222
q2_3 / Q15	On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)	0.72	0.48	0.152
q2_4 / Q24	On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	0.85	0.28	0.323
Fit statistics: $\chi^2_{SB}(2)=1.4$ , $p=0.504$ ; RMSEA=0.0 (90% CI, 0.0; 0.058); CFI=1.0; SRMR=0.007. Coefficient alpha: 0.82				
<b>3. Skill and Technique Acquisition</b>				
q3_1 / Q17	When I have symptoms, I have skills that help me cope	0.82 / 0.83	0.33	0.336
q3_2 / Q2	I am very good at using aids and devices to make my life easier	0.45 / N/A*	0.80	0.053
q3_3 / Q36	I have effective skills that help me handle stress	0.73 / 0.72	0.47	0.182
q3_4 / Q14	I have a very good idea of how to manage my health problems	0.71 / 0.70	0.50	0.168
q3_5 / Q10	I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	0.64 / 0.63	0.59	0.153
Fit statistics: $\chi^2_{SB}(5)=15.1$ , $p=0.010$ ; RMSEA=0.046 (90% CI, 0.021; 0.074); CFI=1.0; SRMR=0.024. Coefficient alpha: 0.77				
Fit statistics: $\chi^2_{SB}(2)=8.2$ , $p=0.016$ ; RMSEA=0.057 (90% CI, 0.021; 0.100); CFI=1.0; SRMR=0.023. Coefficient alpha: 0.79 (model excluding q3_2)				
<b>4. Constructive Attitudes and Approaches</b>				
q4_1 / Q39	If others can cope with problems like mine, I can too	0.68	0.54	0.084
q4_2 / Q18	I try not to let my health problems stop me from enjoying life	0.75	0.43	0.107
q4_3 / Q35	I do not let my health problems control my life	0.85	0.28	0.263
q4_4 / Q28	My health problems do not ruin my life	0.79	0.38	0.128
q4_5 / Q32	I feel I have a very good life even when I have health problems	0.84	0.29	0.181
Fit statistics: $\chi^2_{SB}(5)=14.4$ , $p=0.013$ ; RMSEA=0.045 (90% CI, 0.019; 0.072); CFI=1.0; SRMR=0.021. Coefficient alpha: 0.85				

**Table 15 (continued)** Step 1 of Jöreskog's 3-step procedure; eight one-factor models on heiQ pretests (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>5. Self-Monitoring and Insight</b>				
q5_1 / Q41	With my health in mind, I have realistic expectations of what I can and cannot do	0.58 / 0.59	0.66	0.145
q5_2 / Q4	As well as seeing my doctor, I regularly monitor changes in my health	0.51 / 0.51	0.74	0.094
q5_3 / Q8	I know what things can trigger my health problems and make them worse	0.56 / 0.51	0.69	0.148
q5_4 / Q22	When I have health problems, I have a clear understanding of what I need to do to control them	0.76 / 0.76	0.42	0.257
q5_5 / Q19	I have a very good understanding of when and why I am supposed to take my medication	0.63 / 0.63	0.60	0.095
q5_6 / Q38	I carefully watch my health and do what is necessary to keep as healthy as possible	0.61 / 0.64	0.63	0.191
q5_7 / Q12	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	0.54 / N/A*	0.71	0.102
Fit statistics: $\chi^2_{SB}(14)=90.1$ , $p<0.001$ ; RMSEA=0.076 (90% CI, 0.061; 0.091); CFI=0.97; SRMR=0.057. Coefficient alpha: 0.75				
Fit statistics: $\chi^2_{SB}(9)=32.4$ , $p<0.001$ ; RMSEA=0.052 (90% CI, 0.034; 0.072); CFI=0.99; SRMR=0.036. Coefficient alpha: 0.72 (model excluding q5_7)				
<b>6. Health Service Navigation</b>				
q6_1 / Q21	I communicate very confidently with my doctor about my healthcare needs	0.87 / 0.84	0.25	0.343
q6_2 / Q13	I have very positive relationships with my healthcare professionals	0.80 / 0.82	0.36	0.152
q6_3 / Q25	I confidently give healthcare professionals the information they need to help me	0.62 / N/A*	0.62	0.114
q6_4 / Q27	I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)	0.70 / 0.71	0.52	0.132
q6_5 / Q34	I work in a team with my doctors and other healthcare professionals	0.76 / 0.77	0.43	0.199
Fit statistics: $\chi^2_{SB}(5)=27.0$ , $p<0.001$ ; RMSEA=0.068 (90% CI, 0.004; 0.094); CFI=0.99; SRMR=0.030. Coefficient alpha: 0.82				
Fit statistics: $\chi^2_{SB}(2)=7.7$ , $p=0.022$ ; RMSEA=0.055 (90% CI, 0.018; 0.098); CFI=1.0; SRMR=0.019. Coefficient alpha: 0.82 (model excluding q6_3)				
<b>7. Social Integration and Support</b>				
q7_1 / Q20	I have enough friends who help me cope with my health problems	0.85 / 0.88	0.28	0.291
q7_2 / Q33	I get enough chances to talk about my health problems with people who understand	0.71 / 0.69	0.50	0.133
q7_3 / Q6	If I need help, I have plenty of people I can rely on	0.81 / 0.82	0.34	0.239
q7_4 / Q31	Overall, I feel well looked after by friends or family	0.87 / 0.84	0.24	0.287
q7_5 / Q23	When I feel ill, my family and carers really understand what I am going through	0.75 / N/A*	0.44	0.157
Fit statistics: $\chi^2_{SB}(5)=32.3$ , $p<0.001$ ; RMSEA=0.076 (90% CI, 0.052; 0.102); CFI=0.99; SRMR=0.027. Coefficient alpha: 0.88				
Fit statistics: $\chi^2_{SB}(2)=3.0$ , $p=0.225$ ; RMSEA=0.023 (90% CI, 0.0; 0.072); CFI=1.0; SRMR=0.011. Coefficient alpha: 0.86 (model excluding q7_5)				
<b>8. Emotional Well-Being</b>				
q8_1 / Q42	If I think about my health, I get depressed	0.83	0.31	0.200
q8_2 / Q37	I get upset when I think about my health	0.82	0.33	0.217
q8_3 / Q26	I often feel angry when I think about my health	0.86	0.26	0.245

**Table 15 (continued)** Step 1 of Jöreskog's 3-step procedure; eight one-factor models on heiQ pretests (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
8. Emotional Well-Being (continued)				
q8_4 / Q09	My health problems make me very dissatisfied with my life	0.78	0.39	0.141
q8_5 / Q1	I often worry about my health	0.70	0.52	0.138
q8_6 / Q29	I feel hopeless because of my health problems	0.70	0.51	0.092
Fit statistics: $\chi^2_{SB}(9)=34.5$ , $p<0.001$ ; RMSEA=0.055 (90% CI, 0.036; 0.075); CFI=1.0; SRMR=0.026. Coefficient alpha: 0.89				

\* As this item was excluded in the second step of the model evaluation of the subscale, most factor loadings of the remaining items of the scale changed slightly.

Legend

- Loading: Standardised factor loading
- Error: Error variance
- FSR: Factor score regression coefficient
- $\chi^2_{SB}$ : Satorra-Bentler  $\chi^2$
- RMSEA: Root mean square error of approximation
- 90% CI: 90% Confidence interval
- CFI: Comparative fit index
- SRMR: Standardized root mean square residual

*Step 2 – two-factor models*

The specification of the two-factor models again largely confirmed the hypothesised factor structure of the heiQ (Osborne *et al.*, 2007). Given that as many as 28 pairs of subscales had been specified, this section only provides a summary of the main findings pertaining to those pairs of heiQ subscales that had indicated intra- and/or inter-factor correlated errors or cross-loadings. The interpretation of the MIs was mainly based on qualitative judgement. If the content of respective items was similar the result was retained for later analyses. Otherwise, MIs were ignored unless they indicated substantive problems within the questionnaire.

As shown in Table 16, a correlated error between item “I try not to let my health problems stop me from enjoying life” (q1\_3) and item “I try to make the most of my life” (q4\_2) was suggested. Given that their respective content was similar, it was tested whether model fit improved when allowing for this inter-factor error. It followed that  $\chi^2_{SB}$  decreased significantly when this correlated error was included in the model.<sup>16</sup> As its inclusion did not indicate a large correlated error (0.15) and the fit of the model was acceptable without this modification, it was decided not to allow for this inter-factor correlated error to keep each heiQ subscale unambiguous. Regardless, the information was considered for later analyses.

<sup>16</sup> The chi-square difference test requires an adjustment formula for the Satorra-Bentler chi-square. Details on this formula are provided in Section 5.3.

**Table 16** Summary of results of Step 2 of Jöreskog's 3-step procedure (n=949)

Item #	Two-factor models	MI
Positive and Active Engagement in Life & Constructive Attitudes and Approaches		
q4_2	I try not to let my health problems stop me from enjoying life	Inter-factor TD q4_2 and q1_3
q1_3	I try to make the most of my life	
Fit statistics: $\chi^2_{SB}(34)=149.1$ , $p<0.001$ ; RMSEA=0.060 (90% CI, 0.050; 0.070); CFI=0.99; SRMR=0.038		
Fit statistics: $\chi^2_{SB}(33)=101.0$ , $p<0.001$ ; RMSEA=0.047 (90% CI, 0.036; 0.057); CFI=1.0; SRMR=0.034 (TD q4_2 and q1_3)		
Positive and Active Engagement in Life & Self-Monitoring and Insight		
q5_7	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	Intra-factor TD q5_7 and q5_3
q5_3	I know what things can trigger my health problems and make them worse	
q5_5	I have a very good understanding of when and why I am supposed to take my medication	Intra-factor TD q5_5 and q5_4
q5_4	When I have health problems, I have a clear understanding of what I need to do to control them	
Fit statistics: $\chi^2_{SB}(53)=250.9$ , $p<0.001$ ; RMSEA=0.063 (90% CI, 0.055; 0.071); CFI=0.98; SRMR=0.061		
Fit statistics: $\chi^2_{SB}(43)=167.2$ , $p<0.001$ ; RMSEA=0.055 (90% CI, 0.047; 0.064); CFI=0.98; SRMR=0.052 (excluding item q5_7)		
Fit statistics: $\chi^2_{SB}(42)=160.9$ , $p<0.001$ ; RMSEA=0.055 (90% CI, 0.046; 0.064); CFI=0.99; SRMR=0.048 (excluding q5_7, and TD q5_5 and q5_4)		
Health-Directed Behaviour & Self-Monitoring and Insight		
q5_7	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	Intra-factor TD q5_7 and q5_3
q5_3	I know what things can trigger my health problems and make them worse	
q5_6	I carefully watch my health and do what is necessary to keep as healthy as possible	Cross-loading on HDB
Fit statistics: $\chi^2_{SB}(43)=219.4$ , $p<0.001$ ; RMSEA=0.066 (90% CI, 0.057; 0.075); CFI=0.98; SRMR=0.064		
Fit statistics: $\chi^2_{SB}(34)=140.3$ , $p<0.001$ ; RMSEA=0.057 (90% CI, 0.048; 0.068); CFI=0.98; SRMR=0.053 (excluding item q5_7)		
Fit statistics: $\chi^2_{SB}(33)=105.5$ , $p<0.001$ ; RMSEA=0.048 (90% CI, 0.038; 0.059); CFI=0.99; SRMR=0.046 (excluding q5_7 and LX q5_6 on HDB)		
Skill and Technique Acquisition & Self-Monitoring and Insight		
q3_3	I have effective skills that help me handle stress	Intra-factor TD q3_3 and q3_1
q3_1	When I have symptoms, I have skills that help me cope	
q5_7	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	Intra-factor TD q5_7 and q5_3
q5_3	I know what things can trigger my health problems and make them worse	
Fit statistics: $\chi^2_{SB}(53)=261.0$ , $p<0.001$ ; RMSEA=0.064 (90% CI, 0.057; 0.072); CFI=0.98; SRMR=0.054		
Fit statistics: $\chi^2_{SB}(52)=215.5$ , $p<0.001$ ; RMSEA=0.058 (90% CI, 0.050; 0.066); CFI=0.98; SRMR=0.051 (TD q3_3 and q3_1)		
Fit statistics: $\chi^2_{SB}(42)=144.8$ , $p<0.001$ ; RMSEA=0.051 (90% CI, 0.042; 0.060); CFI=0.99; SRMR=0.043 (TD q3_3 and q3_1, and excluding q5_7)		

**Table 16 (continued)** Summary of results of Step 2 of Jöreskog's 3-step procedure (n=949)

Item #	Two-factor models	MI
Constructive Attitudes and Approaches & Health Service Navigation		
q6_3	I confidently give healthcare professionals the information they need to help me  Fit statistics: $\chi^2_{SB}(34)=152.7$ , $p<0.001$ ; RMSEA=0.061 (90% CI, 0.051; 0.071); CFI=0.99; SRMR=0.055 Fit statistics: $\chi^2_{SB}(26)=106.2$ , $p<0.001$ ; RMSEA=0.057 (90% CI, 0.046; 0.069); CFI=0.99; SRMR=0.052 (excluding q6_3)	Regular large standardised residuals
Self-Monitoring and Insight & Health Service Navigation		
q6_3	I confidently give healthcare professionals the information they need to help me	Regular large standardised residuals
q5_7	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	Intra-factor TD q5_7 and q5_3
q5_3	I know what things can trigger my health problems and make them worse	
Fit statistics: $\chi^2_{SB}(53)=243.3$ , $p<0.001$ ; RMSEA=0.062 (90% CI, 0.054; 0.070); CFI=0.98; SRMR=0.058 Fit statistics: $\chi^2_{SB}(43)=174.9$ , $p<0.001$ ; RMSEA=0.057 (90% CI, 0.048; 0.066); CFI=0.98; SRMR=0.052 (excluding q6_3) Fit statistics: $\chi^2_{SB}(34)=106.4$ , $p<0.001$ ; RMSEA=0.047 (90% CI, 0.037; 0.058); CFI=0.99; SRMR=0.040 (excluding q6_3 and q5_7)		

Legend

MI:	Suggested modification index for an item pair
TD:	Correlated error theta-delta (could be either intra- or inter-item TD)
HDB:	heiQ subscale Health-Directed Behaviour
LX:	Lambda X; cross-loading
$\chi^2_{SB}$ :	Satorra-Bentler $\chi^2$
RMSEA:	Root mean square error of approximation
90% CI:	90% Confidence interval
CFI:	Comparative fit index
SRMR:	Standardized root mean square residual

A further correlated error was observed for item “When I have symptoms, I have skills that help me cope” (q3\_1) and item “I have effective skills that help me handle stress” (q3\_3). Although this intra-factor error fitted relatively well with the respective content of the items, it was only suggested in one of the two-factor models. Hence, the information was retained for further analyses but it was not given much importance.

Some more general observations were made for combinations of subscales involving either Self-Monitoring and Insight or Health Service Navigation.

In Self-Monitoring and Insight, most models suggested a correlated error between item q5\_3 and item q5\_7. Given that it had already been decided in the previous one-factor models to discard the latter item from this heiQ subscale, these results confirmed this modification.



Furthermore, two models (only one example is shown in Table 16) suggested a correlated error between items “When I have health problems, I have a clear understanding of what I need to do to control them” (q5\_4) and “I have a very good understanding of when and why I am supposed to take my medication” (q5\_5). In spite of small but statistically significant improvements, this modification was ignored as the respective contents of these items were not sufficiently similar. Finally, the combination of subscales Self-Monitoring and Insight, and Health-Directed Behaviour alluded to a potential cross-loading of item “I carefully watch my health and do what is necessary to keep as healthy as possible” (q5\_6) on Health-Directed Behaviour. Although the content of the item corresponded with the theme of this subscale, the modification did not lead to a substantial loading (0.28). In view of all items of Health-Directed Behaviour having loadings of >0.70, the loading of 0.28 was considered to be an insufficient indicator of this heiQ subscale.

In two-factor models that included subscale Health Service Navigation respective model fit generally improved when item q6\_3 was excluded (not all examples are shown in Table 16). Given that the one-factor model of this subscale had suggested removal of this item, these results of the two-factor models confirmed that this adjustment had been justified.

### *Step 3 – full heiQ model*

In the factor analysis of the full heiQ model the most restricted model was specified first. That is, neither intra- nor inter-factor correlated errors were allowed. Items were also defined to be unifactorial, i.e. no cross-loadings were permitted. For completeness, two full models were specified. The first model included all original 42 heiQ variables to be able to compare results with the heiQ validation (Osborne *et al.*, 2007), whereas the second model included 38 items. The latter model was the set of items that was used in the present thesis, i.e. the original 42 heiQ items excluding the four items that had been found to be problematic in the previous one- and two-factor models. The pretest sample (n=949) was further re-validated separately for each of the three randomised groups heiQ-PP, heiQ-PPT, and heiQ-PPR as well as for the dataset of retrospective pretests (heiQ-PPR Retro) as it was important for later analyses that the factor structure held for each of these datasets separately (see Chapter 5).

The fit statistics were largely the same across the full models including either 42 or 38 items.  $\chi^2_{SB}$  was large in both models which could mainly be attributed to the sample size as well as size of the models. Given that the original validation had suggested similar results (Osborne *et al.*, 2007) and the remaining fit statistics indicated good fit, less attention was paid to the significant  $\chi^2_{SB}$  statistic. The fit indices of the 42-item heiQ were:  $\chi^2_{SB(791)}=2280.7$  ( $p<0.001$ ), RMSEA=0.045 (90% CI, 0.042-0.047), CFI=0.98, SRMR=0.060 which was almost identical to

the fit indices of the reduced 38-item heiQ:  $\chi^2_{SB}(637)=1814.1$  ( $p<0.001$ ), RMSEA=0.044 (90% CI, 0.042-0.047), CFI=0.99, and SRMR=0.058. Furthermore, with only few exceptions, the majority of items had large factor loadings, i.e. they were strong indicators of their respective factor. Given space constraints, only the results of the factor analysis of the 38 heiQ items are shown in Table 17. The results of the factor analysis of the original 42 items are provided in Appendix 15.

**Table 17** Step 3 of Jöreskog's 3-step procedure; full model of the 38 heiQ items (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>1. Positive and Active Engagement in Life</b>				
q1_1 / Q11	I am doing interesting things in my life	0.81	0.34	0.158
q1_2 / Q3	Most days I am doing some of the things I really enjoy	0.70	0.51	0.108
q1_3 / Q7	I try to make the most of my life	0.77	0.42	0.123
q1_4 / Q16	I have plans to do enjoyable things for myself during the next few days	0.74	0.45	0.128
q1_5 / Q30	I feel like I am actively involved in life	0.83	0.32	0.176
<b>2. Health-Directed Behaviour</b>				
q2_1 / Q40	I walk for exercise, for at least 15 minutes per day, most days of the week	0.79	0.38	0.184
q2_2 / Q5	I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	0.76	0.42	0.157
q2_3 / Q15	On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)	0.71	0.50	0.102
q2_4 / Q24	On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	0.88	0.23	0.287
<b>3. Skill and Technique Acquisition</b>				
q3_1 / Q17	When I have symptoms, I have skills that help me cope	0.76	0.42	0.152
q3_3 / Q36	I have effective skills that help me handle stress	0.72	0.48	0.110
q3_4 / Q14	I have a very good idea of how to manage my health problems	0.77	0.41	0.140
q3_5 / Q10	I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	0.64	0.60	0.094
<b>4. Constructive Attitudes and Approaches</b>				
q4_1 / Q39	If others can cope with problems like mine, I can too	0.68	0.54	0.062
q4_2 / Q18	I try not to let my health problems stop me from enjoying life	0.77	0.41	0.082
q4_3 / Q35	I do not let my health problems control my life	0.84	0.30	0.172
q4_4 / Q28	My health problems do not ruin my life	0.77	0.41	0.084
q4_5 / Q32	I feel I have a very good life even when I have health problems	0.85	0.27	0.141
<b>5. Self-Monitoring and Insight</b>				
q5_1 / Q41	With my health in mind, I have realistic expectations of what I can and cannot do	0.57	0.68	0.047
q5_2 / Q4	As well as seeing my doctor, I regularly monitor changes in my health	0.51	0.74	0.032

**Table 17 (continued)** Step 3 of Jöreskog's 3-step procedure; full model of the 38 heiQ items (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>5. Self-Monitoring and Insight (continued)</b>				
q5_3 / Q8	I know what things can trigger my health problems and make them worse	0.50	0.75	0.041
q5_4 / Q22	When I have health problems, I have a clear understanding of what I need to do to control them	0.79	0.37	0.102
q5_5 / Q19	I have a very good understanding of when and why I am supposed to take my medication	0.61	0.63	0.029
q5_6 / Q38	I carefully watch my health and do what is necessary to keep as healthy as possible	0.64	0.59	0.073
<b>6. Health Service Navigation</b>				
q6_1 / Q21	I communicate very confidently with my doctor about my healthcare needs	0.82	0.32	0.238
q6_2 / Q13	I have very positive relationships with my healthcare professionals	0.80	0.36	0.142
q6_4 / Q27	I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)	0.73	0.46	0.144
q6_5 / Q34	I work in a team with my doctors and other healthcare professionals	0.78	0.39	0.208
<b>7. Social Integration and Support</b>				
q7_1 / Q20	I have enough friends who help me cope with my health problems	0.87	0.25	0.281
q7_2 / Q33	I get enough chances to talk about my health problems with people who understand	0.72	0.48	0.118
q7_3 / Q6	If I need help, I have plenty of people I can rely on	0.81	0.35	0.196
q7_4 / Q31	Overall, I feel well looked after by friends or family	0.85	0.28	0.200
<b>8. Emotional Well-Being</b>				
q8_1 / Q42	If I think about my health, I get depressed	0.83	0.31	0.188
q8_2 / Q37	I get upset when I think about my health	0.81	0.34	0.190
q8_3 / Q26	I often feel angry when I think about my health	0.85	0.27	0.214
q8_4 / Q09	My health problems make me very dissatisfied with my life	0.79	0.38	0.138
q8_5 / Q1	I often worry about my health	0.69	0.52	0.126
q8_6 / Q29	I feel hopeless because of my health problems	0.71	0.50	0.085

Fit statistics:  $\chi^2_{SB}(637)=1814.1$ ,  $p<0.001$ ; RMSEA=0.044 (90% CI, 0.042;0.047); CFI=0.99; SRMR=0.058

Legend:

Loading:	Standardised factor loading
Error:	Error variance
FSR:	Factor score regression coefficient
$\chi^2_{SB}$ :	Satorra-Bentler $\chi^2$
RMSEA:	Root mean square error of approximation
90% CI:	90% Confidence interval
CFI:	Comparative fit index
SRMR:	Standardized root mean square residual

The results of the factor analyses of heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro are presented in Table 18. It was observed that  $\chi^2_{SB}$  was generally smaller than the one in

the previous factor analysis on all pretest heiQs which could be attributed to the decrease in sample size. All remaining fit statistics of each group were largely the same when compared with the full pretest dataset, with not only the results of the factor analyses of the actual pretest heiQs being similar (heiQ-PP; heiQ-PPT; heiQ-PPR) but the dataset of retrospective pretests (heiQ-PPR Retro) also showing comparable fit statistics. Hence, in spite of collecting these data at the end of self-management courses, the factor structure was similar to that of the actual pretests. Across all four datasets the RMSEA was <0.05, the CFI was well above the cut-off point of 0.95, and the SRMR was no larger than 0.075.

**Table 18** Full model of the 38-item scale on heiQ pretest data of heiQ-PP (n=331), heiQ-PPT (n=304), and heiQ-PPR / heiQ-PPR Retro (n=314)

	$\chi^2_{SB}$	RMSEA (90% CI)	CFI	SRMR
Pretests, heiQ-PP	$\chi^2_{SB}(637)=1082.1, p<0.001$	0.046 (90% CI, 0.041-0.051)	0.99	0.069
Pretests, heiQ-PPT	$\chi^2_{SB}(637)=1015.5, p<0.001$	0.044 (90% CI, 0.039-0.049)	0.99	0.075
Pretests, heiQ-PPR	$\chi^2_{SB}(637)=1114.6, p<0.001$	0.049 (90% CI, 0.044-0.054)	0.98	0.067
Retrospective pretests, heiQ-PPR	$\chi^2_{SB}(637)=1075.4, p<0.001$	0.047 (90% CI, 0.042-0.052)	0.99	0.066

Legend:

- $\chi^2_{SB}$ : Satorra-Bentler  $\chi^2$
- RMSEA: Root mean square error of approximation
- 90% CI: 90% Confidence interval
- CFI: Comparative fit index
- SRMR: Standardized root mean square residual

In spite of these satisfactory fit statistics across all four datasets, the LISREL output indicated that the Phi matrix was non-positive definite in the model of retrospective pretests (heiQ-PPR Retro). Non-positive definiteness can occur as a result of, for example, linear dependency of items (collinearity), pairwise deletion of missing data, or start values (Wothke, 1993). While it has been suggested that non-positive definiteness of matrices such as the Phi matrix can occur within correctly specified models (Rigdon, 1997), this result may allude to a potential problem within the dataset of heiQ-PPR Retro.

#### 4.4.4 Summary

The 3-step procedure (Jöreskog, 1993) largely confirmed the hypothesised factor structure of the heiQ (Osborne *et al.*, 2007). After deleting four items, all one-factor models indicated excellent to acceptable model fit of each subscale. Results further suggested that all items were unifactorial. Neither inter-factor correlated errors nor cross-loadings were large enough

to be considered problematic. The full model again confirmed the heiQ factor structure. While both the model based on the original 42 items and the model based on the reduced 38-item scale showed satisfactory fit statistics, the latter model was considered more suitable for the present thesis as subsequent analyses required excellent psychometric properties, so that potential changes in model fit could solely be attributed to bias rather than problems with the measurement instrument. Hence, four items were deleted from the heiQ on the basis of inferior psychometric performance for the analyses in this thesis. However, a further re-validation study would be necessary to support a recommendation that they be permanently deleted from the inventory as the first validation (Osborne et al., 2007) had not found these items to be problematic.

# Chapter 5

A model of  
measurement  
invariance to detect  
response shift

## 5 A model of measurement invariance to detect response shift

### 5.1 Introduction

The analyses of Chapter 3 had shown that the magnitude of actual mean change scores, i.e. change based on pretest-posttest comparisons, was influenced by the design of the posttest heiQ. That is, the cognitive task people performed at posttest appeared to influence their ratings of actual posttest levels. Differences were particularly pronounced between subjects who provided posttest levels only (heiQ-PP) and those who provided retrospective pretest levels in addition to posttest levels (heiQ-PPR), with the latter group showing significantly higher actual posttest levels than the former. In contrast, when comparing change scores derived from actual pretest-posttest data (heiQ-PPR) with those derived from retrospective pretest-posttest data (heiQ-PPR Retro), only small group differences were observed.

Following from these findings the analyses of the present chapter were aimed at exploring which of the four change measures – three based on actual (heiQ-PP; heiQ-PPT; heiQ-PPR) and one based on retrospective pretest-posttest data (heiQ-PPR Retro) – is the most valid representation of outcomes of self-management interventions. In this context, threats to the validity of change scores are understood as the influence of biases on scores. Considering that outcomes derived from participant self-report are susceptible to a range of biases (see Section 1.2.4.3), this chapter investigated response shift effects in the samples consisting of actual pretest-posttest data. To explore whether response shift bias was apparent in these groups, the *response shift model* based on factor analysis that was mentioned in Section 1.2.4.4 was applied (Oort, 2005b; Oort *et al.*, 2005). Based on the discussions of Section 3.7, it was expected that less response shift was detectable in heiQ-PPR. Moreover, considering that response shift – per definition – cannot occur in scores that were assessed at the same time (Howard & Dailey, 1979), the heiQ items of the sample of retrospective pretest-posttest data were examined for *measurement invariance*. By applying essentially the same model as for the actual pretest-posttest data, it was aimed to confirm that the items of this dataset (heiQ-PPR Retro) were invariant.

The statistical models that were applied in this chapter assume knowledge of factor analysis and SEM. Given that an introduction to these statistical methods was provided in Chapter 4, the following section is limited to a general description of the models of the present chapter.

The analyses of Chapter 5 are based on a concept known as *measurement invariance* which is generally applied in situations involving the analyses of *multiple groups* (Byrne *et al.*, 1989; Cheung & Rensvold, 2002; Mulaik, 1972). When comparing multiple groups, it is assumed that the psychometric properties of the instrument used to collect data are invariant across these groups. Hence, it has to be ruled out that group differences are not due to specific

attributes of the groups before making any judgment on group differences (Gregorich, 2006; Meredith, 1993), i.e. it is only permissible to compare groups' means if the derived scores are based on a questionnaire with identical psychometric properties (Byrne *et al.*, 1989; Cheung & Rensvold, 2002; Steenkamp & Baumgartner, 1998). These properties can be investigated by testing for invariance (McGaw & Jöreskog, 1971) in a) the factor structure, b) relationships between factors, and c) the factor loadings across groups (Mulaik, 1972). Furthermore, these parameters are presented in a hierarchical order with *configural* and *scalar invariance* being considered the minimum requirements for the comparison of the factor means of two or more groups (Bollen, 1989; Gregorich, 2006; Meredith, 1993; Steenkamp & Baumgartner, 1998). More details on this hierarchy follow in Section 5.3.

In the same manner as for multiple populations, the concept of measurement invariance can be applied to *repeated measures*. Instead of comparing different groups, however, the same participants are assessed over time. As briefly mentioned in Section 1.2.4.4, this method can detect *gamma* (Golembiewsky *et al.*, 1976) and *beta change* (Schmitt, 1982), with the former being represented by differences in the factor structure and the latter being represented by differences in factor loadings and/or factor variances. The *response shift model* (Oort, 2005b; Oort *et al.*, 2005) is a further extension of these models for repeated measures. By matching the different parameter estimates of the measurement model with the types of response shift (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999), the model purports to assess whether an item has been recalibrated, reconceptualised, and/or reprioritised. In addition to detecting the response shift parameters, the model can be used to quantify true change which is represented by the differences in common factor means. The response shift model consists of the following four steps (Oort, 2005b; Oort *et al.*, 2005):<sup>17</sup>

- Step 1 – Specification of the measurement model;
- Step 2 – Overall test of response shift bias (item intercepts, factor loadings, and residual variances are constrained to be equal across occasions);
- Step 3 – Response shift detection (all non-tenable constraints of Step 2 are relaxed);
- Step 4 – Assessment of true change and other types of change (factor variances, factor correlations, and residual correlations are tested for invariance).

Although both models, i.e. the model of measurement invariance for multiple groups (cross-sectional) and for repeated measures (longitudinal) are similar, there is a difference in their specification. While most parameters are specified in the same way, in repeated measures it is additionally important to allow for across-occasion correlations as the two datasets are

---

<sup>17</sup> The steps are explained in more detail in Section 5.3 as they pertain to the present thesis.



related (Jöreskog, 2002-2005). As mentioned in Section 4.3.2, it is necessary to allow for correlated errors between the same item over time (*Theta-delta-epsilon matrix*), and to allow for across-occasion correlations of the same factor (*Phi matrix*). Hence, the specified models had to account for these across-occasion correlations.

Mainly drawing from ideas of Oort and colleagues (2005), the present chapter is structured in a way that at first an adjusted model of the 4-step procedure is presented (see Section 5.3). This model was then applied to the three randomised groups of actual pretest-posttest data to investigate whether response shift bias had a possible confounding effect on the results of these group-level data (see Section 5.4 to 5.6). Further, the model was applied to the dataset of heiQ-PPR Retro to examine whether the heiQ items were invariant between retrospective pretest and posttest data (see Section 5.7). The chapter concludes with the assessment of true change across all four datasets (see Section 5.8).

## 5.2 Aims

The aims of the present chapter were:

- 5.a To assess whether response shift could be detected in actual pretest-posttest data;
- 5.b To assess whether the magnitude of response shift differed across the three datasets (heiQ-PP; heiQ-PPT; heiQ-PPR);
- 5.c To test whether the model parameters were invariant when analysing the dataset of retrospective pre- and posttests (heiQ-PPR Retro).

## 5.3 The assessment of invariance of ordinal data in repeated measures

The statistical model that was applied in this chapter was derived from the 4-step procedure introduced in Section 5.1 (Oort, 2005b; Oort *et al.*, 2005). The following modifications were undertaken: a) Step 2 was extended to the inclusion of all response shift parameters, i.e. parameters within the *Phi* and *Theta-delta matrix* were constrained, and as a consequence, Step 4 consisted of the assessment of true change only. This modification was applied as it was deemed important that a test of overall response shift included each potential response shift parameter. And b) the hierarchy of measurement invariance, which had not been strictly included in the original model (Oort, 2005b), was added. While this hierarchy is not essential for the purpose of detecting group-level response shift, it is a prerequisite of Step 4 as not all invariance constraints are critical to ensure an unbiased comparison of factor means (Bollen,

1989; Gregorich, 2006; Meredith, 1993; Steenkamp & Baumgartner, 1998). Finally, c) given that the original response shift model was developed for continuous data (Oort, 2005b; Oort *et al.*, 2005), the model of the present chapter had to be extended to be adequate for ordinal data (Jöreskog, 2002-2005; Millsap & Yun-Tein, 2004). More details on the respective rationale for each of these modifications are provided in later parts of this section.

The modified 4-step procedure was defined as follows:

- Step 1 – Configural invariance (baseline model without any constraints)
- Step 2 – Complete invariance (constraints are imposed on all response shift parameters)
- Step 3 – Response shift detection (all non-tenable constraints of Step 2 are relaxed);
- Step 4 – Assessment of true change

The four steps are described in detail hereafter:

#### Step 1 – Configural invariance (baseline model without any constraints)

This first step consisted of the specification of a baseline model which served as a starting point for the systematic comparison of models of Steps 2 and 3 (Steenkamp & Baumgartner, 1998). Given that the samples of actual pretest-posttest data were fitted to the pre-defined heiQ factor structure (see Section 4.4), this first step entailed the test of configural invariance (Steenkamp & Baumgartner, 1998). This comprised a validation of pretest and posttest heiQ, assuming identical factor patterns across the two measurement occasions (van de Vliert *et al.*, 1985). This baseline model was aimed at confirming that the samples could generally be fitted to the hypothesised measurement model. In the rather unlikely event of not finding satisfactory model fit, the analyses would have to be stopped here. It would then have to be concluded that the reconceptualisation of the heiQ items had been so strong from pretest to posttest that the data were not suitable for further analyses (Oort, 2005b; van de Vliert *et al.*, 1985). Given that all pretest samples had been validated in Section 4.4, these analyses were merely a validation of the three posttest samples.

Expressed in LISREL notation Step 1 tested the following research question:

- $\text{Patt}(\Lambda_{x(\text{Pre})}) = \text{Patt}(\Lambda_{x(\text{Post})})?$

The expression  $\text{Patt}(\Lambda_x)$  represents the pattern matrix of the factor loadings.

An important characteristic of these baseline models was that no constraints were imposed on the parameters, i.e. all parameters were estimated freely. The only constraints that were stipulated were the ones necessary for model identification:

- (1) Scaling of each respective latent variable by setting its variance to one (Oort, 2005b);
- (2) Invariance of the respective threshold of each individual item over time (Jöreskog, 2002-2005; Millsap & Yun-Tein, 2004);
- (3) Fixing of all factor means to zero (Oort, 2005b).

For clarification, these constraints need to be further elaborated before proceeding to Step 2:

In analyses that involve latent variables it is always necessary to standardise each factor for identification purposes (Jöreskog, 2002-2005). This can be achieved either a) by defining a reference variable, i.e. the loading of one item per factor is set to one (Byrne, 1998; Jöreskog & Sörbom, 1996-2001), or b) by setting the factor variance to one (Oort, 2005b). In analyses that involve tests of measurement invariance, however, it is more complex to choose a reference variable as it has to be ensured that the variable itself is invariant (Kline, 2005). In the present models it was therefore decided to choose the second option of standardising the factors as shown in constraint 1). Moreover, as will be explained in Step 4, factor variances are not relevant for response shift detection (Oort, 2005b). Hence, through fixing the factor variances, no important information was lost with regard to the aim of this chapter.

Given that ordinal data were modelled in this thesis, i.e. their scaling is not comparable to metric data, it was essential that the underlying continua of the observed variables were used (Jöreskog, 2002-2005). As explained in Section 4.3.3, the underlying variables are assigned a scale by establishing thresholds (Olsson, 1979) which are a measurement property of the items (Muthén & Christoffersson, 1981). Hence, apart from standardising the factors, it was necessary to fix these thresholds to be equal over time as specified in constraint 2). This ensured that the underlying continua were on identical scales (Jöreskog, 2002-2005). As a result, the input matrices were based on these thresholds. Moreover, the input matrices were based on covariances instead of polychoric correlations to allow for the analysis of mean structures. By using Jöreskog's alternative parameterisation (see Section 4.3.3), the first two thresholds were fixed at zero and one respectively, so that means and standard deviations of each item could be estimated (Jöreskog, 2002-2005).

Finally, when comparing the latent means of several groups, it is common that a reference group is chosen. By fixing its factor means to zero, the means of the remaining groups can be evaluated relative to this reference group. Moreover, it is assumed that the item intercepts are invariant across groups (Byrne, 1998; Jöreskog & Sörbom, 1996-2001; Schumacker &

Lomax, 2004). However, given that the test of invariance of the intercepts was one of the research questions of the present chapter, invariance of the intercepts was not assumed but needed to be tested in subsequent steps. Therefore, the baseline model did not impose any constraints on the item intercepts but in constraint 3) all factor means were fixed at zero, a necessary and sufficient condition for the identification of this model (Oort, 2005b).

### Step 2 – Complete invariance (constraints are imposed on all response shift parameters)

In Step 2 all relevant response shift parameters were constrained across occasions. Instead of successively imposing constraints on the model (forward search), each model was fully constrained (backward search) to test for overall response shift. While the power to detect response shift bias was high in this backward search, the approach came at the expense of potentially interpreting all observed changes as response shifts, i.e. including those that may not have been response shifts (Oort *et al.*, 2005). In view of the large number of models that were investigated in this chapter, i.e. four datasets times eight heiQ subscales, it was not tenable to additionally execute a full forward search on all 32 models. Instead, an alternative control was used. By testing each final model against its baseline model it was ensured that the final, more constrained model fitted equally well compared with its baseline.

Once Step 2 was finalised, the results were interpreted as follows: a) if the full response shift model was not significantly worse than the baseline model, the response shift detection procedure was stopped concluding that either no response shift had occurred or that these response shifts had not been strong enough to influence the group-level data, or b) if the full response shift model was significantly worse than the baseline model, it was concluded that at least one type of response shift had occurred. How many parameters were affected and what type of response shifts had occurred was then examined in Step 3.

Before describing each constraint in detail, it shall be briefly discussed how differences between models were judged in terms of significance. Given that each response shift model was nested within its respective baseline model (Gregorich, 2006; Steenkamp & Baumgartner, 1998), the  $\chi^2$  difference test was applied (Oort, 2005b; Oort *et al.*, 2005). Assuming that the baseline model was specified correctly (Yuan & Bentler, 2004), a significant difference was interpreted as an indication of response shift. As mentioned in Section 4.4.3, this difference test needs to be adjusted when using the  $\chi^2_{SB}$  (Satorra & Bentler, 2001), as was applied in the present analyses (see Appendix 16 for this adjustment formula). Given that this test again depends on the sample size, further indices were used to judge differences between models (Schmitt, 1982; Steenkamp & Baumgartner, 1998; Yuan & Bentler, 2004). Apart from SRMR, RMSEA, and CFI (see Section 4.3.4), the expected cross-

validation index (ECVI) was used, a statistic appropriate for the comparison of models within a single sample (Browne & Cudeck, 1989; Cudeck & Browne, 1983; Oort *et al.*, 2005).

The following constraints were explored in Step 2 of the present model:

- $\Lambda_{x(\text{Pre})} = \Lambda_{x(\text{Post})}$ ?
- $\tau_{x(\text{Pre})} = \tau_{x(\text{Post})}$ ?
- $\text{Diag}(\theta_{\delta(\text{Pre})}) = \text{Diag}(\theta_{\delta(\text{Post})})$ ?
- $\Phi_{(\text{Pre})} = \Phi_{(\text{Post})}$ ?
- $\theta_{\delta(\text{Pre})} = \theta_{\delta(\text{Post})}$ ?

The above expressions represent the following parameters of the models:  $\Lambda_x$  describes the magnitude of the factor loadings,  $\tau_x$  are the intercepts of the observed variables,  $\text{Diag}(\theta_{\delta})$  are the variances of the residuals,  $\Phi$  are the factor correlations, and  $\theta_{\delta}$  are the error correlations.

As mentioned in the introduction to this section and as illustrated above, Step 2 included all possible response shift parameters. In contrast to the original approach that only constrained factor loadings, intercepts, and error variances in Step 2 (Oort, 2005b; Oort *et al.*, 2005), the models of this chapter also included factor and error correlations. The rationale for including additional constraints in Step 2 of the present models shall be explained in more detail:

Firstly, Oort and colleagues (2005) had not included factor correlations in Step 2 but these were investigated in Step 4 where they were interpreted as higher level reconceptualisations or reprioritisations. While this interpretation is largely consistent with past research, in which factor correlations were defined as gamma change (Schmitt, 1982), it has been suggested that gamma change needs to be ruled out before testing for any of the other types of change (Randolph & Elloy, 1989; Schmitt, 1982; van de Vliert *et al.*, 1985). Although this constraint is not relevant for an unbiased comparison of the factor means, it was considered an important response shift parameter and therefore an important part of Step 2.

Secondly, the error correlations, i.e. the off-diagonals of the Theta-delta matrix (Jöreskog & Sörbom, 1993), had also been excluded from the original model. Potential changes in these parameters were interpreted as lower level reconceptualisations or reprioritisations and were again tested in Step 4 (Oort, 2005b; Oort *et al.*, 2005). These correlations, however, mean that the error terms of items covary, i.e. they share variance that cannot be explained by their latent variable. If this covariance is reasonably large, it might allude to a problem with the construct validity of the measurement instrument (Hair *et al.*, 2006). Consequently, a change in these parameters from pretest to posttest was considered to be an important aspect of the full response shift model given its potential impact on the construct validity of the heiQ.

### Step 3 – Response shift detection (all non-tenable constraints of Step 2 are relaxed)

Step 3 of the model was applied when the model of Step 2 was significantly worse than its baseline model. A significantly worse model indicates that at least some of the invariance constraints are not tenable (Byrne, 2001) which then had to be investigated systematically. This procedure was generally stopped when a) no further single modification index could be identified that suggested a significant improvement to the model and b) none of the standardised residuals alluded to problem items (Oort *et al.*, 2005). At this stage it was also verified that the final model was not significantly worse than its baseline model.

Considering that each invariance constraint has a unique interpretation, the implications of the model modifications are described hereafter. In view of the hierarchy of these constraints, these are provided in order of their importance as they pertain to an unbiased comparison of factor means (Gregorich, 2006; Steenkamp & Baumgartner, 1998):

#### (1) *Configural invariance* ( $\text{Patt}(\Lambda_{x(\text{Pre})}) = \text{Patt}(\Lambda_{x(\text{Post})})$ )

This type of invariance is the highest level of invariance and means that the factor patterns of the pretests and posttests should be identical. Given that this elementary condition was tested in Step 1, it is only named here for completeness.

#### (2) *Metric invariance* ( $\Lambda_{x(\text{Pre})} = \Lambda_{x(\text{Post})}$ )

This constraint tests the equality of the factor loadings and is the second most important constraint of invariance tests. Given that the loadings relate latent and observed variables, it is essential that these are stable across conditions (Steenkamp & Baumgartner, 1998). If loadings were not stable, a change in the latent variable would have a different implication for its indicator variable at pretest compared with its implication at posttest, rendering across-occasion comparisons meaningless. Hence, configural and metric invariance are regarded as the minimum requirements for the comparison of scores on items across groups (Bollen, 1989; Gregorich, 2006; Steenkamp & Baumgartner, 1998). In the backward search that was applied in the present analyses, this meant that the constraints on factor loadings were not relaxed until all other response shift parameters had been investigated.

Where  $\Lambda_{x(\text{Pre})} \neq \Lambda_{x(\text{Post})}$ , this was interpreted as a reprioritisation of items within a factor. That is, a particular item changed in importance of defining the latent variable because subjects assigned more or less importance to this item as opposed to other items within the same factor. Alternatively, this change could also be interpreted in terms of the latent variable, with

the latent variable being reprioritised in a way that its influence on respective items changed from pretest to posttest (Oort, 2005b).

(3) *Scalar invariance* ( $\tau_{x(\text{Pre})} = \tau_{x(\text{Post})}$ )

While being a necessary condition, metric invariance is still not a sufficient condition for an unbiased comparison of factor means. Only if item intercepts are also equal across groups, a change in the mean of an observed variable can be attributed to a change in the mean of the latent variable (Gregorich, 2006; Steenkamp & Baumgartner, 1998).

Where  $\tau_{x(\text{Pre})} \neq \tau_{x(\text{Post})}$ , this was interpreted as a uniform recalibration of items (Oort, 2005b). This type of response shift manifests itself in the mean parameter of the measured variables, i.e. a change in the item intercepts. These intercepts relate latent variables to their respective indicator variables (Bollen, 1989). Hence, a change in the mean of a measured variable can be fully explained by a change in the mean of the underlying factor. In the case of a scale recalibration, it would not be possible to fully attribute a change in the mean of the affected item to a change in the mean of their latent variable, with the change being (partially) caused by a renewed judgment of the response scale (Oort, 2005b). These invariance constraints on the intercepts were not relaxed until after the other parameters had been explored, with the exception of factor loadings, given that metric invariance precedes scalar invariance in the hierarchy of measurement invariance.

In sum, *configural*, *metric*, and *scalar invariance* are the essential invariance constraints that must be fulfilled to ensure an unbiased comparison of factor means. While these conditions should optimally be met by all items of a given latent variable, it is sufficient if at least two items per factor fulfil these requirements. That is, two items per factor need to be invariant to establish partial invariance (Byrne *et al.*, 1989; Steenkamp & Baumgartner, 1998). Hence, comparisons of factor means based on a reduced number of invariant items are defensible (Gregorich, 2006).

(4) *Miscellaneous constraints* ( $\text{Diag}(\theta_{\delta(\text{Pre})}) = \text{Diag}(\theta_{\delta(\text{Post})})$ ,  $\Phi_{(\text{Pre})} = \Phi_{(\text{Post})}$ ,  $\theta_{\delta(\text{Pre})} = \theta_{\delta(\text{Post})}$ )

Although the final set of invariance constraints is not critical for an unbiased comparison of factor means, these parameters are still important for the detection of response shift. Given that these constraints are not part of the hierarchy of measurement invariance, the order of testing the constraints is arbitrary and largely depends on the individual research interest (Bollen, 1989; Steenkamp & Baumgartner, 1998).

The remaining response shift parameters were interpreted as follows:

Where  $\text{Diag}(\theta_{\delta(\text{Pre})}) \neq \text{Diag}(\theta_{\delta(\text{Post})})$ , this was interpreted as a non-uniform recalibration. Similar to the concept of item intercepts, a change in item variances is generally caused by a change in the variances of the latent variable. If such change cannot be fully attributed to changes in the Phi matrix but rather to changes in the Theta-delta matrix, it can be concluded that a non-uniform recalibration occurred (Oort, 2005b).

Where  $\Phi_{(\text{Pre})} \neq \Phi_{(\text{Post})}$ , higher-level reconceptualisations or reprioritisations were inferred. If factors had different interrelationships at posttest compared with pretest, this was interpreted as a reconceptualisation of the meaning of the underlying construct and/or a reprioritisation of the respective latent variables. This type of response shift was one of the main reasons why the models of pretest and posttest data in the present chapter were specified as two X-measurement models rather than a longitudinal model (Oort, 2005b).

Where  $\theta_{\delta(\text{Pre})} \neq \theta_{\delta(\text{Post})}$ , lower-level reconceptualisations or reprioritisations were concluded. As described in Step 2, these correlations mean that the error terms of items covary, i.e. they share variance that cannot be explained by their factor. Given that a reasonably large error correlation might allude to problems with the construct validity of the measurement instrument (Hair *et al.*, 2006), a potential change in this parameter alludes to changes in the construct validity of the instrument over time.

#### Step 4 – Assessment of true change

The first three steps of this 4-step procedure were sufficient for the detection of response shift. In contrast, in Step 4 true change in factor means and true change in factor variances could be assessed. While the latter provides information on potential changes in the degree of homogeneity of the group (Jöreskog, 2002-2005; Oort, 2005b), this test was omitted as it was not relevant for this chapter. Hence, only true changes in factor means were assessed.

In LISREL notation the final model tested the following research question:

- $\kappa_{(\text{Pre})} = \kappa_{(\text{Post})}$ ?

$\kappa$  represents the common factor mean at pretest and at posttest, respectively.

Where  $\kappa_{(\text{Pre})} \neq \kappa_{(\text{Post})}$ , this was interpreted as true change in the means from pretest to posttest (Oort, 2005b). Given that self-management courses aim to induce change in the participants, it was expected that differences would be observed when comparing the pretest factor means with the posttest factor means. Hence, Step 4 was not an invariance test as such but it assessed the amount of change in participants. Given that the different types of response shifts had already been accounted for in Step 3, the size of ‘true change’ could be quantified.



‘Observed change’ has been defined as ‘true change’ plus ‘response shift’ (Visser *et al.*, 2005); hence, in this final step of the model – if significant response shifts are found – the magnitude of the overall response shift could be assessed.

The four steps of the new model are again summarised in Table 19.

**Table 19** 4-step procedure for the test of measurement invariance

Steps	Parameters		Interpretation	Comment
Step 1 – Configural invariance	$\text{Patt}(\Lambda_x)$	Factor patterns	Reconceptualisation	This invariance is essential for the next steps.
Step 2 – Full response shift model of complete invariance	$\Lambda_x$	Factor loadings	Reprioritisation	This step tests for overall response shift in the model.
	$\tau_x$	Intercepts	Recalibration (uniform)	
	$\text{Diag}(\theta_\delta)$	Residual variances	Recalibration (non-uniform)	
	$\Phi$	Factor covariances	Higher-level reconceptualisation	
Step 3 – Invariance of all tenable response shift constraints	$\Lambda_x$	Factor loadings	Reprioritisation	Only the tenable constraints from Step 2 are kept, i.e. those that are not tenable indicate response shift.
	$\tau_x$	Intercepts	Recalibration (uniform)	
	$\text{Diag}(\theta_\delta)$	Residual variances	Recalibration (non-uniform)	
	$\Phi$	Factor covariances	Higher-level reconceptualisation	
Step 4 – Assessment of true change	$\kappa$	Factor means	True mean change	This step assesses true change in factor means and factor variances.
	$\text{Diag}(\Phi)$	Factor variances	True change in group’s variances	

The application of Steps 1 to 4 to heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro

In the following four sections, the new model was applied to the three randomised groups and the sample of retrospective pretest-posttest data. Given that the effective sample sizes were too small to include all 38 heiQ items in a single model (Tanaka, 1987), the analyses were performed on the heiQ subscale-level. This is one possibility to circumvent the sample size problem (Millsap & Hartog, 1988). Given that analyses on one-factor models are not suitable to test configural invariance or invariance of factor correlations, additional models

needed to be specified that combined several subscales at a time. As will be explained in detail in Section 5.4.1, the factor combinations 1,2,3,5 and 4,6,7,8 respectively, were chosen for this exercise. While Positive and Active Engagement in Life, Health-Directed Behaviour, Skill and Technique Acquisition, and Self-Monitoring and Insight were defined in one model, Constructive Attitudes and Approaches, Health Service Navigation, Social Integration and Support, and Emotional Well-Being were combined in the second four-factor model.

## 5.4 Response shift – heiQ-PP

### 5.4.1 Specific methods

The eight one-factor models were specified in the first part of the analyses. Given that these models render the exploration of factor patterns as well as factor covariances impossible, the analyses were limited to the investigation of invariance of error variances, error covariances, item intercepts, and factor loadings. In LISREL notation these tests were:

- $\text{Diag}(\theta_{\delta(\text{Pre})}) = \text{Diag}(\theta_{\delta(\text{Post})})?$
- $\theta_{\delta(\text{Pre})} = \theta_{\delta(\text{Post})}?$
- $\tau_{x(\text{Pre})} = \tau_{x(\text{Post})}?$
- $\Lambda_{x(\text{Pre})} = \Lambda_{x(\text{Post})}?$

After specifying unconstrained baseline models of each of the one-factor models, the fully constrained response shift model was investigated for each of the heiQ subscales (Step 2). If either of these response shift models was significantly worse than its baseline model, Step 3 was applied to successively explore the above parameters. In accordance with the hierarchy of the invariance constraints, all non-tenable error variances and error covariances were de-constrained first. If model fit was still not satisfactory, then the constraints on intercepts, and finally those on the factor loadings, were relaxed. In some cases, parameters were proposed after having de-constrained some of the parameters that were more important in view of the hierarchy of measurement invariance. For example, after having de-constrained some factor loadings, some more intercepts may have been suggested. These were then modified if they improved model fit significantly.

As will be seen in subsequent analyses (see Sections 5.4.2, 5.5.2, 5.6.2, 5.7.2), the results of the one-factor models regularly indicated response shifts in the heiQ subscales 1,2,3,5, while subscales 4,6,7,8 appeared to be largely free of response shift. Consequently, it was decided to use these factor combinations to specify the four-factor models. The rationale for choosing these combinations was twofold: 1) the first combination of subscales was chosen to explore

whether observed response shifts had been a result of response shifts within subscales or whether the interaction of different subscales had influenced the results. Given that these heiQ subscales had regularly shown response shifts in the one-factor models, this four-factor model had the highest potential to detect further response shifts if they existed. And 2) the aim of the model including subscales 4,6,7,8 was to confirm that no response shift had occurred in any of these four heiQ subscales given that hardly any response shift had been detected in their respective one-factor models. Hence, if invariance of factor patterns and factor covariances could be established in addition to the other parameters, results would become more robust. In summary, these four-factor models investigated invariance of factor patterns ( $\text{Patt}(\Lambda_x)$ ) and factor covariances ( $\Phi_{(\text{Pre})}$ ) in addition to the parameters that had been explored in the one-factor models.

### 5.4.2 Results

As presented in Table 20, the specification of the eight baseline models (BL) of heiQ-PP suggested satisfactory fit statistics across all subscales (Step 1). Apart from Health Service Navigation, and Social Integration and Support all models had a non-significant  $\chi^2_{\text{SB}}$ . Further, RMSEA, CFI, and SRMR indicated good fit of the models. As a result, none of the baseline models required modification.

The response shift models (RS) of Step 2 indicated a significant difference to the baseline models in Health-Directed Behaviour, Skill and Technique Acquisition, and Self-Monitoring and Insight, while the remaining subscales appeared to be free of response shift. Therefore, Step 3 only needed to be performed on three subscales. After successively de-constraining the parameters, one constraint on the error variance of item 2\_3 was not tenable in Health-Directed Behaviour, indicating a non-uniform recalibration. Given that no further constraints on error variances needed to be relaxed, the constraints on the intercepts were examined. Initially, all items appeared to meet scalar invariance; hence, factor loadings were explored which suggested a reprioritisation of item 2\_4. Once this constraint was relaxed, the intercept of item 2\_2 indicated to be non-invariant (uniform recalibration). After de-constraining the three parameters, the adjusted model showed a similar fit compared with the baseline model. In contrast, in Skill and Technique Acquisition only two error variances (item 3\_1; item 3\_5) were non-invariant, each indicating non-uniform recalibration. Finally, in Self-Monitoring and Insight, one error variance and one loading were not invariant. While results suggested a non-uniform recalibration of item 5\_1, item 5\_5 had changed in priority within this subscale. After allowing for non-invariance of the aforementioned parameters, the difference between each subscale's final model and its respective baseline was non-significant.

**Table 20** Fit indices of the response shift detection procedure, one-factor models of heiQ-PP

Step	Constr. parameter	$\chi^2_{SB}$	df	p-value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1	41.1	29	NS	0.036	(0.000;0.059)	0.48	0.997	0.039
2	RS	60.7	43	0.039	0.035	(0.009;0.055)	0.51	0.995	0.061
1	BL 2	16.2	15	NS	0.016	(0.000;0.056)	0.29	1.000	0.033
2*	RS	53.8	26	0.001	0.057	(0.035;0.079)	0.46	0.991	0.105
3*	TD2_3	45.1	25	0.008	0.049	(0.025;0.072)	0.41	0.993	0.096
3*	<b>LX2_4</b>	33.2	24	NS	0.034	(0.000;0.060)	0.34	0.997	0.062
3*	<b>TX2_2</b>	29.3	23	NS	0.029	(0.000;0.057)	0.33	0.998	0.061
1	BL 3	20.1	15	NS	0.032	(0.000;0.065)	0.28	0.997	0.035
2*	RS	56.8	26	<0.001	0.060	(0.039;0.081)	0.42	0.984	0.060
3*	TD3_1	42.0	25	0.018	0.045	(0.019;0.069)	0.33	0.991	0.056
3*	TD3_5	35.5	24	NS	0.038	(0.000;0.063)	0.30	0.994	0.054
1	BL 4	30.0	29	NS	0.010	(0.000;0.044)	0.44	1.000	0.034
2	RS	36.7	43	NS	0.000	(0.000;0.028)	0.23	1.000	0.040
1	BL 5	63.7	47	NS	0.033	(0.000;0.052)	0.69	0.994	0.060
2*	RS	102.6	64	0.002	0.043	(0.027;0.058)	0.80	0.986	0.078
3*	TD5_1	94.7	63	0.006	0.039	(0.021;0.055)	0.75	0.988	0.077
3*	<b>LX5_5</b>	82.7	62	0.041	0.032	(0.007;0.049)	0.69	0.992	0.069
1	BL 6	31.7	15	0.007	0.058	(0.029;0.084)	0.44	0.994	0.047
2	RS	49.3	26	0.004	0.052	(0.029;0.074)	0.48	0.992	0.066
1	BL 7	34.6	15	0.003	0.063	(0.035;0.091)	0.48	0.994	0.043
2	RS	43.5	26	0.017	0.045	(0.019;0.068)	0.47	0.995	0.050
1	BL 8	59.0	47	NS	0.028	(0.000;0.048)	0.60	0.998	0.035
2	RS	82.3	64	NS	0.029	(0.000;0.047)	0.67	0.997	0.074

\* Scaled chi-square difference significant at the  $p < 0.05$  level

Legend

Constr. parameter: Constrained parameter  
 BL 1-8: Baseline model of subscales 1-8  
 RS: Response shift model / fully constrained model  
 TD: Theta-delta = error variances or covariances  
**TX:** **Tau X (item intercepts), relevant for an unbiased comparison of factor means**  
**LX:** **Lambda-X = factor loadings, relevant for an unbiased comparison of factor means**

The respective baseline models (BL) of each of the four-factor models showed satisfactory fit indices (see Table 21). Apart from showing a highly significant  $\chi^2_{SB}$ , which is likely due to the size of the model rather than misspecification or misfit, all remaining fit statistics suggested good fit of these models. The factor combination of subscales 1,2,3,5, however, suggested a loading of item 5\_6 on subscale Health-Directed Behaviour in the pretest dataset. Given that this cross-loading had been suggested in the heiQ re-validation (see Section 4.4.3) and the

content of the item (“I carefully watch my health and do what is necessary to keep as healthy as possible”) had elements of the Health-Directed Behaviour construct, the baseline model was adjusted by allowing for this cross-loading in the pretest as well as the posttest model. Given that model fit improved significantly, the adjusted model was used in Steps 2 and 3.

Step 2 (RS) of factor combination 1,2,3,5 indicated significantly worse model fit compared with its baseline model (see Table 21). After successively relaxing some of the constraints in Step 3, it was found that three error variances and two factor loadings were not invariant. These had been detected in the one-factor models, i.e. items 2\_3, 3\_1, and 5\_1 indicated a non-uniform recalibration and items 2\_4 and 5\_5 suggested a reprioritisation within their respective heiQ subscale. After allowing for non-invariance of these five parameters, the adjusted model had a similar fit to the baseline model.

The factor combination 4,6,7,8 did not find any evidence of the presence of response shift, supporting the results of the one-factor models that these subscales were free of this bias.

**Table 21** Fit indices of the response shift detection procedure, four-factor models of heiQ-PP

Step	Constr. parameter	$\chi^2_{SB}$	df	<i>p</i> -value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1235	870.8	618	<0.001	0.035	(0.030;0.041)	6.53	0.991	0.070
1*	Modified	845.4	616	<0.001	0.034	(0.028;0.039)	6.39	0.992	0.066
2*	RS	997.9	676	<0.001	0.038	(0.033;0.043)	6.81	0.989	0.080
3*	TD3_1	983.1	675	<0.001	0.037	(0.032;0.042)	6.72	0.989	0.080
3*	TD2_3	973.1	674	<0.001	0.037	(0.032;0.042)	6.68	0.990	0.080
3*	LX2_4	961.5	673	<0.001	0.036	(0.031;0.041)	6.63	0.990	0.075
3*	TD5_1	948.0	672	<0.001	0.035	(0.030;0.040)	6.56	0.990	0.075
3*	LX5_5	939.3	671	<0.001	0.035	(0.029;0.040)	6.52	0.991	0.074
1	BL4678	848.5	618	<0.001	0.034	(0.028;0.039)	6.40	0.993	0.069
2	RS	910.3	677	<0.001	0.032	(0.027;0.038)	6.46	0.993	0.075

\* Scaled chi-square difference significant at the *p*<0.05 level

Legend

BL 1235 / BL 4678: Baseline model combining subscales 1,2,3,5 and 4,6,7,8 respectively  
 Modified: Cross-loading of item 5\_6 on heiQ subscale 2  
 RS: Response shift model

### 5.4.3 Summary

At the heiQ subscale-level, only a small number of items were found that were non-invariant in heiQ-PP. Three heiQ subscales comprising a total of seven items indicated some form of response shift. With regard to an unbiased comparison of the factor means – which requires

metric and scalar invariance – even fewer items were affected. Two items of Health-Directed Behaviour and one item of Self-Monitoring and Insight would have to be excluded from the calculation of the factor means. Given that the calculation of the factor means was based on the findings of the one-factor models, these results were relevant for Step 4 of the model.

In the four-factor models of heiQ-PP, a total of five items were detected that indicated either non-uniform recalibration or reprioritisation. In view of metric and scalar invariance, two items were affected, i.e. one item of Health-Directed Behaviour and one item of Self-Monitoring and Insight would not be eligible to be included in further mean comparisons. Given that in this thesis the calculation of factor means was based on the results of the one-factor models, the results of the four-factor models only served for the detection of further response shifts. That is, they were used to test configural invariance and invariance of factor correlations (see Section 5.3) but they had no implications for later analyses.

## **5.5 Response shift – heiQ-PPT**

### **5.5.1 Specific methods**

Given that the procedure of the analyses has been presented in the previous Section 5.4.1, no additional explanations are provided for the analyses of heiQ-PPT. Hence, the results of the eight one-factor and the two four-factor models are described hereafter.

### **5.5.2 Results**

As presented in Table 22, the baseline models (BL) of group heiQ-PPT indicated satisfactory fit indices across all subscales. Apart from three subscales, all models had a non-significant  $\chi^2_{SB}$ . Further, RMSEA, CFI, and SRMR indicated generally good fit of all models. Therefore, none of the baseline models required modification.

The response shift model (RS) resulted in only one heiQ subscale that showed a significant  $\chi^2_{SB}$  difference test. While seven subscales appeared to be free of response shift bias, the intercept of item 1\_4 (uniform recalibration) and the loadings of items 1\_1, 1\_2, and 1\_4 (reprioritisation) were found to be non-invariant in Positive and Active Engagement in Life. Furthermore, heiQ-PPT was the only dataset that suggested a correlated error, i.e. the errors of items 1\_2 and 1\_3 were found to correlate at posttest but not at actual pretest, suggesting a lower-level reconceptualisation. After allowing for the free estimation of these parameters, the difference between the final and the baseline model was non-significant.

**Table 22** Fit indices of the response shift detection procedure, one-factor models of heiQ-PPT

Step	Constr. parameter	$\chi^2_{SB}$	df	<i>p</i> -value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1	41.4	29	NS	0.038	(0.000;0.062)	0.55	0.996	0.053
2*	RS	79.8	43	<0.001	0.053	(0.035;0.071)	0.72	0.989	0.137
3*	<b>TX1_4</b>	75.1	42	0.001	0.051	(0.032;0.069)	0.70	0.990	0.136
3*	<b>LX1_1</b>	63.4	41	0.014	0.043	(0.020;0.062)	0.64	0.993	0.112
3*	TD(Post) 1_2F/1_3F	58.9	40	0.027	0.040	(0.014;0.060)	0.61	0.994	0.116
3*	<b>LX1_2</b>	50.0	39	NS	0.031	(0.000;0.053)	0.56	0.997	0.094
3*	<b>LX1_4</b>	39.1	38	NS	0.010	(0.000;0.042)	0.49	1.000	0.060
1	BL 2	17.0	15	NS	0.021	(0.000;0.060)	0.30	0.999	0.031
2	RS	26.1	26	NS	0.003	(0.000;0.046)	0.29	1.000	0.075
1	BL 3	36.0	15	0.002	0.068	(0.040;0.097)	0.41	0.982	0.044
2	RS	54.8	25	<0.001	0.061	(0.038;0.083)	0.46	0.975	0.072
1	BL 4	35.0	29	NS	0.026	(0.000;0.054)	0.55	0.999	0.037
2	RS	41.7	43	NS	0.000	(0.000;0.037)	0.25	1.000	0.046
1	BL 5	45.2	47	NS	0.000	(0.000;0.035)	0.40	1.000	0.054
2	RS	66.8	64	NS	0.012	(0.000;0.037)	0.65	0.998	0.074
1	BL 6	20.7	15	NS	0.035	(0.000;0.069)	0.35	0.998	0.027
2	RS	38.0	26	NS	0.039	(0.000;0.064)	0.42	0.996	0.049
1	BL 7	30.3	15	0.011	0.058	(0.027;0.088)	0.41	0.994	0.044
2	RS	37.8	26	NS	0.038	(0.000;0.064)	0.39	0.996	0.062
1	BL 8	67.9	47	0.025	0.038	(0.014;0.057)	0.79	0.997	0.040
2	RS	84.5	64	0.044	0.033	(0.006;0.050)	0.79	0.997	0.060

\* Scaled chi-square difference significant at the  $p < 0.05$  level

#### Legend

For an extensive legend refer to Table 20

As shown in Table 23 the four-factor models of heiQ-PPT suggested satisfactory fit statistics of the baseline models (BL). Apart from each showing a highly significant  $\chi^2_{SB}$ , all remaining fit indices were satisfactory. In Step 2, the response shift model of factor combination 1,2,3,5 (RS) showed a significantly worse fit compared with its baseline model. The three factor loadings that had already been detected in the one-factor models were again non-invariant. When comparing the final model of this four-factor model with its baseline, it was found that it was still significantly worse. Given that no further single constraint could be identified that was strong enough to significantly improve the fit of the final model, the analyses were stopped. In contrast, model 4,6,7,8 again was found to be free of response shift bias.

**Table 23** Fit indices of the response shift detection procedure, four-factor models of heiQ-PPT

Step	Constr. parameter	$\chi^2_{SB}$	df	<i>p</i> -value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1235	884.2	618	<0.001	0.038	(0.032;0.043)	7.32	0.988	0.072
2*	RS	1028.0	677	<0.001	0.041	(0.036;0.046)	7.73	0.984	0.090
3*	LX1_1	1010.2	676	<0.001	0.040	(0.035;0.046)	7.64	0.985	0.086
3*	LX1_2	994.9	675	<0.001	0.040	(0.034;0.045)	7.56	0.985	0.084
3*	LX1_4	983.4	674	<0.001	0.039	(0.034;0.044)	7.50	0.986	0.080
1	BL 4678	782.5	618	<0.001	0.030	(0.023;0.036)	6.61	0.994	0.060
2	RS	838.4	677	<0.001	0.028	(0.021;0.034)	6.60	0.994	0.067

\* Scaled chi-square difference significant at the  $p < 0.05$  level

#### Legend

BL 1235 / BL 4678: Baseline model combining subscales 1,2,3,5 and 4,6,7,8 respectively  
 RS: Response shift model

### 5.5.3 Summary

Only few response shifts were detected in heiQ-PPT, with a total of five items of subscale Positive and Active Engagement in Life being affected. As a result, full metric and scalar invariance could be established in seven heiQ subscales, while partial invariance could be established in Positive and Active Engagement in Life. To ensure an unbiased comparison of the factor means in this subscale, three items needed to be excluded from the calculation of the factor means. In the four-factor models, the three factor loadings of Positive and Active Engagement in Life again proved to be non-invariant.

## 5.6 Response shift – heiQ-PPR

### 5.6.1 Specific methods

As per the previous analyses on pretest-posttest data, the findings of the one- and four-factor models of group heiQ-PPR are presented hereafter.

### 5.6.2 Results

As presented in Table 24, the fit indices of all baseline models (BL) of heiQ-PPR indicated satisfactory model fit. Some subscales suggested slightly worse model fit compared with the



results of heiQ-PP and heiQ-PPT with five subscales showing a significant  $\chi^2_{SB}$ . However, given that all other fit statistics indicated that the fit of the baseline models was adequate, no further adjustments appeared necessary.

**Table 24** Fit indices of the response shift detection procedure, one-factor models of heiQ-PPR

Step	Constr. parameter	$\chi^2_{SB}$	df	p-value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1	46.6	29	0.021	0.044	(0.018;0.067)	0.54	0.995	0.050
2*	RS	89.3	43	<0.001	0.059	(0.041;0.076)	0.70	0.987	0.071
3*	<b>TX1_3</b>	66.5	42	0.009	0.043	(0.022;0.062)	0.58	0.993	0.060
1	BL 2	34.5	15	0.003	0.064	(0.036;0.093)	0.41	0.992	0.048
2*	RS	60.0	26	<0.001	0.065	(0.043;0.086)	0.53	0.987	0.057
3*	TD2_3	54.2	25	<0.001	0.061	(0.039;0.083)	0.48	0.989	0.061
3*	<b>TX2_4</b>	50.7	24	0.001	0.058	(0.036;0.080)	0.45	0.990	0.054
3*	TD2_2	43.2	23	0.007	0.053	(0.028;0.077)	0.42	0.992	0.052
1	BL 3	14.9	15	NS	0.000	(0.000;0.053)	0.21	1.000	0.038
2	RS	31.6	26	NS	0.026	(0.000;0.054)	0.30	0.996	0.059
1	BL 4	34.5	29	NS	0.025	(0.000;0.052)	0.49	0.999	0.031
2	RS	55.1	43	NS	0.030	(0.000;0.051)	0.55	0.997	0.071
1	BL 5	75.5	47	0.005	0.044	(0.024;0.062)	0.81	0.987	0.058
2*	RS	111.9	64	<0.001	0.049	(0.033;0.064)	0.88	0.977	0.077
3*	TD5_5	105.7	63	<0.001	0.047	(0.030;0.062)	0.85	0.980	0.073
3*	<b>TX5_4</b>	100.3	62	0.002	0.044	(0.028;0.060)	0.83	0.982	0.070
3*	<b>TX5_5</b>	95.2	61	0.003	0.042	(0.025;0.058)	0.81	0.984	0.066
1	BL 6	25.1	15	0.049	0.046	(0.004;0.077)	0.38	0.996	0.040
2	RS	36.8	26	NS	0.036	(0.000;0.062)	0.40	0.996	0.057
1	BL 7	20.4	15	NS	0.034	(0.000;0.068)	0.34	0.998	0.028
2	RS	26.2	26	NS	0.005	(0.000;0.045)	0.33	1.000	0.041
1	BL 8	99.3	47	<0.001	0.060	(0.043;0.076)	0.84	0.991	0.049
2	RS	111.8	64	<0.001	0.049	(0.033;0.064)	0.84	0.991	0.073

\* Scaled chi-square difference significant at the  $p < 0.05$  level

Legend

For an extensive legend refer to Table 20

After constraining all relevant response shift parameters in Step 2 (RS), three subscales indicated a significant decline in model fit. In Positive and Active Engagement in Life, one intercept was non-invariant, suggesting a uniform recalibration in item 1\_3. Moreover, three items were non-invariant in Health-Directed Behaviour (non-uniform recalibration of items 2\_2 and 2\_3; uniform recalibration of item 2\_4). Finally, in Self-Monitoring and Insight, a non-uniform recalibration was observed in item 5\_5, and uniform recalibrations were found in

items 5\_4 and 5\_5. After allowing for non-invariance of these parameters, the respective difference between baseline and final model of the three subscales was non-significant.

Model fit of the four-factor models was again satisfactory when no constraints were imposed on the respective models (see Table 25). Despite a significant  $\chi^2_{SB}$ , the remaining goodness-of-fit indices indicated good fit of the baseline models (BL).

The results of the response shift models (RS) were similar to those of the previous datasets. Factor combination 4,6,7,8 again showed robust results in the sense that no item indicated any form of response shift. In contrast, factor combination 1,2,3,5 indicated a significantly worse fit compared with its baseline model. Response shift was detected in five items. Apart from one item intercept that had not been detected in the one-factor model of Positive and Active Engagement in Life, all remaining parameters had already been detected in the one-factor models. After allowing for non-invariance of the aforementioned parameters, no further single item showed response shift that was strong enough to improve model fit significantly. Hence, Step 3 of the analyses of model 1,2,3,5 had to be stopped with a final model that was still significantly worse than its baseline model.

**Table 25** Fit indices of the response shift detection procedure, four-factor models of heiQ-PPR

Step	Constr. parameter	$\chi^2_{SB}$	df	p-value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1235	931.1	618	<0.001	0.040	(0.035;0.045)	7.41	0.986	0.067
2*	RS	1092.9	677	<0.001	0.044	(0.039;0.049)	7.88	0.982	0.079
3*	TD5_5	1074.2	676	<0.001	0.043	(0.039;0.048)	7.78	0.983	0.078
3*	TD2_3	1053.3	675	<0.001	0.042	(0.037;0.047)	7.65	0.984	0.078
3*	TX5_5	1040.5	674	<0.001	0.042	(0.037;0.047)	7.60	0.984	0.079
3*	LX1_3	1034.9	673	<0.001	0.041	(0.036;0.046)	7.58	0.984	0.079
3*	TX1_4	1027.1	672	<0.001	0.041	(0.036;0.046)	7.56	0.985	0.078
1	BL 4678	804.5	618	<0.001	0.031	(0.025;0.037)	6.50	0.993	0.056
2	RS	873.8	677	<0.001	0.031	(0.024;0.036)	6.62	0.993	0.065

\* Scaled chi-square difference significant at the  $p < 0.05$  level

Legend

BL 1235 / BL 4678: Baseline model combining subscales 1,2,3,5 and 4,6,7,8 respectively

### 5.6.3 Summary

A total of six items indicated at least one type of response shift in the one-factor models of heiQ-PPR. In contrast to the previous two groups, however, all factor loadings were invariant over time. Hence, full metric invariance could be established for all eight subscales. In the context of an unbiased comparison of factor means, this meant that only scalar invariance

had to be established. Accordingly, two items of Self-Monitoring and Insight, and one of each Positive and Active Engagement in Life, and Health-Directed Behaviour had to be excluded from the calculation of factor means. The analyses of the four-factor models detected three items that were not suitable for the comparison of factor means in model 1,2,3,5. The results of the four-factor models, however, had no implications for the calculation of factor means (see Section 5.4.3).

## **5.7 Measurement invariance – heiQ-PPR Retro**

### **5.7.1 Specific methods**

In contrast to the previous three analyses based on actual pretest-posttest data, this section investigated measurement invariance in the sample of retrospective pretest-posttest data. As discussed in Section 1.2.4.4, the collection of retrospective pretests has been proposed as a possible remedy to circumvent the influence of response shift bias which might threaten the validity of change scores derived from actual pretest-posttest data (Howard & Dailey, 1979; Terborg *et al.*, 1980). Because respective ratings of posttest and retrospective pretest levels are provided in close proximity, it has been argued that the two ratings are provided from the same perspective (Howard, Schmeck *et al.*, 1979). As a consequence, items should not be affected by recalibration, reconceptualisation, or reprioritisation. While it has been highlighted that a comparison of the two scores is only valid if both are affected by response shift in the same way (Oort *et al.*, 2003), it is the basic assumption of this method that change scores derived from retrospective pretest-posttest data are not confounded by response shift bias.

Although the present dataset contained retrospective rather than actual pretests, the model of the previous three sections (see Section 5.4.1) could be applied to explore measurement invariance in these data. Following from the above introduction, it was expected that these data would show that items were invariant across the two measurement occasions.

### **5.7.2 Results**

As shown in Table 26, the baseline models (BL) of most subscales indicated slightly worse fit than the baseline models of the previous three samples of actual pretest-posttest data. While most fit indices were satisfactory, the SRMR was generally larger than in the other datasets. While the SRMR was still acceptable in seven heiQ subscales, it was deemed too large in the baseline model of Positive and Active Engagement in Life (BL 1). After testing several alternatives, the best fitting model was obtained when item 1\_1 “I am doing interesting things

in my life” was excluded. The baseline model of this subscale was therefore re-specified, with this modified model being used in all subsequent steps.

**Table 26** Fit indices of the response shift detection procedure, one-factor models of heiQ-PPR Retro

Step	Constr. parameter	$\chi^2_{SB}$	df	p-value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1	46.8	29	0.020	0.044	(0.018;0.067)	0.70	0.996	0.096
1*	Modified**	24.9	15	NS	0.046	(0.000;0.077)	0.38	0.996	0.085
2*	RS	31.3	26	NS	0.026	(0.000;0.054)	0.42	0.998	0.118
3*	LX1_5	28.0	25	NS	0.020	(0.000;0.051)	0.40	0.999	0.110
3*	TX1_4	24.1	24	NS	0.003	(0.000;0.046)	0.23	1.000	0.103
1	BL 2	20.4	15	NS	0.034	(0.000;0.067)	0.35	0.998	0.071
2*	RS	65.7	26	<0.001	0.070	(0.049;0.091)	0.80	0.988	0.138
3*	TX2_3	24.1	25	NS	0.000	(0.000;0.043)	0.18	1.000	0.101
3*	TX2_4	19.8	24	NS	0.000	(0.000;0.036)	0.18	1.000	0.090
3*	LX2_3	16.8	23	NS	0.000	(0.000;0.029)	0.18	1.000	0.075
1	BL 3	18.1	15	NS	0.026	(0.000;0.062)	0.32	0.998	0.032
2*	RS	46.9	26	0.007	0.051	(0.026;0.074)	0.49	0.988	0.063
3*	LX3_4	30.5	25	NS	0.027	(0.000;0.055)	0.38	0.997	0.054
1	BL 4	29.5	29	NS	0.008	(0.000;0.056)	0.62	1.000	0.052
2*	RS	66.3	43	0.013	0.042	(0.020;0.061)	0.91	0.995	0.099
3*	TX4_2	51.3	42	NS	0.027	(0.000;0.049)	0.77	0.998	0.086
3*	TX4_1	44.0	41	NS	0.015	(0.000;0.043)	0.71	0.999	0.073
1	BL 5	66.7	47	0.031	0.037	(0.012;0.056)	1.07	0.994	0.075
2*	RS	113.8	64	<0.001	0.050	(0.035;0.065)	1.39	0.984	0.095
3*	TX5_3	103.5	63	<0.001	0.045	(0.029;0.061)	1.31	0.987	0.084
3*	TX5_5	99.4	62	0.002	0.044	(0.027;0.059)	1.29	0.988	0.077
3*	LX5_6	89.9	61	0.010	0.039	(0.020;0.055)	1.21	0.991	0.075
3*	LX5_1	85.4	60	0.017	0.037	(0.016;0.054)	1.17	0.992	0.075
1	BL 6	15.4	15	NS	0.010	(0.000;0.055)	0.43	1.000	0.046
2	RS	23.7	26	NS	0.000	(0.000;0.040)	0.17	1.000	0.086
1	BL 7	29.9	15	0.012	0.056	(0.025;0.086)	0.51	0.996	0.044
2	RS	36.7	26	NS	0.036	(0.000;0.062)	0.56	0.997	0.066
1	BL 8	75.6	47	0.005	0.044	(0.024;0.062)	1.05	0.995	0.073
2*	RS	119.3	64	<0.001	0.053	(0.038;0.067)	1.39	0.991	0.102
3*	TD8_3	111.8	63	<0.001	0.050	(0.034;0.065)	1.30	0.992	0.097
3*	TD8_4	100.9	62	0.001	0.045	(0.028;0.060)	1.19	0.994	0.095
3*	TD8_5	94.2	61	0.004	0.042	(0.024;0.058)	1.14	0.995	0.091

\* Scaled chi-square difference significant at the  $p < 0.05$  level

\*\* Model excluding item q1\_1

**Legend**

For an extensive legend refer to Table 20

In Step 2<sup>18</sup>, the models of heiQ-PPR Retro indicated significantly worse model fit compared with respective baseline models in five subscales. That is, in Health-Directed Behaviour, Skill and Technique Acquisition, Constructive Attitudes and Approaches, Self-Monitoring and Insight, and Emotional Well-Being, a significant  $\chi^2_{SB}$  difference test was observed, with non-invariant items being found in Constructive Attitudes and Approaches, and Emotional Well-Being for the first time. Moreover, while the model of Positive and Active Engagement in Life did not indicate a significantly worse fit, SRMR moved from 0.085 to an unacceptable value of 0.118. Consequently, this heiQ subscale was also investigated in Step 3 regardless of the non-significant  $\chi^2_{SB}$  difference test, leading to a total of six heiQ subscales needing further adjustments because the models of Step 2 were worse than respective baseline models.

In detail, the following items were non-invariant between posttest and retrospective pretest across subscales. In subscale Positive and Active Engagement in Life, item 1\_5 showed a non-invariant loading and item 1\_4 indicated a non-invariant intercept. After relaxing the two constraints no further single parameter was found that led to a significant improvement of the model. Hence, despite the SRMR of the final model being too large, the re-specification was stopped here. In contrast, after de-constraining the intercepts of items 2\_3 and 2\_4, and the factor loading of item 2\_3, model fit of subscale Health-Directed Behaviour was again close to the baseline model. In Skill and Technique Acquisition, the loading of item 3\_4 was non-invariant, while in Constructive Attitudes and Approaches, items 4\_1 and 4\_2 each showed a non-invariant intercept. Some more constraints were found in subscale Self-Monitoring and Insight, with non-tenable constraints on the intercepts of items 5\_3 and 5\_5, as well as the factor loadings of 5\_1 and 5\_6. Finally, in Emotional Well-Being, the error variances of items 8\_3, 8\_4, and 8\_5 were non-invariant.

The two four-factor models indicated good fit at baseline (see Table 27). With each showing  $RMSEA < 0.05$ ,  $CFI > 0.98$ , and  $SRMR \sim 0.06$ , none of the baseline models (BL) needed further adjustments. In contrast to all other datasets, however, both models indicated non-positive definite sample covariance matrices. Hence, at least some of the elements of the respective covariance matrix did not meet certain conditions that are necessary for further mathematical operations. To circumvent non-positive definite matrices, a ridge correction can be applied (Wothke, 1993). Given that in the present analyses LISREL 8.72 automatically applied this procedure and provided all parameter estimates, results still appeared robust. Furthermore, in model 1,2,3,5 the Phi matrix was non-positive definite. Given that this violation of positive definiteness might occur despite a correctly specified model, it has been suggested that this warning message can be ignored (Rigdon, 1997).

---

<sup>18</sup> In the present analyses of heiQ-PPR Retro, "RS" reflects a "Fully constrained model" rather than a "Response shift model".

When moving to Step 2 (RS), both models indicated a significantly worse model fit compared with their respective baseline model. In model 1,2,3,5, the constraints on the intercepts of items 1\_2, 1\_4, 2\_3, 2\_4, 3\_4, 5\_3, 5\_5 and the constraints on the loadings of items 2\_3 and 3\_4 were not tenable. In model 4,6,7,8, the error variances of items 4\_4, 8\_3, 8\_4, 8\_5 and the intercept of item 4\_1 were non-invariant. With the exception of the intercepts of items 1\_2 and 3\_4, and the error variance of item 4\_4, these parameters had already been suggested in the one-factor models. After de-constraining the parameters, the final model of 1,2,3,5 had a similar fit to its baseline model, while model 4,6,7,8 was slightly worse than its baseline.

**Table 27** Fit indices of the response shift detection procedure, four-factor models of heiQ-PPR Retro

Step	Constr. parameter	$\chi^2_{SB}$	df	p-value	RMSEA	RMESA (90% CI)	ECVI	CFI	SRMR
1	BL 1235	993.6	618	<0.001	0.044	(0.039;0.049)	5.63	0.985	0.060
2*	RS	1217.9	677	<0.001	0.051	(0.046;0.055)	6.23	0.979	0.080
3*	TX2_3	1131.0	676	<0.001	0.046	(0.042;0.051)	5.90	0.982	0.077
3*	TX1_4	1119.8	675	<0.001	0.046	(0.041;0.051)	5.86	0.982	0.074
3*	TX1_2	1103.2	674	<0.001	0.045	(0.040;0.050)	5.81	0.983	0.072
3*	TX5_3	1088.9	673	<0.001	0.044	(0.040;0.049)	5.75	0.984	0.069
3*	TX5_5	1082.6	672	<0.001	0.044	(0.039;0.049)	5.74	0.984	0.069
3*	LX3_4	1072.2	671	<0.001	0.044	(0.039;0.049)	5.71	0.984	0.068
3*	TX3_4	1064.3	670	<0.001	0.043	(0.038;0.048)	5.69	0.984	0.068
3*	TX2_4	1059.2	669	<0.001	0.043	(0.038;0.048)	5.68	0.985	0.068
3*	LX2_3	1046.8	668	<0.001	0.043	(0.038;0.047)	5.64	0.985	0.066
1	BL 4678	701.4	618	<0.001	0.021	(0.011;0.028)	8.74	0.998	0.063
2*	RS	813.5	677	<0.001	0.025	(0.018;0.032)	9.30	0.996	0.081
3*	TD8_3	803.1	676	<0.001	0.025	(0.017;0.031)	9.24	0.996	0.080
3*	TX4_1	794.4	675	<0.001	0.024	(0.016;0.030)	9.17	0.996	0.077
3*	TD8_4	788.6	674	<0.001	0.023	(0.015;0.030)	9.12	0.997	0.077
3*	TD8_5	781.0	673	<0.001	0.023	(0.014;0.029)	9.06	0.997	0.077
3*	TD4_4	771.9	672	<0.001	0.022	(0.013;0.029)	8.95	0.997	0.076

\* Scaled chi-square difference significant at the  $p < 0.05$  level

Legend

BL 1235 / BL 4678: Baseline model combining subscales 1,2,3,5 and 4,6,7,8 respectively

### 5.7.3 Summary

In the sample of retrospective pretest-posttest data, the 4-step procedure (Oort, 2005b; Oort *et al.*, 2005) detected several non-invariant items. With five subscales and a total of 14 items showing non-invariance, substantially more items were non-invariant compared with items of the previous datasets. These findings had further implications for the calculation of the factor

means. While partial metric and scalar invariance could be established in all subscales, items of the following heiQ subscales had to be excluded to ensure an unbiased comparison of the factor means: one item of Skill and Technique Acquisition, two items of each Positive and Active Engagement in Life, Health-Directed Behaviour, and Constructive Attitudes and Approaches, and four items of Self-Monitoring and Insight. In contrast to the three datasets containing actual pre- and posttests, the analyses of the four-factor models of heiQ-PPR Retro not only suggested a significantly worse model fit when constraining all parameters of model 1,2,3,5 but some items of model 4,6,7,8 were also non-invariant in this dataset.

## 5.8 Factor means of heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro

In this final section of the chapter, the results of Step 4 of all four datasets are summarised. As mentioned in Section 5.3, the first three steps of the 4-step procedure were sufficient for the detection of non-invariant items. In contrast, this final step provides information on ‘true’ change, i.e. change scores derived from the factor means of each of these datasets. Given that the previous steps had selected those items that could not be used in the comparison of factor means as they would have biased results, this final step was based on those items that had shown to be invariant between actual pretest and posttest, and retrospective pretest and posttest data, respectively. As a consequence, a comparison of change scores across the four datasets was limited given that some of these did not include all heiQ items.

**Table 28** Change scores across heiQ-PP (n=331), heiQ-PPT (n=304), heiQ-PPR (n=314), and heiQ-PPR Retro (n=314) derived from a means model (SEM) based on only those items that met the condition of scalar and metric invariance (Step 4)

heiQ subscale	heiQ-PP Mean	heiQ-PPT Mean	heiQ-PPR Mean	heiQ-PPR Retro Mean
1. Positive and Active Engagement in Life	0.36*	0.35* <sup>†</sup>	0.83* <sup>†</sup>	0.75* <sup>†</sup>
2. Health-Directed Behaviour	0.36* <sup>†</sup>	0.41*	0.47* <sup>†</sup>	0.56* <sup>†</sup>
3. Skill and Technique Acquisition	0.66*	0.82*	1.11*	1.00* <sup>†</sup>
4. Constructive Attitudes and Approaches	0.23*	0.30*	0.60*	0.44* <sup>†</sup>
5. Self-Monitoring and Insight	0.41* <sup>†</sup>	0.43*	0.73* <sup>†</sup>	0.90* <sup>†</sup>
6. Health Service Navigation	0.26*	0.18*	0.41*	0.35*
7. Social Integration and Support	0.13*	0.23*	0.34*	0.26*
8. Emotional Well-Being	0.25*	0.22*	0.23*	0.31*

\* Significant, i.e. the change score is more than twice its standard error (Bollen, 1989)

<sup>†</sup> The mean change scores of these heiQ subscales are based on a reduced number of items, i.e. items that had not met the condition of metric and scalar invariance were excluded from Step 4 to ensure an unbiased comparison of factor means.

As shown in Table 28, the derived change scores based on factor means refer to only those items that met the condition of metric and scalar invariance as only invariant items allow for an unbiased comparison of factor means. The factor means of those subscales that needed to be based on a reduced number of items are marked with †. These were: Health-Directed Behaviour, and Self-Monitoring and Insight of heiQ-PP; Positive and Active Engagement in Life of heiQ-PPT; Positive and Active Engagement in Life, Health-Directed Behaviour, and Self-Monitoring and Insight of heiQ-PPR, and subscales Positive and Active Engagement in Life, Health-Directed Behaviour, Skill and Technique Acquisition, Constructive Attitudes and Approaches, and Self-Monitoring and Insight of heiQ-PPR Retro.

Despite being partially based on a different set of items, the overall conclusions about group differences were similar to the results of Chapter 3 where change scores had been derived from arithmetic rather than factor means. Although all change scores based on factor means were significant (see Table 28), some large group differences were observed. In particular, when comparing mean change scores of heiQ-PP with those of heiQ-PPR change scores of the latter group were again substantially larger than those of the former group in most heiQ subscales. Furthermore, the comparison of change scores based on factor means of group heiQ-PPT with those of group heiQ-PPR also indicated some large group differences, while change scores based on factor means of heiQ-PP and those of heiQ-PPT generally did not differ substantially. Finally, only small differences between change scores of heiQ-PPR and those of heiQ-PPR Retro were observed.

## 5.9 Discussion

### *Response shift in heiQ-PP, heiQ-PPT, and/or heiQ-PPR*

The first part of this chapter was aimed at investigating whether response shift bias had a confounding effect on the results of evaluations of self-management programs. In particular, it was explored whether the three datasets consisting of actual pre- and posttests (heiQ-PP; heiQ-PPT; heiQ-PPR) were confounded by response shift. Considering the validity concerns about change scores that are derived from actual pretest-posttest data (see Section 1.2.4.4), it was expected that this bias could be detected in all three datasets. However, following from the findings of Chapter 3 (see Section 3.7) it was further expected that the change scores of heiQ-PPR were less affected by response shift compared with those of the other two groups.

Contrary to these expectations, however, none of the analyses detected many non-invariant items in any of the datasets, i.e. less than one fifth of items were non-invariant. Following the definitions proposed in the response shift model that was applied in this study (Oort, 2005b;



Oort *et al.*, 2005), factorial invariance of items between two measurement occasions can be interpreted as absence of group-level response shifts. The only response shifts that could be detected occurred in items belonging to subscales Positive and Active Engagement in Life, Health-Directed Behaviour, Skill and Technique Acquisition, and Self-Monitoring and Insight. In contrast, the items of the remaining four heiQ subscales were free of response shift bias. With reference to the continuum of self-report outcomes that was proposed in Section 1.2.3.3 (see also Figure 8), the degree of appraisal involved in answering items from the former heiQ subscales did not differ from that hypothesised for items from the latter heiQ subscales. Both groups of subscales, i.e. the group of subscales that contained non-invariant items and the group of subscales whose items were invariant, included subscales that had been placed at different points along the continuum. Consequently, the findings of this chapter could not be explained by the continuum of self-report outcomes.

Moreover, while non-invariant items were only detected in specific subscales, i.e. in Positive and Active Engagement in Life, Health-Directed Behaviour, Skill and Technique Acquisition, and Self-Monitoring and Insight, different heiQ items within these subscales were affected in different datasets. That is, the lack of invariance of items was not consistent across datasets, suggesting that response shift bias was not a stable characteristic of certain heiQ items.

The analyses further indicated that the hypothesis of explaining the high posttest levels of heiQ-PPR with a diminished influence of response shift bias was unfounded. Considering that a relatively small number of questionnaire items were found to be non-invariant across datasets, response shift bias could not explain the group differences that had been observed in Chapter 3.

Finally, the last step of the 4-step procedure supported the findings of Chapter 3 in that large group differences in respective change scores existed. After excluding non-invariant items, the analyses suggested that change scores derived from heiQ-PPR were generally larger than those derived from heiQ-PP or heiQ-PPT. Hence, regardless of the method used to assess change scores, i.e. factor means or arithmetic means, differences between the three groups were substantial.

The following interpretations of the findings of the present chapter are proposed:

- a) In view of the overall aim of this thesis, the results of the present chapter suggested that a relatively small number of heiQ items were reconceptualised, reprioritised, and/or recalibrated from actual pretest to posttest across groups. Between 13-18% of the items were affected by response shift, whereas only 8-11% had to be excluded to ensure an unbiased comparison of respective factor means. That none of the heiQ datasets was substantially influenced by response shift suggests that the measurement of outcomes of

self-management programs using actual pretest and actual posttest data is defensible. Consequently, the validity concerns about change scores derived from traditional pretest-posttest data (Hill & Betz, 2005) could not be supported by the heiQ data.

- b) It can further be inferred that, despite the design of the posttest heiQ having an influence on the magnitude of reported actual posttest levels, it did not affect the conceptualisation, prioritisation, and calibration of items. Although the baseline models of heiQ-PPR were marginally worse, the psychometric properties of the heiQ across all pre- and posttests was satisfactory. Again, despite some large differences in the factor means between the three datasets, the majority of heiQ items were invariant despite the cognitive tasks that participants performed at posttest.
- c) The findings of this chapter also provided strong support for the robustness of the heiQ. The psychometric properties of this questionnaire appeared robust across datasets and seemed only marginally influenced by the test situation.
- d) In view of these results, some questions remain unanswered. It was unexpected that only little indication of the presence of response shift was found. While only few studies have explored response shift in participants of self-management programs, one study suggests that up to 70% of participants may experience a response shift (Osborne *et al.*, 2006). One possible reason why these findings differed from the results of this chapter is that the two studies used different methods, i.e. the subject-level analyses applied in the cited study (Osborne *et al.*, 2006) are likely to be more sensitive to detect response shift than the group-level analyses applied in this thesis. Given that group-level analyses are only able to detect response shift bias if it occurred in a relatively large number of participants (Oort, 2005a, 2005b), response shifts at the individual level may have been obscured in these analyses. Hence, more individualised approaches may have been more suitable to detect response shift bias (Donaldson, 2005). However, for the purpose of this research it remains that it could be demonstrated that, at a group level, response shift bias only had minor confounding effects on the results of evaluations of self-management programs.
- e) The somewhat contrasting findings of the present chapter compared with the findings of Chapter 3 require further discussion. While the analyses of Chapter 3 demonstrated that the design of the posttest heiQs influenced the level of the posttest ratings, Chapter 5 suggested that the heiQ performed similarly well across datasets, suggesting that hardly any group differences existed with regard to the psychometric properties of the heiQ. As a result, it has to be concluded that the group differences observed in Chapter 3 are not related to group-level response shifts but to the effect of the cognitive task participants performed at posttest and/or potential other bias.

- f) Finally, the heiQ items may be robust against confounding through response shift. Given that these items had been written in a way that they were not vulnerable to response shift (Osborne *et al.*, 2007), it is plausible that response shift is not a threat to the validity of change scores when using this measurement instrument.

#### *Measurement invariance in the dataset of retrospective pretests and actual posttests*

In contrast to the results of the samples of actual pretest-posttest data, the analyses of the retrospective pretest-posttest data showed unexpectedly more items that were not invariant. When applying the interpretation of the different parameters in terms of the types of response shift (Oort, 2005b; Oort *et al.*, 2005), results suggested more 'response shift' than in the previous samples. Given that the posttest and the retrospective pretest data were assessed at the same time and as a result, change scores should be free of response shift (Howard & Dailey, 1979), alternative explanations why relatively many heiQ items were found to be non-invariant need to be discussed. The following explanations are proposed:

- a) Existing theory defines response shift as a change in perspective due to a significant event in a person's life (Schwartz *et al.*, 2006). As a result, data obtained prior to such an event may not be comparable to those collected after the event. In contrast, when pretest and posttest data are collected simultaneously, with pretest scores being collected in retrospect, it can be assumed that both ratings are provided from the same perspective (Howard & Dailey, 1979). While this is the essential assumption of using retrospective pretest data to derive change scores, it is possible that the two data are not affected by response shift in the same way (Oort *et al.*, 2003). It is therefore possible that the dataset of heiQ-PPR Retro may have been hampered by those participants who attended to the retrospective pretest questions from a different perspective to the one underlying their ratings of the actual posttest questions.
- b) In heiQ-PPR Retro it was further observed that both four-factor models had resulted in a non-positive definite sample covariance matrix, and the model combining heiQ subscales 1,2,3,5 had additionally resulted in a non-positive definite Phi matrix. Although LISREL provided an output that appeared robust, these non-positive definite matrices allude to potential problems within the sample of retrospective pretest-posttest data. For example, the specific design of this posttest heiQ may have resulted in *linear dependencies* in the dataset. Such linear dependencies may have led to instability in the respective estimates that in turn caused the apparent non-invariance between the retrospective pretests and the posttests. While a linear dependency between the two scores may not be detrimental, and the analyses suggested that the scores were two separate constructs, research into

the quality of the obtained scores seems necessary. If a linear dependency existed, some of the items may need to be removed (Wothke, 1993) or the design of the questionnaire may have to be modified, such as collecting the two data on separate pages.

- c) A further aspect of the design of heiQ-PPR may relate to a special form of dependency of the retrospective pretest scores. Given that posttest ratings were provided first, this may have restricted answers to the retrospective pretests. Instead of optimising the responses (Krosnick, 1991, 1999; Krosnick & Alwin, 1987), participants may have responded to the retrospective pretest questions in a way that is consistent with the theories that were introduced in Section 1.2.4.3, for example, *effort justification* (Aronson & Mills, 1959; Hill & Betz, 2005), *implicit theory of stability or change* (Ross, 1989; Schwarz *et al.*, 1998), or *social desirability* (Crowne & Marlowe, 1964; Paulhus, 1991). Being aware that they were providing a measure of change, participants may have deliberately provided ratings at or below their actual posttest levels and as a result, they may have restricted the range of response options to these retrospective pretests. This may be a further reason why several items were not invariant between posttest and retrospective pretest data.

### *Concluding remarks*

The analyses of the present chapter led to the following results with regard to the research questions posed in Section 1.3:

- III. Can response shift be detected in actual pretest-posttest data when applying a model of measurement invariance?
- The 4-step procedure detected a relatively small number of questionnaire items that were not invariant in the datasets of actual pre- and posttests. Generally less than 10% of the heiQ items had to be excluded to ensure an unbiased comparison of the factor means of the actual pretests and posttests. As a consequence, the significantly higher posttest levels of heiQ-PPR could not be explained by response shift bias.
- IV. Are the model parameters invariant across retrospective pretests and actual posttests?
- In contrast to the previous analyses, more items were found to be non-invariant in the sample of retrospective pretest-posttest data. While it is possible that response shift affected the two scores in different ways (Oort *et al.*, 2003), it may be more likely that the specific design of heiQ-PPR influenced answers to these items in a way that it caused non-invariance of some items. For example, possible linear dependencies in the dataset may have led to instability in the respective estimates that in turn caused

the apparent non-invariance between the retrospective pretest data and the actual posttest data. Moreover, given that posttest ratings were provided first, this may have restricted answers to the retrospective pretests which again may be a reason for the observed non-invariance of some of the heiQ items.

The analyses of this chapter provided evidence that suggests answers to the above research questions. Moreover, the analyses of the datasets consisting of actual pretests and posttests confirmed the quality of the data obtained via the heiQ. That is, regardless of the design of the posttest heiQ, all samples showed robust psychometric properties of this questionnaire. In contrast, in the dataset consisting of retrospective pretests and actual posttests more heiQ items were found to be non-invariant, suggesting that change scores derived from these data may be less robust.

Despite these findings, the analyses of the present chapter did not provide an explanation for the group differences in the four samples observed in both Chapters 3 and 5. Regardless of the method used to derive change scores, i.e. arithmetic means (Chapter 3) or factor means (Chapter 5), some large group differences were observed. Hence, different tasks at posttest did not alter the psychometric performance of the heiQ but they changed the magnitude of actual posttest scores. Consequently, the possible explanations for these group differences remain similar to the ones proposed in Chapter 3. Observed differences could be due to:

- *Satisficing*;
- *Social desirability*;
- *Effort justification bias*;
- *Implicit theory of stability or change*;
- *Response shift* (subject-level rather than group-level analyses need to be applied).

Moreover, the following issues relating to heiQ-PPR Retro need further investigation:

- *Linear dependency between retrospective pretests and actual posttests*;
- *The design of heiQ-PPR restricted the response options for retrospective pretests*;

Further investigation of most of these explanations is beyond the scope of this thesis and has to be left for future research. However, considering that it had been identified that socially desirable responding is another potentially important confounder of scores in the present research (see Section 1.2.4), the influence of this bias on the change scores of the four heiQ datasets is investigated in the following Chapter 6.

# Chapter 6

Change scores  
mediated by social  
desirability

## **6 Change scores mediated by social desirability**

### **6.1 Introduction**

Based on the conclusions of Chapters 3 and 5, this final data analysis chapter of the thesis was aimed at investigating the potential influence of social desirability bias on scores of the four datasets heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro. It was explored whether observed differences between the three samples of actual pretest-posttest data could be explained by a mediating effect of social desirability and whether this bias had influenced change scores derived from retrospective pretest-posttest data (heiQ-PPR Retro). To assess the possibility of social desirability bias, the MC-C scale (Reynolds, 1982) was added to previous one-factor models. Although it was assumed that a similar proportion of subjects across the randomised groups had a propensity to provide socially desirable answers, it was expected that the effect of social desirability on the scores would differ across samples. That is, it was expected that the influence of social desirability bias was strongest in heiQ-PPR and heiQ-PPR Retro because of the simultaneous provision of retrospective pretest and actual posttest data at posttest.

### **6.2 Aims**

The aims of the chapter were:

- 6.a To validate Reynolds' short form MC-C (Reynolds, 1982);
- 6.b To investigate whether social desirability substantially contributed to variation in change scores derived from actual pretest-posttest data (heiQ-PP, heiQ-PPT, heiQ-PPR);
- 6.c To investigate whether social desirability substantially contributed to variation in change scores derived from retrospective pretest-posttest data (heiQ-PPR Retro).

### **6.3 The factor structure of the short form MC-C of the Marlowe-Crowne scale**

#### **6.3.1 Specific methods**

As described in Section 1.2.4.5, research has suggested that the original MC scale (Crowne & Marlowe, 1960) is multi-dimensional. While solutions range from two factors to multiple factors (Ballard, 1992; Barger, 2002; Crino *et al.*, 1983), a two-factor solution is a common representation of this scale (Millham, 1974; Paulhus, 1984; Ramanaiah *et al.*, 1977). The

extent to which these results are applicable to the present sample, however, is questionable. Firstly, previous research was frequently based on data derived from students (Ballard, 1992; Barger, 2002; Millham, 1974; Paulhus, 1984; Ramanaiah *et al.*, 1977). Given that social desirability bias has been found to be associated with age and gender – with older women being more prone to socially desirable responses (Ray, 1988; Visser *et al.*, 1989) – previous findings may not hold for the present sample that predominantly consisted of elderly women (see Section 3.3). Secondly, factor analyses on the MC short forms were generally aimed at confirming/rejecting a one-factor solution rather than exploring potential multi-factor solutions of these social desirability measures (see Section 1.2.4.5). For these reasons, the MC-C was factor-analysed before embarking upon the analyses.

The present section therefore starts with a description of the preparation of the MC-C data for later modelling. Further details on the MC-C scale were provided earlier in this thesis (see Section 2.3.2).

Before analysing the MC-C data, some preparatory steps were undertaken. Firstly, in line with the preparation of the heiQ data, each case with more than 50% missing items (>6 items missing) was deleted. Secondly, due to the alternate keying of the MC items, it could easily be detected if participants exhibited an acquiescent response style. That is, respondents who had provided either only 'true' or only 'false' answers were discarded as it could be assumed that they had filled out the MC-C scale regardless of the content of individual items. Once this preparation was finalised, all remaining missing values were replaced using the EM Algorithm (see Section 2.3 for a review of the heiQ data preparation).

Given that no previous study was found that utilised MC-C data from a sample with similar characteristics to the present sample, the analyses were approached in an exploratory way. The data were first analysed in CEFA (Browne *et al.*, 2004) which is the recommended program for unrestricted factor analyses (McDonald, 2005). As the MC-C was assumed to measure one underlying construct, i.e. social desirability, all multi-factor structures were analysed with oblique rotation to allow for correlations between the factors. For this GEOMIN was used (Browne, 2001; McDonald, 2005).<sup>19</sup> Due to the scaling of the MC-C, the input matrix was based on polychoric correlations. With these correlations, CEFA defaulted to the ordinary least squares method for the estimation of the parameters (Browne *et al.*, 2004). Eigenvalues and a scree plot served as a guideline to establish the number of underlying factors. Once the preliminary factor structure was determined, it was again tested in LISREL (Jöreskog & Sörbom, 1996-2001), using RML to estimate the model parameters (Jöreskog, 2002-2005).

---

<sup>19</sup> As recommended by Browne (2001), several rotation methods were employed to test whether the same factor solution evolved across different methods. In the present case, oblique CF-Varimax and Infomax were applied with both resulting in the same factor structure.



Given that subsequent analyses were performed on the randomised groups, the present analyses were carried out on the full dataset as well as on the three samples separately. This was necessary to ensure that no systematic differences existed between the groups. Given that the datasets heiQ-PPR and heiQ-PPR Retro were obtained from the same participants, only one factor analysis needed to be carried out.

### 6.3.2 Results

After discarding those cases with excess missing items and those who had suggested an obvious acquiescent response style, the final dataset consisted of n=908 participants. Hence, about 5% of cases were lost when adding the MC-C scale to the analyses. The final sample sizes of the randomised groups were n=318 for heiQ-PP, n=291 for heiQ-PPT, and n=299 for heiQ-PPR/heiQ-PPR Retro. Given that sample sizes of n>100 subjects (in the case of low reliability of the mediating variable n>200) have been regarded as sufficient to perform SEM analyses with mediating variables (Frazier *et al.*, 2004; Hoyle & Kenny, 1999), the size of these final samples were deemed appropriate for the present analyses.

The CEFA analyses suggested that a one-factor solution did not fit the data well. With two eigenvalues clearly above one (3.4 and 1.9, respectively) and two further eigenvalues at 1.1, factor solutions ranging from two to four factors were further explored. While the fit statistics improved substantially in all multi-factor solutions, the models beyond two factors were not superior to the two-factor solution and each suggested that one factor consisted of one item only. Therefore, the following two-factor solution appeared most suitable for the MC-C: items 1,2,3,4,6,8,11,12 loaded on factor 1 and items 5,6,7,9,10,13 loaded on factor 2. That is, with the exception of item 6 that indicated small loadings on both factors, the remaining 12 items indicated unambiguous factor loadings on either factor 1 or factor 2, respectively. Given that subsequent confirmatory analyses in LISREL supported a factor solution with item 6 loading on factor 1, it was decided to allocate this item to the first factor.

The final factor solution reflected the scaling direction of the MC-C, with factor 1 consisting of those items that were keyed 'false' such as "I sometimes feel resentful when I don't get my way". Following the design and content of these items, this factor may describe some form of 'defensiveness' or 'denial' mechanism. In contrast, factor 2 consists of items that are keyed 'true' such as "No matter who I am talking to, I'm always a good listener". This factor may reflect 'self-presentation' or 'impression management' as suggested in previous studies (see Section 1.2.4.5). To better distinguish between the two social desirability factors, they will be referred to as 'defensiveness' and 'self-presentation' in the present chapter.

Given that CEFA only provides a limited range of fit statistics, these analyses were mainly used for establishing the factor structure of the MC-C. For the evaluation of model fit as well as the presentation of the results, the LISREL output was used which is presented hereafter. The CEFA results are provided in Appendix 17.

After establishing which items loaded on which factor, all analyses were again run in LISREL. These analyses confirmed that the two-factor solution was superior to any other combination. Given that model fit could again be improved significantly when allowing for a correlated error between items 1 “It is sometimes hard for me to go on with my work if I am not encouraged” and 3 “On a few occasions, I have given up doing something because I thought too little of my ability”, this correlated error was included in the final model. Table 29 shows the results based on the full sample (n=908). Apart from a significant  $\chi^2_{SB}$ , all remaining fit statistics indicated a satisfactory fit of the model:  $\chi^2_{SB}(63)=118.9$  ( $p<0.001$ ), RMSEA=0.031 (90% CI, 0.023;0.040), CFI=0.99, and SRMR=0.066. In contrast to the fit statistics, some of the factor loadings and both coefficient alphas, however, were below the recommended cut-off values of 0.50 for the factor loadings and 0.70 for coefficient alpha, respectively (Hair *et al.*, 2006). It was further found that ‘self-presentation’ and ‘defensiveness’ correlated moderately (0.44).

**Table 29** Confirmatory Factor Analysis of the MC-C, full sample (n=908)

		Standardised factor loading	Error variance
<u>Self-presentation</u>			
1	It is sometimes hard for me to go on with my work if I am not encouraged.	0.47	0.78
2	I sometimes feel resentful when I don't get my way.	0.75	0.44
3	On a few occasions, I have given up doing something because I thought too little of my ability.	0.50	0.75
4	There have been times when I felt like rebelling against people in authority even though I knew they were right.	0.60	0.64
6	There have been occasions when I took advantage of someone.	0.42	0.82
8	I sometimes try to get even rather than forgive and forget.	0.53	0.72
11	There have been times when I was quite jealous of the good fortune of others.	0.56	0.68
12	I am sometimes irritated by people who ask favours of me.	0.45	0.80
<u>Defensiveness</u>			
5	No matter who I'm talking to, I'm always a good listener.	0.56	0.69
7	I'm always willing to admit it when I make a mistake.	0.62	0.62
9	I am always courteous, even to people who are disagreeable.	0.71	0.50
10	I have never been irked when people expressed ideas very different from my own.	0.45	0.80
13	I have never deliberately said something that hurt someone's feelings.	0.38	0.86

Fit statistics:  $\chi^2_{SB}(63)=118.9$ ,  $p<0.001$ ; RMSEA=0.031 (90% CI, 0.023;0.040); CFI=0.99; SRMR=0.066. Coefficient alpha: factor 1=0.65; factor 2=0.54; Phi=0.44 (standardised correlation between factor 1 and 2)

Given that the two-factor solution yielded the best possible model fit for the MC-C, the model was then tested on each randomised group separately. As shown in Table 30, Table 31, and Table 32, the two-factor solution again indicated good fit. While the fit statistics were similar across groups, the MC-C performed best in heiQ-PP. Not only the  $\chi^2_{SB}$  was non-significant, but all remaining fit indices indicated good fit: RMSEA=0.023 (90% CI, 0.0;0.043), CFI=0.99, and SRMR=0.079. In contrast, the fit statistics of the remaining two samples indicated slightly worse model fit with  $\chi^2_{SB}(63)=90.0$  ( $p<0.015$ ), RMSEA=0.038 (90% CI, 0.018;0.056), CFI=0.98, SRMR=0.098 for heiQ-PPT and  $\chi^2_{SB}(63)=90.0$  ( $p<0.015$ ), RMSEA=0.038 (90% CI, 0.018;0.056), CFI=0.98, SRMR=0.098 for heiQ-PPR/heiQ-PPR Retro.

Similar to the analyses on the full sample, some small factor loadings and coefficient alphas  $<0.70$  were observed in all datasets. Finally, the magnitude of the factor correlations (Phi,  $\phi$ ) differed slightly across datasets. While 'self-presentation' and 'defensiveness' showed similar correlations in heiQ-PP and heiQ-PPR/heiQ-PPR Retro ( $\phi=0.48$  and  $\phi=0.52$ , respectively), this correlation was smaller in heiQ-PPT ( $\phi=0.29$ ). However, all correlations were significant at the  $p=0.05$  level.

**Table 30** Confirmatory Factor Analysis of the MC-C, heiQ-PP (n=318)

		Standardised factor loading	Error variance
<u>Self-presentation</u>			
1	It is sometimes hard for me to go on with my work if I am not encouraged.	0.44	0.81
2	I sometimes feel resentful when I don't get my way.	0.76	0.43
3	On a few occasions, I have given up doing something because I thought too little of my ability.	0.51	0.74
4	There have been times when I felt like rebelling against people in authority even though I knew they were right.	0.54	0.71
6	There have been occasions when I took advantage of someone.	0.33	0.89
8	I sometimes try to get even rather than forgive and forget.	0.48	0.77
11	There have been times when I was quite jealous of the good fortune of others.	0.51	0.74
12	I am sometimes irritated by people who ask favours of me.	0.37	0.87
<u>Defensiveness</u>			
5	No matter who I'm talking to, I'm always a good listener.	0.54	0.71
7	I'm always willing to admit it when I make a mistake.	0.64	0.59
9	I am always courteous, even to people who are disagreeable.	0.70	0.51
10	I have never been irked when people expressed ideas very different from my own.	0.48	0.77
13	I have never deliberately said something that hurt someone's feelings.	0.42	0.83

Fit statistics:  $\chi^2_{SB}(63)=73.7$ ,  $p=NS$ ; RMSEA=0.023 (90% CI, 0.0;0.043); CFI=0.99; SRMR=0.079. Coefficient alpha: factor 1=0.59; factor 2=0.56; Phi=0.48

**Table 31** Confirmatory Factor Analysis of the MC-C, heiQ-PPT (n=291)

		Standardised factor loading	Error variance
<u>Self-presentation</u>			
1	It is sometimes hard for me to go on with my work if I am not encouraged.	0.51	0.74
2	I sometimes feel resentful when I don't get my way.	0.70	0.52
3	On a few occasions, I have given up doing something because I thought too little of my ability.	0.47	0.78
4	There have been times when I felt like rebelling against people in authority even though I knew they were right.	0.68	0.54
6	There have been occasions when I took advantage of someone.	0.35	0.88
8	I sometimes try to get even rather than forgive and forget.	0.53	0.72
11	There have been times when I was quite jealous of the good fortune of others.	0.57	0.68
12	I am sometimes irritated by people who ask favours of me.	0.52	0.74
<u>Defensiveness</u>			
5	No matter who I'm talking to, I'm always a good listener.	0.51	0.74
7	I'm always willing to admit it when I make a mistake.	0.68	0.54
9	I am always courteous, even to people who are disagreeable.	0.72	0.48
10	I have never been irked when people expressed ideas very different from my own.	0.40	0.84
13	I have never deliberately said something that hurt someone's feelings.	0.38	0.86

Fit statistics:  $\chi^2_{SB}(63)=90.0$ ,  $p=0.015$ ; RMSEA=0.038 (90% CI, 0.018;0.056); CFI=0.98; SRMR=0.098. Coefficient alpha: factor 1=0.65; factor 2=0.52; Phi=0.29

**Table 32** Confirmatory Factor Analysis of the MC-C, heiQ-PPR / heiQ-PPR Retro (n=299)

		Standardised factor loading	Error variance
<u>Self-presentation</u>			
1	It is sometimes hard for me to go on with my work if I am not encouraged.	0.49	0.76
2	I sometimes feel resentful when I don't get my way.	0.79	0.38
3	On a few occasions, I have given up doing something because I thought too little of my ability.	0.54	0.71
4	There have been times when I felt like rebelling against people in authority even though I knew they were right.	0.58	0.67
6	There have been occasions when I took advantage of someone.	0.53	0.72
8	I sometimes try to get even rather than forgive and forget.	0.55	0.69
11	There have been times when I was quite jealous of the good fortune of others.	0.60	0.64
12	I am sometimes irritated by people who ask favours of me.	0.46	0.79
<u>Defensiveness</u>			
5	No matter who I'm talking to, I'm always a good listener.	0.67	0.55
7	I'm always willing to admit it when I make a mistake.	0.54	0.71
9	I am always courteous, even to people who are disagreeable.	0.70	0.52
10	I have never been irked when people expressed ideas very different from my own.	0.47	0.78
13	I have never deliberately said something that hurt someone's feelings.	0.33	0.89

Fit statistics:  $\chi^2_{SB}(63)=96.6$ ,  $p=0.004$ ; RMSEA=0.042 (90% CI, 0.024;0.059); CFI=0.98; SRMR=0.096. Coefficient alpha: factor 1=0.68; factor 2=0.53; Phi=0.52

### 6.3.3 Summary

The factor analysis of the MC-C suggested that this scale is a two-factor measure of social desirability. While the fit statistics indicated best model fit of this solution, the specification of two factors was further supported by the relatively small correlation of the two factors. Hence, the MC-C scale appears to measure two unique aspects of social desirability bias – ‘self-presentation’ and ‘defensiveness’ – which is largely consistent with previous research in this area (Millham, 1974; Paulhus, 1984; Ramanaiah *et al.*, 1977). Considering that a short form of the MC scale not only showed acceptable fit but also indicated that a two-factor solution is a good representation of the measured construct, confirms that the MC-C scale (Reynolds, 1982) is a valid alternative to the original MC scale (Crowne & Marlowe, 1960).

## 6.4 Social desirability – heiQ-PP

### 6.4.1 Specific methods

Based on the above results, the following analyses included the MC-C scale as a two-factor measure of social desirability. The current section provides details on the specification of the models that were used in the analyses of the present chapter.

The following models included social desirability as a mediating variable between predictor (=actual/retrospective pretest) and outcome (=posttest). While alternative conceptualisations of ‘mediation’ exist (Collins *et al.*, 1998; MacKinnon *et al.*, 2000), the analyses were based on a framework proposed by Kenny and colleagues (Baron & Kenny, 1986; Judd & Kenny, 1981; Kenny *et al.*, 1998). That is, a *mediator* is defined as a variable that influences the mechanism in which effects occur. Given this direct influence on the mechanism, a mediator is part of the causal chain of a model, i.e. it follows the predictor but precedes the outcome. In the case of *perfect mediation*, a mediator is the actual mechanism through which an effect occurs, i.e. the original predictor exerts no effect on the outcome anymore once the mediator is included in the model (Baron & Kenny, 1986; Judd & Kenny, 1981). In contrast, when the path between predictor and outcome variable decreases but remains significant, the effect is referred to as *partial mediation* (Judd & Kenny, 1981; Lehmann *et al.*, 2001).

Derived from the above theory, social desirability was specified as a *partial mediator* in the present models. While it was expected that the inclusion of this variable would improve the prediction of the posttest scores, it was assumed that the pretest would remain a significant predictor of the posttest scores, rendering the specification of social desirability as a perfect mediator inappropriate.

To establish whether a variable is a potential mediator between a predictor and an outcome variable, the following conditions need to be established (Baron & Kenny, 1986; Judd & Kenny, 1981):

- (1) The mediator and the predictor must correlate, i.e. the predictor must affect the mediating variable for the latter to be a mediator between predictor and outcome. This can be tested by regressing the mediator on the predictor variable.
- (2) The predictor must also affect the outcome, i.e. it has to be established that a relationship between the two variables exists before testing for potential mediating effects of a third variable. This can be achieved by regressing the outcome on the predictor variable.<sup>20</sup>
- (3) The mediator must affect the outcome, i.e. it needs to be established that the regression of outcome on mediator is significant. In this model it is also tested whether the original predictor still affects the outcome. That is, the path from predictor to outcome should be larger in Step 2 than in Step 3, i.e. when the mediator is included in the model. In the case of perfect mediation, this path should be non-significant in Step 3.

While the above steps are necessary conditions to establish mediation, they are not sufficient conditions (Little *et al.*, 2007). It is further necessary to ensure that the mediational effect is significant, i.e. the statistical significance of the product of the paths from predictor to mediator, and from mediator to outcome needs to be established (Shrout & Bolger, 2002). To determine its significance, the following formula was applied to calculate the standard error  $SE_M$  of the mediation (Little *et al.*, 2007; Sobel, 1982):<sup>21</sup>

$$SE_M = \sqrt{(\gamma_{11}^2 SE_{\beta_{21}}^2 + \beta_{21}^2 SE_{\gamma_{11}}^2)},$$

where  $\gamma_{11}$  is the effect of the predictor on the mediator,  $\beta_{21}$  is the effect of the mediator on the outcome, and  $SE_{\beta_{21}}$  and  $SE_{\gamma_{11}}$  respectively, are their standard errors.

Once conditions 1) to 3) as well as statistical significance are established, it can be inferred that the hypothesised mediator has a true mediating effect in the model. In terms of practical implications of the mediating effect, however, it is useful to interpret the mediational effect in the overall context of the model (Little *et al.*, 2007). For this, the ratio  $P_M$  of mediation to total effect can be calculated, i.e. it is assessed what proportion of the total effect is actually being mediated. In the present study  $P_M$  was not only useful for judging the practical significance of

---

<sup>20</sup> While the need for the second condition has been questioned by several authors (Collins *et al.*, 1998; MacKinnon *et al.*, 2000), others recommended differentiating between distal and proximal effects. That is, for proximal effects, as was the case in the present situation, condition 2) is regarded as conceptually useful (Shrout & Bolger, 2002).

<sup>21</sup> The correct formula includes the sum of the squared errors of the two paths (Baron & Kenny, 1986). However, given that this term is very small, it is frequently omitted (Little *et al.*, 2007; Sobel, 1982).

the mediator in the model, but also to compare the influence of social desirability across the four datasets.  $P_M$  was calculated as follows (MacKinnon *et al.*, 1995):

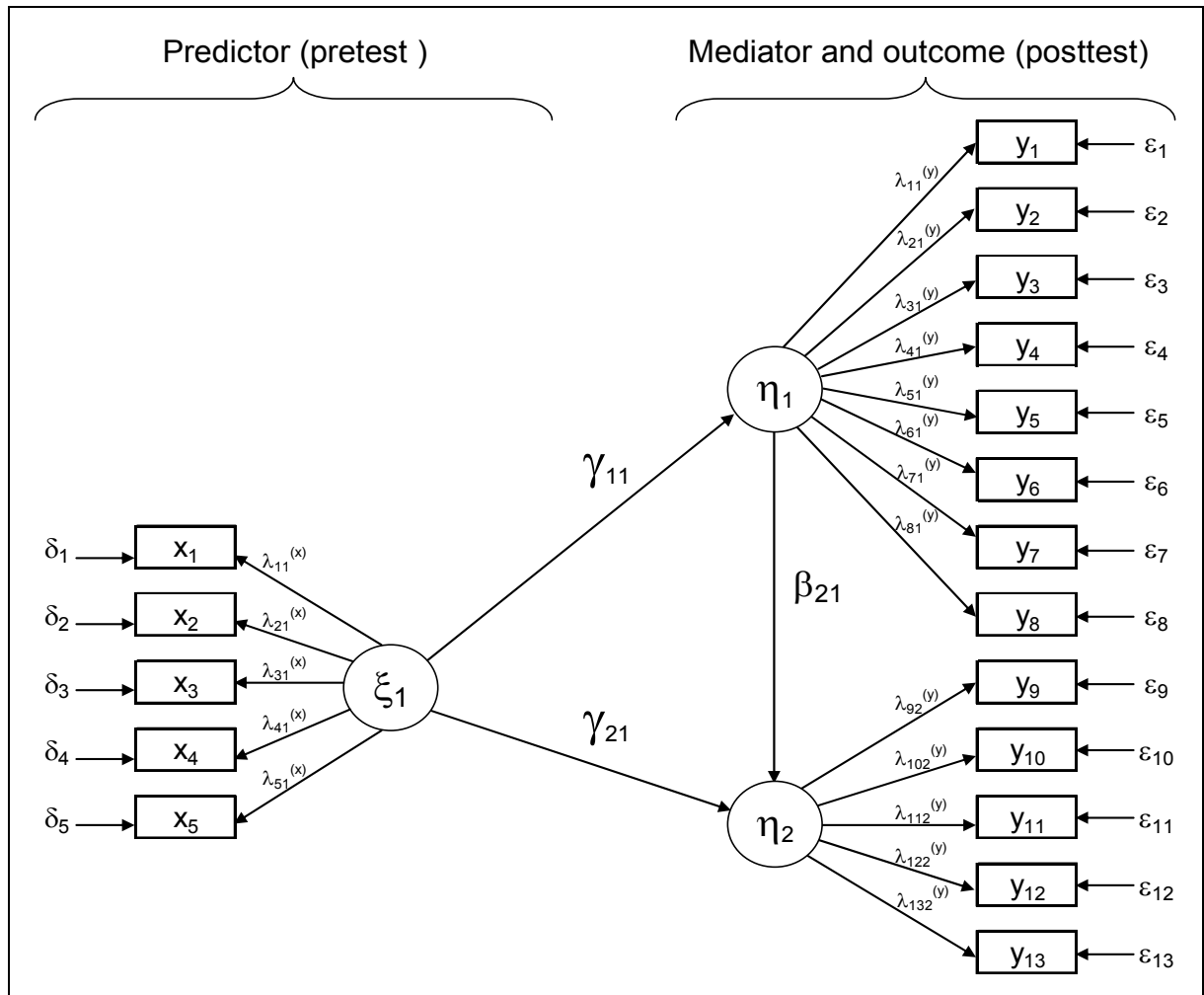
$$P_M = \gamma_{11}\beta_{21} / (\gamma_{11}\beta_{21} + \gamma_{21}),$$

where the mediational effect, i.e. the product of  $\gamma_{11}$  and  $\beta_{21}$ , was divided by the total effect. The total effect can be defined as a) the original effect that the predictor had on the outcome before the mediator was included in the model (Shrout & Bolger, 2002) or b) as the indirect effect plus the direct effect  $\gamma_{21}$  of predictor on the outcome in the mediational model (Bollen, 1987; MacKinnon *et al.*, 1995).

Based on the above elaborations, five steps were undertaken to test whether ‘defensiveness’ and/or ‘self-presentation’ partially mediated the relationship between actual/retrospective pretest and posttest data in the present models:

- (1) The correlation between the respective pretest data and ‘social desirability’ was tested by defining all variables as exogenous variables. That is, the heiQ pretests, and the factors ‘defensiveness’ and ‘self-presentation’ were specified in the X-measurement model. The resulting *Phi matrix* was used to interpret the significance of the correlations.
- (2) The path between respective pretests and posttests was investigated by estimating the regression coefficient from the X-measurement model to the Y-measurement model. The significance of this path was then assessed through the *Gamma matrix*.
- (3) In addition to the previous step, Step 3 included the mediator as a second endogenous variable, i.e. both posttest and mediator were regressed on the respective pretest. The model further included a path from the mediator to the posttest. Apart from judging the statistical significance of the above paths it was further ensured that the path from pretest to posttest was smaller in Step 3 than in Step 2. In view of the underlying assumptions of the models it was expected that this path from pretest to posttest would be reduced, once social desirability was included, but that it would still remain significant.
- (4) Once conditions 1) to 3) were met, the statistical significance of the mediational effect was tested by applying the formula described above.
- (5) While the previous four steps are necessary as well as sufficient conditions to establish mediation, it was further useful to estimate the proportion mediated in each respective model. That is, the calculation of  $P_M$  facilitated the comparison of all potential mediating effects across the four heiQ datasets.

The mediational model applied in this chapter is visualised in Figure 35. In this example, pretest and posttest data are presented as the latent variables  $\xi_1$  and  $\eta_2$ , with each being determined by five observed variables. Social desirability is shown as the latent variable  $\eta_1$ . Social desirability and the posttest are regressed on the pretest ( $\gamma_{11}$  and  $\gamma_{21}$ ), and the posttest is regressed on social desirability ( $\beta_{21}$ ).



**Figure 35** Structural equation model including 'defensiveness' as the mediating variable – illustrating Positive and Active Engagement in Life (see Table 14 for a legend of the LISREL notation)

To prepare the data for the analyses, all variables had to be converted into matrix-form. Similar to the analyses in Chapter 5, it was again essential to assign equal thresholds to each item across the respective pretest-posttest dataset to ensure that the variables were on the same scale (Jöreskog, 2002-2005). The application of RML for the parameter estimation further required the computation of the asymptotic covariance matrix (see Section 4.3.3).



## 6.4.2 Results

After carrying out Step 1, results suggested that only the first factor of social desirability was related to the pretests. That is, 'defensiveness' (SD1) correlated significantly with the pretest data across all heiQ subscales. These effects ranged from 0.24 to 0.39, i.e. they were small to medium effects (Cohen, 1988; Shrout & Bolger, 2002). In contrast, none of the subscales indicated an association between 'self-presentation' (SD2) and the pretests in heiQ-PP (see Table 33). As a consequence of the analyses of Step 1, only the 'defensiveness' factor of social desirability was explored as a potential partial mediator in heiQ-PP, while the 'self-presentation' factor of this bias could be ruled out as a partial mediator in this dataset.

**Table 33** Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PP

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
SD1-SD2	Cov	0.247*	0.254*	0.246*	0.248*	0.249*	0.240*	0.250*	0.251*
	(SE)	(0.058)	(0.060)	(0.058)	(0.058)	(0.058)	(0.058)	(0.059)	(0.059)
	Corr	0.480	0.482	0.477	0.475	0.478	0.477	0.480	0.473
SD1-pre	Cov	0.192*	0.208*	0.255*	0.360*	0.312*	0.199*	0.264*	0.233*
	(SE)	(0.071)	(0.071)	(0.057)	(0.084)	(0.092)	(0.057)	(0.079)	(0.069)
	Corr	0.242	0.255	0.375	0.388	0.310	0.271	0.280	0.280
SD2-pre	Cov	0.015	0.089	0.098	0.032	0.155	0.019	0.009	-0.032
	(SE)	(0.068)	(0.068)	(0.061)	(0.076)	(0.098)	(0.066)	(0.070)	(0.065)
	Corr	0.020	0.116	0.153	0.037	0.156	0.026	0.010	-0.041

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

### Legend

PAE:	Positive and Active Engagement in Life
HDB:	Health-Directed Behaviour
STA:	Skill and Technique Acquisition
CAA:	Constructive Attitudes and Approaches
SMI:	Self-Monitoring and Insight
HSN:	Health Service Navigation
SIS:	Social Integration and Support
EWB:	Emotional Well-Being
SD 1:	Factor 1 'defensiveness'
SD 2:	Factor 2 'self-presentation'
Cov:	Covariance
SE:	Standard error
Corr:	Correlation

As expected, in Step 2 it was found that all direct paths from pre- to posttest were significant. While subscale Social Integration and Support showed the strongest association between the two scores, all heiQ subscales showed substantial paths from predictor to outcome (see Table 34).

**Table 34** Regression of the posttests on the pretests (Gamma matrix), heiQ-PP

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.829*	0.648*	0.610*	0.672*	0.593*	0.743*	0.774*	0.758*
	(SE)	(0.069)	(0.044)	(0.065)	(0.066)	(0.071)	(0.053)	(0.044)	(0.049)
	stand.	0.776	0.744	0.626	0.718	0.665	0.777	0.808	0.761

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

For an extensive legend refer to Table 33

Finally, Table 35 presents the associations between all variables once 'defensiveness' (SD1) was included in the models. As observed previously in Step 1, all paths between pretest and 'defensiveness' were significant. Once the pretest data were controlled for 'defensiveness', Emotional Well-Being was the only subscale that showed a significant association between 'defensiveness' and the posttest scores. Hence, the relationship between pre- and posttests was potentially mediated in only one of the eight heiQ subscales.

**Table 35** Regression of 'defensiveness' (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PP

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.826*	0.654*	0.606*	0.698*	0.575*	0.757*	0.765*	0.696*
	(SE)	(0.072)	(0.049)	(0.075)	(0.076)	(0.077)	(0.058)	(0.046)	(0.052)
	stand.	0.772	0.753	0.618	0.752	0.642	0.788	0.796	0.700
pre-SD1	Path	0.169*	0.172*	0.301*	0.229*	0.166*	0.192*	0.160*	0.191*
	(SE)	(0.059)	(0.057)	(0.067)	(0.052)	(0.049)	(0.053)	(0.047)	(0.056)
	stand.	0.252	0.259	0.388	0.399	0.326	0.283	0.284	0.290
SD1-post	Path	0.026	-0.038	0.024	-0.141	0.128	-0.060	0.070	0.322*
	(SE)	(0.107)	(0.088)	(0.101)	(0.123)	(0.132)	(0.102)	(0.092)	(0.099)
	stand.	0.016	-0.029	0.019	-0.087	0.073	-0.043	0.041	0.213

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

For an extensive legend refer to Table 33

Given that 'defensiveness' was found to be associated with the pretest-posttest data of the Emotional Well-Being subscale, Steps 4 and 5 were performed on this subscale only. Hence, the potential reduction in the path between pretest and posttest, the size of the mediated effect, the statistical significance of this effect, and the proportion mediated were calculated for Emotional Well-Being. The following results were obtained:

### Emotional Well-Being

- Reduction in the path pretest to posttest (the model without the mediator compared with the model including the mediator):  $0.758 - 0.696 = 0.062$ ;
- Size of the mediated effect:  $0.191 * 0.322 = 0.062$ ;
- $SE_M = \sqrt{(0.191^2 0.099^2 + 0.322^2 0.056^2)} = 0.026$ ;
- $P_M = 0.062 / (0.696 + 0.062) = 0.082$ .

It was observed that the path between pretests and posttests decreased in the expected direction once 'defensiveness' was included in the model. Given that this mediational effect was significant, i.e. the effect was more than twice its standard error (Bollen, 1989), there was sufficient evidence that 'defensiveness' operated as a partial mediator between pretest and posttest in Emotional Well-Being. The practical significance of the effect, however, was small. The results indicated that 'defensiveness' contributed only 8.2% of the total variation in change scores in heiQ-PP ( $P_M=0.082$ ).

### **6.4.3 Summary**

The exploration of social desirability as a potential mediator of the effect of pretest to posttest in heiQ-PP provided little evidence that social desirability had any impact on the scores. Firstly, the 'self-presentation' factor of social desirability could be ruled out as a potential mediator which is largely consistent with the literature. While the notion of 'defence' and 'self-protection' was introduced as one critical aspect of the approval motive (Crowne & Marlowe, 1964), later research suggested that subjects' motivations to present themselves in a socially desirable fashion was linked more strongly to defensiveness rather than self-presentation (Millham, 1974; Millham & Kellogg, 1980). Secondly, despite the significant association of 'defensiveness' with all pretests, this factor of social desirability exerted little influence on the posttests once the pretests were controlled for this variable. Only subscale Emotional Well-Being suggested that 'defensiveness' operated as a true, albeit minor mediator.

In sum, given that only little influence of social desirability had been expected in heiQ-PP, these findings confirmed that social desirability bias as measured by the MC-C was unlikely to threaten the validity of change scores derived from this research design.

## 6.5 Social desirability – heiQ-PPT

### 6.5.1 Specific methods

The procedure of the analyses of this section was identical to the one of the previous Section 6.4. Hence, no further explanations are presented and the results of the one-factor models of heiQ-PPT are presented hereafter.

### 6.5.2 Results

The association between the pretests and the two factors of social desirability is presented in Table 36. Again, the heiQ pretests correlated more often with ‘defensiveness’ (SD1) than with ‘self-presentation’ (SD2). Four of the eight heiQ subscales showed weak and significant correlations between pretests and ‘defensiveness’, and two further subscales also had weak non-significant correlations (Self-Monitoring and Insight; Social Integration and Support). In contrast, only two heiQ subscales showed a significant, albeit negative correlation between the pretests and ‘self-presentation’. In view of these overall small associations between heiQ pretests and ‘self-presentation’, this component of social desirability was again excluded from further analyses. Hence, only the mediational effect of ‘defensiveness’ was tested in the models of heiQ-PPT.

**Table 36** Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PPT

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
SD1-SD2	Cov	0.141*	0.142*	0.141*	0.137*	0.148*	0.141*	0.144*	0.134*
	(SE)	(0.058)	(0.058)	(0.058)	(0.056)	(0.061)	(0.058)	(0.059)	(0.057)
	Corr	0.284	0.288	0.286	0.288	0.281	0.289	0.288	0.270
SD1-pre	Cov	0.189*	0.073	0.197*	0.147*	0.166	0.017	0.140	0.394*
	(SE)	(0.079)	(0.066)	(0.077)	(0.049)	(0.088)	(0.070)	(0.074)	(0.080)
	Corr	0.203	0.088	0.230	0.263	0.162	0.021	0.157	0.418
SD2-pre	Cov	-0.172*	-0.042	-0.092	-0.073	-0.138	0.049	-0.030	-0.189*
	(SE)	(0.081)	(0.072)	(0.076)	(0.045)	(0.106)	(0.077)	(0.075)	(0.086)
	Corr	-0.188	-0.051	-0.105	-0.134	-0.125	0.061	-0.034	-0.196

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

#### Legend

For an extensive legend refer to Table 33

The results of the regression of posttest on pretest data (Step 2) are presented in Table 37. While all paths were again significant, the path from pretest to posttest in Skill and Technique

Acquisition was substantially smaller when compared with the remaining subscales and also when compared with the results obtained for heiQ-PP (see Table 34).

**Table 37** Regression of the posttests on the pretests (Gamma matrix), heiQ-PPT

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.642*	0.535*	0.275*	0.761*	0.537*	0.791*	0.716*	0.652*
	(SE)	(0.057)	(0.066)	(0.068)	(0.056)	(0.081)	(0.057)	(0.046)	(0.043)
	stand.	0.691	0.615	0.306	0.723	0.544	0.754	0.788	0.773

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

For an extensive legend refer to Table 33

The results of the models with the ‘defensiveness’ (SD1) factor of social desirability bias as a partial mediator are shown in Table 38. In addition to the four subscales that had already shown a significant correlation between the pretests and ‘defensiveness’, Self-Monitoring and Insight showed a significant association in Step 3, i.e. one of the subscales that had indicated a weak non-significant correlation in Step 1. Of these five significant subscales, however, only two showed a significant path from ‘defensiveness’ to the posttests once the heiQ pretests were controlled for this factor of social desirability. That is, Positive and Active Engagement in Life, and Emotional Well-Being showed a significant, albeit small effect of ‘defensiveness’ on scores of heiQ-PPT.

**Table 38** Regression of ‘defensiveness’ (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPT

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.614*	0.523*	0.257*	0.742*	0.530*	0.787*	0.695*	0.597*
	(SE)	(0.060)	(0.067)	(0.072)	(0.062)	(0.082)	(0.058)	(0.047)	(0.049)
	stand.	0.662	0.602	0.286	0.705	0.537	0.751	0.767	0.710
pre-SD1	Path	0.112*	0.054	0.131*	0.237*	0.080*	0.014	0.089	0.218*
	(SE)	(0.045)	(0.047)	(0.049)	(0.076)	(0.040)	(0.053)	(0.047)	(0.042)
	stand.	0.207	0.091	0.232	0.270	0.166	0.023	0.159	0.427
SD1-post	Path	0.248*	0.190	0.135	0.080	0.100	0.239*	0.232*	0.246*
	(SE)	(0.123)	(0.105)	(0.141)	(0.088)	(0.172)	(0.107)	(0.095)	(0.110)
	stand.	0.144	0.131	0.084	0.067	0.049	0.139	0.144	0.150

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

For an extensive legend refer to Table 33

In addition to Positive and Active Engagement in Life, and Emotional Well-Being, two further subscales (Health Service Navigation; Social Integration and Support) showed a significant path from 'defensiveness' to posttest. That is, although these heiQ subscales had not shown a significant association between heiQ pretests and 'defensiveness', they had a significant association between 'defensiveness' and posttests once the pretests were controlled. Given that the path from the heiQ pretests to 'defensiveness' had approached significance in Social Integration and Support, it was explored whether further tests of statistical significance would support a partial mediational effect of 'defensiveness' in this subscale. In contrast, in Health Service Navigation the heiQ pretests and 'defensiveness' had clearly not been associated. Hence, a mediational effect of 'defensiveness' was ruled out for this subscale.

Following from the above results, further analyses were carried out for Positive and Active Engagement in Life, Social Integration and Support, and Emotional Well-Being. The effect on the pretest-posttest path, the size of the mediated effect, the statistical significance of the effect, and the proportion mediated in the three heiQ subscales were as follows:

#### Positive and Active Engagement in Life

- Reduction in the path pretest to posttest:  $0.642 - 0.614 = 0.028$ ;
- Size of the mediated effect:  $0.112 * 0.248 = 0.028$ ;
- $SE_M = \sqrt{(0.112^2 * 0.123^2 + 0.248^2 * 0.045^2)} = 0.018$ ;
- $P_M = 0.028 / (0.614 + 0.028) = 0.044$ .

#### Social Integration and Support

- Reduction in the path pretest to posttest:  $0.716 - 0.695 = 0.021$ ;
- Size of the mediated effect:  $0.089 * 0.232 = 0.021$ ;
- $SE_M = \sqrt{(0.089^2 * 0.095^2 + 0.232^2 * 0.047^2)} = 0.014$ ;
- $P_M = 0.021 / (0.695 + 0.021) = 0.029$ .

#### Emotional Well-Being

- Reduction in the path pretest to posttest:  $0.652 - 0.597 = 0.055$ ;
- Size of the mediated effect:  $0.218 * 0.246 = 0.054$ ;
- $SE_M = \sqrt{(0.218^2 * 0.110^2 + 0.246^2 * 0.042^2)} = 0.026$ ;
- $P_M = 0.054 / (0.597 + 0.054) = 0.083$ .

It was found that the paths between heiQ pretests and posttests decreased in the expected direction across all three heiQ subscales once 'defensiveness' was included in the models. In both subscales Positive and Active Engagement in Life, and Social Integration and Support the mediational effect was small and also non-significant, i.e. each respective effect was less than twice its standard error (Bollen, 1989). Consequently, only in subscale Emotional Well-Being the partial mediational effect of 'defensiveness' was found to be significant. However, similar to the findings for heiQ-PP, the proportion of this effect was small, with this factor of social desirability contributing only 8.3% of the total variation in change scores in heiQ-PPT ( $P_M=0.083$ ).

### **6.5.3 Summary**

Compared with the results of heiQ-PP, the analyses of the present section initially appeared to show a stronger association of the change scores of heiQ-PPT with the 'defensiveness' component of social desirability, with three heiQ subscales indicating a potential mediational effect. However, after controlling the pretests for 'defensiveness', only Emotional Well-Being remained in which 'defensiveness' operated as a partial mediator between heiQ pretests and posttests. In a similar manner to heiQ-PP, this effect was again small, explaining less than 10% of the total effect. In addition, it was found that 'defensiveness' was associated with the heiQ posttests in subscale Health Service Navigation. This association, however, was not of mediational nature given that this heiQ subscale had not shown a significant correlation of 'defensiveness' and the pretests.

In sum, apart from a minor mediational effect in Emotional Well-Being, the 'defensiveness' factor of social desirability did not explain the association between the pretests and posttests in heiQ-PPT. These findings are again in line with prior expectations that social desirability bias would not explain change scores in this dataset.

## **6.6 Social desirability – heiQ-PPR**

### **6.6.1 Specific methods**

The procedure of the analyses of this section was identical to the ones of the previous two sections (see Sections 6.4 and 6.5).

## 6.6.2 Results

The results of Step 1 were largely similar to those obtained for heiQ-PP. All heiQ pretests correlated significantly with the ‘defensiveness’ factor of social desirability (SD1), whereas none of the subscales showed an association between ‘self-presentation’ (SD2) and the heiQ pretests (see Table 39). Identical to the previous sections, the ‘self-presentation’ component of social desirability was therefore excluded from all subsequent analyses.

**Table 39** Covariance between SD1, SD2, and the pretests (Phi matrix), heiQ-PPR

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
SD1-SD2	Cov	0.272*	0.280*	0.280*	0.266*	0.284*	0.285*	0.280*	0.276*
	(SE)	(0.060)	(0.061)	(0.062)	(0.059)	(0.062)	(0.063)	(0.062)	(0.061)
	Corr	0.526	0.527	0.521	0.527	0.521	0.511	0.521	0.521
SD1-pre	Cov	0.352*	0.236*	0.207*	0.401*	0.340*	0.568*	0.387*	0.288*
	(SE)	(0.096)	(0.099)	(0.059)	(0.073)	(0.078)	(0.130)	(0.094)	(0.075)
	Corr	0.285	0.190	0.277	0.432	0.397	0.355	0.308	0.300
SD2-pre	Cov	-0.064	-0.003	0.041	0.025	0.068	0.148	0.088	0.040
	(SE)	(0.096)	(0.100)	(0.059)	(0.075)	(0.077)	(0.140)	(0.099)	(0.084)
	Corr	-0.061	-0.002	0.061	0.032	0.088	0.107	0.078	0.047

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

### Legend

For an extensive legend refer to Table 33

Table 40 shows the association of heiQ pretests and posttests. As expected, all paths were significant. Similar to the findings of heiQ-PPT, this association was substantially smaller in Skill and Technique Acquisition, while it was also relatively small in subscale Self-Monitoring and Insight.

**Table 40** Regression of the posttests on the pretests (Gamma matrix), heiQ-PPR

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.701*	0.587*	0.325*	0.709*	0.432*	0.620*	0.793*	0.640*
	(SE)	(0.071)	(0.063)	(0.074)	(0.083)	(0.085)	(0.065)	(0.052)	(0.048)
	stand.	0.683	0.645	0.352	0.650	0.454	0.672	0.789	0.778

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

### Legend

For an extensive legend refer to Table 33



Finally, Table 41 shows the results of the models including ‘defensiveness’ (SD1). It was found that none of the subscales showed a significant path from the mediating variable to the posttests. Hence, once the pretests were controlled for ‘defensiveness’, this factor exerted no influence on the posttests, rendering further investigations of this dataset unnecessary.

**Table 41** Regression of ‘defensiveness’ (SD1) and the posttests on the pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPR

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
pre-post	Path	0.717*	0.579*	0.299*	0.700*	0.386*	0.622*	0.776*	0.629*
	(SE)	(0.073)	(0.066)	(0.077)	(0.092)	(0.091)	(0.074)	(0.057)	(0.049)
	stand.	0.699	0.635	0.322	0.644	0.405	0.676	0.770	0.763
pre-SD1	Path	0.149*	0.091*	0.233*	0.281*	0.292*	0.151*	0.152*	0.193*
	(SE)	(0.039)	(0.040)	(0.067)	(0.048)	(0.061)	(0.034)	(0.037)	(0.050)
	stand.	0.286	0.181	0.277	0.429	0.396	0.359	0.305	0.298
SD1-post	Path	-0.108	0.088	0.123	0.026	0.148	-0.019	0.123	0.066
	(SE)	(0.134)	(0.135)	(0.094)	(0.134)	(0.112)	(0.169)	(0.123)	(0.072)
	stand.	-0.055	0.049	0.111	0.015	0.115	-0.009	0.061	0.052

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

For an extensive legend refer to Table 33

### 6.6.3 Summary

‘Defensiveness’ clearly did not operate as a partial mediator between pretests and posttests in heiQ-PPR. Despite significant associations between ‘defensiveness’ and pretests across all heiQ subscales, ‘defensiveness’ did not exert any influence on the posttests once the pretests were controlled. In contrast to the previous two datasets, it had been expected that social desirability would explain some variation in the change scores of heiQ-PPR; however, the ‘defensiveness’ component of social desirability as measured by the MC-C scale had to be ruled out as an explanatory variable for the obtained scores in this dataset.

## 6.7 Social desirability – heiQ-PPR Retro

### 6.7.1 Specific methods

Details on the procedure of the analyses of this section were provided in Section 6.4.1. The results of heiQ-PPR Retro are therefore presented hereafter. In contrast to the previous three sections, the present dataset contained retrospective instead of actual pretest data.

## 6.7.2 Results

Results were again similar to the previous observations in that the pretests – which in this case were retrospective pretests – were associated with ‘defensiveness’ in most subscales, whereas none of the subscales showed a significant relationship between the pretests and ‘self-presentation’ (see Table 42). Consequently, all subsequent models were reduced to the exploration of the effect of ‘defensiveness’ on the retrospective pretest-posttest data.

**Table 42** Covariance between SD1, SD2, and the retrospective pretests (Phi matrix), heiQ-PPR Retro

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
SD1-SD2	Cov	0.266*	0.278*	0.278*	0.278*	0.285*	0.281*	0.279*	0.274*
	(SE)	(0.059)	(0.061)	(0.061)	(0.061)	(0.062)	(0.061)	(0.061)	(0.061)
	Corr	0.523	0.527	0.524	0.520	0.520	0.523	0.520	0.522
SD1-retropre	Cov	0.254*	0.138	0.245*	0.460*	0.320*	0.459*	0.489*	0.295*
	(SE)	(0.067)	(0.071)	(0.063)	(0.101)	(0.060)	(0.123)	(0.095)	(0.068)
	Corr	0.308	0.157	0.319	0.387	0.470	0.310	0.389	0.346
SD2-retropre	Cov	-0.045	-0.058	0.033	0.182	0.063	0.164	0.206	-0.001
	(SE)	(0.062)	(0.068)	(0.064)	(0.095)	(0.046)	(0.131)	(0.107)	(0.071)
	Corr	-0.062	-0.075	0.048	0.169	0.146	0.123	0.181	-0.002

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

### Legend

retropre: retrospective pretest data

For an overview of the remaining abbreviations refer to the extensive legend provided in Table 33

The paths between the retrospective pretest and the posttest data are presented in Table 43. As observed in heiQ-PPT and heiQ-PPR, the association between the pretests and posttests was smallest in Skill and Technique Acquisition.

**Table 43** Regression of the posttests on the retrospective pretests (Gamma matrix), heiQ-PPR Retro

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
retropre-post	Path	0.622*	0.703*	0.473*	0.723*	0.624*	0.714*	0.817*	0.692*
	(SE)	(0.057)	(0.048)	(0.066)	(0.049)	(0.090)	(0.049)	(0.030)	(0.035)
	stand.	0.629	0.716	0.446	0.760	0.543	0.798	0.919	0.804

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

### Legend

retropre: retrospective pretest data

For an overview of the remaining abbreviations refer to the extensive legend provided in Table 33

Given that most pretests had been associated with ‘defensiveness’, the next models tested the relationship between ‘defensiveness’ and posttests once the pretests were controlled for this component of social desirability. Identical to the analyses on heiQ-PPR, no subscale showed a significant association between the heiQ posttests and ‘defensiveness’ (see Table 44). Hence, further exploration of these data was unnecessary as social desirability could be ruled out as a potential partial mediator between the retrospective pretests and the posttests.

**Table 44** Regression of ‘defensiveness’ (SD1) and the posttests on the retrospective pretests (Gamma matrix), and regression of the posttests on SD1 (Beta matrix), heiQ-PPR Retro

		PAE	HDB	STA	CAA	SMI	HSN	SIS	EWB
retropre- post	Path	0.633*	0.691*	0.454*	0.719*	0.600*	0.708*	0.843*	0.691*
	(SE)	(0.057)	(0.049)	(0.069)	(0.058)	(0.102)	(0.052)	(0.034)	(0.040)
	stand.	0.642	0.705	0.428	0.756	0.513	0.793	0.944	0.805
retropre - SD1	Path	0.218*	0.099	0.250*	0.199*	0.428*	0.127*	0.188*	0.253*
	(SE)	(0.056)	(0.057)	(0.066)	(0.041)	(0.093)	(0.035)	(0.037)	(0.056)
	stand.	0.298	0.139	0.313	0.386	0.470	0.304	0.386	0.344
SD1-post	Path	-0.057	0.118	0.077	0.019	0.075	0.035	-0.123	-0.001
	(SE)	(0.084)	(0.073)	(0.098)	(0.119)	(0.105)	(0.127)	(0.073)	(0.061)
	stand.	-0.042	0.086	0.058	0.011	0.058	0.016	-0.067	-0.001

\* Significant, i.e. the effect is more than twice its standard error (Bollen, 1989)

Legend

retropre: retrospective pretest data

For an overview of the remaining abbreviations refer to the extensive legend provided in Table 33

### 6.7.3 Summary

The results of this section were almost identical to heiQ-PPR. While the pretests – in this case retrospective pretests – were generally related to ‘defensiveness’, none of the posttests showed a significant association with ‘defensiveness’ once the heiQ pretests were controlled. Given that both heiQ-PPR and heiQ-PPR Retro were based on the same posttests, these results may not be unexpected when considering the somewhat similar association of their respective pretests (actual/retrospective) with the mediator (see Table 39 and Table 42), and similar paths from actual/retrospective pretest to posttest (see Table 40 and Table 43).

In view of the nature of retrospective pretests, some observations are noteworthy. Given that actual posttests and retrospective pretests were provided simultaneously, it could have been expected that the two scores would have had a stronger association than actual pretests and posttests (heiQ-PPR). However, this was not the case (see Table 40 and Table 43). Further, the overall lack of association of social desirability with these scores was unexpected, again

given that the ratings of actual posttest and retrospective pretest levels had been provided simultaneously. While retrospective pretest questions have been criticised for their potential vulnerability to social desirability bias (see Section 1.2.4.4), the observed findings of these analyses do not support this criticism which is in line with previous research (Howard *et al.*, 1981; Sprangers, 1989; Terborg *et al.*, 1980).

## 6.8 Discussion

### *Social desirability across heiQ-PP, heiQ-PPT, heiQ-PPR, and heiQ-PPR Retro*

Based on the findings of Chapters 3 and 5, the analyses of the present chapter were aimed at investigating the potential mediating effect of social desirability on the relationship between pretests and posttests in the four heiQ datasets. When modelling data of the MC-C scale it was observed that the 'self-presentation' component of social desirability could immediately be ruled out as a mediating variable, as it generally failed to show an association with the heiQ pretest data, a necessary condition for being a mediator (Baron & Kenny, 1986; Judd & Kenny, 1981). In contrast, the 'defensiveness' component of social desirability was generally related to the heiQ pretests. Consequently, this factor of social desirability was considered as a partial mediator across all datasets. However, the analyses indicated that 'defensiveness' did not account for any variation between actual/retrospective pretests and posttests in the heiQ subscales, with the exception of a minor effect in Emotional Well-Being in both heiQ-PP and heiQ-PPT. It was concluded that social desirability, as measured by the MC-C scale, had not influenced change scores across datasets.

Before moving to the final conclusions of the thesis, several aspects of the present analyses are discussed briefly. While no effect of social desirability on the relationship between pretest and posttest scores was found, the following alternative explanations are proposed:

- a) The MC-C scale was used to explore a potential mediating effect of social desirability on the pretest-posttest scores across the four heiQ datasets. While the original 33-item scale is one of the most widely used scales to assess social desirability – and there is sufficient support in the literature that the short form MC-C is a valid alternative to the full MC scale (see Section 1.2.4.5) – it is possible that the analyses were hampered by a suboptimal performance of this shortened measure. Despite acceptable fit indices, low reliability and some small factor loadings of the MC-C (see Section 6.3) may have limited the power of the analyses to detect mediational effects of social desirability.
- b) Alternatively, it is plausible that the assessment of change as measured by the heiQ was not vulnerable to socially desirable responses. In a similar manner to previous arguments

in the context of response shift (see Section 5.9), it is possible that social desirability did not exert any influence on scores because the heiQ items were written in a way that they discourage response styles (Osborne *et al.*, 2007).

- c) Furthermore, the potential co-existence of equivalent models needs to be acknowledged (Frazier *et al.*, 2004; MacCallum *et al.*, 1993). For example, it would have been plausible to define ‘defensiveness’ as a predictor of both pretest and posttest scores. However, in view of the research questions the present model seemed to make the most theoretical sense (Little *et al.*, 2007). That is, the path between pretest and posttest was understood as the primary path in the model, and social desirability was defined as a response style that potentially partially mediated the relationship between pretest and posttest data.
- d) Considering alternative explanations for the observed differences between the datasets – which were discussed in detail in Sections 3.7 and 5.9 – it is also possible that social desirability is not the only mediating variable. The present analyses may have resulted in biased estimates because additional mediators were not included (Judd & Kenny, 1981). However, given that an exploration of most of the explanations for the findings is beyond the scope of the present thesis, the investigation of further mediating variables needs to be left for future research.
- e) Finally, it is plausible that a model of moderated mediation (Baron & Kenny, 1986; Shrout & Bolger, 2002) may have been more appropriate to model the heiQ data, with variables such as age, gender, or education operating as moderating variables. For example, it is possible that the observed effects of the questionnaire design could have been explained by a mediating effect of socially desirable responding of older participants but not of their younger counterparts. The respective sample size of the present datasets, however, did not allow for such modelling; hence, these aspects have to be left for future research.

### *Concluding remarks*

The analyses of the present chapter led to the following results with regard to the research questions posed in Section 1.3:

- V. Can bias through social desirability be detected in change scores derived from actual pretest-posttest data?
  - Apart from a small significant association of the ‘defensiveness’ component of social desirability in heiQ subscale Emotional Well-Being of both heiQ-PP and heiQ-PPT, there was no indication of a mediational effect of social desirability in actual pretest-

posttest data of heiQ-PP, heiQ-PPT, or heiQ-PPR. That is, once actual pretest scores were controlled, social desirability did not account for any variance in posttest scores.

VI. Can bias through socially desirable responses be detected in change scores derived from retrospective pretest-posttest data?

- Identical to the previous analyses social desirability did not account for any variance in posttest scores once retrospective pretest scores were controlled. Hence, this bias could be ruled out as a potential confounder of change scores based on retrospective pretest-posttest data.

The present chapter provided evidence that social desirability bias did not operate as a substantive mediating variable in the heiQ datasets. Alternative explanations must therefore be responsible for the observed effects of the design of the posttest questionnaire that led to different ratings of actual posttest levels, as well as some observed differences in change scores based on retrospective as opposed to actual pretest data. Given that these alternative explanations were discussed in detail in the previous analysis chapters (see Sections 3.7 and 5.9), they shall not be repeated here. The present chapter concludes with the finding that the observed differences in change scores cannot be explained by a mediational effect of social desirability bias as measured by the MC-C scale. A final review of the findings of the present thesis is provided in the following Chapter 7.

# Chapter 7

Summary,  
conclusions, and  
future directions

## **7 Summary, conclusions, and future directions**

### **7.1 Introduction**

In consideration of the growing burden of chronic disease in both developed and developing countries, self-management interventions are an increasingly important part of chronic care (see Sections 1.2.1 and 1.2.2). Despite a large number of trials, current evidence about the effectiveness of these programs is still inconclusive. Several trials suggested clinical benefits for conditions such as diabetes and hypertension, whereas small effects were reported for conditions such as arthritis. However, a review of published meta-analyses and a systematic review of individual trials (see Sections 1.2.3.1 and 1.2.3.2) suggested that observed effects may not only be related to the disease groups but potentially to the types of outcomes that were assessed, with evaluations based on self-report outcomes generally showing small and largely inconsistent results (see Section 1.2.3.3). Given that the measurement of self-report outcomes is complex (Schwartz & Rapkin, 2004) and scores are susceptible to a range of biases (Cronbach, 1946; Paulhus, 1991; Podsakoff *et al.*, 2003; Webb *et al.*, 1966), this thesis investigated the validity of the traditional pretest-posttest design across different approaches to gathering outcomes data. In addition to exploring whether the design of the questionnaire influenced participants' ratings at posttest, systematic analyses were carried out to examine whether biases such as response shift or social desirability were present.

This final chapter provides a summary of the main findings of the thesis. In addition to an overview of Chapters 3, 5 and 6, the implications of these findings for the measurement of outcomes of self-management interventions are discussed. After considering the strengths and limitations of this research, the thesis concludes with recommendations and directions for future research.

### **7.2 Summary of the findings**

To investigate a range of aspects pertaining to the validity of scores derived from participant self-report, the following analyses were undertaken:

1. The first analyses of the thesis (Chapter 3) explored a) whether the design of the posttest questionnaires (heiQ-PP; heiQ-PPT; heiQ-PPR) influenced conclusions about program effectiveness and b) whether change scores derived from retrospective pretest-posttest data (heiQ-PPR Retro) led to a different set of conclusions when compared with change scores derived from actual pretest-posttest data (heiQ-PPR).



2. The review of common biases in program evaluations suggested that response shift, i.e. a change in perspective as a result of an intervention, may be a threat to the validity of change scores based on actual pre- and posttests in the evaluation of self-management interventions (see Section 1.2.4.4). As a consequence, Chapter 5 explored the influence of response shift bias on actual pretest-posttest data (heiQ-PP; heiQ-PPT; heiQ-PPR). Further, it was explored whether questionnaire items were invariant in the dataset of retrospective pretests and actual posttests (heiQ-PPR Retro).
3. The literature review further suggested that social desirability bias might pose a threat to the validity of change scores derived from participant self-report outcomes, particularly when change was derived from retrospective pretest-posttest data (see Section 1.2.4.5). As a consequence, analyses of Chapter 6 assessed whether this bias mediated change scores derived from actual/retrospective pretest-posttest data.

### *Summary of the findings of Chapter 3*

The findings of Chapter 3 suggested significant differences in actual posttest levels in six of eight heiQ subscales. Given that randomisation at study onset resulted in groups with similar scores at the beginning of the self-management interventions (=actual pretest), observed differences could be attributed to the design of the posttest questionnaires (see Section 3.4). Chapter 3 demonstrated that the design of the posttest questionnaires had a significant influence on mean change scores, with differences between groups being so substantial that it would result in different conclusions about program effectiveness. When applying an alternative method of presenting change by grouping participants into 'decline', 'no change', and 'improvement', differences between the groups were still present but less pronounced and conclusions about program effectiveness would not differ substantially across groups (see Section 3.5).

Participants who had been allocated to group heiQ-PPR provided self-rated pretest levels in retrospect, i.e. at the end of a self-management intervention (=retrospective pretest). That is, in addition to actual pretest levels that had been provided at the start of the intervention, a second set of pretest scores was available from this group. When comparing actual with retrospective pretest data, significant albeit small differences between the two scores were observed in three subscales, with retrospective pretests being lower than actual pretests. In view of the overall magnitude of the differences, these findings suggested that change based on either actual or retrospective pretest data would lead to similar conclusions about program effectiveness. These observations applied to both mean change scores and the alternative method of presenting change by grouping participants into 'decline', 'no change', and

'improvement'. It was further observed that the slightly larger mean change scores based on retrospective pretests compared with those based on actual pretests were generally related to smaller proportions of participants in the 'decline' category, rather than larger proportions in the 'improvement' category (see Section 3.6).

#### *Summary of the findings of Chapter 5*

Based on the findings of Chapter 3, subsequent analyses were aimed at exploring whether response shift bias could be detected in the datasets consisting of actual pre- and posttests. The application of a factor-analytic model to investigate whether group-level response shifts had a confounding effect on derived change scores revealed that a relatively small number of questionnaire items were affected by response shifts, with different items being non-invariant across datasets. That is, there was no pattern of response shift across designs; the lack of factorial invariance did not seem to be a characteristic of certain items (see Sections 5.4, 5.5, and 5.6). Although about 10% of questionnaire items across datasets indicated response shifts, i.e. these items needed to be excluded to ensure an unbiased comparison of the factor means, partial metric and scalar invariance could be established in each subscale across all datasets. Hence, group-level response shifts were not strong enough in any of the datasets to threaten the validity of comparing actual pretest with posttest data (see Section 5.8).

When applying the factor-analytic model on the sample of retrospective pretest-posttest data, more questionnaire items were found that lacked factorial invariance (see Section 5.7). About one third of items indicated some form of non-invariance of which almost all items had to be excluded from the calculation of change scores to ensure an unbiased comparison of the factor means. While items of a total of five heiQ subscales were affected, metric and scalar invariance could still be established in all subscales, permitting an unbiased comparison of the factor means in this dataset (see Section 5.8).

#### *Summary of the findings of Chapter 6*

Given that the present data did not support a possible confounding effect of response shift bias and could not explain the differences between the groups observed in Chapter 3, the final analyses of the present thesis explored the role of social desirability bias across the four datasets. Apart from a small significant association of the 'defensiveness' component of social desirability with change scores in heiQ subscale Emotional Well-Being in heiQ-PP and heiQ-PPT, these analyses demonstrated that there was no indication of a mediational effect of social desirability in any of the datasets (see Sections 6.4 to 6.7).

### 7.3 Conclusions

This thesis was aimed at exploring the validity of the traditional pretest-posttest design to assess program outcomes using the self-report inventory heiQ. To the author's knowledge this is the first study in the self-management setting to systematically investigate different approaches to gathering self-report outcomes data and apply advanced group-level statistical models to explore common biases that potentially threaten the validity of change scores.

The analyses demonstrated that contrasting approaches to gathering self-report outcomes data had a significant influence on participants' self-rated levels at actual posttest. That is, the collection of 'direct change' questions (=transition questions) in addition to actual posttest data resulted in significantly higher posttest levels in two subscales compared with posttest data derived from a traditional posttest questionnaire. This effect was significantly stronger in a questionnaire design that collected retrospective pretests simultaneous to collecting actual posttests, with significantly higher posttest levels in six of eight subscales compared with data derived from a traditional posttest design. In particular, the latter observation is critical for program evaluators who consider applying retrospective pretest data as a substitute for actual pretests when measuring change given that obtained change scores are influenced by this research strategy.

While a large body of literature exists comparing actual with retrospective pretests, this thesis has highlighted a dimension of deriving change scores from retrospective pretest data that to date appears to have been underestimated. Rather than focusing on differences between actual and retrospective pretests, this thesis has provided new evidence that the process of collecting retrospective pretests at posttest can have a substantial influence on reported posttest levels and in turn on the magnitude of change scores. That is, an additional task at posttest that may highlight to participants that the researcher is interested in 'change scores' may strongly influence their thinking and consequently their ratings at posttest. Given that a direct measure of change in form of transition questions appeared to have less influence on the ratings, the nature of the cognitive task and potentially those aspects of this task that highlight to participants what the researcher is seeking appear to be important aspects to be able to explain observed group differences.

While results suggested that providing retrospective pretest data led to significantly higher posttest levels, it remained uncertain whether change scores derived from these data were confounded because of the collection of retrospective pretest data simultaneous to collecting posttest data (see Section 3.7), or whether the simultaneous collection of these data led to a more accurate reflection of change. As described in Section 1.2.4.4, the original rationale for this type of questionnaire design is that it is used to circumvent response shift bias, i.e. by

providing 'pretest' and posttest data at the end of an intervention it is assumed that respondents provide both data from the same perspective. While it is hoped that both data are provided from the same frame of reference, i.e. the respondent's frame of reference after the intervention, it cannot be ruled out that participants provide their posttest levels relative to their retrospective pretest levels, as the analyses of the thesis suggest. Therefore, it remains to be explored whether these posttest scores are a more valid reflection of people's posttest levels with regard to deriving change scores from these data. As a consequence, subsequent analyses attended to the following questions:

- Do group-level response shifts have confounding effects on mean change scores based on actual pretest-posttest data across datasets, and are potential confounding effects alleviated in the dataset that collected retrospective pretest in addition to posttest data?
- Are items in the dataset of retrospective pretest-posttest data invariant?

As summarised in Section 7.2, across datasets of actual pre- and posttests only a relatively small number of questionnaire items were affected by group-level response shifts, i.e. only a few questionnaire items were non-invariant. In contrast, more items were found to be non-invariant in the dataset of retrospective pre- and posttests. These findings were unexpected and have several implications for program evaluation and future research:

Firstly, the finding that the majority of heiQ items were invariant between actual pretest and posttest across designs suggests that the data gathered using the heiQ are not confounded by response shift bias. Given that the majority of items were not only found to be invariant across actual pretest-posttest data but also across different designs indicates that the psychometric properties of the heiQ are robust. Given that at least partial factorial invariance needs to be established to ensure an unbiased comparison of factor means, these findings suggest that the heiQ can be recommended as a robust instrument to measure outcomes of self-management programs. In view of the slightly worse performance of the dataset of retrospective pre- and posttests, it remains to be investigated whether retrospective pretest data can be recommended for use in evaluations of self-management interventions.

Secondly, while these findings underscore the quality of the heiQ when used on actual pretest-posttest data, analyses could not provide any evidence regarding the validity of each of the datasets. Given that large differences in mean posttest scores had been observed, it was anticipated that sophisticated factor-analytic modelling would (partly) explain differences between the datasets. Surprisingly, all datasets appeared to perform similarly well. That is, despite differences in mean change scores, the present analyses suggested that the validity concerns about change scores derived from traditional pretest-posttest data (see Section 1.2.4.4) seemed to be unfounded in the present datasets.

Thirdly, in the datasets of actual pre- and posttests only a small number of questionnaire items were found to be non-invariant, i.e. only a few items indicated response shift bias. As discussed in Section 5.9, along with a small number of items affected by response shift, there was no pattern across designs. The lack of factorial invariance did not seem to be a stable characteristic of certain items. However, it was found that response shifts seemed to occur in items of some subscales but not in items of other subscales. For example, none of the items of Constructive Attitudes and Approaches, Health Service Navigation, Social Integration and Support, and Emotional Well-Being seemed to be affected by response shift, whereas non-invariant items generally belonged to one of the remaining four heiQ subscales. Despite this somewhat consistent observation across subscales, those subscales located to the 'evaluative' pole of the continuum as introduced in Section 1.2.3.3 (see also discussion in Section 5.9) were not noticeably more affected by response shift bias than others. Although more items were found to be non-invariant in the dataset of retrospective pre- and posttests, it was again not possible to explain these findings by this continuum.

Finally, although only a small number of items indicated response shifts, it had to be ensured that these group-level response shifts did not pose a threat to the validity of comparing factor means. Because of the small number of items affected by group-level response shifts in the datasets of actual pre- and posttests, (partial) factorial invariance could be established in all subscales across all datasets. Hence, an unbiased comparison of factor means was possible. Although more items were found to be non-invariant in the dataset of retrospective pre- and posttests, (partial) factorial invariance could again be established in all subscales, ensuring an unbiased comparison of factor means. As a result, the application of a factor-analytic model to detect group-level response shifts did not provide evidence of a possible confounding effect on results of evaluations of self-management interventions.

Given that observed differences between the datasets did not relate to confounding effects through response shifts, the final analyses of the thesis attended to the following question:

- Are mean change scores across datasets related to a confounding effect through social desirability bias?

The application of a mediational model based on structural equation modeling did not provide evidence that social desirability, as measured by the MC-C, had a confounding effect on self-report data derived from participants of self-management courses. This is again an important finding for program evaluators. As group-based health programs and potential attachment of participants to course leaders may encourage response styles such as social desirability, it is important that potential effects of this bias could be ruled out in the present research. Hence, the study provides strong evidence that the measurement of outcomes of self-management interventions is not affected by social desirability in assessments applying the heiQ.

In summary, the group-level analyses in this research suggested that contrasting approaches to gathering outcomes data from participants of self-management interventions resulted in significantly different ratings of actual posttest levels. However, the observed change scores could neither be explained by confounding effects due to response shift nor social desirability biases. The thesis must conclude that the nature of the cognitive task at posttest had a significant influence on participants' self-rated posttest levels, whereas group-level response shift and social desirability biases could be ruled out as a possible explanation for observed differences between the groups. Hence, this study established that future research is needed on the influence of a cognitive task on obtained results as well as the aspect of the task that may highlight to participants that the researcher is interested in an assessment of 'change'.

In spite of the need for further research, based on the findings of the present thesis, some recommendations are provided for future evaluations of self-management interventions:

Given that in this thesis the posttest heiQs were randomly distributed across courses, it can be assumed that the obtained change scores would have been identical across groups if the same method had been applied. To achieve an equivalent interpretation of obtained scores when applying any of these questionnaires, one possibility is to modify the definition of what constitutes 'improvement' across the four methods analogous to the concept of a minimal important difference (Guyatt *et al.*, 2002; Jaeschke *et al.*, 1989; Juniper *et al.*, 1994). Based on the findings, it is proposed that such a minimal important difference be set lower for mean change scores derived from a traditional pretest-posttest design (heiQ-PP), whereas such a threshold would need to be higher when change scores were derived from a design such as heiQ-PPR, given that this design had resulted in significantly larger change scores. In view of the heiQ subscales it may further be appropriate to adjust these thresholds according to each subscale. Finally, if change scores were based on factor means (see Chapter 5) – instead of arithmetic means (see Chapter 3) – these thresholds would again require adjustment.

The previous suggestions only attended to the practical significance of the results as they pertain to mean change scores. As observed in Chapter 3, when using an alternative method of reporting change, conclusions about program effectiveness were less affected by the questionnaire design. That is, the categorisation of participants into 'decline', 'no change', or 'improvement' led to largely similar conclusions across datasets. Hence, this method may be the preferred way of reporting results as it appears robust to design effects. If an approach such as this were to be adopted, it would still be useful to determine a benchmark referring to what constitutes a successful intervention. Akin to a minimal important difference it may be useful to adjust the proportion of participants in the 'improvement' category according to the method used. For example, it may suffice to achieve one third of participants across heiQ subscales in the 'improvement' category if results were derived from a questionnaire design

such as heiQ-PP, whereas 40% of participants may be needed in the 'improvement' category for an intervention to be considered successful if data were derived from a questionnaire design such as heiQ-PPR (see Section 3.5).

## 7.4 Strengths

Although some strengths and limitations of this study have been described throughout the thesis, they are summarised in the following two sections. First, a list of the main strengths of the thesis is provided:

- a) This thesis explored the measurement of outcomes of self-management interventions by applying a unique research design. To the author's knowledge, no comparable research has been conducted in the field of chronic disease self-management interventions that investigated the influence of the questionnaire design on program outcomes and resulting conclusions about program effectiveness. To ensure that the comparison of scores was free of bias through potential intra-group effects, the questionnaires were randomly distributed within self-management courses. The findings of this research provide important insight into the measurement of change. In particular, the finding that change scores were significantly influenced by the cognitive tasks participants performed at posttest suggests that more attention must be paid to this issue to ensure that outcomes can be interpreted correctly.
- b) Apart from the unique research design, advanced statistical techniques were applied to answer the research questions of this thesis (see Section 1.3). With samples of  $n > 300$  participants per group it was possible to apply SEM, a statistical method appropriate for multivariate data analysis at a group level (Bollen & Arminger, 1991). This technique is particularly useful when modelling observed variables that are determined by underlying latent constructs, as is the case with heiQ data. A main advantage of this technique is that through using latent variables in the regression analyses, obtained parameter estimates are assumed to be largely free of measurement error (Judd & Kenny, 1981).
- c) Because of the application of SEM, it was further possible to treat the data in a way that was appropriate for the ordinal scaling of the questionnaire. The moment matrix based on polychoric correlations was used in combination with its asymptotic covariance matrix. While the sample sizes did not allow for WLS, RML was applied for the estimation of the model parameters. RML has been described as the best alternative to WLS if data are ordinal but sample sizes are too small for WLS (Jöreskog, 2002-2005).

- d) The application of SEM further provided the opportunity to test measurement invariance of the heiQ items across actual/retrospective pretest and posttest data, and at the same time explore several types of response shift. While the test for measurement invariance is an advanced statistical technique, the response shift model is not only one of the most recently developed approaches to detect response shift (Oort, 2005b; Oort *et al.*, 2005), but in the present thesis the 4-step procedure was further extended to the hierarchy of measurement invariance and to the analysis of ordinal data. The factor-analytic model described in Chapter 5 therefore advances current methods to detect response shift at a group level.
- e) To the author's knowledge, this thesis also provides the first test of the influence of social desirability bias as a partial mediating variable in the measurement of outcomes of self-management interventions. While two studies included scores on social desirability as covariates (Glasgow *et al.*, 1992; Vlaeyen *et al.*, 1996), no further study was found that collected data on social desirability bias and included this variable in statistical modelling. Again, an advanced statistical technique was applied in this thesis that allowed for the analysis of complex relationships, such as including social desirability as a mediational variable (Baron & Kenny, 1986; Holmbeck, 1997).
- f) Finally, this thesis proposed a robust method of presenting change across participants of self-management courses as an alternative to using mean change scores. This method is similar to NNT analyses (see Section 3.5) and consists of grouping subjects into 'decline', 'no change', or 'improvement' categories on the basis of an individual ES of 0.5. Two advantages of this method are apparent: 1) program evaluations that are based on mean change scores derived from participant self-report may be vulnerable to a range of biases and other threats to the validity of obtained scores. Therefore, an alternative way of presenting only those participants that received substantial benefits from attending an intervention and comparing these with people who experienced substantial 'decline' may provide a clearer indication of the value of a program. 2) Further, results suggested that this method is robust to differences between questionnaire designs. This may be another indication of the superiority of this method to present outcomes of such interventions.

## 7.5 Limitations

The main weaknesses of the present thesis are as follows:

- a) This research relied on one data source only. That is, only quantitative data derived from participant self-report were used to answer the research questions. In contrast, further



data would have been valuable for triangulating the findings (Webb *et al.*, 1966). When considering the observations made in the present thesis, a useful approach may have been to obtain qualitative data on the appraisal processes of the participants of the three randomised groups. However, within the time and financial constraints of the thesis, it is assumed that appropriate methods were chosen to answer the research questions.

- b) The present study relied on data derived from only one self-report inventory – the heiQ – to measure outcomes of self-management interventions. As a result, the findings have limited generalisability to evaluations using other instruments. However, due to the dearth of alternative tools that were specifically designed to assess outcomes of such programs (Osborne *et al.*, 2007), these findings are highly relevant for current and future users of the heiQ. It remains that neither response shift nor social desirability bias appear to confound results when using the heiQ to assess program outcomes.
- c) While it can be assumed that the study samples were representative of self-management course participants in Australia, it is possible that the samples were heterogeneous, with some effects not detected because they only applied to a subgroup of people whose outcomes were hidden in the group-level data. For example, research has suggested that outcomes differ across age (Fu *et al.*, 2003; Lorig *et al.*, 1985; Nolte *et al.*, 2007). Applied to this thesis this may mean that, for example in Chapter 5, the analyses obscured response shift effects in certain subgroups. That is, it is possible that more young people experience response shifts, whereas their older counterparts may accept their chronic disease as part of ageing (O'Boyle *et al.*, 2000). Consequently, a treatment-induced response shift may not occur in older course participants whose data may have obscured those of younger subjects (Howard, Ralph *et al.*, 1979; Sprangers, 1989). For social desirability an alternative model may have been more suitable, such as a model of moderated mediation (Baron & Kenny, 1986; Shrout & Bolger, 2002). However, given that larger sample sizes are needed to model any of the above suggestions, these types of analyses must be left for future research.
- d) In the context of response shift, it is also plausible that its current operationalisation and the group-level approach do not fully capture this bias: 1) the match between types of response shifts (Schwartz & Sprangers, 1999; Sprangers & Schwartz, 1999) and the parameters of a factor model (Oort, 2005b; Oort *et al.*, 2005) may be problematic. That is, while the definitions of *reconceptualisations* and *reprioritisations* seem plausible, volatile factor patterns and/or unstable performances of items may also be an indication of poor quality of the measurement instrument rather than response shift. 2) *Recalibration*, on the other hand, results from a change in the definition of a scale. In addition to the suggested interpretation of recalibration (see Section 1.2.4.4), it is however possible that a renewed

judgment manifests itself in a subject's position on the scale relative to the pretest levels. This may result in confounding of the magnitude of change scores, while items may still be invariant across measurement occasions. 3) The factor-analytic approach has been criticised to be insensitive to detect response shifts at a subject level. Alternative methods such as those allowing for within-subject models have been proposed and may be more suitable for researchers interested in detecting and quantifying the direction as well as magnitude of response shift bias (Donaldson, 2005).

While the first two aspects regarding the appropriateness of the factor-analytic method warrant further research, it is assumed that the method was suitable to detect response shifts if they had occurred in a sufficiently large proportion of course participants. Hence, for the purpose of the present thesis that was aimed at detecting response shift bias in view of its potential confounding effect on results of evaluations of self-management interventions, it is assumed that the application of group-level analyses were appropriate. Considering that evaluations are aimed at assessing changes at a group level, individual response shifts were not critical, unless accumulated individual response shifts had a confounding effect on the group-level results of these evaluations. As a consequence, Donaldson's (2005) critique of the approach does not seem to invalidate the application of such group-level analyses to detect the influence of response shift bias on the results of program evaluations but it rather extends the range of possible methods to detect response shift and further the understanding of this phenomenon.

- e) In the context of social desirability bias, it is possible that the MC-C scale was not a strong indicator of this response style. With satisfactory goodness-of-fit indices but low reliability and some small factor loadings, the MC-C had rather suboptimal properties. Hence, a social desirability measure with stronger psychometric properties may have led to different results.
- f) Finally, it is possible that mediators were omitted in the analyses of Chapter 6 (Judd & Kenny, 1981). If this was the case such models are considered misspecified (Shrout & Bolger, 2002). In view of the explanations for the findings (see Sections 3.7, 5.9 and 6.8), it was, however, beyond the scope of this thesis to model further variables.

## **7.6 Recommendations for future research**

Following from the analyses of this thesis the findings of Chapters 3, 5 and 6 warrant further investigation in future studies.

The analyses of Chapter 3 provided important insight into the performance of different ways of measuring change as well as the influence of presenting results on the conclusions about program effectiveness. While it was found that subjects' ratings of their actual posttest levels were significantly influenced by the presence of a second cognitive task at posttest, reasons for these higher levels could not be identified in this thesis. It is therefore proposed that further research attends to respondents' cognitive processes when providing answers to a range of questions as well as those aspects of the task that highlight to respondents what the researcher is seeking. In particular, qualitative analyses on possible confounding effects on scores through *satisficing*, *effort justification bias*, and *implicit theory of stability or change* are recommended to find explanations for observed differences between the traditional pretest-posttest design and those involving additional tasks – particularly a retrospective pretest – at posttest.

The group-level *response shift* model as applied in Chapter 5 provided little evidence of the presence of response shifts in the present data. Regardless of potential limitations of this method as discussed in Sections 5.9 and 7.5, it is assumed that the present analyses were appropriate to rule out reconceptualisations, reprioritisations, and recalibrations of most heiQ items between actual pretest and posttest. While it is reassuring that evaluations of self-management programs based on actual pretest-posttest data do not seem to be confounded by group-level response shifts, it remains to be determined whether group-levels analyses obscure potential response shift effects at the individual level. Hence, a more individualistic approach may deliver further insight into the response shift phenomenon (Donaldson, 2005). Future models might also include response shift as a mediating variable, with response shift measured by questionnaires such as the 'heiQ Perspective' (Osborne *et al.*, 2007). Finally, it was proposed that *recalibration* may be understood as a change in participants' relative position on a given scale from actual pretest to posttest. Therefore, further research into the influence of this type of response shift on the overall magnitude of change appears useful. For example, it is possible that this type of recalibration might explain observed differences in mean change scores across the three questionnaire designs.

The analyses of Chapter 5 further suggested that several heiQ items were non-invariant between retrospective pretest and posttest. Hence, before the application of this design can be recommended for the evaluation of self-management programs, further research appears necessary. While qualitative data may provide important insight into participants' perspective when providing answers to respective questions, further analyses with regard to a potential linear dependency between posttest and retrospective pretest data seems useful. Finally, it is necessary to investigate whether retrospective pretest data relative to posttest data are a more accurate reflection of change or if retrospective pretest data are confounded because

of *recall bias* and/or participants engaging in an *implicit theory of stability or change* when responding to retrospective pretest questions.

The results of the analyses of Chapter 6 suggested that *social desirability* could be ruled out as a potential explanation for observed differences between groups. Although it is assumed that the analyses were appropriate to detect this response style, it remains that the MC-C scale had suboptimal psychometric properties. It may therefore be useful to carry out the same analyses but using a different measure of social desirability. Apart from applying one of the already existing scales as referred to in Section 1.2.4.5, it may also be useful to develop a refined or new measure of social desirability through advanced statistical techniques such as those applied for the development of the heiQ (see Section 2.3.2).

In conclusion, this thesis has provided important insight into the measurement of outcomes of chronic disease self-management programs. While validity concerns about the traditional pretest-posttest method in view of confounding effects on results through response shift and social desirability biases could not be supported, the thesis has highlighted that the cognitive task participants perform when providing data at posttest – in particular when providing retrospective pretest data – significantly influenced the ratings of their actual posttest levels. Considering that past research has predominantly focused on other aspects of validity such as applying control group designs to circumvent common threats to internal and external validity, the thesis suggests that more attention must be paid to other issues such as the influence of the questionnaire design on results. This thesis concludes that further research, in particular into the influence of cognitive tasks on obtained scores, is important to improve the interpretation of outcomes of self-management interventions.

## References

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: John Wiley & Sons.
- Ahmed, S, Mayo, NE, 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, 611-627.
- Albrecht, GL, & Devlieger, PJ. (1999). The disability paradox: high quality of life against all odds. *Soc Sci Med*, 48, 977-988.
- Allison, PD. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112, 545-557.
- Andrews, P, & Meyer, RG. (2003). Marlowe-Crowne Social Desirability Scale and Short Form C: forensic norms. *Journal of Clinical Psychology*, 59(4), 483-492.
- Arbuckle, JL. (1996). Full information estimation in the presence of incomplete data. In G Marcoulides & R Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Armenakis, AA, & Zmud, RW. (1979). Interpreting the measurement of change in organizational research. *Personnel Psychology*, 32, 709-724.
- Aronson, E, & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology*, 59, 177-181.
- Astin, JA, Beckner, W, Soeken, K, Hochberg, MC, & Berman, B. (2002). Psychological interventions for rheumatoid arthritis: a meta-analysis of randomized controlled trials. *Arthritis Rheum*, 47(3), 291-302.
- Australian Bureau of Statistics. (2004). Scenarios for Australia's aging population. Retrieved November 27, 2006, from <http://www.abs.gov.au/Ausstats/abs@.nsf/0/95560b5d7449b135ca256e9e001fd879?OpenDocument#>.
- Australian Institute of Health and Welfare. (2002). Chronic diseases and associated risk factors in Australia, 2001. Canberra.
- Ballard, R. (1992). Short forms of the Marlowe-Crowne Social Desirability Scale. *Psychol Rep*, 71, 1155-1160.
- Ballard, R, Crino, MD, & Rubinfeld, S. (1988). Social desirability response bias and the Marlowe-Crowne Social Desirability Scale. *Psychol Rep*, 63, 227-237.
- Bandura, A. (1997). *Self-efficacy - the exercise of control*. New York: W.H. Freeman and Company.
- Barger, SD. (2002). The Marlowe-Crowne affair: short forms, psychometric structure, and social desirability. *J Pers Assess*, 79(2), 286-305.
- Barlow, JH, Turner, AP, & Wright, CC. (2000). A randomized controlled study of the Arthritis Self-Management Programme in the UK. *Health Educ Res*, 15(6), 665-680.
- Barlow, JH, Wright, CC, Sheasby, J, Turner, AP, & Hainsworth, J. (2002). Self-management approaches for people with chronic conditions: a review. *Patient Educ Couns*, 48(2), 177-187.

- Baron, RM, & Kenny, DA. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, 51(6), 1173-1182.
- Batterham, R, Southern, D, Appleby, N, Elsworth, G, Fabris, S, Dunt, D, et al. (2002). Construction of a GP integration model. *Soc Sci Med*, 54(8), 1225-1241.
- Baumgartner, H, & Steenkamp, J-BEM. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156.
- Beaglehole, R, & Yach, D. (2003). Globalisation and the prevention and control of non-communicable disease: the neglected chronic diseases of adults. *Lancet*, 362, 903-908.
- Bentler, PM. (1977). Factor simplicity index and transformations. *Psychometrika*, 42(2), 277-295.
- Bentler, PM. (1990). Comparative fit indexes in structural models. *Psychol Bull*, 107(2), 238-246.
- Bentler, PM, & Dudgeon, P. (1996). Covariance structure analysis: statistical practice, theory, and directions. *Annu Rev Psychol*, 47(1), 563-592.
- Bentler, PM, & Yuan, K-H. (1999). Structural Equation Modeling with small samples: test statistics. *Multivariate Behavioral Research*, 34(2), 181-197.
- Billiet, JB, & McClendon, MJ. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, 7(4), 608-628.
- Bland, JM, & Altman, DG. (1994a). Statistics notes: regression towards the mean. *BMJ*, 308, 1499.
- Bland, JM, & Altman, DG. (1994b). Statistics notes: some examples of regression towards the mean. *BMJ*, 309, 780.
- Boesen, EH, Ross, L, Frederiksen, K, Thomsen, BL, Dahlstrøm, K, Schmidt, G, et al. (2005). Psychoeducational intervention for patients with cutaneous malignant melanoma: a replication study. *J Clin Oncol*, 23(6), 1270-1277.
- Bollen, KA. (1987). Total, direct, and indirect effects in Structural Equation Models. *Sociological Methodology*, 17, 37-69.
- Bollen, KA. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Bollen, KA, & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology*, 21, 235-262.
- Bollen, KA, & Long, JS. (1993). Introduction. In K Bollen & J Long (Eds.), *Testing Structural Equation Models*. Newbury Park, London, New Delhi: Sage Publications.
- Boomsma, A, & Hoogland, JJ. (2001). The robustness of LISREL modeling revisited. In R Cudeck, S Du Toit & D Sörbom (Eds.), *Structural Equation Modeling: present and future* (pp. 139-168). Lincolnwood, IL: Scientific Software International.
- Brady, TJ. (1997). Do common arthritis self-efficacy measures really measure self-efficacy? *Arthritis Care Res*, 10(1), 1-8.

- Breetvelt, IS, & Van Dam, FSAM. (1991). Underreporting by cancer patients: the case of response-shift. *Soc Sci Med*, 32(9), 981-987.
- Brown, MB, & Forsythe, AB. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364-367.
- Brown, RL. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods. *Structural Equation Modeling*, 1, 287-316.
- Browne, MW. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111-150.
- Browne, MW, & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24(4), 445-455.
- Browne, MW, Cudeck, R, Tateneni, K, & Mels, G. (2004). CEFA: Comprehensive Exploratory Factor Analysis, Version 2.00 [Computer software and manual]. Retrieved June 8, 2006, from <http://quantrm2.psy.ohio-state.edu/browne/>.
- Browne, MW, & Du Toit, SHC. (1991). Models for learning data. In M Collins & J Horn (Eds.), *Best methods for the analysis of change*. Washington, DC: American Psychological Association.
- Browne, MW, MacCallum, RC, Kim, CT, Andersen, BL, & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, 7(4), 403-421.
- Burckhardt, CS, Mannerkorpi, K, Hedenberg, L, & Bjelle, A. (1994). A randomized, controlled clinical trial of education and physical training for women with fibromyalgia. *J Rheumatol*, 21, 714-720.
- Byrne, BM. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Byrne, BM. (2001). Structural equation modeling with AMOS, EQS, and LISREL: comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*, 1, 55-86.
- Byrne, BM, Shavelson, RJ, & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol Bull*, 105, 456-466.
- Campbell, DR, & Stanley, JC. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally College Publisher.
- Carroll, JB. (1978). How shall we study individual differences in cognitive abilities? Methodological and theoretical perspectives. *Intelligence*, 2, 87-115.
- Cattell, RB. (1988). The meaning and strategic use of factor analysis. In J Nesselroade & R Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 131-203). New York: Plenum Press.
- Cheung, GW, & Rensvold, RB. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.
- Child, D. (1990). *Essentials of factor analysis* (2nd ed.). London: Cassell.

- Chodosh, J, Morton, SC, Mojica, W, Maglione, M, Suttorp, MJ, Hilton, L, et al. (2005). Meta-analysis: chronic disease self-management programs for older adults. *Ann Intern Med*, 143(6), 427-438.
- Choi, BCK, & Pak, AWP. (2005). A catalog of biases in questionnaires. Retrieved March 15, 2005, from [http://www.cdc.gov/pcd/issues/2005/jan/04\\_0050.htm](http://www.cdc.gov/pcd/issues/2005/jan/04_0050.htm).
- Clark, NM, Janz, NK, Becker, MH, Schork, MA, Wheeler, J, Liang, J, et al. (1992). Impact of self-management education on the functional health status of older adults with heart disease. *Gerontologist*, 32, 438-443.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychol Bull*, 112, 155-159.
- Cohen, JL, van Houten Sauter, S, DeVellis, RF, & McEvoy DeVellis, B. (1986). Evaluation of arthritis self-management courses led by laypersons and by professionals. *Arthritis Rheum*, 29(3), 388-393.
- Collins, LM, Graham, JW, & Flaherty, BP. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, 33(2), 295-312.
- Cook, TD, & Campbell, DT. (1979). *Quasi-experimentation: design & analysis issues for field settings*. Chicago: Rand McNally College Publishing Company.
- Cote, JA, & Buckley, R. (1987). Estimating trait, method, and error variance: generalizing across 70 construct validation studies. *Journal of Marketing Research*, 24, 315-318.
- Crino, MD, Svoboda, M, Rubinfeld, S, & White, MC. (1983). Data on the Marlowe-Crowne and Edwards Social Desirability scales. *Psychol Rep*, 53, 963-968.
- Cronbach, LJ. (1946). Response sets and test validity. *Educ Psycho Measure*, 6, 475-494.
- Cronbach, LJ. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(8), 297-334.
- Cronbach, LJ, & Furby, L. (1970). How should we measure "change" - or should we? *Psychol Bull*, 74(1), 68-80.
- Crowne, DP, & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349-354.
- Crowne, DP, & Marlowe, D. (1964). *The approval motive: studies in evaluative dependence*. New York, London, Sydney: John Wiley & Sons.
- Cudeck, R, & Browne, MW. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, 18(2), 147-167.
- Curran, PJ, West, SG, & Finch, JF. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16-29.
- Daltroy, LH, Larson, MG, Eaton, HM, Phillips, CB, & Liang, MH. (1999). Discrepancies between self-reported and observed physical function in the elderly: the influence of response shift and other factors. *Soc Sci Med*, 48(11), 1549-1561.



- DeMaio, T.J. (1984). Social desirability and survey measurement: a review. In C Turner & E Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 257-282). New York: Russell Sage.
- Dempster, AP, Laird, NM, & Rubin, DB. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.
- Deshields, T, Tait, R, Gfeller, J, & Chibnall, J. (1995). Relationship between social desirability and self-report in chronic pain patients. *Clin J Pain*, 11, 189-193.
- Devos-Comby, L, Cronan, T, & Roesch, SC. (2006). Do exercise and self-management interventions benefit patients with osteoarthritis of the knee? A metaanalytic review. *J Rheumatol*, 33(4), 744-756.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9(3), 327-346.
- Donaldson, GW. (2005). Structural equation models for quality of life response shifts: promises and pitfalls. *Qual Life Res*, 14, 2345-2351.
- Edwards, AL. (1957). *The social desirability variable in personality assessment and research*. New York: Holt, Rinehart and Winston, Inc.
- Edwards, AL, & Walsh, JA. (1964). Response sets in standard and experimental personality scales. *American Educational Research Journal*, 1, 52-61.
- Enders, CK. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1), 128-141.
- Eysenck, HJ. (1994). Meta-analysis and its problems. *BMJ*, 309, 789-792.
- Fan, X, Thompson, B, & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6, 56-83.
- Fayers, PM, & Hand, DJ. (1997). Factor analysis, causal indicators and quality of life. *Qual Life Res*, 6, 139-150.
- Ferrando, PJ, & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. *Psicológica*, 21, 301-323.
- Ferrando, PJ, Lorenzo-Seva, U, & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research*, 38(3), 353-374.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications.
- Fischer, DG, & Fick, C. (1993). Measuring social desirability: short forms of the Marlowe-Crowne Social Desirability Scale. *Educ Psycho Measure*, 53, 417-424.
- Fraboni, M, & Cooper, D. (1989). Further validation of three short forms of the Marlowe-Crowne scale of Social Desirability. *Psychol Rep*, 65, 595-600.

- Franke, RH. (1979). The Hawthorne experiments: re-view. *American Sociological Review*, 44(5), 861-867.
- Franke, RH, & Kaul, JD. (1978). The Hawthorne experiments: first statistical interpretation. *American Sociological Review*, 43, 623-643.
- Frasure-Smith, N, Lespérance, F, Juneau, M, Talajic, M, & Bourassa, MG. (1999). Gender, depression, and one-year prognosis after myocardial infarction. *Psychosomatic Medicine*, 61, 26-37.
- Frazier, PA, Tix, AP, & Barron, KE. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1), 115-134.
- Fries, JF, Lorig, K, & Holman, HR. (2003). Patient self-management in arthritis? Yes! *J Rheumatol*, 30(6), 1130-1132.
- Fu, D, Fu, H, McGowan, P, Shen, Y-E, Zhu, L, Yang, H, et al. (2003). Implementation and quantitative evaluation of chronic disease self-management programme in Shanghai, China: randomized controlled trial. *Bulletin of the World Health Organization*, 81(3), 174-182.
- Gerbing, DW, & Anderson, JC. (1993). Monte Carlo evaluations of goodness-of-fit indices for Structural Equation Models. In K Bollen & J Long (Eds.), *Testing Structural Equation Models*. Newbury Park, London, New Delhi: Sage Publications.
- Gibson, PC, Powell, H, Coughlan, J, Wilson, AJ, Abramson, M, Haywood, P, et al. (2002). Self-management education and regular practitioner review for adults with asthma. *The Cochrane Database of Systematic Reviews*(3, Art. No.: CD001117. DOI: 10.1002/14651858.CD001117.).
- Glasgow, RE, Toobert, DJ, Hampson, SE, Brown, JE, Lewinsohn, PM, & Donnelly, J. (1992). Improving self-care among older patients with type II diabetes: the "Sixty something..." study. *Patient Educ Couns*, 19, 61-74.
- Goeppinger, J, Arthur, MW, Baglioni Jr., AJ, Brunk, SE, & Brunner, CM. (1989). A reexamination of the effectiveness of self-care education for persons with arthritis. *Arthritis Rheum*, 32(6), 706-716.
- Gold, MS, & Bentler, PM. (2000). Treatment of missing data: a Monte Carlo comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling*, 7(3), 319-355.
- Gold, MS, Bentler, PM, & Kim, KH. (2003). A comparison of maximum-likelihood and asymptotically distribution-free methods of treating incomplete nonnormal data. *Structural Equation Modeling*, 10(1), 47-79.
- Golembiewsky, R, Billingsley, K, & Yeager, S. (1976). Measuring change and persistence in human affairs: types of change generated by OD designs. *J Appl Behav Sci*, 12, 133-157.
- Gregorich, SE. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med Care*, 44(11), S78-S94.
- Griffiths, C, Motlib, J, Azad, A, Ramsay, J, Eldridge, S, Feder, G, et al. (2005). Randomised controlled trial of a lay-led self-management programme for Bangladeshi patients with chronic disease. *British Journal of General Practice*, 55, 831-837.

- Guterman, S. (2005). U.S. and German case studies in chronic care management: an overview. *Health Care Financing Review*, 27(1), 1-8.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.
- Guyatt, GH, Osoba, D, Wu, AW, Wyrwich, KW, & Norman, GR. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*, 77(4), 371-383.
- Haas, M, Group, E, Muench, J, Kraemer, D, Brummel-Smith, K, Sharma, R, et al. (2005). Chronic disease self-management program for low back pain in the elderly. *J Manipulative Physiol Ther*, 28(4), 228-237.
- Hair, JF, Black, WC, Babin, BJ, Anderson, RE, & Tatham, RL. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Education, Inc.
- Harman, HH. (1976). *Modern factor analysis* (3rd, revised ed.). Chicago: The University of Chicago Press.
- Hays, RD, Hayashi, T, & Stewart, AL. (1989). A five-item measure of socially desirable response set. *Educ Psycho Measure*, 49, 629-636.
- Hedges, LV, & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Heuts, PHTG, de Bie, R, Drieteelaar, M, Aretz, K, Hopman-Rock, M, Bastiaenen, CHG, et al. (2005). Self-management in osteoarthritis of hip or knee: a randomized clinical trial in a primary healthcare setting. *J Rheumatol*, 32(3), 543-549.
- Hill, LG, & Betz, DL. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation*, 26(4), 501-517.
- Hipp, JR, & Bollen, KA. (2003). Model fit in Structural Equation Models with censored, ordinal, and dichotomous variables: testing vanishing tetrads. *Sociological Methodology*, 33, 267-305.
- Hofstee, WKB, Ten Berge, JMF, & Hendriks, AAJ. (1998). How to score questionnaires. *Personality and Individual Differences*, 25, 897-909.
- Holbrook, AL, Green, MC, & Krosnick, JA. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1), 79-125.
- Holmbeck, GN. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: examples from the child-clinical and pediatric psychology literature. *J Consult Clin Psychol*, 65(4), 599-610.
- Hopman-Rock, M, & Westhoff, MH. (2000). The effects of a health educational and exercise program for older adults with osteoarthritis of the hip or knee. *J Rheumatol*, 27(8), 1947-1954.
- Howard, GS, & Dailey, PR. (1979). Response-shift bias: a source of contamination of self-report measures. *J Appl Psychol*, 64(2), 144-150.
- Howard, GS, Dailey, PR, & Gulanick, NA. (1979). The feasibility of informed pretests in attenuating response-shift bias. *Applied Psy Measure*, 3(4), 481-494.

- Howard, GS, Millham, J, Slaten, S, & O'Donnell, L. (1981). Influence of subject response style effects on retrospective measures. *Applied Psy Measure*, 5(1), 89-100.
- Howard, GS, Ralph, KM, Gulanick, NA, Maxwell, SE, Nance, SW, & Gerber, SK. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psy Measure*, 3, 1-23.
- Howard, GS, Schmeck, RR, & Bray, JH. (1979). Internal invalidity in studies employing self-report instruments: a suggested remedy. *Journal of Educational Measurement*, 16(2), 129-135.
- Hoyle, RH, & Kenny, DA. (1999). Sample size, reliability, and tests of statistical mediation. In R Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195-222). Thousand Oaks, CA: Sage.
- Hu, LT, & Bentler, PM. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55.
- Hu, L-T, & Bentler, PM. (1995). Evaluating model fit. In R Hoyle (Ed.), *Structural Equation Modeling: concepts, issues, and applications* (pp. 76-99). Thousand Oaks: Sage Publications.
- Ismail, K, Winkley, K, & Rabe-Hesketh, S. (2004). Systematic review and meta-analysis of randomised controlled trials of psychological interventions to improve glycaemic control in patients with type 2 diabetes. *Lancet*, 363(9421), 1589-1597.
- Jackson, DN, & Messick, SJ. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educ Psycho Measure*, 21, 771-790.
- Jaeschke, R, Singer, J, & Guyatt, GH. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*, 10, 407-415.
- Jordan, JE, & Osborne, RH. (2007). Chronic disease self-management education programs: challenges ahead. *MJA*, 186(2), 84-87.
- Jöreskog, KG. (1969). A general approach to confirmatory factor analysis. *Psychometrika*, 34, 183-202.
- Jöreskog, KG. (1979). Basic ideas of factor and component analysis. In K Jöreskog & D Sörbom (Eds.), *Advances in factor analysis and structural equation models* (pp. 5-20). Cambridge, Mass: Abt Books.
- Jöreskog, KG. (1990). New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24(4), 387-404.
- Jöreskog, KG. (1993). Testing Structural Equation Models. In K Bollen & J Long (Eds.), *Testing Structural Equation Models*. Newbury Park, London, New Delhi: Sage Publications.
- Jöreskog, KG. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59, 381-389.
- Jöreskog, KG. (2002-2005). Structural Equation Modeling with ordinal variables using LISREL. Retrieved November 28, 2006, from <http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf>.

- Jöreskog, KG, & Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research*, 36(3), 347-387.
- Jöreskog, KG, & Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS command language*. Lincolnwood, IL: Scientific Software International, Inc.
- Jöreskog, KG, & Sörbom, D. (1996-2001). *LISREL 8: User's reference guide* (2nd ed.). Lincolnwood, IL: Scientific Software International, Inc.
- Jöreskog, KG, & Sörbom, D. (1996-2002). *PRELIS 2: User's reference guide*. Lincolnwood, IL: Scientific Software International, Inc.
- Jöreskog, KG, & Yang, F. (1996). Nonlinear structural equation models: the Kenny-Judd Model with interaction effects. In G Marcoulides & R Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques* (pp. 57-88). Mahwah, NJ: Lawrence Erlbaum Associates.
- Judd, CM, & Kenny, DA. (1981). Process analysis: estimating mediation in treatment evaluations. *Eval Rev*, 5, 602-619.
- Juniper, EF, Guyatt, GH, Willan, A, & Griffith, LE. (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol*, 47(1), 81-87.
- Kagawa-Singer, M. (1993). Redefining health: living with cancer. *Soc Sci Med*, 37(3), 295-304.
- Kaiser, HF. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31-36.
- Kalliopuska, M. (1992). Social desirability related to social class among adults. *Psychol Rep*, 70(3, Pt 1), 808-810.
- Kazis, ES, Anderson, JJ, & Meenan, RF. (1989). Effect sizes for interpreting changes in health status. *Med Care*, 27(Suppl), S178-S189.
- Keefe, FJ, Caldwell, DS, Williams, DA, Gil, KM, Mitchell, D, Robertson, C, et al. (1990). Pain coping skills training in the management of osteoarthritic knee pain: a comparative study. *Behav Ther*, 21, 49-62.
- Kenny, DA, Kashy, DA, & Bolger, N. (1998). Data analysis in social psychology. In D Gilbert, S Fiske & G Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 233-265). New York: McGraw-Hill.
- Kline, RB. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Knipschild, P. (1994). Systematic reviews: some examples. *BMJ*, 309, 719-721.
- Komarahadi, FL, Maurischat, C, Harter, M, & Bengel, J. (2004). Zusammenhänge von Depressivität und Ängstlichkeit mit sozialer Erwünschtheit bei chronischen Schmerzpatienten. *Der Schmerz*, 18(38-44).
- Kozma, A, & Stones, MJ. (1987). Social desirability in measures of subjective well-being: a systematic evaluation. *Journal of Gerontology*, 42(1), 56-59.
- Krosnick, JA. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl Cogn Psychol*, 5, 213-236.

- Krosnick, JA. (1999). Survey research. *Ann Rev Psychol*, 50, 537-567.
- Krosnick, JA, & Alwin, DF. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 201-219.
- Lam, TCM, & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation*, 24(1), 65-80.
- Lawrence, DM. (2005). Editorial - Chronic disease care: rearranging the deck chairs. *Ann Intern Med*, 6(143), 458-459.
- Leake, R, Friend, R, & Wadhwa, N. (1999). Improving adjustment to chronic illness through strategic self-presentation: an experimental study on a renal dialysis unit. *Health Psychol*, 18, 54-62.
- Lee, S-Y, Song, X-Y, Skevington, S, & Hao, Y-T. (2005). Application of structural equation models to quality of life. *Structural Equation Modeling*, 12(3), 435-453.
- LeFort, SM, Gray-Donald, K, Rowat, KM, & Jeans, ME. (1998). Randomized controlled trial of a community-based psychoeducation program for the self-management of chronic pain. *Pain*, 74(2-3), 297-306.
- Lehmann, D, McDonald, R, Cote, J, Heath, T, Irwin, J, & Ambler, T. (2001). Methodological and statistical concerns of the experimental behavioral researcher. *Journal of Consumer Psychology*, 10(1/2), 89-100.
- Leite, WL, & Beretvas, SN. (2005). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of desirable responding. *Educ Psycho Measure*, 65(1), 140-154.
- Lentz, TF. (1938). Acquiescence as a factor in the measurement of personality. *Psychol Bull*, 35, 659.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.
- Little, RJA, & Rubin, DB. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, 18, 292-326.
- Little, RJA, & Rubin, DB. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Little, TD, Card, NA, Bovaird, JA, Preacher, KJ, & Crandall, CS. (2007). Structural equation modeling of mediation and moderation with contextual factors. In T Little, J Bovaird & N Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 207-230). Mahwah, NJ: Lawrence Erlbaum Associates.
- Loevinger, J. (1959). Theory and techniques of assessment. *Ann Rev Psychol*, 10, 287-316.
- Loftus, EF, Smith, KD, Klinger, MR, & Fiedler, J. (1991). Memory and mismemory for health events. In J Tanur (Ed.), *Questions about questions: inquiries into the cognitive bases of surveys*. New York: Russell Sage.
- Loo, R, & Loewen, P. (2004). Confirmatory factor analyses of scores from full and short versions of the Marlowe-Crowne Social Desirability scale. *Journal of Applied Social Psychology*, 34(11), 2343-2352.

- Loo, R, & Thorpe, K. (2000). Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne Social Desirability Scale. *The Journal of Social Psychology*, 140(5), 628-635.
- Lorig, K. (2001). *Patient education: a practical approach* (3rd ed.). Thousand Oaks, CA: Sage Publications.
- Lorig, K, Feigenbaum, P, Regan, C, Ung, E, Chastain, RL, & Holman, HR. (1986). A comparison of lay-taught and professional-taught arthritis self-management courses. *J Rheumatol*, 13(4), 763-767.
- Lorig, K, González, VM, & Ritter, P. (1999). Community-based Spanish language arthritis education program: a randomized trial. *Med Care*, 37(9), 957-963.
- Lorig, K, & Holman, H. (1993). Arthritis self-management studies: a twelve-year review. *Health Educ Q*, 20(1), 17-28.
- Lorig, K, Lubeck, D, Kraines, RG, Seleznick, M, & Holman, HR. (1985). Outcomes of self-help education for patients with arthritis. *Arthritis Rheum*, 28(6), 680-685.
- Lorig, K, Seleznick, M, Lubeck, D, Ung, E, Chastain, RL, & Holman, HR. (1989). The beneficial outcomes of the arthritis self-management course are not adequately explained by behavior change. *Arthritis Rheum*, 32(1), 91-95.
- Lorig, KR. (1982). Arthritis self-management: a patient education program. *Rehabilitation Nursing*, 16-20.
- Lorig, KR. (2003). Editorial - Self-management education: more than a nice extra. *Med Care*, 41(6), 699-701.
- Lorig, KR, González, VM, & Laurent, DD. (1999). *The Chronic Disease Self-Management Program: leaders manual*. Stanford University, Palo Alto.
- Lorig, KR, Mazonson, PD, & Holman, HR. (1993). Evidence suggesting that health education for self-management in patients with chronic arthritis has sustained health benefits while reducing health care costs. *Arthritis Rheum*, 36(4), 439-446.
- Lorig, KR, Ritter, P, Stewart, AL, Sobel, DS, Brown, BW, Bandura, A, et al. (2001). Chronic Disease Self-Management Program - 2-year health status and health care utilization outcomes. *Med Care*, 39(11), 1217-1223.
- Lorig, KR, Ritter, PL, & González, VM. (2003). Hispanic chronic disease self-management - a randomized community-based outcome trial. *Nurs Res*, 52(6), 361-369.
- Lorig, KR, Sobel, DS, Ritter, PL, Laurent, D, & Hobbs, M. (2001). Effect of a self-management program on patients with chronic disease. *Eff Clin Pract*, 4(6), 256-262.
- Lorig, KR, Sobel, DS, Stewart, AL, Brown, BW, Bandura, A, Ritter, P, et al. (1999). Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization. *Med Care*, 37(1), 5-14.
- Lydick, E, & Epstein, RS. (1993). Interpretation of quality of life changes. *Qual Life Res*, 2, 221-226.
- MacCallum, RC, Browne, MW, & Sugawara, HM. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130-149.

- MacCallum, RC, Wegener, DT, Uchino, BN, & Fabrigar, LR. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychol Bull*, 114, 185-199.
- MacKinnon, DP, Krull, JL, & Lockwood, CM. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4), 173-181.
- MacKinnon, DP, Warsi, G, & Dwyer, JH. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, 30, 41-62.
- Marcoulides, GA, & Schumacker, RE. (1996a). *Advanced Structural Equation Modeling - issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Marcoulides, GA, & Schumacker, RE. (1996b). Introduction. In G Marcoulides & R Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Marsh, HW, Balla, JR, & Hau, K. (1996). An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. In G Marcoulides & R Schumacker (Eds.), *Advanced structural equation modeling: issues and techniques* (pp. 315-353). Mahwah, NJ: Lawrence Erlbaum Associates.
- Marsh, HW, Balla, JR, & McDonald, RP. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychol Bull*, 103, 391-410.
- Marsh, HW, Hau, K-T, & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers of overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320-341.
- McCrae, RR, & Costa, PT. (1983). Social desirability scales: more substance than style. *J Consult Clin Psychol*, 51(6), 882-888.
- McDonald, RP. (1999). *Test theory: a unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- McDonald, RP. (2005). Semiconfirmatory Factor Analysis: the example of anxiety and depression. *Structural Equation Modeling*, 12(1), 163-172.
- McGaw, B, & Jöreskog, KG. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical*, 24, 154-168.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525-543.
- Millham, J. (1974). Two components of need for approval score and their relationship to cheating following success and failure. *Journal of Research in Personality*, 8(4), 378-392.
- Millham, J, & Kellogg, RW. (1980). Need for social approval: Impression management or self-deception? *Journal of Research in Personality*, 14(4), 445-457.
- Millsap, RE, & Hartog, SB. (1988). Alpha, beta, and gamma change in evaluation research: a structural equation approach. *J Appl Psychol*, 73, 574-584.
- Millsap, RE, & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479-515.



- Moorman, RH, & Podsakoff, PM. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology*, 65, 131-149.
- Morf, ME, & Jackson, DN. (1972). An analysis of two response styles: true responding and item endorsement. *Educ Psycho Measure*, 32(2), 329-353.
- Mulaik, SA. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
- Mullen, PD, Laville, EA, Biddle, AK, & Lorig, K. (1987). Efficacy of psychoeducational interventions on pain, depression, and disability in people with arthritis: a meta-analysis. *J Rheumatol*, 14(Suppl 15), 33-39.
- Murray, CJL, & Lopez, AD. (1996). *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Boston, MA: Harvard School of Public Health.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- Muthén, B, & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, 46, 407-419.
- Muthén, B, Kaplan, D, & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3), 431-462.
- Muthén, LK, & Muthén, BO. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599-620.
- Nederhof, A. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology*, 15, 263-280.
- Newman, S, Steed, L, & Mulligan, K. (2004). Self-management interventions for chronic illness. *Lancet*, 364(9444), 1523-1537.
- Nolte, S, Elsworth, GR, Sinclair, AJ, & Osborne, RH. (2007). The extent and breadth of benefits from participating in chronic disease self-management courses: a national patient-reported outcomes survey. *Patient Educ Couns*, 65(3), 351-360.
- Norman, G. (2003). Hi! How are you? Response shift, implicit theories and differing epistemologies. *Qual Life Res*, 12, 239-249.
- Norman, GR. (2005). The relation between the minimally important difference and patient benefit. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 2, 69-73.
- Norman, GR, & Streiner, DL. (2000). *Biostatistics - The bare essentials* (2nd ed.). Hamilton, Ontario: BC Decker.
- Nunnally, JC, & Bernstein, IH. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- O'Boyle, CA, McGee, HM, & Browne, JP. (2000). Measuring response shift using the Schedule for Evaluation of Individual Quality of Life. In C Schwartz & M Sprangers (Eds.), *Adaptation to changing health: response shift in quality-of-life research*. Washington, DC: American Psychological Association.

- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443-460.
- Oort, FJ. (2005a). Towards a formal definition of response shift (in reply to G.W. Donaldson). *Qual Life Res*, 14(10), 2353-2355.
- Oort, FJ. (2005b). Using structural equation modeling to detect response shift and true change. *Qual Life Res*, 14(3), 587-598.
- Oort, FJ, Visser, MRM, & Sprangers, M. (2003). Incorporating the Then Test into the structural equation modeling (SEM) approach to response shift detection. *Qual Life Res*, 12(7), 784.
- Oort, FJ, Visser, MRM, & Sprangers, MAG. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Qual Life Res*, 14(3), 599-609.
- Osborne, R, & Whitfield, K. (2004). NAMCIG PROJECT # 9, Final report, development and implementation of a national self-management quality and monitoring system.
- Osborne, RH, Elsworth, GE, & Whitfield, K. (2007). The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. *Patient Educ Couns*, 66, 192-201.
- Osborne, RH, Hawkins, M, & Sprangers, MAG. (2006). Change of perspective: a measurable and desired outcome of chronic disease self-management intervention programs that violates the premise of preintervention/postintervention assessment. *Arthritis Care Res*, 55(3), 458-465.
- Osborne, RH, Spinks, JM, & Wicks, IP. (2004). Patient education and self-management programs in arthritis. *MJA*, 180, S23-S26.
- Osteoporosis Victoria. (2001). *The Osteoporosis Prevention and Self Management Course - leaders manual*. Elsternwick: Osteoporosis Victoria, A division of Arthritis Victoria.
- Paulhus, DL. (1984). Two-component models of socially desirable responding. *J Pers Soc Psychol*, 46, 598-609.
- Paulhus, DL. (1991). Measurement and control of response bias. In J Robinson, P Shaver & L Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). New York: Academic Press.
- Pearson, RW, Ross, M, & Dawes, RM. (1992). Personal recall and the limits of retrospective questions in surveys. In J Tanur (Ed.), *Questions about questions: inquiries into the cognitive bases of surveys*. New York: Russell Sage.
- Pittman, PM, Arnold, SB, & Schlette, S. (2005). Care management in Germany and the U.S.: an expanded laboratory. *Health Care Financing Review*, 27(1), 9-18.
- Podsakoff, PM, MacKenzie, SB, Lee, J-Y, & Podsakoff, NP. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J Appl Psychol*, 88(5), 879-903.
- Ramanaiah, NV, Schill, T, & Leung, LS. (1977). A test of the hypothesis about the two-dimensional nature of the Marlowe-Crowne social desirability scale. *Journal of Research in Personality*, 11(2), 251-259.

- Randolph, WA, & Elloy, DF. (1989). How can OD consultants and researchers assess gamma change? A comparison of two analytical procedures. *J Manage*, 15, 633-648.
- Ray, JJ. (1988). Lie scales and the elderly. *Personality and Individual Differences*, 9(2), 417-418.
- Redman, BK. (2001). *The practice of patient education* (9th ed.). St. Louis, Missouri: Mosby.
- Rees, J, Waldron, D, O'Boyle, CA, & MacDonagh, RP. (2002). Response shift in individualized quality of life in patients with advanced prostate cancer. *Clinical Therapeutics*, 24(Suppl 2), 33-34.
- Reynolds, WM. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 38(1), 119-125.
- Rigdon, E. (1997). Not positive definite matrices - causes and cures. Retrieved March 22, 2006, from <http://www2.gsu.edu/~mkteer/npdmatri.html>.
- Robinette, RL. (1991). The relationship between the Marlowe-Crowne Form C and the validity scales of the MMPI. *Journal of Clinical Psychology*, 47(3), 396-399.
- Rorer, LG. (1965). The great response-style myth. *Psychol Bull*, 63, 129-156.
- Rorer, LG, & Goldberg, LR. (1965). Acquiescence and the vanishing variance component. *J Appl Psychol*, 49, 422-430.
- Rosnow, RL, & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: general procedures for research consumers. *Psychological Methods*, 1, 331-340.
- Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychol Rev*, 96, 341-357.
- Ross, M, & MacDonald, TK. (1997). How can we be sure? Using truth criteria to validate memories. In M Myslobodsky (Ed.), *Mythomanias* (pp. 181-202). Hillsdale, NJ: Erlbaum.
- Rubin, DB. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rust, J, & Golombok, S. (1999). *Modern psychometrics - the science of psychological assessment* (2nd ed.). London: Routledge.
- Sachs, L. (2002). *Angewandte Statistik* (10th ed.). Berlin, Heidelberg, New York: Springer-Verlag.
- Sackett, DL. (1994). The Cochrane Collaboration. *ACP Journal Club*, 120(Suppl 3), A-11.
- Samsa, G, Edelman, D, Rothman, ML, Williams, GR, Lipscomb, J, & Matchar, D. (1999). Determining clinically important differences in health status measures: a general approach with illustration to the Health Utility Index Mark II. *Pharmacoeconomics*, 15, 141-155.
- Satorra, A, & Bentler, PM. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *1988 Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308-313.

- Satorra, A, & Bentler, PM. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A von Eye & C Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage Publications.
- Satorra, A, & Bentler, PM. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.
- Schafer, JL, & Graham, JW. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147-177.
- Schmitt, N. (1982). The use of analysis of covariance structure to assess beta and gamma change. *Multivariate Behavioral Research*, 17(3), 343-358.
- Schmitt, N, Pulakos, E, & Lieblein, A. (1984). Comparison of three techniques to assess group-level beta and gamma change. *Applied Psy Measure*, 8, 249-260.
- Scholten, C, Brodowicz, T, Graninger, W, Gardavsky, I, Pils, K, Pesau, B, et al. (1999). Persistent functional and social benefit 5 years after a multidisciplinary arthritis training program. *Arch Phys Med Rehabil*, 80, 1282-1287.
- Schumacker, RE, & Lomax, RG. (2004). *A beginner's guide to Structural Equation Modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Schwartz, CE, Bode, R, Repucci, N, Becker, J, Sprangers, MA, & Fayers, PM. (2006). The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Qual Life Res*, 15(9), 1533-1550.
- Schwartz, CE, & Rapkin, BD. (2004). Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. Retrieved April 15, 2006, from <http://www.hqlo.com/content/2/1/16>.
- Schwartz, CE, & Sprangers, MAG. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Soc Sci Med*, 48(11), 1531-1548.
- Schwarz, N, Groves, RM, & Schuman, H. (1998). Survey methods. In D Gilbert, S Fiske & G Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 143-179). New York: McGraw-Hill.
- Schwarz, N, & Oyserman, D. (2001). Asking questions about behavior: cognition, communication and questionnaire construction. *American Journal of Evaluation*, 22(2), 127-160.
- Schwarz, N, & Strack, F. (1985). Cognitive and affective processes in judgements of subjective well-being: a preliminary model. In H Brandstatter & E Kirchler (Eds.), *Economic psychology* (pp. 439-447). Linz, Austria: R. Tauner.
- Shrout, PE, & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, 7(4), 422-445.
- Skeff, KM, Stratos, GA, & Bergen, MR. (1992). Evaluation of a medical faculty development program: a comparison of traditional pre/post and retrospective pre/post self-assessment ratings. *Eval Health Prof*, 15, 351-366.
- Sobel, ME. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S Leinhardt (Ed.), *Sociological methodology* (pp. 290-312). San Francisco: Jossey-Bass.

- Solomon, DH, & Lee, TH. (2003). Patient self-management in arthritis? Yes, more research! *J Rheumatol*, 30(6), 1133-1134.
- Solomon, DH, Warsi, A, Brown-Stevenson, T, Farrell, M, Gauthier, S, Mikels, D, et al. (2002). Does self-management education benefit all populations with arthritis? A randomized controlled trial in a primary care physician network. *J Rheumatol*, 29(2), 362-368.
- Sommer, R. (1968). Hawthorne dogma. *Psychol Bull*, 70, 592-595.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Br J Math Stat Psychol*, 27, 229-239.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201-292.
- Sprangers, M. (1989). Response-shift bias in program evaluation. *Imp Assess Bull*, 7, 153-166.
- Sprangers, M, & Hoogstraten, J. (1988). On delay and reassessment of retrospective preratings. *J Exper Educ*, 56, 148-153.
- Sprangers, M, & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *J Appl Psychol*, 74, 265-272.
- Sprangers, MA, Van Dam, FS, Broersen, J, Lodder, L, Wever, L, Visser, MR, et al. (1999). Revealing response shift in longitudinal research on fatigue: the use of the thentest approach. *Acta Oncol*, 38(6), 709-718.
- Sprangers, MAG, & Schwartz, CE. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med*, 48(11), 1507-1515.
- Statistisches Bundesamt. (2000, 12.10.00). Statistisches Jahrbuch 2000 für das In- und Ausland erschienen: Leben in Deutschland, unser Platz in Europa und der Welt. Retrieved May 31, 2004, from <http://www.destatis.de/presse/deutsch/pm2000/p3640221.htm>.
- Steenkamp, J-BEM, & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78-90.
- Steiger, JH. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.
- Stieglitz, RD. (1990). Validationsstudien zum retrospektiven Vortest in der Therapieforschung. *Zeitschrift für Klinische Psychologie*, 19, S144-150.
- Strahan, R, & Gerbasi, KC. (1972). Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 28, 191-193.
- Strong, K, Mathers, C, Leeder, S, & Beaglehole, R. (2005). Chronic diseases 1 - Preventing chronic diseases: how many lives can we save? *Lancet*, 366(1578-1582).
- Stucki, G, Daltroy, L, Katz, JN, Johannesson, M, & Liang, MH. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol*, 49(7), 711-717.

- Swerissen, H, Belfrage, J, Weeks, A, Jordan, L, Walker, C, Furler, J, et al. (2006). A randomised control trial of a self-management program for people with a chronic illness from Vietnamese, Chinese, Italian and Greek backgrounds. *Patient Educ Couns*, 64(1-3), 360-368.
- Taal, E, Riemsma, RP, Brus, HL, Seydel, ER, Rasker, JJ, & Wiegman, O. (1993). Group education for patients with rheumatoid arthritis. *Patient Educ Couns*, 20(2-3), 177-187.
- Taal, E, Riemsma, RP, Kirwan, JR, & Rasker, JJ. (2004). What are the real effects of arthritis self-management education programs on pain and disability? Comment on the article by Warsi et al. *Arthritis Rheum*, 50(3), 1012-1013.
- Tanaka, JS. (1987). "How big is big enough?": sample size and goodness of fit in Structural Equation Models with latent variables. *Child Development*, 58(1), 134-146.
- Tanaka, JS. (1993). Multifaceted conceptions of fit in Structural Equation Models. In K Bollen & J Long (Eds.), *Testing Structural Equation Models*. Newbury Park, London, New Delhi: Sage Publications.
- Terborg, JR, & Davis, GA. (1982). Evaluation of a new method of assessing change to planned job redesign as applied to Hackman and Oldham's job characteristic model. *Organizational Behavior and Human Performance*, 29, 112-128.
- Terborg, JR, Howard, GS, & Maxwell, SE. (1980). Evaluating planned organizational change: a method for assessing alpha, beta, and gamma change. *Acad Manage Rev*, 5, 109-121.
- The Cochrane Collaboration. (2007). Cochrane Reviews. Retrieved January 3, 2007, from <http://www.cochrane.org/reviews/>.
- Thompson, RC, & Hunt, JG. (1996). Inside the black box of alpha, beta, and gamma change: using a cognitive-processing model to assess attitude structure. *Acad Manage Rev*, 21, 655-690.
- Tourangeau, R, & Rasinski, KA. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychol Bull*, 103, 299-314.
- Trochim, WM, & Linton, R. (1986). Conceptualization for evaluation and planning. *Evaluation and Program Planning*, 9, 289-308.
- Trochim, WMK, Milstein, B, Wood, BJ, Jackson, S, & Pressler, V. (2004). Setting objectives for community and systems change: an application of concept mapping for planning a statewide health improvement initiative. *Health Promot Pract*, 5(1), 8-19.
- van de Vliert, E, Huismans, SE, & Stok, JJJ. (1985). The criterion approach to unraveling beta and alpha change. *Acad Manage Rev*, 10, 269-274.
- Visser, AP, Breemhaar, B, & Kleijnen, JGVM. (1989). Social desirability and program evaluation in health care. *Imp Assess Bull*, 7, 99-112.
- Visser, MRM, Oort, FJ, & Sprangers, MAG. (2005). Methods to detect response shift in quality of life data: a convergent validity study. *Qual Life Res*, 14, 629-639.
- Vlaeyen, JW, Teeken-Gruben, NJ, Goossens, ME, Rutten-van Mülken, MPMH, Pelt, RAGB, van Eek, H, et al. (1996). Cognitive-educational treatment of fibromyalgia: a randomized clinical trial, I: clinical effects. *J Rheumatol*, 23, 1237-1245.

- Von Korff, M, Glasgow, RE, & Sharpe, M. (2002). ABC of psychological medicine: organising care for chronic illness. *BMJ*, 325(7355), 92-94.
- Von Korff, M, Moore, JE, Lorig, K, Cherkin, DC, Saunders, K, González, VM, et al. (1998). A randomized trial of a lay person-led self-management group intervention for back pain patients in primary care. *Spine*, 23(23), 2608-2615.
- Wagner, EH, Davis, C, Schaefer, J, Von Korff, M, & Austin, B. (1999). A survey of leading chronic disease management programs: are they consistent with the literature? *Managed Care Quarterly*, 7(3), 56-66.
- Walter, SD. (2001). Number needed to treat (NNT): Estimation of a measure of clinical benefit. *Statist Med*, 20(24), 3947-3962.
- Warsi, A, LaValley, MP, Wang, PS, Avorn, J, & Solomon, DH. (2003). Arthritis self-management education programs: a meta-analysis of the effect on pain and disability. *Arthritis Rheum*, 48(8), 2207-2213.
- Warsi, A, Wang, PS, LaValley, MP, Avorn, J, & Solomon, DH. (2004). Self-management education programs in chronic disease: a systematic review and methodological critique of the literature. *Arch Intern Med*, 164(15), 1641-1649.
- Webb, EJ, Campbell, DT, Schwartz, RD, & Sechrest, L. (1966). *Unobtrusive measures: nonreactive measures in the social sciences*. Chicago: Rand McNally.
- Weingarten, SR, Henning, JM, Badamgarav, E, Knight, K, Hasselblad, V, Gano Jr, A, et al. (2002). Interventions used in disease management programmes for patients with chronic illness - which ones work? Meta-analysis of published reports. *BMJ*, 325(7370), 925-932.
- West, SG, Finch, JF, & Curran, PJ. (1995). Structural equation models with nonnormal variables - problems and remedies. In R Hoyle (Ed.), *Structural Equation Modeling: concepts, issues, and applications* (pp. 56-75). Thousand Oaks: Sage Publications.
- Wilson, IB, & Cleary, PD. (1995). Linking clinical variables with health-related quality of life - a conceptual model of patient outcomes. *JAMA*, 273(1), 59-65.
- World Health Organization. (2002). *Innovative care for chronic conditions: building blocks for action*. Geneva: World Health Organization.
- World Health Organization. (2003). *Adherence to long-term therapies: evidence for action*. Geneva: World Health Organization.
- World Health Organization. (2006). Facts related to chronic diseases. Retrieved November 27, 2006, from <http://www.who.int/dietphysicalactivity/publications/facts/chronic/en/>.
- World Health Organization, & Public Health Agency of Canada. (2005). Preventing chronic diseases: a vital investment: WHO global report. Retrieved April 15, 2006, from [http://www.who.int/chp/chronic\\_disease\\_report/en/](http://www.who.int/chp/chronic_disease_report/en/).
- Worth, H. (2002). Effects of patient education in asthma and COPD - what is provable? *Medizinische Klinik*, 97(2, Suppl), 20-24.
- Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K Bollen & J Long (Eds.), *Testing Structural Equation Models*. Newbury Park, London, New Delhi: Sage Publications.

- Wyrwich, KW, Bullinger, M, Aaronson, N, Hays, RD, Patrick, DL, Symonds, T, et al. (2005). Estimating clinically significant differences in quality of life outcomes. *Qual Life Res*, 14, 285-295.
- Yach, D, Hawkes, C, Gould, CL, & Hofman, KJ. (2004). The global burden of chronic diseases - overcoming impediments to prevention and control. *JAMA*, 291, 2616-2622.
- Yuan, K-H, & Bentler, PM. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30, 165-200.
- Yuan, K-H, & Bentler, PM. (2004). On chi-square difference and z-tests in mean and covariance structure analysis when the base model is misspecified. *Educ Psycho Measure*, 64, 737-757.
- Zmud, RW, & Armenakis, AA. (1978). Understanding the measurement of change. *Acad Manage Rev*, 3, 661-669.
- Zook, A, & Sipps, GJ. (1985). Cross-validation of a short form of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 41, 236-238.



## Appendices

**Appendix 1** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>	
				lower	upper	IG	CG
Lorig <i>et al.</i> , 1986, ASMP, led by health professionals (HP)	1) IG, led by HP (n=29)	<u>4 months – HP</u>				<u>n=29</u>	<u>n=29</u>
		Arthritis exercise	0.86	0.32	1.40	0.95	0.11
	2) CG (n=29)	Disability	0.00	-0.51	0.51	0.14	0.14
		Knowledge	0.62	0.09	1.14	1.00	0.42
		Pain	-0.22	-0.73	0.30	0.14	0.38
		Relaxation	0.24	-0.27	0.76	0.03	-0.22
		Visits to physician	-0.34	-0.86	0.18	-0.11	0.21
Lorig <i>et al.</i> , 1986, ASMP, lay-led	1) IG, lay-led (n=27)	<u>4 months – peer</u>				<u>n=27</u>	<u>n=29</u>
		Arthritis exercise	1.03	0.47	1.59	1.11	0.11
	2) CG (n=29)	Disability	0.13	-0.39	0.66	0.25	0.14
		Knowledge	-0.05	-0.58	0.47	0.37	0.42
		Pain	-0.28	-0.81	0.24	0.07	0.38
		Relaxation	0.84	0.29	1.38	0.61	-0.22
		Visits to physician	-0.10	-0.62	0.42	0.07	0.21
Lorig <i>et al.</i> , 1989, ASMP	1) IG (n=501)	<u>4 months</u>				<u>n=501</u>	<u>n=206</u>
		Arthritis exercise	0.54	0.38	0.71	0.48	-0.05
	2) CG (n=206)	Depression	0.11	-0.06	0.27	0.10	-0.01
		Disability	0.05	-0.11	0.21	0.02	-0.03
		Knowledge	0.78	0.62	0.95	0.83	0.04
		Pain	0.15	-0.01	0.31	0.23	0.07
		Relaxation	0.26	0.10	0.43	0.44	-0.02
		Self-management activities	0.57	0.40	0.73	0.49	-0.06
Goepfinger <i>et al.</i> , 1989 <sup>3</sup> , BUOA	1) IG (n=100)	<u>4 months</u>				<u>n=100</u>	<u>n=153</u>
		Depression	0.12	-0.13	0.37	0.12	0.00
	2) CG (n=153)	Disability	0.02	-0.23	0.27	0.00	-0.02
		Helplessness	0.45	0.19	0.70	0.60	0.16
		Pain	0.08	-0.17	0.34	0.21	0.13
		Knowledge	0.85	0.59	1.11	0.95	0.10
		Self-care behaviours	0.47	0.21	0.72	0.38	-0.09
Keefe <i>et al.</i> , 1990, ASMP	1) IG (n=36)	<u>Post-treatment</u>				<u>n=36</u>	<u>n=31</u>
		Disability	-0.18	-0.66	0.30	-0.20	-0.03
	2) CG (n=31)	Pain	0.20	-0.28	0.68	-0.01	0.20
		Psychological disability	0.54	0.05	1.03	0.38	-0.16
Taal <i>et al.</i> , 1993, group education for people with RA	1) IG (n=27)	<u>Post-treatment</u>				<u>n=27</u>	<u>n=30</u>
		Anxiety	0.20	-0.32	0.72	0.31	0.13
	2) CG (n=30)	Arthritis impact	0.02	-0.50	0.54	0.09	0.07
		Depression	0.05	-0.47	0.57	0.35	0.36
		Disability	0.42	-0.10	0.95	0.02	-0.43
		Pain	0.18	-0.34	0.70	0.26	0.07

**Appendix 1 (continued)** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>	
				lower	upper	IG	CG
Burckhardt <i>et al.</i> , 1994, Fibromyalgia, six sessions, each 1.5 hours; curriculum included coping strategies, problem-solving, and relaxation	1) IG (n=28)	<u>6 weeks</u>				<u>n=28</u>	<u>n=30</u>
		6-min walking test	0.19	-0.32	0.71	0.05	-0.15
	2) CG (n=30)	Anxiety	0.10	-0.41	0.62	0.11	0.00
		Chair test	0.21	-0.31	0.72	0.10	-0.11
		Depression	0.00	-0.52	0.52	0.12	0.11
		Fatigue	0.14	-0.37	0.66	0.19	0.08
		Fibromyalgia attitudes	0.23	-0.29	0.75	0.30	0.04
		Flexibility	-0.34	-0.86	0.18	-0.41	-0.06
		Job difficulty	0.03	-0.48	0.55	0.26	0.26
		Morning tiredness	-0.12	-0.63	0.40	0.10	0.32
		Overall well-being	0.51	-0.02	1.03	0.70	0.07
		Pain	-0.14	-0.65	0.38	0.07	0.21
		Pain in tender points	0.29	-0.23	0.81	0.31	0.04
		Physical function	0.23	-0.29	0.74	0.26	0.00
		Quality of life	0.59	0.06	1.11	0.18	-0.41
		Self-efficacy, function	0.18	-0.33	0.70	0.13	-0.05
		Self-efficacy, other	0.32	-0.20	0.84	0.44	0.07
		Self-efficacy, pain	0.08	-0.43	0.60	0.13	0.04
		Stiffness	-0.24	-0.76	0.27	-0.15	0.08
Tender points	0.30	-0.22	0.81	0.24	-0.08		
LeFort <i>et al.</i> , 1998, modified ASMP for chronic pain	1) IG (n=52)	<u>6 weeks</u>				<u>n=52</u>	<u>n=50</u>
		Bodily pain	0.57	0.17	0.97	0.47	-0.12
	2) CG (n=50)	Dependency	0.44	0.04	0.83	0.27	-0.18
		Depression	0.22	-0.17	0.61	0.17	-0.04
		Disability	0.30	-0.09	0.69	0.26	-0.03
		General health	0.16	-0.23	0.55	0.17	0.00
		Life satisfaction	0.52	0.13	0.92	0.38	-0.15
		Mental health	0.25	-0.14	0.64	0.39	0.14
		Pain problem severity	0.55	0.15	0.94	0.63	0.10
		Pain quality	0.43	0.04	0.83	0.21	-0.23
		Physical functioning	0.14	-0.25	0.52	0.12	-0.01
		Resourcefulness	0.50	0.10	0.89	0.31	-0.20
		Role behaviours	0.53	0.14	0.93	0.43	-0.12
		Role-emotional	0.17	-0.22	0.56	0.43	0.27
		Role-physical	0.70	0.30	1.10	0.69	-0.10
		Self-efficacy	0.72	0.32	1.12	0.64	-0.11
		Social functioning	0.30	-0.09	0.69	0.28	-0.02
		Uncertainty	0.18	-0.21	0.57	0.17	-0.01
		Vitality	0.74	0.34	1.14	0.58	-0.17
Von Korff <i>et al.</i> , 1998, back pain, four sessions, modelled after Stanford program	1) IG (n=129)	<u>3 months</u>				<u>n=124</u>	<u>n=121</u>
		Disability	0.15	-0.10	0.40	0.48	0.31
	2) CG (n=126)	Interference	0.08	-0.17	0.33	0.75	0.67
		Mental health	0.02	-0.23	0.27	0.25	0.21
		Pain	-0.07	-0.32	0.18	0.75	0.80
		Self-care orientation	0.21	-0.04	0.46	0.63	0.43
Worry	-0.01	-0.26	0.24	0.71	0.65		

**Appendix 1 (continued)** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>	
				lower	upper	IG	CG
Lorig <i>et al.</i> , 1999b, Spanish ASMP	1) IG (n=189)	<u>4 months</u>				<u>n=189</u>	<u>n=97</u>
		Aerobic exercise	0.08	-0.17	0.32	0.23	0.16
	2) CG (n=97)	Depression	0.06	-0.19	0.30	0.26	0.20
		Disability	0.14	-0.10	0.39	0.15	0.00
		General health	0.24	-0.01	0.48	0.41	0.12
		Medication use	-0.29	-0.53	-0.04	0.08	0.38
		Pain	0.37	0.12	0.61	0.35	-0.01
		Range of motion exercise	0.47	0.22	0.71	0.38	-0.10
		Self-efficacy	0.48	0.24	0.73	0.48	-0.02
Visits to physician	-0.08	-0.32	0.17	0.02	0.13		
Lorig <i>et al.</i> , 1999c, CDSMP	1) IG (n=561)	<u>6 months</u>				<u>n=561</u>	<u>n=391</u>
		Aerobic exercise	0.20	0.07	0.33	0.16	-0.02
	2) CG (n=391)	Cognitive symptom management	0.34	0.21	0.47	0.43	0.07
		Communication with physician	0.13	0.00	0.25	0.22	0.09
		Disability	0.08	-0.05	0.21	0.03	-0.05
		Energy/fatigue	0.11	-0.02	0.24	0.13	0.02
		Health distress	0.14	0.01	0.27	0.20	0.06
		Hospital stays	0.02	-0.11	0.15	0.10	0.05
		Nights in hospital	0.20	0.08	0.33	0.07	-0.14
		Pain	0.02	-0.11	0.15	0.12	0.09
		Psychological well-being	0.05	-0.08	0.18	0.10	0.04
		Self-rated health	0.12	-0.01	0.25	0.10	-0.02
		Shortness of breath	-0.03	-0.16	0.09	-0.02	0.02
		Social/role activities limitations	0.14	0.01	0.27	0.06	-0.07
		Stretching / strengthening	0.15	0.02	0.28	0.24	0.09
Visits to physician	0.04	-0.09	0.17	0.14	0.09		
Scholten <i>et al.</i> , 1999, Multi-disciplinary Arthritis Training Program	1) IG (n=38)	<u>Post-treatment</u>				<u>n=38</u>	<u>n=30</u>
		Coping	0.21	-0.27	0.69	0.25	0.03
	2) CG (n=30)	Depression	0.42	-0.07	0.90	0.30	-0.09
		Depression, Beck	0.86	0.36	1.36	0.84	-0.03
		Disability	1.42	0.88	1.95	1.28	0.00
Distraction	0.22	-0.26	0.70	0.18	-0.04		
Barlow <i>et al.</i> , 2000, ASMP	1) IG (n=311)	<u>4 months</u>				<u>n=234</u>	<u>n=189</u>
		Anxiety	0.14	-0.05	0.33	0.21	0.08
	2) CG (n=233)	Cognitive symptom management	0.41	0.21	0.60	0.46	0.04
		Communication with physician	0.19	0.00	0.38	0.23	0.03
		Depression	0.25	0.06	0.44	0.27	0.04
		Dietary habit	0.18	-0.01	0.38	0.09	-0.09
		Fatigue	0.18	-0.02	0.37	0.17	-0.02
		Negative Affect	0.05	-0.14	0.24	0.08	0.03
		Pain	0.03	-0.16	0.22	0.12	0.09

**Appendix 1 (continued)** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>		
				lower	upper	IG	CG	
Barlow <i>et al.</i> , 2000, ASMP (continued)		Physical functioning	-0.04	-0.23	0.15	-0.01	0.03	
		Positive Affect	0.21	0.02	0.41	0.28	0.07	
		Self-efficacy, other symptoms	0.31	0.12	0.50	0.43	0.13	
		Self-efficacy, pain	0.26	0.07	0.45	0.41	0.14	
					<u>n=86</u>	<u>n=78</u>		
		EuroQoL	0.10	-0.21	0.40	0.13	0.03	
		EuroQoL, VAS	-0.05	-0.35	0.26	0.02	0.07	
		Physician visits: other	-0.36	-0.67	-0.05	-0.17	0.19	
		Visits to physician to discuss arthritis	0.04	-0.27	0.35	0.15	0.08	
Hopman-Rock <i>et al.</i> , 2000, ASMP modified for OA	1) IG (n=60)	<u>Post-treatment</u>				<u>n=56</u>	<u>n=49</u>	
		20m walking test	0.16	-0.22	0.54	0.18	0.02	
	2) CG (n=60)	Knee extension, left	0.30	-0.08	0.69	0.15	-0.15	
		Knee extension, right	0.18	-0.21	0.56	0.08	-0.10	
		Knowledge	1.11	0.70	1.52	1.28	0.17	
		Mobility	0.22	-0.16	0.61	-0.09	-0.29	
		Pain	0.43	0.04	0.81	0.10	-0.34	
		Pain intolerance	0.08	-0.31	0.46	0.26	0.21	
		QoL	0.28	-0.11	0.66	0.24	-0.05	
		QoL, VAS	0.35	-0.04	0.73	0.02	-0.37	
		Self-efficacy	0.51	0.12	0.89	0.19	-0.34	
		Stair climbing down	-0.15	-0.53	0.24	0.16	0.18	
		Stair climbing up	-0.07	-0.45	0.31	0.22	0.19	
		Timed up-and-go	0.08	-0.30	0.46	0.14	0.05	
	Toe reaching left	0.18	-0.20	0.57	0.09	-0.09		
	Toe reaching right	0.00	-0.38	0.38	0.09	0.11		
Fu <i>et al.</i> , 2003, CDSMP modified for Chinese culture	1) IG (n=430)	<u>6 months</u>				<u>n=430</u>	<u>n=349</u>	
		Aerobic exercise	0.19	0.05	0.33	0.21	0.02	
	2) CG (n=349)	Cognitive symptom management	0.37	0.22	0.51	0.38	0.05	
		Communication with physician	-0.07	-0.21	0.07	0.04	0.10	
		Depression	0.13	-0.01	0.27	0.28	0.16	
		Disability	0.32	0.18	0.46	0.25	-0.05	
		Energy	-0.04	-0.18	0.11	0.03	0.07	
		Fatigue	0.20	0.06	0.34	0.16	-0.04	
		Health distress	0.26	0.12	0.40	0.26	0.01	
		Hospital stays	0.34	0.20	0.48	0.16	-0.19	
		Illness intrusiveness	0.07	-0.07	0.21	0.00	-0.07	
		Nights in hospital	0.15	0.01	0.30	0.07	-0.11	
		Number of ER visits	0.01	-0.13	0.16	0.05	0.06	
		Pain	0.19	0.05	0.33	0.02	-0.17	
		Self-efficacy, manage disease	0.31	0.17	0.45	0.23	-0.08	
		Self-efficacy, manage disease	0.29	0.15	0.43	0.10	-0.19	
	Self-rated health	0.42	0.28	0.56	0.47	0.05		
	Shortness of breath	0.18	0.04	0.32	-0.03	-0.22		

**Appendix 1 (continued)** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>	
				lower	upper	IG	CG
Fu <i>et al.</i> , 2003, CDSMP modified for Chinese culture (continued)		Social/role activities limitations	0.19	0.05	0.33	0.11	-0.08
		Stretching / strengthening	-0.02	-0.16	0.12	0.08	0.11
		Visits to physician	0.02	-0.12	0.16	0.11	0.11
Lorig <i>et al.</i> , 2003, Spanish CDSMP	1) IG (n=327) 2) CG (n=224)	<u>4 months</u>				<u>n=265</u>	<u>n=178</u>
		Communication with physician	0.34	0.14	0.53	0.49	0.16
		Exercise	0.29	0.10	0.48	0.52	0.30
		Fatigue	0.26	0.07	0.46	0.40	0.11
		Health distress	0.47	0.28	0.66	0.52	0.05
		Hospital stays	-0.03	-0.22	0.16	0.02	0.03
		Mental stress management	0.71	0.51	0.90	0.56	-0.15
		Number of ER visits	0.28	0.09	0.47	0.12	-0.16
		Pain	0.23	0.04	0.42	0.36	0.13
		Role function	0.25	0.06	0.45	0.35	0.10
		Self-efficacy	0.16	-0.03	0.35	0.44	0.27
		Self-rated health	0.48	0.29	0.68	0.52	0.04
		Visits to physician	0.18	-0.01	0.37	0.18	0.02
Boesen <i>et al.</i> , 2005, Psycho-educational intervention for cancer, six 2.5 hour sessions run over six weeks	1) IG (n=112) 2) CG (n=129)	<u>6 months</u>				<u>n=112</u>	<u>n=129</u>
		Anger	0.05	-0.21	0.30	0.20	0.11
		Anxiety	0.02	-0.23	0.27	0.31	0.25
		Avoidance	0.11	-0.15	0.36	-0.25	-0.36
		Behavioural coping	0.28	0.03	0.54	0.16	-0.12
		Cognitive coping	0.38	0.13	0.64	0.00	-0.38
		Confusion	0.13	-0.12	0.39	0.26	0.13
		Depression	0.12	-0.13	0.37	0.21	0.07
		Fatigue	0.29	0.03	0.54	0.21	-0.10
Vigour	0.43	0.18	0.69	0.41	-0.04		
Griffiths <i>et al.</i> , 2005, CDSMP in a south Asian group	1) IG (n=221) 2) CG (n=218)	<u>4 months</u>				<u>n=221</u>	<u>n=218</u>
		Anxiety	0.02	-0.16	0.21	0.11	0.08
		Communication with physician	0.09	-0.10	0.27	0.32	0.23
		Depression	0.04	-0.15	0.22	0.07	0.04
		EuroQoL	0.00	-0.19	0.18	0.02	0.03
		Fatigue	-0.01	-0.20	0.18	0.12	0.13
		Pain	0.03	-0.16	0.21	0.24	0.22
		Self-efficacy	0.13	-0.06	0.32	0.36	0.24
		Self-management behaviour	0.19	0.00	0.38	0.57	0.38
		Shortness of breath	0.10	-0.08	0.29	0.15	0.05
Visits to physician	0.01	-0.18	0.20	0.17	0.15		
Haas <i>et al.</i> , 2005, CDSMP low back pain	1) IG (n=60) 2) CG (n=49)	<u>6 months</u>				<u>n=60</u>	<u>n=49</u>
		Disability	0.26	-0.12	0.64	0.41	0.13
		Disability days	0.42	0.04	0.80	0.46	-0.04
		General health	-0.21	-0.59	0.17	-0.05	0.16
		Pain	0.03	-0.34	0.41	0.27	0.23
Pain days	0.10	-0.27	0.48	0.22	0.12		

**Appendix 1 (continued)** Systematic review of self-management interventions that were based on or similar to the Stanford programs

Study / intervention	Type of RCT	Outcome measures	Effect size <sup>1</sup> Cohen's d	95% CI		Effect size separately <sup>2</sup>	
				lower	upper	IG	CG
Haas <i>et al.</i> , 2005, CDSMP low back pain (continued)		Mental health	0.47	0.08	0.85	0.38	-0.12
		Self-efficacy, other symptoms	0.12	-0.25	0.50	0.02	-0.11
		Self-efficacy, pain	-0.04	-0.42	0.34	0.02	0.06
		Vitality	0.22	-0.16	0.60	0.17	-0.06
Heuts <i>et al.</i> , 2005, ASMP modified for OA	1) IG (n=132)	<u>3 months</u>				<u>n=132</u>	<u>n=141</u>
		Functional status	0.19	-0.05	0.42	0.17	-0.03
	2) CG (n=141)	General health	0.09	-0.15	0.33	-0.11	-0.19
		Health change	0.10	-0.14	0.33	0.21	0.10
		Pain, hip	-0.02	-0.26	0.22	0.08	0.10
		Pain, knee	0.25	0.01	0.49	0.28	0.00
		Pain-related fear	0.42	0.18	0.66	0.30	-0.13
		Physical functioning	0.08	-0.16	0.31	-0.06	-0.12
Self-efficacy	0.05	-0.18	0.29	0.10	0.04		

<sup>1</sup> Effect sizes (ES) were based on Cohen's d (Cohen, 1988).

<sup>2</sup> ES were calculated separately for the intervention and the control group. ES refer to the follow-up score at time X minus baseline score divided by the SD at baseline of the respective group.

<sup>3</sup> Goepfinger *et al.* (1989) only reported pooled SD of respective pretest scores. Hence, where ES are reported separately, each group's change score is divided by this pooled SD. This differs from the other studies in which the individual change scores are divided by each group's individual SD.

Legend

ASMP:	Arthritis Self-Management Program
BUOA:	Bone Up On Arthritis
CDSMP:	Chronic Disease Self-Management Program
CG:	Control group
ER:	Emergency room
ES:	Effect size
HP:	Health professional
IG:	Intervention group
OA:	Osteoarthritis
QoL	Quality of life
RA:	Rheumatoid arthritis

**Instructions**

Please indicate how strongly you disagree or agree with the following statements by checking the response which best describes you now.

**Example**

Ms. Jane Citizen has answered these questions in the following way:

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

Questions:

1. I am doing some of my hobbies

2. I have a plan to do physical activity

For Question 1, Jane's answer shows that right now she agrees that she has been doing some of her hobbies lately.

For Question 2, Jane agrees slightly with the statement that right now she has a plan to do physical activity.

**Please answer the following questions:**

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

1 On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)

2 I am very good at using aids and devices to make my life easier

3 Most days I am doing some of the things I really enjoy

4 As well as seeing my doctor, I regularly monitor changes in my health

5 I often worry about my health

6 If I need help, I have plenty of people I can rely on

7 I try to make the most of my life

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

8	I know what things can trigger my health problems and make them worse	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
9	My health problems make me very dissatisfied with my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
10	I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
11	I am doing interesting things in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
12	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
13	I have very positive relationships with my healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
14	I have a very good idea of how to manage my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
15	I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
16	I have plans to do enjoyable things for myself during the next few days	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
17	When I have symptoms, I have skills that help me cope	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
18	I try not to let my health problems stop me from enjoying life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
19	I have a very good understanding of when and why I am supposed to take my medication	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
20	I have enough friends who help me cope with my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
21	I communicate very confidently with my doctor about my healthcare needs	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
22	When I have health problems, I have a clear understanding of what I need to do to control them	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
23	When I feel ill, my family and carers really understand what I am going through	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
24	On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>



Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

25 I confidently give healthcare professionals the information they need to help me

26 I often feel angry when I think about my health

27 I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)

28 My health problems do not ruin my life

29 I feel hopeless because of my health problems

30 I feel like I am actively involved in life

31 Overall, I feel well looked after by friends or family

32 I feel I have a very good life even when I have health problems

33 I get enough chances to talk about my health problems with people who understand

34 I work in a team with my doctors and other healthcare professionals

35 I do not let my health problems control my life

36 I have effective skills that help me handle stress

37 I get upset when I think about my health

38 I carefully watch my health and do what is necessary to keep as healthy as possible

39 If others can cope with problems like mine, I can too

40 I walk for exercise, for at least 15 minutes per day, most days of the week

41 With my health in mind, I have realistic expectations of what I can and cannot do

42 If I think about my health, I get depressed

**Some details about yourself**

Today's date (complete)  /  / 2005

43 What is your age?

44 What is your sex?  
 Female  Male

45 What is your home postcode?

46 Including you, how many people aged 18 and over live in your household?

47 Are you Aboriginal or Torres Strait Islander origin?  
 Yes  No

48 In which country were you born?  
 (please specify)

49 Do you speak a language other than English at home? (please specify)

50 What is the highest level of education you have completed?

None, or some Primary School  
 Primary School  
 High School to year 8  
 High School to year 12  
 TAFE / Trade  
 University, or above

51 Do you have any long-standing illness, disability or infirmity that has troubled you over time or is likely to affect you over a period in the future?  
 (Please tick (✓) **all** that apply.)

Osteoarthritis  Asthma  
 Rheumatoid Arthritis  Diabetes  
 Fibromyalgia  Cancer  
 Osteoporosis  Depression  
 Coronary heart disease  Other

(please specify)

52 Please list your **main** health problem  
 (list **one** only)

53 Have you ever taken part in any health education courses or rehabilitation programs? (e.g. program for arthritis, asthma, diabetes, etc.)  
 If YES, please specify:

Type	Year
<input type="text"/>	<input type="text"/>
<input type="text"/>	<input type="text"/>

54 Which of the following statements best describes your smoking status? Please tick (✓) **all** that apply.

a  I smoke daily      b  I smoke occasionally  
 c  I don't smoke      d  I have never smoked

55 Are you currently in paid employment?  
 Please tick (✓) **one**.

a  Yes, full-time employed  
 b  Yes, part-time employed  
 c  No, unemployed  
 d  No, home duties  
 e  No, retired / pensioner

Other (please specify)

56 Do you have a Health Care Concession Card?  
 If YES - tick (✓) **any** that apply.

a  No  
 b  Pensioner Concession Card (PCC)  
 c  Commonwealth Health Care Card (HCC)  
 d  Commonwealth Seniors Health Card  
 e  Repatriation Card from the Dept. of Veteran Affairs  
 f  Low Income Health Card (LIC)

Other (please specify)

57 Apart from Medicare, are you currently covered by private health insurance?  
 Yes  No

58 Do you have any plans to lose weight in the near future?  
 Yes  No

59 Please state your current **height** **weight**

a <input type="text"/>	ft/in	b <input type="text"/>	lbs
<input type="text"/>	m/cm	<input type="text"/>	kg

**Thank you for taking the time to complete this questionnaire.  
 Please check that you have answered all of the questions.**

**Instructions**

Please indicate how strongly you disagree or agree with the following statements by checking the response which best describes you now.

**Example**

Ms. Jane Citizen has answered these questions in the following way:

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

Questions:

1. I am doing some of my hobbies

2. I have a plan to do physical activity

For Question 1, Jane's answer shows that right now she agrees that she has been doing some of her hobbies lately.

For Question 2, Jane agrees slightly with the statement that right now she has a plan to do physical activity.

**Please answer the following questions:**

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

1 On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)

2 I am very good at using aids and devices to make my life easier

3 Most days I am doing some of the things I really enjoy

4 As well as seeing my doctor, I regularly monitor changes in my health

5 I often worry about my health

6 If I need help, I have plenty of people I can rely on

7 I try to make the most of my life

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

8	I know what things can trigger my health problems and make them worse	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
9	My health problems make me very dissatisfied with my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
10	I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
11	I am doing interesting things in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
12	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
13	I have very positive relationships with my healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
14	I have a very good idea of how to manage my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
15	I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
16	I have plans to do enjoyable things for myself during the next few days	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
17	When I have symptoms, I have skills that help me cope	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
18	I try not to let my health problems stop me from enjoying life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
19	I have a very good understanding of when and why I am supposed to take my medication	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
20	I have enough friends who help me cope with my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
21	I communicate very confidently with my doctor about my healthcare needs	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
22	When I have health problems, I have a clear understanding of what I need to do to control them	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
23	When I feel ill, my family and carers really understand what I am going through	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
24	On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Check a box by crossing it:

**Right now**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

25 I confidently give healthcare professionals the information they need to help me

26 I often feel angry when I think about my health

27 I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)

28 My health problems do not ruin my life

29 I feel hopeless because of my health problems

30 I feel like I am actively involved in life

31 Overall, I feel well looked after by friends or family

32 I feel I have a very good life even when I have health problems

33 I get enough chances to talk about my health problems with people who understand

34 I work in a team with my doctors and other healthcare professionals

35 I do not let my health problems control my life

36 I have effective skills that help me handle stress

37 I get upset when I think about my health

38 I carefully watch my health and do what is necessary to keep as healthy as possible

39 If others can cope with problems like mine, I can too

40 I walk for exercise, for at least 15 minutes per day, most days of the week

41 With my health in mind, I have realistic expectations of what I can and cannot do

42 If I think about my health, I get depressed

**Instructions**  
 In this section please indicate how strongly you disagree or agree with the following statements by checking the response which best describes your experience of the health education program you have just taken part in.

Check a box by crossing it:

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

- |    |                                                                            |                                                                                                                                                       |
|----|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| 43 | I intend to tell other people that the program is very worthwhile          | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 44 | The program has helped me set goals that are reasonable and within reach   | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 45 | I trust the information and advice I was given in the program              | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 46 | Course leaders were very well organised                                    | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 47 | I feel it was worth my time and effort to take part in the program         | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 48 | Difficult topics and discussions were handled well by my program leaders   | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 49 | I thought the program content was very relevant to my situation            | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 50 | I feel that everyone in the program had the chance to speak if they wanted | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |
| 51 | The people in the group worked very well together                          | <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> |

52 Were there any things which made it hard for you to attend the course? Please tick (✓) **all** that apply.

- a Transport to the venue where the course was held
- b Access to the venue where the course was held (e.g. stairs, wheelchair access)
- c Parking at the venue where the course was held
- d Costs associated with the course
- e Other / further comments (list/specify)

We have almost finished with the survey. Listed below are a number of statements concerning your attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally.

Please tick whether **TRUE** or **FALSE** applies for you.

	TRUE	FALSE
a It is sometimes hard for me to go on with my work if I am not encouraged.	<input type="checkbox"/>	<input type="checkbox"/>
b I sometimes feel resentful when I don't get my way.	<input type="checkbox"/>	<input type="checkbox"/>
c On a few occasions, I have given up doing something because I thought too little of my ability.	<input type="checkbox"/>	<input type="checkbox"/>
d There have been times when I felt like rebelling against people in authority even though I knew they were right.	<input type="checkbox"/>	<input type="checkbox"/>
e No matter who I'm talking to, I'm always a good listener.	<input type="checkbox"/>	<input type="checkbox"/>
f There have been occasions when I took advantage of someone.	<input type="checkbox"/>	<input type="checkbox"/>
g I'm always willing to admit it when I make a mistake.	<input type="checkbox"/>	<input type="checkbox"/>
h I sometimes try to get even rather than forgive and forget.	<input type="checkbox"/>	<input type="checkbox"/>
i I am always courteous, even to people who are disagreeable.	<input type="checkbox"/>	<input type="checkbox"/>
j I have never been irked when people expressed ideas very different from my own.	<input type="checkbox"/>	<input type="checkbox"/>
k There have been times when I was quite jealous of the good fortune of others.	<input type="checkbox"/>	<input type="checkbox"/>
l I am sometimes irritated by people who ask favours of me.	<input type="checkbox"/>	<input type="checkbox"/>
m I have never deliberately said something that hurt someone's feelings.	<input type="checkbox"/>	<input type="checkbox"/>

**Some details about yourself**

Today's date (complete)

53 What is your age?

55 What is your home postcode?

56 Which of the following statements best describes your smoking status? Please tick (✓) **one only**.

- a  I smoke daily
- b  I smoke occasionally
- c  I don't smoke
- d  I have never smoked

54 What is your sex?

- Female
- Male

57 Do you have any plans to lose weight in the near future?

- Yes
- No

58 Please state your current

a height	b weight
<input style="width: 100%; height: 20px;" type="text" value=""/> ft/in	<input style="width: 100%; height: 20px;" type="text" value=""/> lbs
<input style="width: 100%; height: 20px;" type="text" value=""/> m/cm	<input style="width: 100%; height: 20px;" type="text" value=""/> kg

**Thank you for taking the time to complete this questionnaire.  
Please check that you have answered all of the questions.**

**Appendix 4** Posttest heiQ-PPT including the MC-C scale and demographic variables

Page 1

**Instructions**

Please indicate how strongly you disagree or agree with the following statements by checking the response which best describes what your situation has been. Firstly, answer each question according to your situation right now, and then answer the same question regarding whether you have changed since you started the health education program.

**Example**

Ms. Jane Citizen completed a health education program and answered these questions in the following way:

Check a box by crossing it:

	<b>Right now</b>	<b>Compared with before the program</b>
	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>much worse</i> <i>worse</i> <i>the same</i> <i>better</i> <i>much better</i>
Questions:		
1. I am doing some of my hobbies	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>
2. I have a plan to do physical activity	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

For Question 1, Jane's answer shows that right now she agrees that she has been doing some of her hobbies lately. When she compares that with what she did before the program she indicates that she does more of her hobbies now than back then.

For Question 2, Jane agrees slightly with the statement that right now she has a plan to do physical activity. When she compares her current plan to her plan before the program Jane indicates that she had about the same plan to do physical activity.

**Please answer the following questions:**

Check a box by crossing it:

	<b>Right now</b>	<b>Compared with before the program</b>
	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>much worse</i> <i>worse</i> <i>the same</i> <i>better</i> <i>much better</i>
1 On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
2 I am very good at using aids and devices to make my life easier	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
3 Most days I am doing some of the things I really enjoy	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
4 As well as seeing my doctor, I regularly monitor changes in my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
5 I often worry about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
6 If I need help, I have plenty of people I can rely on	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
7 I try to make the most of my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>



Check a box by crossing it:

	<u>Right now</u>	<u>Compared with before the program</u>
	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>much worse</i> <i>worse</i> <i>the same</i> <i>better</i> <i>much better</i>
8 I know what things can trigger my health problems and make them worse	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
9 My health problems make me very dissatisfied with my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
10 I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
11 I am doing interesting things in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
12 I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
13 I have very positive relationships with my healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
14 I have a very good idea of how to manage my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
15 I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
16 I have plans to do enjoyable things for myself during the next few days	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
17 When I have symptoms, I have skills that help me cope	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
18 I try not to let my health problems stop me from enjoying life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
19 I have a very good understanding of when and why I am supposed to take my medication	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
20 I have enough friends who help me cope with my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
21 I communicate very confidently with my doctor about my healthcare needs	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
22 When I have health problems, I have a clear understanding of what I need to do to control them	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
23 When I feel ill, my family and carers really understand what I am going through	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
24 On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Check a box by crossing it:

	<b>Right now</b>	<b>Compared with before the program</b>
	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>much worse</i> <i>worse</i> <i>the same</i> <i>better</i> <i>much better</i>
25 I confidently give healthcare professionals the information they need to help me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
26 I often feel angry when I think about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
27 I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
28 My health problems do not ruin my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
29 I feel hopeless because of my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
30 I feel like I am actively involved in life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
31 Overall, I feel well looked after by friends or family	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
32 I feel I have a very good life even when I have health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
33 I get enough chances to talk about my health problems with people who understand	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
34 I work in a team with my doctors and other healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
35 I do not let my health problems control my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
36 I have effective skills that help me handle stress	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
37 I get upset when I think about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
38 I carefully watch my health and do what is necessary to keep as healthy as possible	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
39 If others can cope with problems like mine, I can too	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
40 I walk for exercise, for at least 15 minutes per day, most days of the week	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
41 With my health in mind, I have realistic expectations of what I can and cannot do	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
42 If I think about my health, I get depressed	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

**Instructions**

In this section please indicate how strongly you disagree or agree with the following statements by checking the response which best describes your experience of the health education program you have recently taken part in.

Check a box by crossing it:

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

- 43 I intend to tell other people that the program is very worthwhile
- 44 The program has helped me set goals that are reasonable and within reach
- 45 I trust the information and advice I was given in the program
- 46 Course leaders were very well organised
- 47 I feel it was worth my time and effort to take part in the program
- 48 Difficult topics and discussions were handled well by my program leaders
- 49 I thought the program content was very relevant to my situation
- 50 I feel that everyone in the program had the chance to speak if they wanted
- 51 The people in the group worked very well together

52 Were there any things which made it hard for you to attend the course? Please tick (✓) **all** that apply.

- a Transport to the venue where the course was held
- b Access to the venue where the course was held (e.g. stairs, wheelchair access)
- c Parking at the venue where the course was held
- d Costs associated with the course
- e Other / further comments (list/specify)

We have almost finished with the survey. Listed below are a number of statements concerning your attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally.

Please tick whether **TRUE** or **FALSE** applies for you.

	TRUE	FALSE
a It is sometimes hard for me to go on with my work if I am not encouraged.	<input type="checkbox"/>	<input type="checkbox"/>
b I sometimes feel resentful when I don't get my way.	<input type="checkbox"/>	<input type="checkbox"/>
c On a few occasions, I have given up doing something because I thought too little of my ability.	<input type="checkbox"/>	<input type="checkbox"/>
d There have been times when I felt like rebelling against people in authority even though I knew they were right.	<input type="checkbox"/>	<input type="checkbox"/>
e No matter who I'm talking to, I'm always a good listener.	<input type="checkbox"/>	<input type="checkbox"/>
f There have been occasions when I took advantage of someone.	<input type="checkbox"/>	<input type="checkbox"/>
g I'm always willing to admit it when I make a mistake.	<input type="checkbox"/>	<input type="checkbox"/>
h I sometimes try to get even rather than forgive and forget.	<input type="checkbox"/>	<input type="checkbox"/>
i I am always courteous, even to people who are disagreeable.	<input type="checkbox"/>	<input type="checkbox"/>
j I have never been irked when people expressed ideas very different from my own.	<input type="checkbox"/>	<input type="checkbox"/>
k There have been times when I was quite jealous of the good fortune of others.	<input type="checkbox"/>	<input type="checkbox"/>
l I am sometimes irritated by people who ask favours of me.	<input type="checkbox"/>	<input type="checkbox"/>
m I have never deliberately said something that hurt someone's feelings.	<input type="checkbox"/>	<input type="checkbox"/>

**Some details about yourself**

Today's date (complete)

53 What is your age?

55 What is your home postcode?

56 Which of the following statements best describes your smoking status? Please tick (✓) **one only**.

- a  I smoke daily
- b  I smoke occasionally
- c  I don't smoke
- d  I have never smoked

54 What is your sex?

- Female
- Male

57 Do you have any plans to lose weight in the near future?

- Yes
- No

58 Please state your current

a height	b weight
<input style="width: 100%; height: 20px;" type="text" value=""/> ft/in	<input style="width: 100%; height: 20px;" type="text" value=""/> lbs
<input style="width: 100%; height: 20px;" type="text" value=""/> m/cm	<input style="width: 100%; height: 20px;" type="text" value=""/> kg

**Thank you for taking the time to complete this questionnaire.  
Please check that you have answered all of the questions.**

**Appendix 5** Posttest heiQ-PPR including the MC-C scale and demographic variables

Page 1

**Instructions**

Please indicate how strongly you disagree or agree with the following statements by checking the response which best describes what your situation has been. Firstly, answer each question according to your situation right now, and then answer the same question regarding your situation before you took part in the health education program.

**Example**

Ms. Jane Citizen completed a health education program and answered these questions in the following way:

Check a box by crossing it:

	<b>Right now</b>	<b>Before the program</b>
Questions:	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>
1. I am doing some of my hobbies	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
2. I have a plan to do physical activity	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

For Question 1, Jane's answer shows that right now she agrees that she has been doing some of her hobbies lately. When she thinks back to what she did before the program she disagrees slightly that she was doing some of her hobbies.

For Question 2, Jane agrees slightly with the statement that right now she has a plan to do physical activity. When she thinks back to what she did before the program Jane also agrees slightly that back then she had a plan to do physical activity.

**Please answer the following questions:**

Check a box by crossing it:

	<b>Right now</b>	<b>Before the program</b>
1 On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
2 I am very good at using aids and devices to make my life easier	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
3 Most days I am doing some of the things I really enjoy	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
4 As well as seeing my doctor, I regularly monitor changes in my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
5 I often worry about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
6 If I need help, I have plenty of people I can rely on	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
7 I try to make the most of my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Check a box by crossing it:

	<b>Right now</b>	<b>Before the program</b>
	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>	<i>strongly disagree</i> <i>disagree</i> <i>disagree slightly</i> <i>agree slightly</i> <i>agree</i> <i>strongly agree</i>
8 I know what things can trigger my health problems and make them worse	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
9 My health problems make me very dissatisfied with my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
10 I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
11 I am doing interesting things in my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
12 I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
13 I have very positive relationships with my healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
14 I have a very good idea of how to manage my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
15 I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
16 I have plans to do enjoyable things for myself during the next few days	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
17 When I have symptoms, I have skills that help me cope	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
18 I try not to let my health problems stop me from enjoying life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
19 I have a very good understanding of when and why I am supposed to take my medication	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
20 I have enough friends who help me cope with my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
21 I communicate very confidently with my doctor about my healthcare needs	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
22 When I have health problems, I have a clear understanding of what I need to do to control them	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
23 When I feel ill, my family and carers really understand what I am going through	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
24 On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Check a box by crossing it:

**Right now**

**Before the program**

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

25	I confidently give healthcare professionals the information they need to help me	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
26	I often feel angry when I think about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
27	I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
28	My health problems do not ruin my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
29	I feel hopeless because of my health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
30	I feel like I am actively involved in life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
31	Overall, I feel well looked after by friends or family	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
32	I feel I have a very good life even when I have health problems	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
33	I get enough chances to talk about my health problems with people who understand	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
34	I work in a team with my doctors and other healthcare professionals	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
35	I do not let my health problems control my life	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
36	I have effective skills that help me handle stress	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
37	I get upset when I think about my health	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
38	I carefully watch my health and do what is necessary to keep as healthy as possible	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
39	If others can cope with problems like mine, I can too	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
40	I walk for exercise, for at least 15 minutes per day, most days of the week	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
41	With my health in mind, I have realistic expectations of what I can and cannot do	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
42	If I think about my health, I get depressed	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

**Instructions**

In this section please indicate how strongly you disagree or agree with the following statements by checking the response which best describes your experience of the health education program you have recently taken part in.

Check a box by crossing it:

*strongly disagree*  
*disagree*  
*disagree slightly*  
*agree slightly*  
*agree*  
*strongly agree*

- 43 I intend to tell other people that the program is very worthwhile
- 44 The program has helped me set goals that are reasonable and within reach
- 45 I trust the information and advice I was given in the program
- 46 Course leaders were very well organised
- 47 I feel it was worth my time and effort to take part in the program
- 48 Difficult topics and discussions were handled well by my program leaders
- 49 I thought the program content was very relevant to my situation
- 50 I feel that everyone in the program had the chance to speak if they wanted
- 51 The people in the group worked very well together

52 Were there any things which made it hard for you to attend the course? Please tick (✓) **all** that apply.

- a Transport to the venue where the course was held
- b Access to the venue where the course was held (e.g. stairs, wheelchair access)
- c Parking at the venue where the course was held
- d Costs associated with the course
- e Other / further comments (list/specify)



We have almost finished with the survey. Listed below are a number of statements concerning your attitudes and traits. Read each item and decide whether the statement is true or false as it pertains to you personally.

Please tick whether **TRUE** or **FALSE** applies for you.

	TRUE	FALSE
a It is sometimes hard for me to go on with my work if I am not encouraged.	<input type="checkbox"/>	<input type="checkbox"/>
b I sometimes feel resentful when I don't get my way.	<input type="checkbox"/>	<input type="checkbox"/>
c On a few occasions, I have given up doing something because I thought too little of my ability.	<input type="checkbox"/>	<input type="checkbox"/>
d There have been times when I felt like rebelling against people in authority even though I knew they were right.	<input type="checkbox"/>	<input type="checkbox"/>
e No matter who I'm talking to, I'm always a good listener.	<input type="checkbox"/>	<input type="checkbox"/>
f There have been occasions when I took advantage of someone.	<input type="checkbox"/>	<input type="checkbox"/>
g I'm always willing to admit it when I make a mistake.	<input type="checkbox"/>	<input type="checkbox"/>
h I sometimes try to get even rather than forgive and forget.	<input type="checkbox"/>	<input type="checkbox"/>
i I am always courteous, even to people who are disagreeable.	<input type="checkbox"/>	<input type="checkbox"/>
j I have never been irked when people expressed ideas very different from my own.	<input type="checkbox"/>	<input type="checkbox"/>
k There have been times when I was quite jealous of the good fortune of others.	<input type="checkbox"/>	<input type="checkbox"/>
l I am sometimes irritated by people who ask favours of me.	<input type="checkbox"/>	<input type="checkbox"/>
m I have never deliberately said something that hurt someone's feelings.	<input type="checkbox"/>	<input type="checkbox"/>

**Some details about yourself**

Today's date (complete)

53 What is your age?

55 What is your home postcode?

56 Which of the following statements best describes your smoking status? Please tick (✓) **one only**.

- a  I smoke daily
- b  I smoke occasionally
- c  I don't smoke
- d  I have never smoked

54 What is your sex?

- Female
- Male

57 Do you have any plans to lose weight in the near future?

- Yes
- No

58 Please state your current

a height	b weight
<input style="width: 100%; height: 20px;" type="text" value=""/> ft/in	<input style="width: 100%; height: 20px;" type="text" value=""/> lbs
<input style="width: 100%; height: 20px;" type="text" value=""/> m/cm	<input style="width: 100%; height: 20px;" type="text" value=""/> kg

**Thank you for taking the time to complete this questionnaire.  
Please check that you have answered all of the questions.**

## Health Education Impact

Please return to the Centre for Rheumatic Diseases

**Course Participation Form**

Course ID					
	OFFICE	USE	ONLY		

Contact Person  Phone

Email

Type of course (please specify)

Date of first session  /  / 2005      Duration of course  weeks

What organisation is running the course?

Where is the course held?     Community Health Centre     Hospital     Arthritis Foundation  
 Church     Other, please specify \_\_\_\_\_

	Course leader 1	Course leader 2
1. Name of course leaders	<input type="text"/>	<input type="text"/>
2. Phone	<input type="text"/>	<input type="text"/>
3. Email address	<input type="text"/>	<input type="text"/>
4. Are you a peer educator or a health professional? (Please state your health profession)	<input type="text"/>	<input type="text"/>
5. In what year did you become a course leader?	<input type="text"/>	<input type="text"/>
6. Where were you trained? (Please specify)	<input type="text"/>	<input type="text"/>
7. Who was your Master Trainer?	<input type="text"/>	<input type="text"/>
8. How many courses have you conducted in the last 12 months?	<input type="text"/>	<input type="text"/>

**Course Participants WITHOUT support persons**

	Participants' initials	Female or Male	Participants' IDs	Session attendance													
				1	2	3	4	5	6	(7)	(8)	(9)	(10)				
1																	
2																	
3																	
4																	
5																	
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	

**Appendix 7** Univariate and multivariate normality tests of the following heiQ data: pretests (n=666), retrospective pretests heiQ-PPR (n=189), posttests (n=603), posttests heiQ-PP (n=244), posttests heiQ-PPT (n=150), and posttests heiQ-PPR (n=209)

**Uni- and multivariate normality heiQ pretests (n=666)**

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1	4.354	1.250	89.928	-0.963	0.441	1.000	25	6.000	90
q1_2	4.556	1.188	98.966	-1.075	0.748	1.000	13	6.000	119
q1_3	5.021	0.932	139.023	-1.439	2.981	1.000	4	6.000	202
q1_4	4.345	1.265	88.631	-0.952	0.278	1.000	23	6.000	88
q1_5	4.197	1.325	81.748	-0.887	0.026	1.000	37	6.000	70
q2_1	3.902	1.669	60.337	-0.392	-1.222	1.000	69	6.000	119
q2_2	4.314	1.518	73.361	-0.758	-0.600	1.000	37	6.000	152
q2_3	4.836	1.111	112.318	-1.454	2.297	1.000	13	6.000	176
q2_4	4.354	1.365	82.347	-0.876	-0.026	1.000	30	6.000	121
q3_1	4.084	1.200	87.846	-0.807	-0.011	1.000	21	6.000	37
q3_2	4.368	1.285	87.689	-0.988	0.289	1.000	25	6.000	93
q3_3	3.950	1.266	80.546	-0.598	-0.436	1.000	27	6.000	40
q3_4	4.371	1.042	108.300	-0.995	0.967	1.000	9	6.000	52
q3_5	4.110	1.177	90.120	-0.785	0.080	1.000	19	6.000	41
q4_1	4.821	0.993	125.300	-1.309	2.160	1.000	5	6.000	143
q4_2	4.796	0.970	127.539	-1.414	2.777	1.000	7	6.000	124
q4_3	4.417	1.235	92.322	-0.896	0.156	1.000	14	6.000	102
q4_4	4.161	1.398	76.802	-0.825	-0.302	1.000	41	6.000	76
q4_5	4.369	1.279	88.191	-1.114	0.496	1.000	28	6.000	77
q5_1	4.734	0.975	125.330	-1.235	1.740	1.000	3	6.000	108
q5_2	4.868	1.011	124.198	-1.464	2.787	1.000	8	6.000	160
q5_3	4.595	1.130	104.968	-1.172	1.105	1.000	8	6.000	110
q5_4	4.471	1.104	104.560	-1.099	0.988	1.000	11	6.000	75
q5_5	5.257	0.847	160.164	-2.050	6.649	1.000	5	6.000	274
q5_6	4.665	0.956	125.876	-0.953	1.021	1.000	2	6.000	99
q5_7	4.700	1.010	120.122	-1.367	2.349	1.000	9	6.000	104
q6_1	4.889	1.052	119.946	-1.348	1.939	1.000	5	6.000	184
q6_2	4.803	1.012	122.473	-1.327	2.363	1.000	9	6.000	144
q6_3	5.120	0.793	166.585	-1.450	4.064	1.000	1	6.000	202
q6_4	4.464	1.235	93.310	-1.143	0.798	1.000	22	6.000	97
q6_5	4.300	1.287	86.212	-0.875	-0.046	1.000	21	6.000	82
q7_1	4.180	1.386	77.806	-0.750	-0.333	1.000	35	6.000	91
q7_2	4.062	1.350	77.625	-0.711	-0.429	1.000	35	6.000	58
q7_3	4.233	1.425	76.639	-0.674	-0.515	1.000	32	6.000	120
q7_4	4.357	1.358	82.786	-1.006	0.192	1.000	36	6.000	102
q7_5	3.844	1.445	68.668	-0.506	-0.775	1.000	57	6.000	57
q8_1	3.590	1.618	57.259	-0.079	-1.254	1.000	80	6.000	85
q8_2	3.462	1.522	58.727	0.065	-1.123	1.000	70	6.000	63
q8_3	3.152	1.619	50.229	0.150	-1.321	1.000	135	6.000	39
q8_4	3.186	1.550	53.061	0.166	-1.161	1.000	114	6.000	40
q8_5	2.641	1.421	47.958	0.701	-0.527	1.000	154	6.000	21
q8_6	3.857	1.567	63.516	-0.331	-1.063	1.000	63	6.000	99

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1	-8.738	0.000	2.047	0.041	80.550	0.000
q1_2	-9.490	0.000	3.071	0.002	99.485	0.000
q1_3	-11.633	0.000	7.158	0.000	186.575	0.000
q1_4	-8.663	0.000	1.404	0.160	77.017	0.000
q1_5	-8.196	0.000	0.230	0.818	67.234	0.000
q2_1	-4.024	0.000	-23.137	0.000	551.500	0.000
q2_2	-7.228	0.000	-4.546	0.000	72.911	0.000
q2_3	-11.714	0.000	6.273	0.000	176.566	0.000
q2_4	-8.120	0.000	-0.053	0.958	65.942	0.000
q3_1	-7.602	0.000	0.029	0.977	57.796	0.000
q3_2	-8.910	0.000	1.452	0.146	81.491	0.000
q3_3	-5.910	0.000	-2.898	0.004	43.324	0.000
q3_4	-8.962	0.000	3.688	0.000	93.927	0.000
q3_5	-7.438	0.000	0.503	0.615	55.578	0.000
q4_1	-10.917	0.000	6.069	0.000	156.023	0.000
q4_2	-11.500	0.000	6.915	0.000	180.073	0.000
q4_3	-8.263	0.000	0.869	0.385	69.036	0.000
q4_4	-7.739	0.000	-1.809	0.070	63.159	0.000
q4_5	-9.746	0.000	2.247	0.025	100.028	0.000
q5_1	-10.488	0.000	5.372	0.000	138.866	0.000
q5_2	-11.768	0.000	6.927	0.000	186.474	0.000
q5_3	-10.104	0.000	4.040	0.000	118.412	0.000
q5_4	-9.648	0.000	3.743	0.000	107.088	0.000
q5_5	-14.445	0.000	9.945	0.000	307.553	0.000
q5_6	-8.672	0.000	3.828	0.000	89.845	0.000
q5_7	-11.246	0.000	6.347	0.000	166.758	0.000
q6_1	-11.136	0.000	5.716	0.000	156.676	0.000
q6_2	-11.018	0.000	6.368	0.000	161.938	0.000
q6_3	-11.695	0.000	8.237	0.000	204.615	0.000
q6_4	-9.928	0.000	3.219	0.001	108.929	0.000
q6_5	-8.109	0.000	-0.162	0.871	65.777	0.000
q7_1	-7.159	0.000	-2.043	0.041	55.433	0.000
q7_2	-6.853	0.000	-2.840	0.005	55.025	0.000
q7_3	-6.551	0.000	-3.641	0.000	56.172	0.000
q7_4	-9.036	0.000	1.031	0.303	82.710	0.000
q7_5	-5.091	0.000	-6.907	0.000	73.633	0.000
q8_1	-0.839	0.402	-26.704	0.000	713.826	0.000
q8_2	0.692	0.489	-16.454	0.000	271.219	0.000
q8_3	1.582	0.114	-40.027	0.000	1604.678	0.000
q8_4	1.753	0.080	-18.530	0.000	346.443	0.000
q8_5	6.768	0.000	-3.758	0.000	59.920	0.000
q8_6	-3.436	0.001	-13.866	0.000	204.059	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.303

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
290.192	84.183	0.000	2407.218	38.479	0.000	8567.471	0.000

**Uni- and multivariate normality heiQ-PPR retrospective pretests (n=189)**

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1r	4.201	1.384	41.716	-0.768	-0.291	1.000	9	6.000	27
q1_2r	4.127	1.307	43.423	-0.484	-0.609	1.000	4	6.000	25
q1_3r	4.630	1.305	48.779	-0.926	0.233	1.000	5	6.000	56
q1_4r	4.159	1.266	45.166	-0.589	-0.239	1.000	6	6.000	23
q1_5r	4.048	1.445	38.508	-0.597	-0.696	1.000	11	6.000	24
q2_1r	3.820	1.669	31.461	-0.308	-1.201	1.000	22	6.000	34
q2_2r	4.021	1.547	35.735	-0.463	-0.947	1.000	13	6.000	34
q2_3r	4.000	1.259	43.678	-0.501	-0.430	1.000	6	6.000	16
q2_4r	4.164	1.333	42.954	-0.673	-0.146	1.000	9	6.000	27
q3_1r	4.085	1.164	48.250	-0.637	-0.294	1.000	3	6.000	11
q3_2r	4.190	1.355	42.518	-0.624	-0.343	1.000	8	6.000	30
q3_3r	3.952	1.289	42.140	-0.527	-0.441	1.000	8	6.000	15
q3_4r	4.270	1.214	48.339	-0.946	0.692	1.000	8	6.000	21
q3_5r	3.873	1.240	42.948	-0.518	-0.336	1.000	8	6.000	11
q4_1r	4.529	1.266	49.201	-0.981	0.327	1.000	4	6.000	39
q4_2r	4.238	1.377	42.323	-0.659	-0.446	1.000	7	6.000	32
q4_3r	4.270	1.439	40.796	-0.808	-0.287	1.000	11	6.000	34
q4_4r	4.206	1.464	39.499	-0.733	-0.422	1.000	12	6.000	34
q4_5r	4.079	1.425	39.353	-0.687	-0.596	1.000	11	6.000	21
q5_1r	4.471	1.146	53.614	-1.041	0.756	1.000	4	6.000	24
q5_2r	4.566	1.217	51.583	-1.051	0.812	1.000	5	6.000	39
q5_3r	4.339	1.243	47.997	-0.903	0.365	1.000	6	6.000	26
q5_4r	4.481	1.094	56.294	-1.023	1.016	1.000	3	6.000	24
q5_5r	5.111	1.023	68.700	-1.763	3.778	1.000	2	6.000	74
q5_6r	4.508	1.156	53.610	-0.855	0.425	1.000	2	6.000	34
q5_7r	4.545	1.059	58.998	-0.988	0.944	1.000	2	6.000	26
q6_1r	4.762	1.168	56.070	-1.086	0.938	1.000	3	6.000	54
q6_2r	4.683	1.146	56.167	-1.131	1.345	1.000	4	6.000	44
q6_3r	4.947	0.955	71.214	-1.338	2.122	2.000	8	6.000	50
q6_4r	4.439	1.264	48.267	-1.066	0.490	1.000	6	6.000	28
q6_5r	4.434	1.268	48.060	-0.909	0.111	1.000	4	6.000	32
q7_1r	4.074	1.566	35.769	-0.527	-0.828	1.000	16	6.000	38
q7_2r	3.868	1.480	35.932	-0.486	-0.849	1.000	16	6.000	19
q7_3r	4.175	1.556	36.881	-0.603	-0.755	1.000	14	6.000	42
q7_4r	4.243	1.478	39.464	-0.677	-0.662	1.000	9	6.000	37
q7_5r	3.772	1.454	35.680	-0.385	-0.983	1.000	14	6.000	15
q8_1r	3.222	1.589	27.879	0.249	-1.107	1.000	30	6.000	18
q8_2r	3.127	1.629	26.388	0.233	-1.263	1.000	37	6.000	14
q8_3r	2.984	1.623	25.284	0.411	-1.167	1.000	39	6.000	13
q8_4r	2.942	1.588	25.461	0.410	-1.126	1.000	40	6.000	11
q8_5r	2.529	1.363	25.516	0.671	-0.474	1.000	51	6.000	4
q8_6r	3.635	1.634	30.587	-0.128	-1.231	1.000	23	6.000	27

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1r	-3.995	0.000	-0.835	0.404	16.660	0.000
q1_2r	-2.664	0.008	-2.357	0.018	12.655	0.002
q1_3r	-4.652	0.000	0.774	0.439	22.245	0.000
q1_4r	-3.182	0.001	-0.638	0.524	10.531	0.005
q1_5r	-3.220	0.001	-2.911	0.004	18.844	0.000
q2_1r	-1.740	0.082	-9.253	0.000	88.647	0.000
q2_2r	-2.558	0.011	-5.069	0.000	32.240	0.000
q2_3r	-2.752	0.006	-1.427	0.154	9.609	0.008
q2_4r	-3.573	0.000	-0.308	0.758	12.860	0.002
q3_1r	-3.409	0.001	-0.846	0.397	12.336	0.002
q3_2r	-3.346	0.001	-1.045	0.296	12.288	0.002
q3_3r	-2.883	0.004	-1.476	0.140	10.491	0.005
q3_4r	-4.730	0.000	1.726	0.084	25.353	0.000
q3_5r	-2.835	0.005	-1.017	0.309	9.074	0.011
q4_1r	-4.866	0.000	0.994	0.320	24.668	0.000
q4_2r	-3.513	0.000	-1.501	0.133	14.594	0.001
q4_3r	-4.169	0.000	-0.821	0.412	18.054	0.000
q4_4r	-3.846	0.000	-1.390	0.164	16.727	0.000
q4_5r	-3.640	0.000	-2.283	0.022	18.458	0.000
q5_1r	-5.092	0.000	1.837	0.066	29.303	0.000
q5_2r	-5.127	0.000	1.931	0.053	30.012	0.000
q5_3r	-4.559	0.000	1.080	0.280	21.949	0.000
q5_4r	-5.023	0.000	2.251	0.024	30.299	0.000
q5_5r	-7.311	0.000	4.693	0.000	75.472	0.000
q5_6r	-4.363	0.000	1.209	0.227	20.501	0.000
q5_7r	-4.893	0.000	2.142	0.032	28.526	0.000
q6_1r	-5.254	0.000	2.134	0.033	32.161	0.000
q6_2r	-5.417	0.000	2.699	0.007	36.629	0.000
q6_3r	-6.105	0.000	3.527	0.000	49.717	0.000
q6_4r	-5.185	0.000	1.343	0.179	28.687	0.000
q6_5r	-4.585	0.000	0.463	0.644	21.233	0.000
q7_1r	-2.882	0.004	-3.915	0.000	23.630	0.000
q7_2r	-2.678	0.007	-4.095	0.000	23.945	0.000
q7_3r	-3.248	0.001	-3.328	0.001	21.623	0.000
q7_4r	-3.592	0.000	-2.682	0.007	20.099	0.000
q7_5r	-2.152	0.031	-5.486	0.000	34.727	0.000
q8_1r	1.415	0.157	-7.286	0.000	55.091	0.000
q8_2r	1.330	0.184	-11.062	0.000	124.132	0.000
q8_3r	2.291	0.022	-8.456	0.000	76.761	0.000
q8_4r	2.288	0.022	-7.637	0.000	63.565	0.000
q8_5r	3.566	0.000	-1.637	0.102	15.397	0.000
q8_6r	-0.739	0.460	-10.059	0.000	101.736	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.161

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
661.653	39.837	0.000	2145.436	16.021	0.000	1843.619	0.000

**Uni- and multivariate normality heiQ actual posttests (n=603)**

**Univariate Summary Statistics for Continuous Variables**

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1f	4.738	1.032	112.786	-1.263	2.070	1.000	9	6.000	118
q1_2f	4.834	0.978	121.419	-1.257	2.129	1.000	4	6.000	134
q1_3f	5.181	0.858	148.253	-1.844	5.577	1.000	4	6.000	220
q1_4f	4.859	0.993	120.219	-1.290	2.192	1.000	5	6.000	145
q1_5f	4.602	1.187	95.188	-1.272	1.392	1.000	18	6.000	109
q2_1f	4.370	1.494	71.802	-0.784	-0.478	1.000	30	6.000	152
q2_2f	4.637	1.309	86.986	-1.116	0.591	1.000	19	6.000	160
q2_3f	5.138	0.807	156.427	-1.608	5.079	1.000	3	6.000	192
q2_4f	4.774	1.132	103.577	-1.341	1.784	1.000	11	6.000	148
q3_1f	4.813	0.882	134.059	-1.348	3.115	1.000	4	6.000	100
q3_2f	4.773	1.028	114.037	-1.496	2.944	1.000	11	6.000	117
q3_3f	4.619	0.990	114.535	-1.094	1.414	1.000	4	6.000	81
q3_4f	4.965	0.822	148.417	-1.396	4.203	1.000	4	6.000	135
q3_5f	4.610	0.961	117.862	-1.275	2.248	1.000	7	6.000	67
q4_1f	5.027	0.955	129.185	-1.704	4.172	1.000	7	6.000	182
q4_2f	4.972	0.979	124.757	-1.768	4.218	1.000	9	6.000	160
q4_3f	4.721	1.144	101.388	-1.339	1.802	1.000	15	6.000	133
q4_4f	4.511	1.301	85.151	-1.123	0.636	1.000	25	6.000	115
q4_5f	4.632	1.183	96.158	-1.208	1.128	1.000	14	6.000	118
q5_1f	5.015	0.757	162.704	-1.547	5.372	1.000	3	6.000	128
q5_2f	5.066	0.872	142.734	-1.533	4.094	1.000	4	6.000	183
q5_3f	4.927	0.944	128.127	-1.529	3.513	1.000	6	6.000	147
q5_4f	4.902	0.890	135.201	-1.366	3.107	1.000	4	6.000	129
q5_5f	5.413	0.692	192.114	-2.025	9.105	1.000	3	6.000	292
q5_6f	4.935	0.850	142.656	-1.181	2.865	1.000	3	6.000	139
q5_7f	4.982	0.814	150.301	-1.543	4.653	1.000	3	6.000	132
q6_1f	5.013	0.956	128.807	-1.309	2.265	1.000	3	6.000	193
q6_2f	4.925	1.005	120.385	-1.260	1.829	1.000	3	6.000	173
q6_3f	5.194	0.713	178.809	-1.128	3.010	2.000	5	6.000	198
q6_4f	4.786	0.989	118.776	-1.326	2.401	1.000	7	6.000	118
q6_5f	4.703	1.100	105.032	-1.220	1.411	1.000	7	6.000	122
q7_1f	4.378	1.328	80.952	-1.004	0.335	1.000	29	6.000	98
q7_2f	4.421	1.173	92.517	-0.989	0.525	1.000	14	6.000	74
q7_3f	4.514	1.298	85.384	-1.018	0.453	1.000	21	6.000	129
q7_4f	4.498	1.342	82.275	-1.039	0.336	1.000	25	6.000	128
q7_5f	4.073	1.366	73.235	-0.666	-0.366	1.000	35	6.000	68
q8_1f	3.818	1.481	63.310	-0.217	-0.984	1.000	43	6.000	81
q8_2f	3.708	1.474	61.759	-0.087	-1.119	1.000	39	6.000	67
q8_3f	3.355	1.511	54.530	0.071	-1.117	1.000	77	6.000	42
q8_4f	3.499	1.486	57.836	0.028	-1.153	1.000	51	6.000	50
q8_5f	2.900	1.389	51.265	0.462	-0.777	1.000	92	6.000	19
q8_6f	4.206	1.466	70.427	-0.578	-0.670	1.000	34	6.000	122

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1f	-10.149	0.000	5.667	0.000	135.114	0.000
q1_2f	-10.119	0.000	5.755	0.000	135.520	0.000
q1_3f	-12.946	0.000	8.930	0.000	247.361	0.000
q1_4f	-10.298	0.000	5.848	0.000	140.250	0.000
q1_5f	-10.200	0.000	4.485	0.000	124.154	0.000
q2_1f	-7.072	0.000	-3.119	0.002	59.746	0.000
q2_2f	-9.299	0.000	2.464	0.014	92.540	0.000
q2_3f	-11.910	0.000	8.619	0.000	216.153	0.000
q2_4f	-10.578	0.000	5.211	0.000	139.053	0.000
q3_1f	-10.613	0.000	6.988	0.000	161.474	0.000
q3_2f	-11.373	0.000	6.802	0.000	175.607	0.000
q3_3f	-9.160	0.000	4.529	0.000	104.426	0.000
q3_4f	-10.869	0.000	7.987	0.000	181.925	0.000
q3_5f	-10.218	0.000	5.928	0.000	139.535	0.000
q4_1f	-12.345	0.000	7.963	0.000	215.810	0.000
q4_2f	-12.626	0.000	8.000	0.000	223.417	0.000
q4_3f	-10.564	0.000	5.241	0.000	139.063	0.000
q4_4f	-9.340	0.000	2.604	0.009	94.009	0.000
q4_5f	-9.840	0.000	3.914	0.000	112.152	0.000
q5_1f	-11.620	0.000	8.806	0.000	212.572	0.000
q5_2f	-11.554	0.000	7.899	0.000	195.895	0.000
q5_3f	-11.531	0.000	7.388	0.000	187.550	0.000
q5_4f	-10.708	0.000	6.980	0.000	163.376	0.000
q5_5f	-13.674	0.000	10.515	0.000	297.562	0.000
q5_6f	-9.684	0.000	6.713	0.000	138.848	0.000
q5_7f	-11.601	0.000	8.328	0.000	203.926	0.000
q6_1f	-10.406	0.000	5.952	0.000	143.717	0.000
q6_2f	-10.135	0.000	5.287	0.000	130.661	0.000
q6_3f	-9.371	0.000	6.875	0.000	135.078	0.000
q6_4f	-10.497	0.000	6.138	0.000	147.853	0.000
q6_5f	-9.907	0.000	4.523	0.000	118.606	0.000
q7_1f	-8.595	0.000	1.570	0.116	76.348	0.000
q7_2f	-8.496	0.000	2.250	0.024	77.253	0.000
q7_3f	-8.687	0.000	2.003	0.045	79.482	0.000
q7_4f	-8.817	0.000	1.577	0.115	80.227	0.000
q7_5f	-6.175	0.000	-2.191	0.028	42.937	0.000
q8_1f	-2.169	0.030	-10.709	0.000	119.378	0.000
q8_2f	-0.878	0.380	-15.368	0.000	236.946	0.000
q8_3f	0.714	0.475	-15.254	0.000	233.186	0.000
q8_4f	0.279	0.781	-17.046	0.000	290.648	0.000
q8_5f	4.465	0.000	-6.582	0.000	63.252	0.000
q8_6f	-5.462	0.000	-5.127	0.000	56.117	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.402

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
417.241	114.354	0.000	2591.410	40.551	0.000	14721.165	0.000



**Uni- and multivariate normality actual posttests heiQ-PP (n=244)**

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1f	4.668	1.074	67.915	-1.258	1.968	1.000	5	6.000	44
q1_2f	4.721	0.988	74.669	-1.277	2.093	1.000	2	6.000	40
q1_3f	5.094	0.877	90.752	-1.810	5.025	1.000	1	6.000	74
q1_4f	4.693	1.046	70.067	-1.075	1.186	1.000	2	6.000	46
q1_5f	4.598	1.105	65.011	-1.368	1.922	1.000	6	6.000	34
q2_1f	4.221	1.502	43.907	-0.567	-0.805	1.000	11	6.000	55
q2_2f	4.471	1.356	51.508	-0.942	0.048	1.000	8	6.000	53
q2_3f	5.102	0.749	106.368	-1.116	2.860	2.000	3	6.000	69
q2_4f	4.660	1.105	65.866	-1.213	1.337	1.000	3	6.000	44
q3_1f	4.680	0.891	82.020	-1.189	2.213	1.000	1	6.000	29
q3_2f	4.643	1.030	70.396	-1.494	2.959	1.000	6	6.000	33
q3_3f	4.447	0.999	69.559	-1.015	1.062	1.000	2	6.000	20
q3_4f	4.902	0.795	96.292	-1.555	4.696	1.000	1	6.000	40
q3_5f	4.512	0.954	73.916	-1.213	1.592	1.000	2	6.000	17
q4_1f	4.906	1.008	76.033	-1.779	4.069	1.000	4	6.000	57
q4_2f	4.939	0.870	88.633	-1.919	5.679	1.000	3	6.000	46
q4_3f	4.631	1.120	64.592	-1.502	2.256	1.000	7	6.000	35
q4_4f	4.475	1.242	56.286	-1.340	1.334	1.000	12	6.000	31
q4_5f	4.602	1.180	60.918	-1.293	1.433	1.000	7	6.000	42
q5_1f	4.877	0.786	96.863	-2.031	6.743	1.000	2	6.000	30
q5_2f	4.975	0.916	84.853	-1.442	3.058	1.000	1	6.000	64
q5_3f	4.836	0.988	76.421	-1.315	2.073	1.000	1	6.000	53
q5_4f	4.832	0.898	84.018	-1.276	2.453	1.000	1	6.000	44
q5_5f	5.389	0.648	129.929	-1.503	5.629	2.000	2	6.000	109
q5_6f	4.803	0.852	88.022	-1.259	2.785	1.000	1	6.000	37
q5_7f	4.848	0.878	86.242	-1.648	4.225	1.000	2	6.000	39
q6_1f	4.988	0.945	82.453	-1.215	1.982	1.000	1	6.000	74
q6_2f	4.861	1.072	70.825	-1.275	1.652	1.000	2	6.000	67
q6_3f	5.143	0.709	113.399	-1.052	2.915	2.000	2	6.000	71
q6_4f	4.730	0.947	77.972	-1.191	1.960	1.000	2	6.000	38
q6_5f	4.566	1.176	60.653	-1.093	0.919	1.000	5	6.000	44
q7_1f	4.361	1.293	52.680	-1.019	0.432	1.000	10	6.000	35
q7_2f	4.430	1.081	64.035	-1.120	1.168	1.000	5	6.000	22
q7_3f	4.496	1.239	56.684	-1.077	0.778	1.000	8	6.000	43
q7_4f	4.504	1.278	55.044	-1.106	0.680	1.000	9	6.000	45
q7_5f	4.033	1.342	46.935	-0.585	-0.339	1.000	13	6.000	28
q8_1f	3.848	1.468	40.960	-0.294	-0.902	1.000	18	6.000	31
q8_2f	3.734	1.393	41.860	-0.048	-0.980	1.000	12	6.000	25
q8_3f	3.336	1.489	35.008	0.095	-1.048	1.000	31	6.000	16
q8_4f	3.574	1.454	38.396	0.019	-1.088	1.000	17	6.000	22
q8_5f	2.934	1.344	34.105	0.561	-0.588	1.000	27	6.000	9
q8_6f	4.123	1.435	44.883	-0.580	-0.576	1.000	15	6.000	39

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1f	-6.579	0.000	3.739	0.000	57.263	0.000
q1_2f	-6.650	0.000	3.870	0.000	59.194	0.000
q1_3f	-8.347	0.000	5.838	0.000	103.760	0.000
q1_4f	-5.873	0.000	2.752	0.006	42.065	0.000
q1_5f	-6.974	0.000	3.690	0.000	62.248	0.000
q2_1f	-3.468	0.001	-4.299	0.000	30.510	0.000
q2_2f	-5.312	0.000	0.295	0.768	28.300	0.000
q2_3f	-6.039	0.000	4.553	0.000	57.195	0.000
q2_4f	-6.413	0.000	2.971	0.003	49.951	0.000
q3_1f	-6.320	0.000	3.988	0.000	55.853	0.000
q3_2f	-7.397	0.000	4.629	0.000	76.151	0.000
q3_3f	-5.622	0.000	2.559	0.010	38.156	0.000
q3_4f	-7.592	0.000	5.683	0.000	89.932	0.000
q3_5f	-6.411	0.000	3.307	0.001	52.039	0.000
q4_1f	-8.261	0.000	5.354	0.000	96.918	0.000
q4_2f	-8.646	0.000	6.118	0.000	112.190	0.000
q4_3f	-7.421	0.000	4.030	0.000	71.318	0.000
q4_4f	-6.876	0.000	2.967	0.003	56.077	0.000
q4_5f	-6.706	0.000	3.102	0.002	54.595	0.000
q5_1f	-8.936	0.000	6.507	0.000	122.190	0.000
q5_2f	-7.223	0.000	4.703	0.000	74.297	0.000
q5_3f	-6.787	0.000	3.849	0.000	60.882	0.000
q5_4f	-6.647	0.000	4.212	0.000	61.918	0.000
q5_5f	-7.426	0.000	6.098	0.000	92.334	0.000
q5_6f	-6.584	0.000	4.494	0.000	63.547	0.000
q5_7f	-7.877	0.000	5.441	0.000	91.642	0.000
q6_1f	-6.419	0.000	3.755	0.000	55.304	0.000
q6_2f	-6.643	0.000	3.381	0.001	55.555	0.000
q6_3f	-5.780	0.000	4.596	0.000	54.527	0.000
q6_4f	-6.329	0.000	3.731	0.000	53.978	0.000
q6_5f	-5.945	0.000	2.322	0.020	40.729	0.000
q7_1f	-5.642	0.000	1.340	0.180	33.623	0.000
q7_2f	-6.055	0.000	2.725	0.006	44.084	0.000
q7_3f	-5.881	0.000	2.067	0.039	38.854	0.000
q7_4f	-5.997	0.000	1.878	0.060	39.494	0.000
q7_5f	-3.570	0.000	-1.198	0.231	14.176	0.001
q8_1f	-1.882	0.060	-5.328	0.000	31.934	0.000
q8_2f	-0.314	0.753	-6.357	0.000	40.508	0.000
q8_3f	0.621	0.535	-7.433	0.000	55.629	0.000
q8_4f	0.121	0.904	-8.189	0.000	67.068	0.000
q8_5f	3.439	0.001	-2.579	0.010	18.480	0.000
q8_6f	-3.541	0.000	-2.501	0.012	18.794	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.274

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
678.618	67.692	0.000	2353.731	22.638	0.000	5094.768	0.000

**Uni- and multivariate normality actual posttests heiQ-PPT (n=150)**

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1f	4.713	0.965	59.837	-1.168	1.882	1.000	1	6.000	24
q1_2f	4.907	0.877	68.495	-1.085	1.865	2.000	4	6.000	34
q1_3f	5.233	0.814	78.698	-1.437	3.430	2.000	3	6.000	61
q1_4f	4.867	0.960	62.084	-1.481	3.046	1.000	1	6.000	32
q1_5f	4.553	1.298	42.960	-1.210	1.012	1.000	7	6.000	31
q2_1f	4.473	1.345	40.745	-0.983	0.219	1.000	5	6.000	32
q2_2f	4.780	1.146	51.077	-1.157	1.134	1.000	2	6.000	41
q2_3f	5.167	0.680	93.100	-0.869	2.575	2.000	1	6.000	45
q2_4f	4.773	1.159	50.423	-1.353	1.566	1.000	2	6.000	37
q3_1f	4.880	0.843	70.914	-1.337	3.674	1.000	1	6.000	28
q3_2f	4.853	0.900	66.023	-1.663	4.763	1.000	2	6.000	27
q3_3f	4.733	0.841	68.948	-1.383	2.830	2.000	6	6.000	16
q3_4f	4.887	0.886	67.518	-1.474	4.335	1.000	2	6.000	31
q3_5f	4.607	0.889	63.455	-1.518	3.750	1.000	2	6.000	12
q4_1f	5.107	0.868	72.044	-1.831	6.257	1.000	2	6.000	48
q4_2f	4.993	0.966	63.319	-1.617	3.983	1.000	2	6.000	44
q4_3f	4.773	1.159	50.423	-1.118	1.152	1.000	3	6.000	43
q4_4f	4.460	1.427	38.288	-1.008	0.263	1.000	10	6.000	36
q4_5f	4.620	1.235	45.808	-1.146	0.895	1.000	4	6.000	33
q5_1f	5.107	0.706	88.573	-0.617	0.664	3.000	4	6.000	42
q5_2f	5.060	0.813	76.236	-1.402	4.483	1.000	1	6.000	42
q5_3f	4.860	0.883	67.443	-1.859	5.508	1.000	2	6.000	24
q5_4f	4.840	0.891	66.566	-1.528	4.163	1.000	2	6.000	26
q5_5f	5.360	0.822	79.908	-2.375	9.549	1.000	2	6.000	73
q5_6f	4.940	0.899	67.289	-1.116	2.516	1.000	1	6.000	40
q5_7f	5.033	0.680	90.697	-1.471	7.729	1.000	1	6.000	30
q6_1f	4.940	0.929	65.160	-1.409	2.940	1.000	1	6.000	37
q6_2f	4.907	0.951	63.205	-1.426	3.007	1.000	1	6.000	36
q6_3f	5.160	0.743	85.106	-1.064	2.322	2.000	1	6.000	48
q6_4f	4.820	0.956	61.758	-1.547	3.629	1.000	2	6.000	28
q6_5f	4.753	1.036	56.202	-1.472	2.722	1.000	2	6.000	28
q7_1f	4.487	1.268	43.345	-1.222	1.068	1.000	7	6.000	24
q7_2f	4.580	1.101	50.959	-1.185	1.188	1.000	2	6.000	21
q7_3f	4.533	1.339	41.451	-1.032	0.288	1.000	5	6.000	34
q7_4f	4.440	1.445	37.642	-1.009	0.103	1.000	9	6.000	34
q7_5f	4.067	1.398	35.619	-0.747	-0.378	1.000	10	6.000	15
q8_1f	3.827	1.558	30.089	-0.269	-1.032	1.000	14	6.000	23
q8_2f	3.727	1.545	29.537	-0.240	-1.159	1.000	14	6.000	16
q8_3f	3.293	1.552	25.982	-0.011	-1.141	1.000	27	6.000	9
q8_4f	3.620	1.509	29.379	-0.115	-1.154	1.000	13	6.000	14
q8_5f	2.913	1.437	24.822	0.360	-0.859	1.000	29	6.000	5
q8_6f	4.433	1.548	35.086	-0.810	-0.505	1.000	9	6.000	45

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1f	-4.993	0.000	3.025	0.002	34.088	0.000
q1_2f	-4.729	0.000	3.010	0.003	31.420	0.000
q1_3f	-5.775	0.000	4.124	0.000	50.354	0.000
q1_4f	-5.895	0.000	3.898	0.000	49.951	0.000
q1_5f	-5.123	0.000	2.061	0.039	30.496	0.000
q2_1f	-4.385	0.000	0.696	0.487	19.713	0.000
q2_2f	-4.960	0.000	2.222	0.026	29.535	0.000
q2_3f	-3.976	0.000	3.585	0.000	28.663	0.000
q2_4f	-5.544	0.000	2.718	0.007	38.118	0.000
q3_1f	-5.497	0.000	4.255	0.000	48.316	0.000
q3_2f	-6.356	0.000	4.755	0.000	63.009	0.000
q3_3f	-5.629	0.000	3.760	0.000	45.822	0.000
q3_4f	-5.877	0.000	4.573	0.000	55.455	0.000
q3_5f	-5.992	0.000	4.294	0.000	54.344	0.000
q4_1f	-6.750	0.000	5.280	0.000	73.433	0.000
q4_2f	-6.244	0.000	4.410	0.000	58.440	0.000
q4_3f	-4.833	0.000	2.244	0.025	28.392	0.000
q4_4f	-4.468	0.000	0.790	0.429	20.592	0.000
q4_5f	-4.924	0.000	1.898	0.058	27.853	0.000
q5_1f	-2.979	0.003	1.543	0.123	11.254	0.004
q5_2f	-5.682	0.000	4.638	0.000	53.792	0.000
q5_3f	-6.812	0.000	5.035	0.000	71.751	0.000
q5_4f	-6.018	0.000	4.495	0.000	56.429	0.000
q5_5f	-7.846	0.000	6.070	0.000	98.414	0.000
q5_6f	-4.828	0.000	3.543	0.000	35.867	0.000
q5_7f	-5.867	0.000	5.680	0.000	66.677	0.000
q6_1f	-5.700	0.000	3.832	0.000	47.175	0.000
q6_2f	-5.746	0.000	3.874	0.000	48.022	0.000
q6_3f	-4.658	0.000	3.397	0.001	33.239	0.000
q6_4f	-6.066	0.000	4.232	0.000	54.706	0.000
q6_5f	-5.869	0.000	3.688	0.000	48.044	0.000
q7_1f	-5.158	0.000	2.136	0.033	31.166	0.000
q7_2f	-5.046	0.000	2.289	0.022	30.703	0.000
q7_3f	-4.550	0.000	0.845	0.398	21.418	0.000
q7_4f	-4.474	0.000	0.425	0.671	20.193	0.000
q7_5f	-3.511	0.000	-1.038	0.299	13.403	0.001
q8_1f	-1.369	0.171	-5.302	0.000	29.982	0.000
q8_2f	-1.227	0.220	-7.071	0.000	51.510	0.000
q8_3f	-0.054	0.957	-6.774	0.000	45.896	0.000
q8_4f	-0.591	0.554	-6.987	0.000	49.164	0.000
q8_5f	1.814	0.070	-3.657	0.000	16.659	0.000
q8_6f	-3.753	0.000	-1.568	0.117	16.544	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.193

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
927.286	50.089	0.000	2205.580	15.611	0.000	2752.666	0.000

**Uni- and multivariate normality actual posttests heiQ-PPR (n=209)**

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
q1_1f	4.837	1.025	68.236	-1.346	2.424	1.000	3	6.000	50
q1_2f	4.914	1.025	69.319	-1.341	2.285	1.000	2	6.000	60
q1_3f	5.244	0.862	87.948	-2.176	7.885	1.000	3	6.000	85
q1_4f	5.048	0.919	79.441	-1.485	3.732	1.000	2	6.000	67
q1_5f	4.641	1.201	55.861	-1.228	1.213	1.000	5	6.000	44
q2_1f	4.469	1.578	40.937	-0.924	-0.362	1.000	14	6.000	65
q2_2f	4.727	1.347	50.732	-1.279	0.997	1.000	9	6.000	66
q2_3f	5.158	0.945	78.898	-2.046	5.911	1.000	3	6.000	78
q2_4f	4.909	1.134	62.608	-1.559	2.859	1.000	6	6.000	67
q3_1f	4.919	0.881	80.669	-1.626	4.513	1.000	2	6.000	43
q3_2f	4.866	1.097	64.135	-1.475	2.372	1.000	3	6.000	57
q3_3f	4.737	1.053	65.048	-1.103	1.360	1.000	2	6.000	45
q3_4f	5.096	0.791	93.152	-1.173	3.517	1.000	1	6.000	64
q3_5f	4.727	1.008	67.781	-1.306	2.466	1.000	3	6.000	38
q4_1f	5.110	0.942	78.446	-1.512	3.072	1.000	1	6.000	77
q4_2f	4.995	1.103	65.479	-1.749	3.350	1.000	4	6.000	70
q4_3f	4.789	1.158	59.813	-1.383	1.942	1.000	5	6.000	55
q4_4f	4.589	1.276	51.991	-0.999	0.191	1.000	3	6.000	48
q4_5f	4.675	1.152	58.681	-1.168	1.013	1.000	3	6.000	43
q5_1f	5.110	0.735	100.465	-1.494	5.769	1.000	1	6.000	56
q5_2f	5.177	0.850	88.001	-1.768	5.801	1.000	2	6.000	77
q5_3f	5.081	0.919	79.947	-1.702	4.998	1.000	3	6.000	70
q5_4f	5.029	0.871	83.458	-1.422	3.552	1.000	1	6.000	59
q5_5f	5.478	0.636	124.517	-1.956	10.376	1.000	1	6.000	110
q5_6f	5.086	0.786	93.568	-1.174	3.609	1.000	1	6.000	62
q5_7f	5.100	0.805	91.564	-1.356	3.539	2.000	5	6.000	63
q6_1f	5.096	0.986	74.738	-1.410	2.426	1.000	1	6.000	82
q6_2f	5.014	0.958	75.657	-1.088	1.102	2.000	5	6.000	70
q6_3f	5.278	0.693	110.089	-1.307	4.166	2.000	2	6.000	79
q6_4f	4.828	1.060	65.831	-1.361	2.283	1.000	3	6.000	52
q6_5f	4.828	1.037	67.285	-1.189	1.221	2.000	12	6.000	50
q7_1f	4.321	1.410	44.299	-0.865	-0.081	1.000	12	6.000	39
q7_2f	4.297	1.311	47.373	-0.744	-0.238	1.000	7	6.000	31
q7_3f	4.522	1.341	48.738	-0.970	0.325	1.000	8	6.000	52
q7_4f	4.531	1.345	48.718	-0.989	0.182	1.000	7	6.000	49
q7_5f	4.124	1.374	43.391	-0.709	-0.335	1.000	12	6.000	25
q8_1f	3.775	1.445	37.764	-0.083	-1.027	1.000	11	6.000	27
q8_2f	3.665	1.520	34.859	0.001	-1.231	1.000	13	6.000	26
q8_3f	3.421	1.511	32.730	0.110	-1.199	1.000	19	6.000	17
q8_4f	3.325	1.497	32.122	0.151	-1.179	1.000	21	6.000	14
q8_5f	2.852	1.411	29.207	0.452	-0.900	1.000	36	6.000	5
q8_6f	4.139	1.433	41.761	-0.454	-0.767	1.000	10	6.000	38

Test of Univariate Normality for Continuous Variables

Variable	Skewness		Kurtosis		Skewness and Kurtosis	
	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
q1_1f	-6.420	0.000	3.938	0.000	56.721	0.000
q1_2f	-6.405	0.000	3.817	0.000	55.592	0.000
q1_3f	-8.662	0.000	6.456	0.000	116.716	0.000
q1_4f	-6.861	0.000	4.851	0.000	70.605	0.000
q1_5f	-6.020	0.000	2.628	0.009	43.145	0.000
q2_1f	-4.866	0.000	-1.194	0.232	25.099	0.000
q2_2f	-6.195	0.000	2.309	0.021	43.716	0.000
q2_3f	-8.363	0.000	5.845	0.000	104.103	0.000
q2_4f	-7.082	0.000	4.282	0.000	68.496	0.000
q3_1f	-7.275	0.000	5.262	0.000	80.623	0.000
q3_2f	-6.828	0.000	3.893	0.000	61.776	0.000
q3_3f	-5.571	0.000	2.825	0.005	39.011	0.000
q3_4f	-5.827	0.000	4.724	0.000	56.262	0.000
q3_5f	-6.287	0.000	3.973	0.000	55.312	0.000
q4_1f	-6.942	0.000	4.434	0.000	67.850	0.000
q4_2f	-7.615	0.000	4.620	0.000	79.332	0.000
q4_3f	-6.540	0.000	3.491	0.000	54.961	0.000
q4_4f	-5.169	0.000	0.690	0.490	27.192	0.000
q4_5f	-5.808	0.000	2.334	0.020	39.182	0.000
q5_1f	-6.888	0.000	5.793	0.000	81.005	0.000
q5_2f	-7.666	0.000	5.805	0.000	92.465	0.000
q5_3f	-7.486	0.000	5.484	0.000	86.117	0.000
q5_4f	-6.663	0.000	4.745	0.000	66.918	0.000
q5_5f	-8.146	0.000	7.019	0.000	115.614	0.000
q5_6f	-5.828	0.000	4.779	0.000	56.806	0.000
q5_7f	-6.452	0.000	4.737	0.000	64.069	0.000
q6_1f	-6.627	0.000	3.939	0.000	59.431	0.000
q6_2f	-5.514	0.000	2.468	0.014	36.502	0.000
q6_3f	-6.292	0.000	5.089	0.000	65.483	0.000
q6_4f	-6.469	0.000	3.815	0.000	56.399	0.000
q6_5f	-5.882	0.000	2.638	0.008	41.560	0.000
q7_1f	-4.617	0.000	-0.113	0.910	21.330	0.000
q7_2f	-4.082	0.000	-0.679	0.497	17.121	0.000
q7_3f	-5.053	0.000	1.023	0.306	26.575	0.000
q7_4f	-5.131	0.000	0.665	0.506	26.767	0.000
q7_5f	-3.917	0.000	-1.074	0.283	16.500	0.000
q8_1f	-0.501	0.616	-6.436	0.000	41.675	0.000
q8_2f	0.006	0.995	-10.856	0.000	117.858	0.000
q8_3f	0.666	0.505	-9.877	0.000	98.001	0.000
q8_4f	0.908	0.364	-9.362	0.000	88.477	0.000
q8_5f	2.627	0.009	-4.852	0.000	30.446	0.000
q8_6f	-2.638	0.008	-3.622	0.000	20.075	0.000

Test of Multivariate Normality for Continuous Variables

Relative Multivariate Kurtosis = 1.218

Skewness			Kurtosis			Skewness and Kurtosis	
Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
696.298	54.557	0.000	2250.975	19.186	0.000	3344.586	0.000

**Appendix 8** Bivariate normality tests of the following heiQ data: pretests (n=666), retrospective pretests heiQ-PPR (n=189), posttests (n=603), posttests heiQ-PP (n=244), posttests heiQ-PPT (n=150), and posttests heiQ-PPR (n=209)

(Only extracts are listed as the presentation otherwise would have exceeded 80 pages. The complete results can be obtained from the researcher.)

**Bivariate normality heiQ pretests (n=666)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs.	Variable	Correlation	Test of Model			Test of Close Fit	
			Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2 vs.	q1_1	0.627 (PC)	64.109	24	0.000	0.042	1.000
q1_3 vs.	q1_1	0.634 (PC)	58.842	24	0.000	0.039	1.000
q1_3 vs.	q1_2	0.554 (PC)	65.648	24	0.000	0.043	1.000
q1_4 vs.	q1_1	0.632 (PC)	94.017	24	0.000	0.055	1.000
q1_4 vs.	q1_2	0.522 (PC)	69.218	24	0.000	0.045	1.000
q1_4 vs.	q1_3	0.551 (PC)	57.057	24	0.000	0.038	1.000
q1_5 vs.	q1_1	0.672 (PC)	85.632	24	0.000	0.052	1.000
q1_5 vs.	q1_2	0.553 (PC)	78.501	24	0.000	0.049	1.000
q1_5 vs.	q1_3	0.598 (PC)	68.719	24	0.000	0.044	1.000
q1_5 vs.	q1_4	0.582 (PC)	63.807	24	0.000	0.042	1.000
q2_1 vs.	q1_1	0.315 (PC)	40.334	24	0.020	0.027	1.000
q2_1 vs.	q1_2	0.287 (PC)	49.892	24	0.001	0.034	1.000
q2_1 vs.	q1_3	0.312 (PC)	52.394	24	0.001	0.035	1.000
q2_1 vs.	q1_4	0.348 (PC)	40.776	24	0.018	0.027	1.000
q2_1 vs.	q1_5	0.362 (PC)	40.427	24	0.019	0.027	1.000
q2_2 vs.	q1_1	0.340 (PC)	63.666	24	0.000	0.042	1.000
q2_2 vs.	q1_2	0.314 (PC)	63.870	24	0.000	0.042	1.000
q2_2 vs.	q1_3	0.296 (PC)	94.579	24	0.000	0.056	1.000
q2_2 vs.	q1_4	0.434 (PC)	73.225	24	0.000	0.046	1.000
q2_2 vs.	q1_5	0.398 (PC)	34.665	24	0.074	0.022	1.000
q2_2 vs.	q2_1	0.608 (PC)	129.478	24	0.000	0.068	1.000
q2_3 vs.	q1_1	0.280 (PC)	31.047	24	0.152	0.018	1.000
q2_3 vs.	q1_2	0.373 (PC)	50.424	24	0.001	0.034	1.000
q2_3 vs.	q1_3	0.320 (PC)	51.676	24	0.001	0.035	1.000
q2_3 vs.	q1_4	0.308 (PC)	43.442	24	0.009	0.029	1.000
q2_3 vs.	q1_5	0.289 (PC)	48.075	24	0.002	0.033	1.000
q2_3 vs.	q2_1	0.588 (PC)	113.026	24	0.000	0.063	1.000
q2_3 vs.	q2_2	0.563 (PC)	85.953	24	0.000	0.052	1.000
q2_4 vs.	q1_1	0.415 (PC)	69.257	24	0.000	0.045	1.000
q2_4 vs.	q1_2	0.424 (PC)	63.848	24	0.000	0.042	1.000
q2_4 vs.	q1_3	0.434 (PC)	112.289	24	0.000	0.062	1.000
q2_4 vs.	q1_4	0.492 (PC)	77.447	24	0.000	0.048	1.000
q2_4 vs.	q1_5	0.467 (PC)	63.815	24	0.000	0.042	1.000
q2_4 vs.	q2_1	0.692 (PC)	90.407	24	0.000	0.054	1.000
q2_4 vs.	q2_2	0.652 (PC)	165.954	24	0.000	0.079	0.999
q2_4 vs.	q2_3	0.601 (PC)	167.833	24	0.000	0.079	0.998
q3_1 vs.	q1_1	0.394 (PC)	75.285	24	0.000	0.047	1.000
q3_1 vs.	q1_2	0.345 (PC)	88.160	24	0.000	0.053	1.000
q3_1 vs.	q1_3	0.389 (PC)	83.477	24	0.000	0.051	1.000
q3_1 vs.	q1_4	0.491 (PC)	84.703	24	0.000	0.052	1.000
q3_1 vs.	q1_5	0.428 (PC)	72.644	24	0.000	0.046	1.000
q3_1 vs.	q2_1	0.297 (PC)	63.209	24	0.000	0.041	1.000
q3_1 vs.	q2_2	0.309 (PC)	87.833	24	0.000	0.053	1.000
q3_1 vs.	q2_3	0.262 (PC)	57.510	24	0.000	0.038	1.000
q3_1 vs.	q2_4	0.450 (PC)	94.631	24	0.000	0.056	1.000
q3_2 vs.	q1_1	0.247 (PC)	48.985	24	0.002	0.033	1.000

Percentage of Tests Exceeding 0.5% Significance Level: 0.0%

Percentage of Tests Exceeding 1.0% Significance Level: 0.0%

Percentage of Tests Exceeding 5.0% Significance Level: 0.0%

**Bivariate normality heiQ-PPR retrospective pretests (n=189)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs. Variable	Correlation	Test of Model			Test of Close Fit	
		Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2r vs. q1_1r	0.704 (PC)	27.808	24	0.268	0.029	1.000
q1_3r vs. q1_1r	0.648 (PC)	38.895	24	0.028	0.057	0.989
q1_3r vs. q1_2r	0.664 (PC)	31.904	24	0.129	0.042	0.998
q1_4r vs. q1_1r	0.746 (PC)	52.363	24	0.001	0.079	0.874
q1_4r vs. q1_2r	0.639 (PC)	44.718	24	0.006	0.068	0.961
q1_4r vs. q1_3r	0.612 (PC)	41.865	24	0.013	0.063	0.978
q1_5r vs. q1_1r	0.724 (PC)	43.460	24	0.009	0.065	0.969
q1_5r vs. q1_2r	0.598 (PC)	30.202	24	0.178	0.037	0.999
q1_5r vs. q1_3r	0.554 (PC)	34.201	24	0.081	0.047	0.997
q1_5r vs. q1_4r	0.728 (PC)	45.028	24	0.006	0.068	0.959
q2_1r vs. q1_1r	0.366 (PC)	17.746	24	0.815	0.000	1.000
q2_1r vs. q1_2r	0.316 (PC)	32.622	24	0.112	0.044	0.998
q2_1r vs. q1_3r	0.329 (PC)	25.997	24	0.353	0.021	1.000
q2_1r vs. q1_4r	0.474 (PC)	40.473	24	0.019	0.060	0.984
q2_1r vs. q1_5r	0.498 (PC)	18.966	24	0.754	0.000	1.000
q2_2r vs. q1_1r	0.470 (PC)	38.385	24	0.032	0.056	0.990
q2_2r vs. q1_2r	0.444 (PC)	23.937	24	0.465	0.000	1.000
q2_2r vs. q1_3r	0.402 (PC)	41.674	24	0.014	0.062	0.979
q2_2r vs. q1_4r	0.642 (PC)	25.168	24	0.397	0.016	1.000
q2_2r vs. q1_5r	0.593 (PC)	35.689	24	0.059	0.051	0.995
q2_2r vs. q2_1r	0.718 (PC)	28.452	24	0.241	0.031	1.000
q2_3r vs. q1_1r	0.378 (PC)	29.315	24	0.208	0.034	0.999
q2_3r vs. q1_2r	0.420 (PC)	28.085	24	0.256	0.030	1.000
q2_3r vs. q1_3r	0.348 (PC)	36.852	24	0.045	0.053	0.993
q2_3r vs. q1_4r	0.454 (PC)	34.246	24	0.080	0.048	0.997
q2_3r vs. q1_5r	0.435 (PC)	24.720	24	0.421	0.013	1.000
q2_3r vs. q2_1r	0.601 (PC)	31.461	24	0.141	0.041	0.999
q2_3r vs. q2_2r	0.684 (PC)	22.813	24	0.531	0.000	1.000
q2_4r vs. q1_1r	0.407 (PC)	14.813	24	0.926	0.000	1.000
q2_4r vs. q1_2r	0.475 (PC)	20.500	24	0.668	0.000	1.000
q2_4r vs. q1_3r	0.471 (PC)	43.555	24	0.009	0.066	0.969
q2_4r vs. q1_4r	0.600 (PC)	46.520	24	0.004	0.070	0.947
q2_4r vs. q1_5r	0.474 (PC)	31.884	24	0.130	0.042	0.998
q2_4r vs. q2_1r	0.633 (PC)	49.230	24	0.002	0.075	0.918
q2_4r vs. q2_2r	0.686 (PC)	24.949	24	0.409	0.014	1.000
q2_4r vs. q2_3r	0.661 (PC)	24.146	24	0.453	0.006	1.000
q3_1r vs. q1_1r	0.479 (PC)	42.316	24	0.012	0.064	0.976
q3_1r vs. q1_2r	0.485 (PC)	38.168	24	0.033	0.056	0.990
q3_1r vs. q1_3r	0.401 (PC)	37.474	24	0.039	0.055	0.992
q3_1r vs. q1_4r	0.600 (PC)	50.092	24	0.001	0.076	0.907
q3_1r vs. q1_5r	0.522 (PC)	26.044	24	0.351	0.021	1.000
q3_1r vs. q2_1r	0.366 (PC)	29.940	24	0.187	0.036	0.999
q3_1r vs. q2_2r	0.431 (PC)	50.996	24	0.001	0.077	0.895
q3_1r vs. q2_3r	0.403 (PC)	29.114	24	0.216	0.034	0.999
q3_1r vs. q2_4r	0.445 (PC)	75.276	24	0.000	0.106	0.329
q3_2r vs. q1_1r	0.378 (PC)	43.790	24	0.008	0.066	0.967
q3_2r vs. q1_2r	0.502 (PC)	33.073	24	0.103	0.045	0.998
q3_2r vs. q1_3r	0.465 (PC)	39.014	24	0.027	0.058	0.988
q3_2r vs. q1_4r	0.392 (PC)	25.668	24	0.370	0.019	1.000
q3_2r vs. q1_5r	0.398 (PC)	38.319	24	0.032	0.056	0.990
q3_2r vs. q2_1r	0.134 (PC)	26.230	24	0.342	0.022	1.000
q3_2r vs. q2_2r	0.250 (PC)	31.005	24	0.154	0.039	0.999
q3_2r vs. q2_3r	0.281 (PC)	24.202	24	0.450	0.007	1.000

Percentage of Tests Exceeding 0.5% Significance Level: 0.0%

Percentage of Tests Exceeding 1.0% Significance Level: 0.0%

Percentage of Tests Exceeding 5.0% Significance Level: 0.0%



**Bivariate normality heiQ actual posttests (n=603)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs.	Variable	Correlation	Test of Model			Test of Close Fit	
			Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2f vs.	q1_1f	0.599 (PC)	58.861	24	0.000	0.049	1.000
q1_3f vs.	q1_1f	0.596 (PC)	58.573	24	0.000	0.049	1.000
q1_3f vs.	q1_2f	0.567 (PC)	33.148	24	0.101	0.025	1.000
q1_4f vs.	q1_1f	0.611 (PC)	43.458	24	0.009	0.037	1.000
q1_4f vs.	q1_2f	0.564 (PC)	51.579	24	0.001	0.044	1.000
q1_4f vs.	q1_3f	0.540 (PC)	30.730	24	0.162	0.022	1.000
q1_5f vs.	q1_1f	0.642 (PC)	62.606	24	0.000	0.052	1.000
q1_5f vs.	q1_2f	0.476 (PC)	54.808	24	0.000	0.046	1.000
q1_5f vs.	q1_3f	0.537 (PC)	47.138	24	0.003	0.040	1.000
q1_5f vs.	q1_4f	0.480 (PC)	35.794	24	0.057	0.029	1.000
q2_1f vs.	q1_1f	0.318 (PC)	44.950	24	0.006	0.038	1.000
q2_1f vs.	q1_2f	0.370 (PC)	44.092	24	0.007	0.037	1.000
q2_1f vs.	q1_3f	0.294 (PC)	46.346	24	0.004	0.039	1.000
q2_1f vs.	q1_4f	0.413 (PC)	46.112	24	0.004	0.039	1.000
q2_1f vs.	q1_5f	0.378 (PC)	44.050	24	0.008	0.037	1.000
q2_2f vs.	q1_1f	0.355 (PC)	45.234	24	0.005	0.038	1.000
q2_2f vs.	q1_2f	0.418 (PC)	42.428	24	0.012	0.036	1.000
q2_2f vs.	q1_3f	0.358 (PC)	51.507	24	0.001	0.044	1.000
q2_2f vs.	q1_4f	0.478 (PC)	52.405	24	0.001	0.044	1.000
q2_2f vs.	q1_5f	0.389 (PC)	48.152	24	0.002	0.041	1.000
q2_2f vs.	q2_1f	0.662 (PC)	48.990	24	0.002	0.042	1.000
q2_3f vs.	q1_1f	0.329 (PC)	44.633	24	0.006	0.038	1.000
q2_3f vs.	q1_2f	0.466 (PC)	42.411	24	0.012	0.036	1.000
q2_3f vs.	q1_3f	0.377 (PC)	62.667	24	0.000	0.052	1.000
q2_3f vs.	q1_4f	0.468 (PC)	39.104	24	0.027	0.032	1.000
q2_3f vs.	q1_5f	0.357 (PC)	36.558	24	0.048	0.029	1.000
q2_3f vs.	q2_1f	0.625 (PC)	38.731	24	0.029	0.032	1.000
q2_3f vs.	q2_2f	0.609 (PC)	45.509	24	0.005	0.039	1.000
q2_4f vs.	q1_1f	0.332 (PC)	50.833	24	0.001	0.043	1.000
q2_4f vs.	q1_2f	0.468 (PC)	32.992	24	0.104	0.025	1.000
q2_4f vs.	q1_3f	0.386 (PC)	61.322	24	0.000	0.051	1.000
q2_4f vs.	q1_4f	0.503 (PC)	72.047	24	0.000	0.058	1.000
q2_4f vs.	q1_5f	0.396 (PC)	55.023	24	0.000	0.046	1.000
q2_4f vs.	q2_1f	0.648 (PC)	70.505	24	0.000	0.057	1.000
q2_4f vs.	q2_2f	0.639 (PC)	63.344	24	0.000	0.052	1.000
q2_4f vs.	q2_3f	0.658 (PC)	60.999	24	0.000	0.051	1.000
q3_1f vs.	q1_1f	0.431 (PC)	38.885	24	0.028	0.032	1.000
q3_1f vs.	q1_2f	0.469 (PC)	37.245	24	0.041	0.030	1.000
q3_1f vs.	q1_3f	0.487 (PC)	54.363	24	0.000	0.046	1.000
q3_1f vs.	q1_4f	0.494 (PC)	76.238	24	0.000	0.060	1.000
q3_1f vs.	q1_5f	0.465 (PC)	40.716	24	0.018	0.034	1.000
q3_1f vs.	q2_1f	0.328 (PC)	46.075	24	0.004	0.039	1.000
q3_1f vs.	q2_2f	0.383 (PC)	82.826	24	0.000	0.064	1.000
q3_1f vs.	q2_3f	0.392 (PC)	46.636	24	0.004	0.040	1.000
q3_1f vs.	q2_4f	0.428 (PC)	75.903	24	0.000	0.060	1.000
q8_2f vs.	q7_4f	0.306 (PC)	49.881	24	0.001	0.042	1.000
q8_2f vs.	q7_5f	0.213 (PC)	55.377	24	0.000	0.047	1.000
<b>q8_2f vs.</b>	<b>q8_1f</b>	<b>0.722 (PC)</b>	<b>240.359</b>	<b>24</b>	<b>0.000</b>	<b>0.122</b>	<b>0.004</b>
<b>W_A_R_N_I_N_G: Underlying bivariate normality may not hold, see BTS-file</b>							
q8_3f vs.	q1_1f	0.344 (PC)	23.660	24	0.481	0.000	1.000
q8_3f vs.	q1_2f	0.305 (PC)	24.443	24	0.436	0.006	1.000
q8_3f vs.	q1_3f	0.268 (PC)	54.332	24	0.000	0.046	1.000
q8_3f vs.	q1_4f	0.254 (PC)	25.102	24	0.400	0.009	1.000

Percentage of Tests Exceeding 0.5% Significance Level: 0.1%

Percentage of Tests Exceeding 1.0% Significance Level: 0.1%

Percentage of Tests Exceeding 5.0% Significance Level: 0.1%

**Bivariate normality heiQ actual posttests heiQ-PP (n=244)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs.	Variable	Correlation	Test of Model			Test of Close Fit	
			Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2f vs.	q1_1f	0.579 (PC)	46.876	24	0.003	0.063	0.992
q1_3f vs.	q1_1f	0.604 (PC)	37.853	24	0.036	0.049	0.999
q1_3f vs.	q1_2f	0.505 (PC)	24.538	24	0.431	0.010	1.000
q1_4f vs.	q1_1f	0.571 (PC)	28.620	24	0.235	0.028	1.000
q1_4f vs.	q1_2f	0.554 (PC)	27.101	24	0.300	0.023	1.000
q1_4f vs.	q1_3f	0.612 (PC)	14.565	24	0.933	0.000	1.000
q1_5f vs.	q1_1f	0.664 (PC)	27.388	24	0.287	0.024	1.000
q1_5f vs.	q1_2f	0.464 (PC)	27.003	24	0.304	0.023	1.000
q1_5f vs.	q1_3f	0.556 (PC)	35.321	24	0.064	0.044	1.000
q1_5f vs.	q1_4f	0.419 (PC)	27.290	24	0.291	0.024	1.000
q2_1f vs.	q1_1f	0.367 (PC)	40.828	24	0.017	0.054	0.998
q2_1f vs.	q1_2f	0.429 (PC)	25.086	24	0.401	0.014	1.000
q2_1f vs.	q1_3f	0.282 (PC)	29.320	24	0.208	0.030	1.000
q2_1f vs.	q1_4f	0.423 (PC)	25.821	24	0.362	0.018	1.000
q2_1f vs.	q1_5f	0.395 (PC)	44.026	24	0.008	0.058	0.996
q2_2f vs.	q1_1f	0.338 (PC)	41.809	24	0.014	0.055	0.998
q2_2f vs.	q1_2f	0.412 (PC)	36.278	24	0.052	0.046	1.000
q2_2f vs.	q1_3f	0.317 (PC)	33.600	24	0.092	0.040	1.000
q2_2f vs.	q1_4f	0.423 (PC)	38.466	24	0.031	0.050	0.999
q2_2f vs.	q1_5f	0.367 (PC)	41.543	24	0.015	0.055	0.998
q2_2f vs.	q2_1f	0.653 (PC)	31.114	24	0.151	0.035	1.000
q2_3f vs.	q1_1f	0.319 (PC)	29.092	19	0.065	0.047	0.998
q2_3f vs.	q1_2f	0.465 (PC)	30.736	19	0.043	0.050	0.997
q2_3f vs.	q1_3f	0.402 (PC)	30.822	19	0.042	0.050	0.997
q2_3f vs.	q1_4f	0.442 (PC)	28.602	19	0.073	0.046	0.998
q2_3f vs.	q1_5f	0.377 (PC)	29.849	19	0.054	0.048	0.998
q2_3f vs.	q2_1f	0.637 (PC)	25.083	19	0.158	0.036	1.000
q2_3f vs.	q2_2f	0.629 (PC)	26.919	19	0.107	0.041	0.999
q2_4f vs.	q1_1f	0.345 (PC)	49.124	24	0.002	0.066	0.987
q2_4f vs.	q1_2f	0.496 (PC)	31.358	24	0.144	0.035	1.000
q2_4f vs.	q1_3f	0.330 (PC)	26.800	24	0.314	0.022	1.000
q2_4f vs.	q1_4f	0.501 (PC)	46.736	24	0.004	0.062	0.992
q2_4f vs.	q1_5f	0.405 (PC)	47.511	24	0.003	0.063	0.991
q2_4f vs.	q2_1f	0.714 (PC)	53.071	24	0.001	0.070	0.972
q2_4f vs.	q2_2f	0.631 (PC)	54.547	24	0.000	0.072	0.964
q2_4f vs.	q2_3f	0.616 (PC)	27.704	19	0.089	0.043	0.999
q3_1f vs.	q1_1f	0.417 (PC)	28.856	24	0.226	0.029	1.000
q3_1f vs.	q1_2f	0.516 (PC)	16.902	24	0.853	0.000	1.000
q3_1f vs.	q1_3f	0.506 (PC)	31.749	24	0.133	0.036	1.000
q3_1f vs.	q1_4f	0.508 (PC)	47.888	24	0.003	0.064	0.990
q3_1f vs.	q1_5f	0.449 (PC)	38.299	24	0.032	0.049	0.999
q3_1f vs.	q2_1f	0.337 (PC)	24.220	24	0.449	0.006	1.000
q8_6f vs.	q6_5f	0.081 (PC)	30.274	24	0.176	0.033	1.000
q8_6f vs.	q7_1f	0.252 (PC)	21.309	24	0.620	0.000	1.000
q8_6f vs.	q7_2f	0.208 (PC)	36.130	24	0.053	0.046	1.000
q8_6f vs.	q7_3f	0.173 (PC)	24.784	24	0.418	0.012	1.000
q8_6f vs.	q7_4f	0.289 (PC)	42.710	24	0.011	0.057	0.997
q8_6f vs.	q7_5f	0.217 (PC)	49.155	24	0.002	0.066	0.987
<b>q8_6f vs.</b>	<b>q8_1f</b>	<b>0.641 (PC)</b>	<b>131.383</b>	<b>24</b>	<b>0.000</b>	<b>0.135</b>	<b>0.005</b>
<b>W_A_R_N_I_N_G: Underlying bivariate normality may not hold, see BTS-file</b>							
q8_6f vs.	q8_2f	0.643 (PC)	91.089	24	0.000	0.107	0.292
q8_6f vs.	q8_3f	0.551 (PC)	70.108	24	0.000	0.089	0.763
q8_6f vs.	q8_4f	0.609 (PC)	62.287	24	0.000	0.081	0.893

Percentage of Tests Exceeding 0.5% Significance Level: 0.1%

Percentage of Tests Exceeding 1.0% Significance Level: 0.1%

Percentage of Tests Exceeding 5.0% Significance Level: 0.1%

**Bivariate normality heiQ actual posttests heiQ-PPT (n=150)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs.	Variable	Correlation	Test of Model			Test of Close Fit	
			Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2f vs.	q1_1f	0.535 (PC)	28.820	19	0.069	0.059	0.951
q1_3f vs.	q1_1f	0.549 (PC)	33.869	19	0.019	0.072	0.872
q1_3f vs.	q1_2f	0.573 (PC)	17.899	15	0.268	0.036	0.981
q1_4f vs.	q1_1f	0.543 (PC)	28.271	24	0.249	0.034	0.997
q1_4f vs.	q1_2f	0.452 (PC)	31.116	19	0.039	0.065	0.921
q1_4f vs.	q1_3f	0.332 (PC)	20.404	19	0.371	0.022	0.996
q1_5f vs.	q1_1f	0.624 (PC)	35.411	24	0.063	0.056	0.976
q1_5f vs.	q1_2f	0.460 (PC)	42.736	19	0.001	0.091	0.626
q1_5f vs.	q1_3f	0.545 (PC)	19.535	19	0.423	0.014	0.997
q1_5f vs.	q1_4f	0.468 (PC)	37.483	24	0.039	0.061	0.962
q2_1f vs.	q1_1f	0.347 (PC)	29.928	24	0.187	0.041	0.994
q2_1f vs.	q1_2f	0.463 (PC)	20.296	19	0.377	0.021	0.996
q2_1f vs.	q1_3f	0.421 (PC)	26.352	19	0.121	0.051	0.973
q2_1f vs.	q1_4f	0.290 (PC)	30.332	24	0.174	0.042	0.994
q2_1f vs.	q1_5f	0.355 (PC)	33.198	24	0.100	0.051	0.986
q2_2f vs.	q1_1f	0.308 (PC)	32.930	24	0.106	0.050	0.987
q2_2f vs.	q1_2f	0.412 (PC)	20.766	19	0.350	0.025	0.995
q2_2f vs.	q1_3f	0.465 (PC)	20.430	19	0.369	0.022	0.996
q2_2f vs.	q1_4f	0.414 (PC)	30.283	24	0.176	0.042	0.994
q2_2f vs.	q1_5f	0.392 (PC)	38.144	24	0.034	0.063	0.957
q2_2f vs.	q2_1f	0.561 (PC)	34.093	24	0.083	0.053	0.982
q2_3f vs.	q1_1f	0.184 (PC)	15.414	19	0.696	0.000	1.000
q2_3f vs.	q1_2f	0.514 (PC)	23.583	15	0.073	0.062	0.911
q2_3f vs.	q1_3f	0.376 (PC)	16.060	15	0.378	0.022	0.991
q2_3f vs.	q1_4f	0.258 (PC)	19.119	19	0.449	0.006	0.998
q2_3f vs.	q1_5f	0.327 (PC)	18.863	19	0.466	0.000	0.998
q2_3f vs.	q2_1f	0.513 (PC)	19.259	19	0.440	0.010	0.998
q2_3f vs.	q2_2f	0.423 (PC)	17.597	19	0.549	0.000	0.999
q2_4f vs.	q1_1f	0.269 (PC)	23.599	24	0.485	0.000	0.999
q2_4f vs.	q1_2f	0.384 (PC)	23.875	19	0.201	0.041	0.987
q2_4f vs.	q1_3f	0.462 (PC)	20.905	19	0.342	0.026	0.995
q2_4f vs.	q1_4f	0.333 (PC)	36.861	24	0.045	0.060	0.967
q2_4f vs.	q1_5f	0.409 (PC)	54.963	24	0.000	0.093	0.620
q2_4f vs.	q2_1f	0.543 (PC)	25.971	24	0.355	0.023	0.999
q2_4f vs.	q2_2f	0.607 (PC)	31.711	24	0.134	0.046	0.991
q2_4f vs.	q2_3f	0.579 (PC)	29.671	19	0.056	0.061	0.941
q3_1f vs.	q1_1f	0.326 (PC)	27.419	24	0.285	0.031	0.998
q3_1f vs.	q1_2f	0.398 (PC)	17.380	19	0.564	0.000	0.999
q3_1f vs.	q1_3f	0.447 (PC)	29.074	19	0.065	0.059	0.948
q3_1f vs.	q1_4f	0.321 (PC)	31.607	24	0.137	0.046	0.991
q3_1f vs.	q1_5f	0.498 (PC)	24.869	24	0.413	0.016	0.999
q3_1f vs.	q2_1f	0.392 (PC)	24.050	24	0.459	0.004	0.999
q3_1f vs.	q2_2f	0.568 (PC)	25.209	24	0.394	0.018	0.999
q3_1f vs.	q2_3f	0.362 (PC)	22.254	19	0.272	0.034	0.992
q3_1f vs.	q2_4f	0.613 (PC)	19.805	24	0.708	0.000	1.000
q3_2f vs.	q1_1f	0.081 (PC)	18.359	24	0.785	0.000	1.000
q3_2f vs.	q1_2f	0.219 (PC)	22.593	19	0.256	0.036	0.991
q3_2f vs.	q1_3f	0.419 (PC)	30.342	19	0.048	0.063	0.932
q3_2f vs.	q1_4f	0.296 (PC)	18.361	24	0.785	0.000	1.000
q3_2f vs.	q1_5f	0.227 (PC)	29.422	24	0.205	0.039	0.995
q3_2f vs.	q2_1f	0.139 (PC)	25.082	24	0.401	0.017	0.999
q3_2f vs.	q2_2f	0.376 (PC)	33.340	24	0.097	0.051	0.985
q3_2f vs.	q2_3f	0.281 (PC)	24.560	19	0.176	0.044	0.984

Percentage of Tests Exceeding 0.5% Significance Level: 0.0%

Percentage of Tests Exceeding 1.0% Significance Level: 0.0%

Percentage of Tests Exceeding 5.0% Significance Level: 0.0%

**Bivariate normality heiQ actual posttests heiQ-PPR (n=209)**

Correlations and Test Statistics

(PE=Pearson Product Moment, PC=Polychoric, PS=Polyserial)

Variable vs.	Variable	Correlation	Test of Model			Test of Close Fit	
			Chi-Squ.	D.F.	P-Value	RMSEA	P-Value
q1_2f vs.	q1_1f	0.652 (PC)	16.438	24	0.872	0.000	1.000
q1_3f vs.	q1_1f	0.624 (PC)	32.890	24	0.106	0.042	0.999
q1_3f vs.	q1_2f	0.612 (PC)	25.615	24	0.373	0.018	1.000
q1_4f vs.	q1_1f	0.695 (PC)	29.553	24	0.200	0.033	1.000
q1_4f vs.	q1_2f	0.630 (PC)	20.458	24	0.670	0.000	1.000
q1_4f vs.	q1_3f	0.592 (PC)	20.532	24	0.666	0.000	1.000
q1_5f vs.	q1_1f	0.650 (PC)	46.308	24	0.004	0.067	0.974
q1_5f vs.	q1_2f	0.505 (PC)	25.155	24	0.397	0.015	1.000
q1_5f vs.	q1_3f	0.528 (PC)	20.005	24	0.697	0.000	1.000
q1_5f vs.	q1_4f	0.568 (PC)	26.603	24	0.323	0.023	1.000
q2_1f vs.	q1_1f	0.239 (PC)	18.170	24	0.795	0.000	1.000
q2_1f vs.	q1_2f	0.247 (PC)	19.477	24	0.726	0.000	1.000
q2_1f vs.	q1_3f	0.223 (PC)	31.759	24	0.133	0.039	0.999
q2_1f vs.	q1_4f	0.462 (PC)	26.391	24	0.334	0.022	1.000
q2_1f vs.	q1_5f	0.372 (PC)	28.705	24	0.231	0.031	1.000
q2_2f vs.	q1_1f	0.398 (PC)	41.033	24	0.017	0.058	0.992
q2_2f vs.	q1_2f	0.413 (PC)	25.534	24	0.377	0.017	1.000
q2_2f vs.	q1_3f	0.325 (PC)	41.382	24	0.015	0.059	0.991
q2_2f vs.	q1_4f	0.563 (PC)	29.837	24	0.190	0.034	1.000
q2_2f vs.	q1_5f	0.417 (PC)	39.316	24	0.025	0.055	0.994
q2_2f vs.	q2_1f	0.718 (PC)	29.508	24	0.202	0.033	1.000
q2_3f vs.	q1_1f	0.415 (PC)	34.828	24	0.071	0.046	0.998
q2_3f vs.	q1_2f	0.438 (PC)	31.188	24	0.148	0.038	1.000
q2_3f vs.	q1_3f	0.364 (PC)	35.355	24	0.063	0.048	0.998
q2_3f vs.	q1_4f	0.607 (PC)	22.241	24	0.565	0.000	1.000
q2_3f vs.	q1_5f	0.375 (PC)	19.265	24	0.738	0.000	1.000
q2_3f vs.	q2_1f	0.665 (PC)	24.579	24	0.429	0.011	1.000
q2_3f vs.	q2_2f	0.690 (PC)	25.702	24	0.368	0.018	1.000
q2_4f vs.	q1_1f	0.341 (PC)	36.046	24	0.054	0.049	0.998
q2_4f vs.	q1_2f	0.486 (PC)	25.266	24	0.391	0.016	1.000
q2_4f vs.	q1_3f	0.373 (PC)	32.939	24	0.105	0.042	0.999
q2_4f vs.	q1_4f	0.597 (PC)	36.077	24	0.054	0.049	0.998
q2_4f vs.	q1_5f	0.371 (PC)	30.831	24	0.159	0.037	1.000
q2_4f vs.	q2_1f	0.644 (PC)	40.851	24	0.017	0.058	0.992
q2_4f vs.	q2_2f	0.656 (PC)	35.631	24	0.060	0.048	0.998
q2_4f vs.	q2_3f	0.740 (PC)	27.641	24	0.275	0.027	1.000
q3_1f vs.	q1_1f	0.513 (PC)	32.810	24	0.108	0.042	0.999
q3_1f vs.	q1_2f	0.448 (PC)	32.764	24	0.109	0.042	0.999
q3_1f vs.	q1_3f	0.495 (PC)	32.822	24	0.108	0.042	0.999
q3_1f vs.	q1_4f	0.573 (PC)	39.995	24	0.021	0.056	0.993
q3_1f vs.	q1_5f	0.470 (PC)	24.215	24	0.449	0.007	1.000
q3_1f vs.	q2_1f	0.266 (PC)	37.299	24	0.041	0.051	0.997
q3_1f vs.	q2_2f	0.350 (PC)	53.131	24	0.001	0.076	0.919
q3_1f vs.	q2_3f	0.413 (PC)	37.294	24	0.041	0.051	0.997
q3_1f vs.	q2_4f	0.429 (PC)	47.172	24	0.003	0.068	0.969
q3_2f vs.	q1_1f	0.259 (PC)	30.995	24	0.154	0.037	1.000
q3_2f vs.	q1_2f	0.381 (PC)	23.756	24	0.476	0.000	1.000
q3_2f vs.	q1_3f	0.308 (PC)	25.727	24	0.367	0.019	1.000
q3_2f vs.	q1_4f	0.337 (PC)	30.799	24	0.160	0.037	1.000
q3_2f vs.	q1_5f	0.234 (PC)	23.714	24	0.478	0.000	1.000
q3_2f vs.	q2_1f	0.166 (PC)	33.600	24	0.092	0.044	0.999
q3_2f vs.	q2_2f	0.217 (PC)	26.777	24	0.315	0.024	1.000
q3_2f vs.	q2_3f	0.295 (PC)	28.907	24	0.224	0.031	1.000

Percentage of Tests Exceeding 0.5% Significance Level: 0.0%

Percentage of Tests Exceeding 1.0% Significance Level: 0.0%

Percentage of Tests Exceeding 5.0% Significance Level: 0.0%

**Appendix 9** Output homogeneity of variances and Brown-Forsythe ANOVA of pretests and posttests across heiQ-PP, heiQ-PPT, and heiQ-PPR

Test of homogeneity of variances, heiQ pretests

	Levene Statistic	df1	df2	Sig.
1. PAE	0.71	2	946	0.491
2. HDB	0.54	2	946	0.586
3. STA	0.24	2	946	0.790
4. CAA	2.16	2	946	0.116
5. SMI	0.03	2	946	0.970
6. HSN	0.15	2	946	0.863
7. SIS	1.32	2	946	0.269
8. EWB	0.72	2	946	0.486

Robust tests of equality of means, heiQ pretests

		Statistic(a)	df1	df2	Sig.
1. PAE	Brown-Forsythe	0.69	2	930	0.503
2. HDB	Brown-Forsythe	1.00	2	945	0.368
3. STA	Brown-Forsythe	0.73	2	940	0.482
4. CAA	Brown-Forsythe	1.82	2	936	0.162
5. SMI	Brown-Forsythe	0.77	2	940	0.464
6. HSN	Brown-Forsythe	0.11	2	938	0.896
7. SIS	Brown-Forsythe	0.78	2	937	0.460
8. EWB	Brown-Forsythe	0.01	2	941	0.993

Test of homogeneity of variances, heiQ posttests

	Levene Statistic	df1	df2	Sig.
1. PAE	0.23	2	946	0.792
2. HDB	4.38	2	946	<b>0.013*</b>
3. STA	0.64	2	946	0.530
4. CAA	0.38	2	946	0.681
5. SMI	0.70	2	946	0.496
6. HSN	0.41	2	946	0.663
7. SIS	1.41	2	946	0.244
8. EWB	0.59	2	946	0.555

Robust tests of equality of means, heiQ posttests

		Statistic(a)	df1	df2	Sig.
1. PAE	Brown-Forsythe	7.13	2	945	<b>0.001*</b>
2. HDB	Brown-Forsythe	4.41	2	934	<b>0.012*</b>
3. STA	Brown-Forsythe	10.88	2	946	<b>0.000*</b>
4. CAA	Brown-Forsythe	3.81	2	942	<b>0.023*</b>
5. SMI	Brown-Forsythe	12.12	2	946	<b>0.000*</b>
6. HSN	Brown-Forsythe	4.11	2	934	<b>0.017*</b>
7. SIS	Brown-Forsythe	0.74	2	938	0.479
8. EWB	Brown-Forsythe	0.07	2	937	0.928

\* significant at the  $p=.05$  level; a: asymptotically F distributed; df: degrees of freedom

PAE: Positive and Active Engagement in Life  
HDB: Health-Directed Behaviour  
STA: Skill and Technique Acquisition  
CAA: Constructive Attitudes and Approaches

SMI: Self-Monitoring and Insight  
HSN: Health Service Navigation  
SIS: Social Integration and Support  
EWB: Emotional Well-Being

**Appendix 10** Output homogeneity of variances and Brown-Forsythe ANOVA of change scores across heiQ-PP, heiQ-PPT, and heiQ-PPR

Test of homogeneity of variances, change scores

	Levene Statistic	df1	df2	Sig.
1. PAE (posttest - pretest)	2.16	2	946	0.116
2. HDB (posttest - pretest)	1.56	2	946	0.210
3. STA (posttest - pretest)	5.11	2	946	<b>0.006*</b>
4. CAA (posttest - pretest)	3.19	2	946	<b>0.042*</b>
5. SMI (posttest - pretest)	2.80	2	946	0.061
6. HSN (posttest - pretest)	2.15	2	946	0.117
7. SIS (posttest - pretest)	1.84	2	946	0.160
8. EWB (posttest - pretest)	0.75	2	946	0.474

Robust tests of equality of means, change scores

		Statistic(a)	df1	df2	Sig.
1. PAE (posttest - pretest)	Brown-Forsythe	11.34	2	905	<b>0.000*</b>
2. HDB (posttest - pretest)	Brown-Forsythe	3.20	2	926	<b>0.041*</b>
3. STA (posttest - pretest)	Brown-Forsythe	5.28	2	906	<b>0.005*</b>
4. CAA (posttest - pretest)	Brown-Forsythe	10.79	2	918	<b>0.000*</b>
5. SMI (posttest - pretest)	Brown-Forsythe	8.42	2	926	<b>0.000*</b>
6. HSN (posttest - pretest)	Brown-Forsythe	4.59	2	919	<b>0.010*</b>
7. SIS (posttest - pretest)	Brown-Forsythe	3.10	2	930	<b>0.046*</b>
8. EWB (posttest - pretest)	Brown-Forsythe	0.08	2	936	0.924

\* significant at the  $p=.05$  level; a: asymptotically F distributed; df: degrees of freedom

PAE: Positive and Active Engagement in Life

HDB: Health-Directed Behaviour

STA: Skill and Technique Acquisition

CAA: Constructive Attitudes and Approaches

SMI: Self-Monitoring and Insight

HSN: Health Service Navigation

SIS: Social Integration and Support

EWB: Emotional Well-Being

**Appendix 11** Chi-square significance tests ('decline', 'no change', 'improvement') across heiQ-PP, heiQ-PPT, and heiQ-PPR

Positive and Active Engagement in Life

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	15.046(a)	2	<b>0.001*</b>
N of Valid Cases	949		

Health-Directed Behaviour

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	8.111(a)	2	<b>0.017*</b>
N of Valid Cases	949		

Skill and Technique Acquisition

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	7.879(a)	2	<b>0.019*</b>
N of Valid Cases	949		

Constructive Attitudes and Approaches

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	18.315(a)	2	<b>0.000*</b>
N of Valid Cases	949		

Self-Monitoring and Insight

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	16.351(a)	2	<b>0.000*</b>
N of Valid Cases	949		

Health Service Navigation

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	6.919(a)	2	<b>0.031*</b>
N of Valid Cases	949		

Social Integration and Support

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	9.475(a)	2	<b>0.009*</b>
N of Valid Cases	949		

Emotional Well-Being

	Value	df	Asymp. sig. (2-sided)
Pearson Chi-Square	.089(a)	2	0.957
N of Valid Cases	949		

---

\* significant at the  $p=.05$  level; a: 0 cells (.0%) have expected count less than 5; df: degrees of freedom

**Appendix 12** Paired t-tests of heiQ-PPR and heiQ-PPR Retro

	Mean	Standard deviation	t	df	Sig. (2-tailed)
Positive and Active Engagement in Life	-0.10	0.84	-2.08	313	<b>0.038*</b>
Health-Directed Behaviour	-0.29	0.96	-5.32	313	<b>0.000*</b>
Skill and Technique Acquisition	0.03	0.89	0.60	313	0.549
Constructive Attitudes and Approaches	-0.04	0.88	-0.78	313	0.437
Self-Monitoring and Insight	-0.12	0.71	-2.94	313	<b>0.003*</b>
Health Service Navigation	0.02	0.89	0.30	313	0.764
Social Integration and Support	0.02	0.91	0.33	313	0.745
Emotional Well-Being	-0.09	0.94	-1.65	313	0.099

---

\* significant at the  $p=.05$  level; t: t-value; df: degrees of freedom



**Appendix 13** Chi-square significance tests ('decline', 'no change', 'improvement') across heiQ-PPR and heiQ-PPR Retro

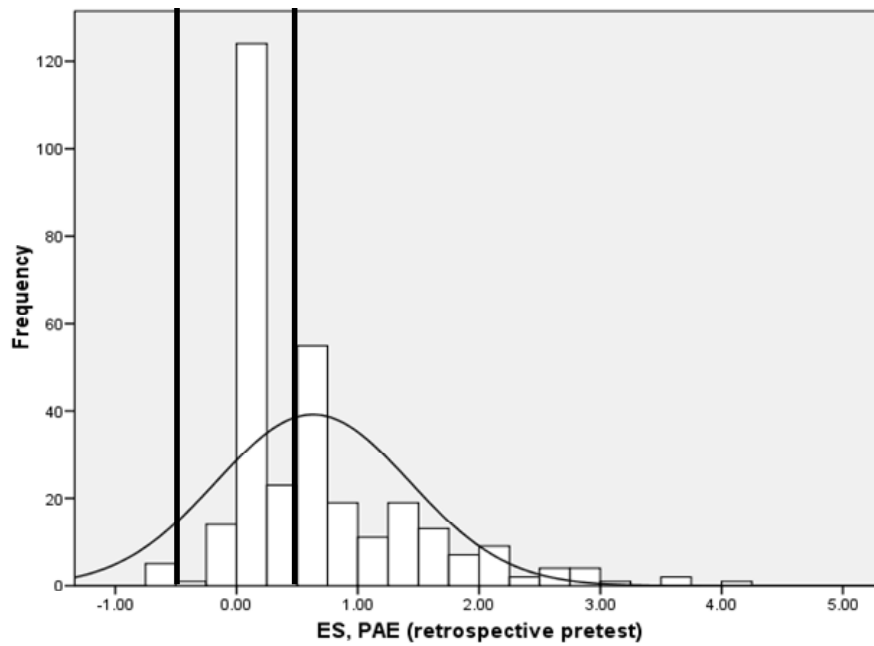
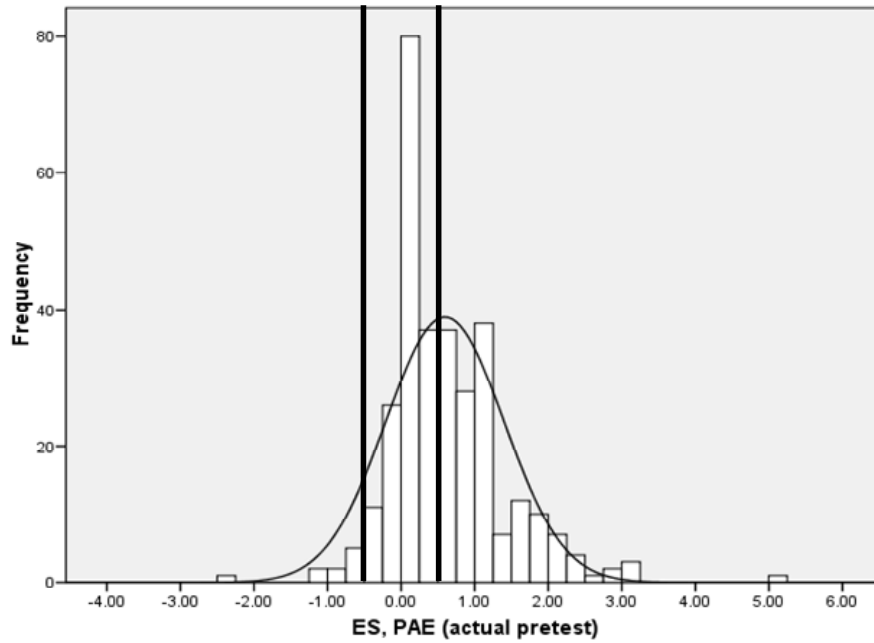
Positive and Active Engagement in Life			
Pearson Chi-Square	Value .057(a)	df 1	Asymp. Sig. (2-sided) 0.811
Health-Directed Behaviour			
Pearson Chi-Square	Value 4.695(a)	df 1	Asymp. Sig. (2-sided) <b>0.030*</b>
Skill and Technique Acquisition			
Pearson Chi-Square	Value 2.192(a)	df 1	Asymp. Sig. (2-sided) 0.139
Constructive Attitudes and Approaches			
Pearson Chi-Square	Value 1.692(a)	df 1	Asymp. Sig. (2-sided) 0.193
Self-Monitoring and Insight			
Pearson Chi-Square	Value .006(a)	df 1	Asymp. Sig. (2-sided) 0.936
Health Service Navigation			
Pearson Chi-Square	Value 41.837(a)	df 1	Asymp. Sig. (2-sided) <b>0.000*</b>
Social Integration and Support			
Pearson Chi-Square	Value 15.320(a)	df 1	Asymp. Sig. (2-sided) <b>0.000*</b>
Emotional Well-Being			
Pearson Chi-Square	Value 3.446(a)	df 1	Asymp. Sig. (2-sided) 0.063

---

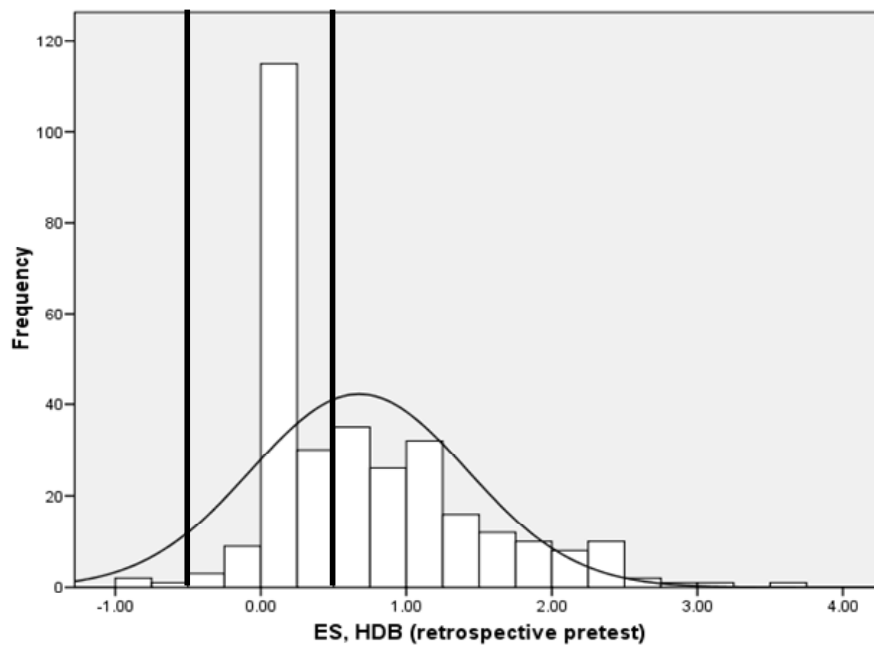
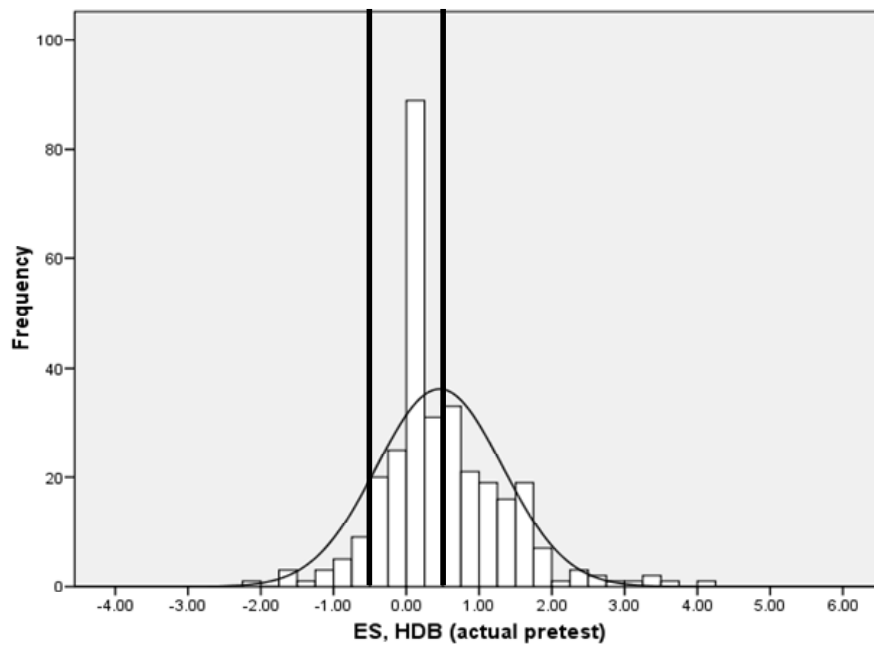
\* significant at the  $p=.05$  level; a: 0 cells (.0%) have expected count less than 5; df: degrees of freedom

**Appendix 14** Histograms of actual (heiQ-PPR) and retrospective (heiQ-PPR Retro) change including proportions of people in the categories 'decline', 'no change', or 'improvement'

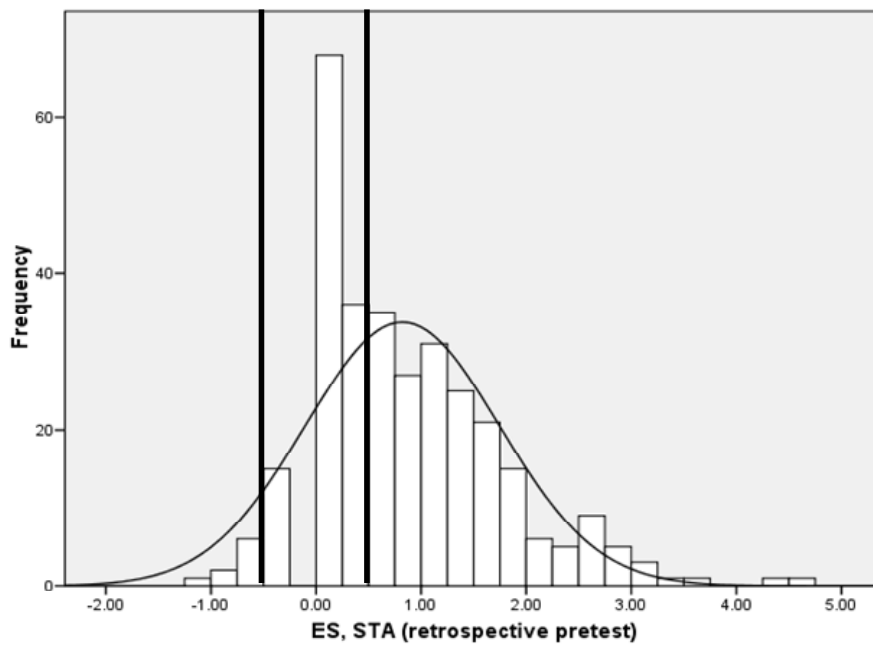
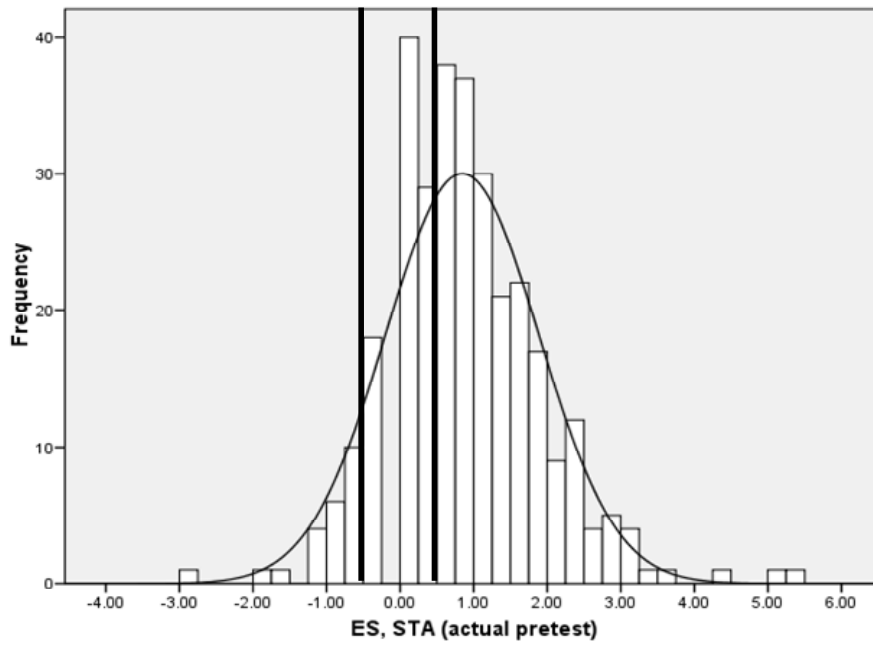
PAE=Positive and Active Engagement in Life (ES derived from actual versus retrospective pretests)



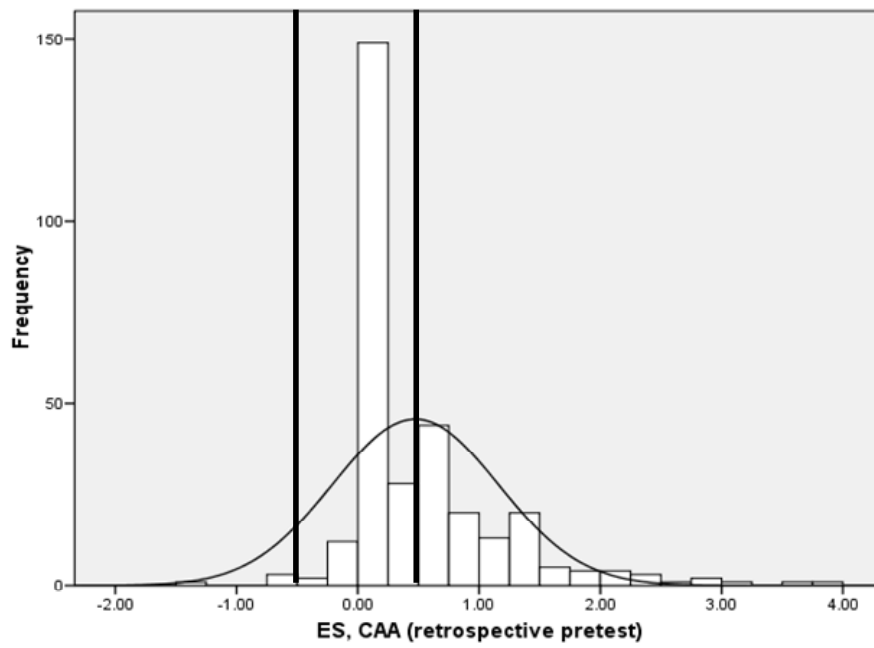
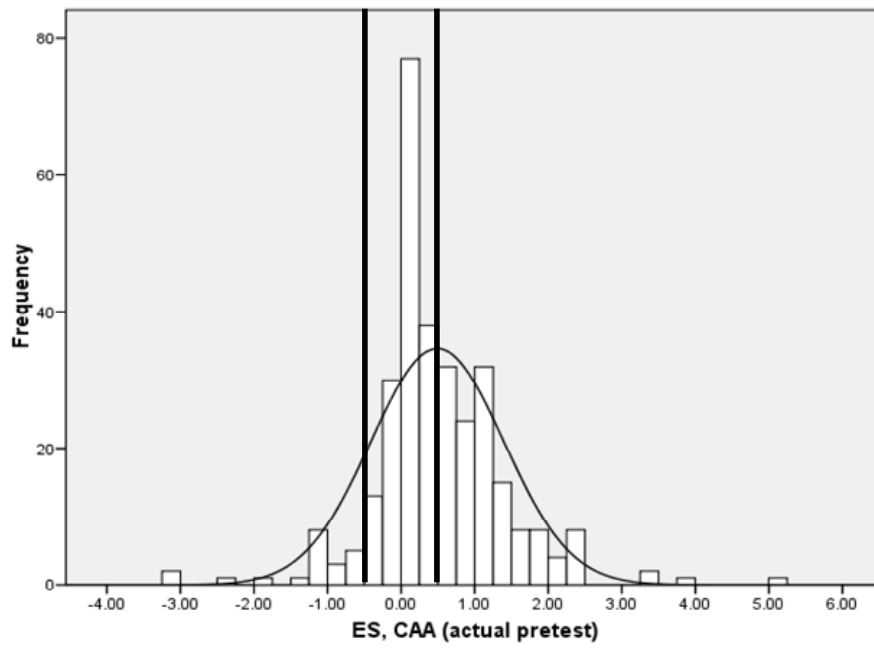
HDB=Health-Directed Behaviour (ES derived from actual versus retrospective pretests)



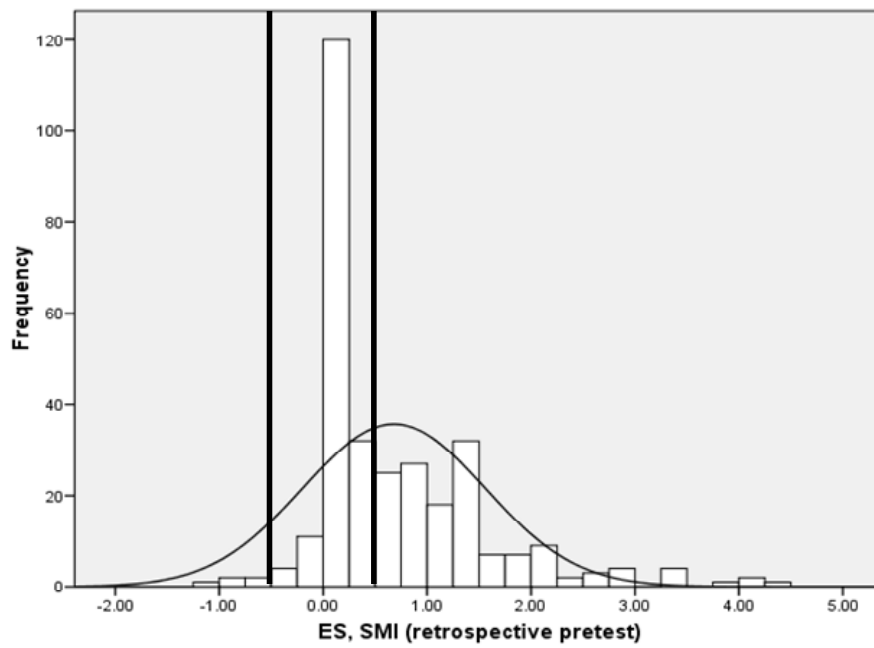
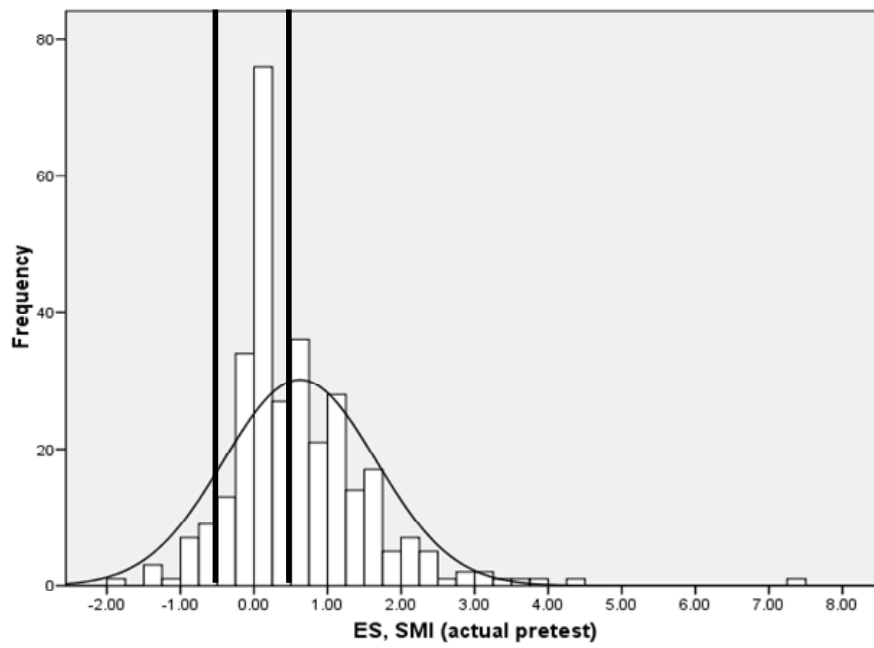
STA=Skill and Technique Acquisition (ES derived from actual versus retrospective pretests)



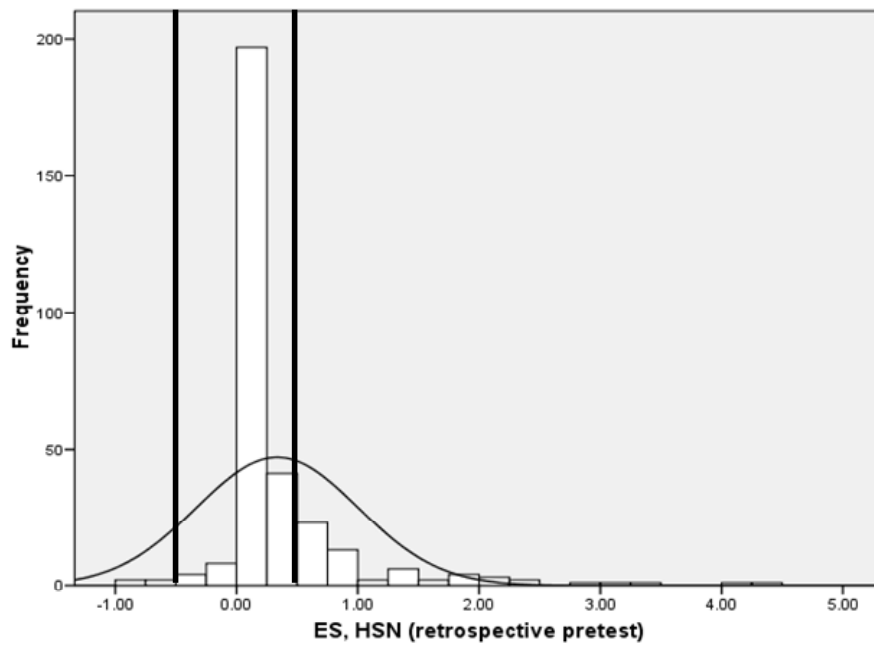
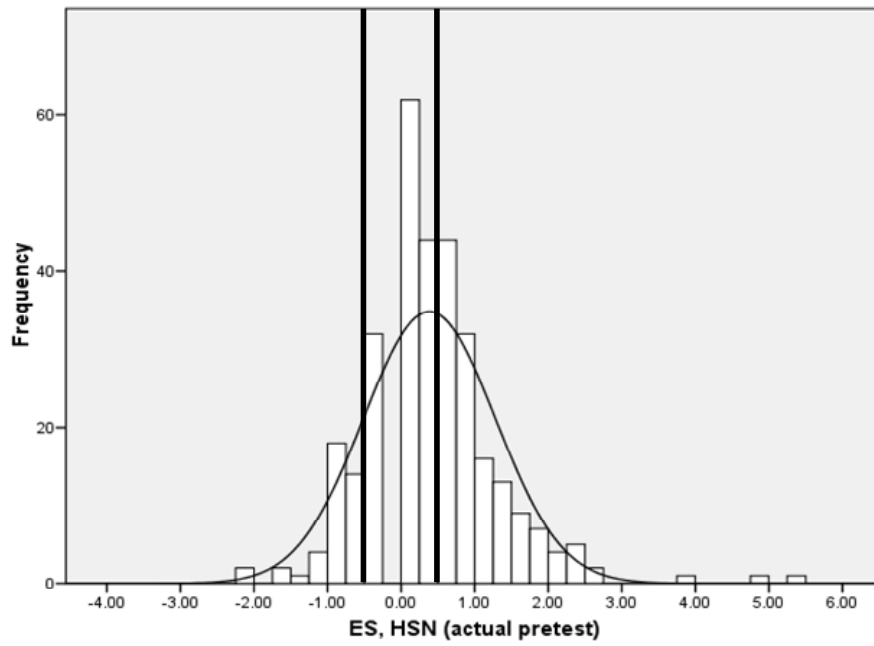
CAA=Constructive Attitudes and Approaches (ES derived from actual versus retrospective pretests)



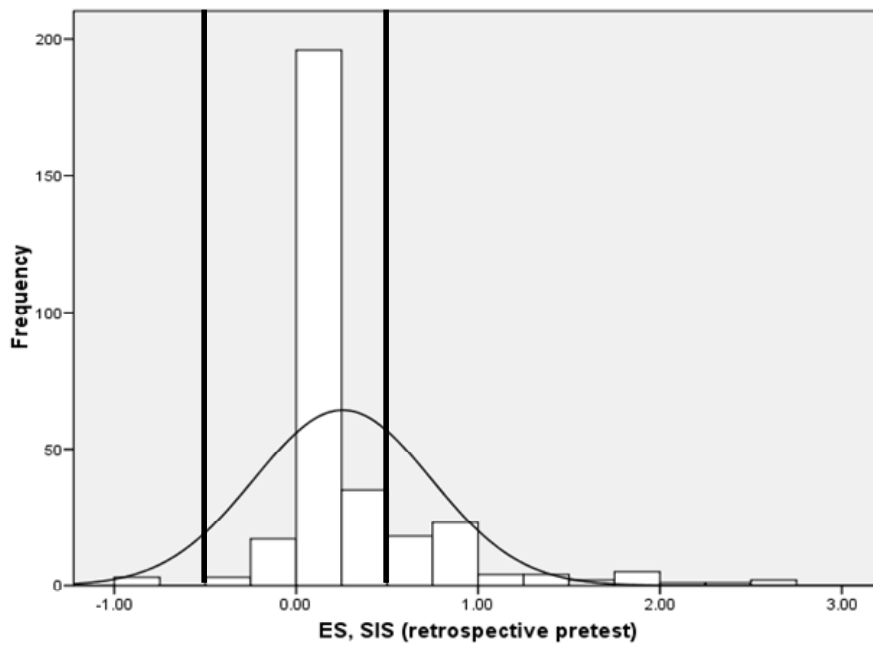
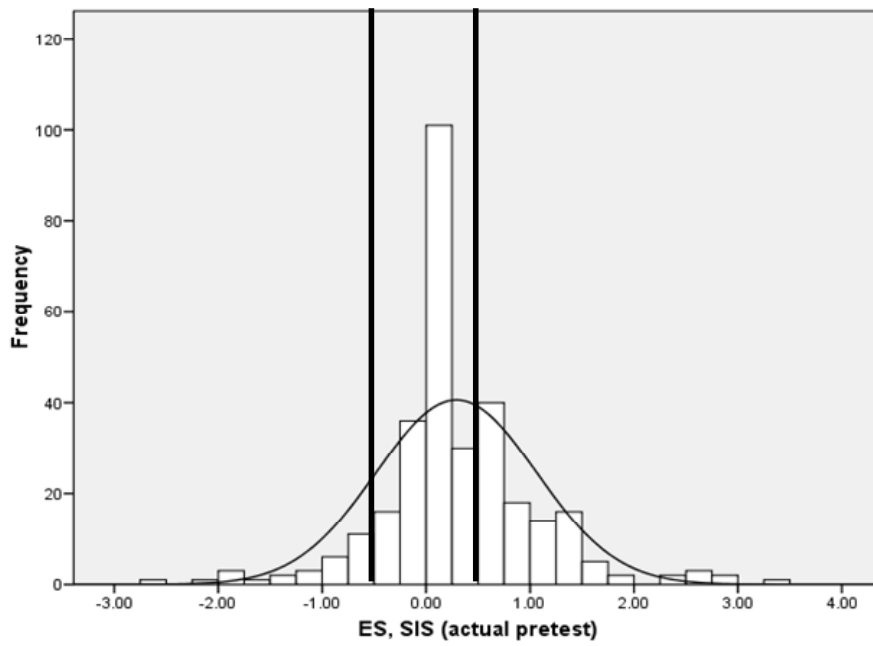
SMI=Self-Monitoring and Insight (ES derived from actual versus retrospective pretests)



HSN=Health Service Navigation (ES derived from actual versus retrospective pretests)

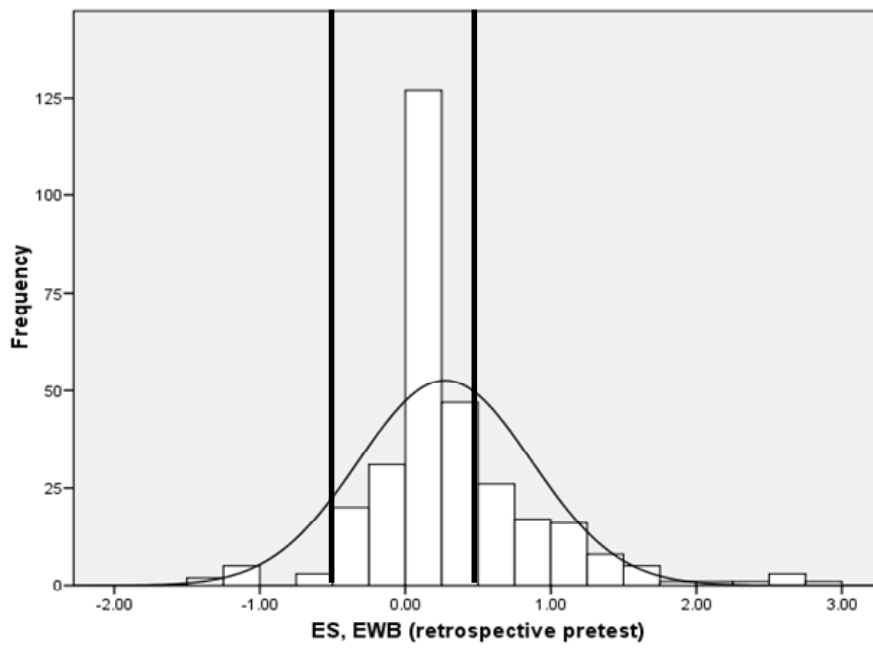
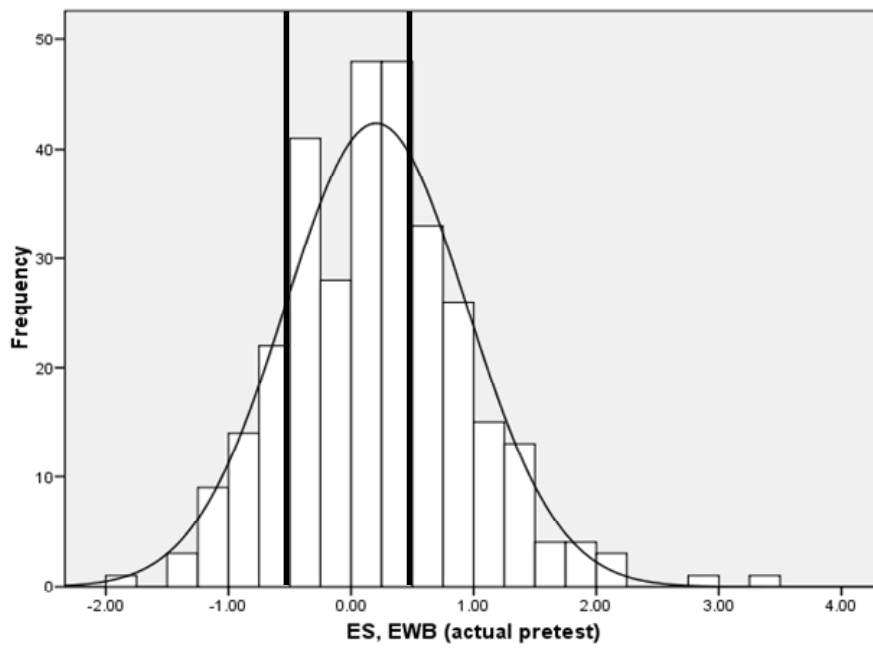


SIS=Social Integration and Support (ES derived from actual versus retrospective pretests)





EWB=Emotional Well-Being (ES derived from actual versus retrospective pretests)



**Appendix 15** Step 3 of Jöreskog's 3-step procedure; full model of the 42 heiQ items (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>1. Positive and Active Engagement in Life</b>				
q1_1 / Q11	I am doing interesting things in my life	0.81	0.34	0.158
q1_2 / Q3	Most days I am doing some of the things I really enjoy	0.70	0.51	0.109
q1_3 / Q7	I try to make the most of my life	0.77	0.41	0.124
q1_4 / Q16	I have plans to do enjoyable things for myself during the next few days	0.74	0.45	0.128
q1_5 / Q30	I feel like I am actively involved in life	0.83	0.32	0.176
<b>2. Health-Directed Behaviour</b>				
q2_1 / Q40	I walk for exercise, for at least 15 minutes per day, most days of the week	0.79	0.38	0.184
q2_2 / Q5	I do at least one type of physical activity every day for at least 30 minutes (e.g., walking, gardening, housework, golf, bowls, dancing, Tai Chi, swimming)	0.76	0.42	0.157
q2_3 / Q15	On most days of the week, I do at least one activity to improve my health (e.g., walking, relaxation, exercise)	0.71	0.50	0.103
q2_4 / Q24	On most days of the week, I set aside time for healthy activities (e.g., walking, relaxation, exercise)	0.88	0.23	0.287
<b>3. Skill and Technique Acquisition</b>				
q3_1 / Q17	When I have symptoms, I have skills that help me cope	0.76	0.43	0.140
q3_2 / Q02	I am very good at using aids and devices to make my life easier	0.47	0.78	0.033
q3_3 / Q36	I have effective skills that help me handle stress	0.72	0.49	0.103
q3_4 / Q14	I have a very good idea of how to manage my health problems	0.77	0.41	0.131
q3_5 / Q10	I have effective ways to prevent my symptoms (e.g., discomfort, pain and stress) from limiting what I can do in my life	0.64	0.59	0.091
<b>4. Constructive Attitudes and Approaches</b>				
q4_1 / Q39	If others can cope with problems like mine, I can too	0.68	0.54	0.061
q4_2 / Q18	I try not to let my health problems stop me from enjoying life	0.77	0.41	0.081
q4_3 / Q35	I do not let my health problems control my life	0.84	0.30	0.172
q4_4 / Q28	My health problems do not ruin my life	0.77	0.41	0.084
q4_5 / Q32	I feel I have a very good life even when I have health problems	0.86	0.27	0.141
<b>5. Self-Monitoring and Insight</b>				
q5_1 / Q41	With my health in mind, I have realistic expectations of what I can and cannot do	0.57	0.68	0.045
q5_2 / Q4	As well as seeing my doctor, I regularly monitor changes in my health	0.53	0.72	0.032
q5_3 / Q8	I know what things can trigger my health problems and make them worse	0.52	0.73	0.043
q5_4 / Q22	When I have health problems, I have a clear understanding of what I need to do to control them	0.79	0.37	0.099

**Appendix 15 (continued)** Step 3 of Jöreskog's 3-step procedure; full model of the 42 heiQ items (n=949)

Item # / # on heiQ	Item	Loading	Error	FSR
<b>5. Self-Monitoring and Insight (continued)</b>				
q5_5 / Q19	I have a very good understanding of when and why I am supposed to take my medication	0.62	0.62	0.029
q5_6 / Q38	I carefully watch my health and do what is necessary to keep as healthy as possible	0.64	0.60	0.070
q5_7 / Q12	I know when my lifestyle (e.g., exercise, diet, stress) is creating health problems for me	0.48	0.77	0.027
<b>6. Health Service Navigation</b>				
q6_1 / Q21	I communicate very confidently with my doctor about my healthcare needs	0.84	0.29	0.256
q6_2 / Q13	I have very positive relationships with my healthcare professionals	0.79	0.37	0.127
q6_3 / Q25	I confidently give healthcare professionals the information they need to help me	0.63	0.61	0.104
q6_4 / Q27	I get my needs met from available healthcare resources (e.g., doctors, hospitals and community services)	0.72	0.48	0.129
q6_5 / Q34	I work in a team with my doctors and other healthcare professionals	0.77	0.41	0.187
<b>7. Social Integration and Support</b>				
q7_1 / Q20	I have enough friends who help me cope with my health problems	0.84	0.29	0.209
q7_2 / Q33	I get enough chances to talk about my health problems with people who understand	0.73	0.47	0.108
q7_3 / Q6	If I need help, I have plenty of people I can rely on	0.80	0.36	0.165
q7_4 / Q31	Overall, I feel well looked after by friends or family	0.87	0.25	0.204
q7_5 / Q23	When I feel ill, my family and carers really understand what I am going through	0.75	0.44	0.115
<b>8. Emotional Well-Being</b>				
q8_1 / Q42	If I think about my health, I get depressed	0.83	0.31	0.188
q8_2 / Q37	I get upset when I think about my health	0.81	0.34	0.190
q8_3 / Q26	I often feel angry when I think about my health	0.85	0.27	0.214
q8_4 / Q09	My health problems make me very dissatisfied with my life	0.79	0.38	0.138
q8_5 / Q1	I often worry about my health	0.69	0.52	0.125
q8_6 / Q29	I feel hopeless because of my health problems	0.71	0.50	0.085

Fit statistics:  $\chi^2_{SB}(791)=2280.7$ ,  $p<0.001$ ; RMSEA=0.045 (90% CI, 0.042;0.047); CFI=0.98; SRMR=0.060

**Legend:**

Loading:	Standardised factor loading
Error:	Error variance
FSR:	Factor score regression coefficient
$\chi^2_{SB}$ :	Satorra-Bentler $\chi^2$
RMSEA:	Root mean square error of approximation
90% CI:	90% Confidence interval
CFI:	Comparative fit index
SRMR:	Standardized root mean square residual

**Appendix 16** Formula of the Satorra-Bentler scaled difference chi-square test statistic (Satorra & Bentler, 2001)

Difference test statistic:  $T_d = T_0 - T_1$

$\chi^2_{SB}$  difference test:  $T_{SB\ d} = T_d / c_{SB\ d}$ , where

$c_{SB\ d} = (r_0 c_{SB\ 0} - r_1 c_{SB\ 1}) / m$ , with scaling corrections:

$c_{SB\ 0} = T_0 / T_{SB\ 0}$  and  $c_{SB\ 1} = T_1 / T_{SB\ 1}$ , and

$m = r_0 - r_1$  (degrees of freedom of the  $\chi^2$  distribution)

Legend

- $M_0$ : Model 0, e.g. baseline model
- $M_1$ : Model 1, e.g. response shift model
- $T_0$ : unscaled chi-square at 0
- $T_{SB\ 0}$ : scaled chi-square at 0
- $T_1$ : unscaled chi-square at 1
- $T_{SB\ 1}$ : scaled chi-square at 1
- $r_0$ : degrees of freedom at 0
- $r_1$ : degrees of freedom at 1

**Appendix 17** Exploratory Factor Analysis (CEFA) results of the MC-C scale

```

o-----o
| CEFA: Comprehensive Exploratory Factor Analysis |
|
|           Release Version 2.00
|           October 2004
|
|           Mathematical Specification:
|           Michael W. Browne, Robert Cudeck,
|           Krishna Tateneni, and Gerhard Mels.
|
|           Programming:
|           Krishna Tateneni, Gerhard Mels,
|           Robert Cudeck, and Michael W. Browne.
|
o-----o
  
```

Date: 2007-10-22  
 Time: 23:32:18

```

o=====o
| Details of Analysis |
o=====o
  
```

Data file: C:\Documents and Settings\noltes\Thesis\Analysis\Chapter 6\EFA  
 - SD\CEFA\MC-C\Geomin OLS - poly 2 factors\SD syntax.inp

Number of observations : 908  
 Number of variables : 13  
 Number of factors : 2

Polychoric correlation matrix to be analysed:  
 - Discrepancy function automatically set to OLS  
 - Standard errors unavailable

Discrepancy function : OLS  
 Dispersion matrix : Correlations  
 Max EFA iterations : 50

Rotation type : Oblique  
 Sort columns using : Descending sums of squares

Rotation Criterion : GEOMIN  
 Optional parameter : 0.100E-01  
 Row weights : None

Rotation convergence : 0.100E-05

```

o=====o
| Estimated Polychoric Correlation Matrix |
o=====o
  
```

Var1 Var2 Var3 Var4 Var5 Var6 Var7 Var8 Var9 Var10 Var11 Var12 Var13

Var1	1.000											
Var2	0.411	1.000										
Var3	0.491	0.377	1.000									
Var4	0.293	0.468	0.339	1.000								
Var5	0.006	0.111	0.014	0.133	1.000							
Var6	0.083	0.265	0.147	0.321	0.147	1.000						
Var7	0.106	0.278	0.046	0.268	0.385	0.356	1.000					
Var8	0.208	0.386	0.263	0.289	0.085	0.292	0.188	1.000				
Var9	0.056	0.180	0.056	0.183	0.397	0.204	0.428	0.296				

1.000								
Var10	0.004	0.203	0.034	0.111	0.275	0.090	0.231	0.035
0.292	1.000							
Var11	0.346	0.397	0.366	0.253	0.080	0.252	0.258	0.308
0.102	0.255	1.000						
Var12	0.142	0.365	0.208	0.259	0.139	0.156	0.139	0.210
0.102	0.244	0.300	1.000					
Var13	0.009	0.052	0.046	0.053	0.149	0.076	0.137	0.199
0.345	0.274	-0.016	0.007	1.000				

Eigenvalues of Sample Correlation Matrix:

0.36E+01	0.19E+01	0.11E+01	0.11E+01	0.90E+00	0.78E+00	0.73E+00
0.62E+00	0.56E+00	0.51E+00	0.49E+00	0.44E+00	0.38E+00	

\*\*\*\*\*  
 \* Exploratory Factor Analysis Details \*  
 \*\*\*\*\*

o=====o  
 | Noniterative Unique Variances, Communalities, and SMCs |  
 o=====o

Variable	Unique Variance	Communality	SMC
Var1	0.723	0.277	0.458
Var2	0.607	0.393	0.458
Var3	0.710	0.290	0.458
Var4	0.715	0.285	0.458
Var5	0.641	0.359	0.458
Var6	0.821	0.179	0.458
Var7	0.586	0.414	0.458
Var8	0.783	0.217	0.458
Var9	0.558	0.442	0.458
Var10	0.842	0.158	0.458
Var11	0.709	0.291	0.458
Var12	0.916	0.084	0.458
Var13	0.890	0.110	0.458

o=====o  
 | OLS Unrotated Factor Loadings |  
 o=====o

	Fac1	Fac2
Var1	0.455	-0.395
Var2	0.693	-0.231
Var3	0.490	-0.410
Var4	0.569	-0.133
Var5	0.334	0.438
Var6	0.439	0.077
Var7	0.513	0.361
Var8	0.507	-0.040
Var9	0.469	0.527
Var10	0.339	0.294
Var11	0.555	-0.199
Var12	0.420	-0.074
Var13	0.213	0.309

```

o=====o
| OLS Unique Variances and Communalities |
o=====o

```

Variable	Unique Variance	Communality
Var1	0.637	0.363
Var2	0.467	0.533
Var3	0.591	0.409
Var4	0.659	0.341
Var5	0.697	0.303
Var6	0.801	0.199
Var7	0.607	0.393
Var8	0.741	0.259
Var9	0.502	0.498
Var10	0.799	0.201
Var11	0.653	0.347
Var12	0.818	0.182
Var13	0.859	0.141

```

o=====o
| OLS Discrepancy Function Details |
o=====o

```

F: 0.23695649

```

o=====o
| Corrected ML Discrepancy Function Details |
o=====o

```

F1: 0.48266435  
F2: -0.00055268  
F: 0.48211167

```

o=====o
| Measures of Fit |
o=====o

```

Sample discrepancy function value : 0.48211167

Population discrepancy function value, Fo  
Bias adjusted point estimate : 0.424  
90 percent confidence interval : ( 0.354; 0.501)

Root mean square error of approximation  
Steiger-Lind RMSEA = SQRT(Fo/DF)  
Point estimate : 0.089  
90 percent confidence interval : ( 0.082; 0.097)

Expected cross-validation index  
Point estimate (modified AIC) : 0.566  
90 percent confidence interval : ( 0.496; 0.644)  
ECVI (modified AIC) for the saturated model : 0.201

Chi-square test statistic : 437.275

Exceedance Probabilities  
Perfect fit (Ho: RMSEA = 0.0) : 0.000  
Close fit (Ho: RMSEA <= 0.05) : 0.000

Multiplier for obtaining test statistic : 907.0  
Degrees of freedom : 53  
Effective number of parameters : 38

\*\*\*\*\*  
 \* Rotation Details \*  
 \*\*\*\*\*

o=====o  
 | Row Weights (Not Necessarily Used!) |  
 o=====o

Kaiser weights, Cureton-Mulaik weights, and  
 their products (final Cureton-Mulaik weights)

Var1	1.660	0.981	1.629
Var2	1.369	0.361	0.495
Var3	1.564	0.970	1.517
Var4	1.712	0.199	0.340
Var5	1.817	0.931	1.692
Var6	2.244	0.118	0.264
Var7	1.595	0.888	1.416
Var8	1.965	0.025	0.050
Var9	1.417	0.987	1.399
Var10	2.228	0.981	2.186
Var11	1.696	0.403	0.684
Var12	2.345	0.118	0.277
Var13	2.667	0.875	2.333

o=====o  
 | GEOMIN Rotated Factor Matrix |  
 o=====o

	Fac1	Fac2
Var1	0.633	-0.189
Var2	0.706	0.069
Var3	0.672	-0.189
Var4	0.540	0.111
Var5	-0.043	0.562
Var6	0.291	0.259
Var7	0.151	0.564
Var8	0.427	0.175
Var9	0.001	0.706
Var10	0.062	0.426
Var11	0.575	0.042
Var12	0.382	0.105
Var13	-0.048	0.387

GEOMIN Criterion: 1.078065689014995

o=====o  
 | Factor Correlations |  
 o=====o

	Fac1	Fac2
Fac1	1.000	
Fac2	0.309	1.000

o=====o  
 | CEFA Completed |  
 o=====o