

Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation

A Practical Guide to Analysis and Interpretation

Stephen J. Walters

School of Health and Related Research, University of Sheffield, UK



A John Wiley and Sons, Ltd., Publication

Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation

STATISTICS IN PRACTICE

Advisory Editor

Stephen Senn

University of Glasgow, UK

Founding Editor

Vic Barnett

Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceuticals; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.

Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation

A Practical Guide to Analysis and Interpretation

Stephen J. Walters

School of Health and Related Research, University of Sheffield, UK



A John Wiley and Sons, Ltd., Publication

This edition first published 2009
© 2009 John Wiley & Sons, Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloguing-in-Publication Data

Walters, Stephen John.

Quality of life outcomes in clinical trials and health-care evaluation : a practical guide to analysis and interpretation / Stephen Walters.

p. ; cm.

Includes bibliographical references and index.

ISBN 978-0-470-75382-8 (cloth)

1. Clinical trials. 2. Quality of life. 3. Outcome assessment (Medical care) I. Title.

[DNLM: 1. Cross-Sectional Studies. 2. Randomized Controlled Trials as Topic. 3. Outcome Assessment (Health Care) 4. Quality of Life. 5. Research Design. WA 950 W235q 2009]

R853.C55W35 2009

610.72'4 – dc22

2009024883

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-75382-8

Set in 10/12pt Times Roman by Laserwords Pvt Ltd, Chennai, India

Printed and bound in the United Kingdom by Antony Rowe Ltd, Chippenham, Wiltshire

Contents

Preface	xi
1 Introduction	1
Summary	1
1.1 What is quality of life?	1
1.2 Terminology	2
1.3 History	2
1.4 Types of quality of life measures	4
1.5 Why measure quality of life?	10
1.6 Further reading	11
2 Measuring quality of life	13
Summary	13
2.1 Introduction	13
2.2 Principles of measurement scales	13
2.2.1 Scales and items	13
2.2.2 Constructs and latent variables	14
2.3 Indicator and causal variables	15
2.3.1 Indicator variables	15
2.3.2 Causal variables	15
2.3.3 Why do we need to worry about the distinction between indicator and causal items?	16
2.3.4 Single-item versus multi-item scales	16
2.4 The traditional psychometric model	16
2.4.1 Psychometrics and QoL scales	17
2.5 Item response theory	17
2.5.1 Traditional scales versus IRT	18
2.6 Clinimetric scales	18
2.7 Measuring quality of life: Indicator or causal items	19
2.8 Developing and testing questionnaires	19
2.8.1 Specify the research question and define the target population	19
2.8.2 Identify concepts	20

2.8.3	Create instrument	21
2.8.4	Assess measurement properties	26
2.8.5	Modify instrument	30
2.9	Further reading	30
3	Choosing a quality of life measure for your study	31
	Summary	31
3.1	Introduction	31
3.2	How to choose between instruments	31
3.3	Appropriateness	33
3.4	Acceptability	33
3.5	Feasibility	34
3.6	Validity	35
3.6.1	Tests for criterion validity	35
3.6.2	Tests for face and content validity	36
3.6.3	Tests for construct validity	36
3.7	Reliability	38
3.7.1	Repeatability reliability	40
3.7.2	Graphical methods for assessing reliability between two repeated measurements	40
3.7.3	Internal reliability or internal consistency reliability	42
3.8	Responsiveness	44
3.8.1	Floor and ceiling effects	44
3.9	Precision	49
3.10	Interpretability	51
3.11	Finding quality of life instruments	53
4	Design and sample size issues: How many subjects do I need for my study?	55
	Summary	55
4.1	Introduction	55
4.2	Significance tests, <i>P</i> -values and power	56
4.3	Sample sizes for comparison of two independent groups	58
4.3.1	Normally distributed continuous data – comparing two means	58
4.3.2	Transformations	61
4.3.3	Comparing two groups with continuous data using non-parametric methods	61
4.3.4	Dichotomous categorical data – comparing two proportions	63
4.3.5	Ordered categorical (ordinal) data	66
4.4	Choice of sample size method with quality of life outcomes	69
4.5	Paired data	70
4.5.1	Paired continuous data – comparison of means	70
4.5.2	Paired binary data – comparison of proportions	72
4.6	Equivalence/non-inferiority studies	73
4.6.1	Continuous data – comparing the equivalence of two means	74
4.6.2	Binary data – comparing the equivalence of two proportions	75
4.7	Unknown standard deviation and effect size	75

4.7.1	Tips on obtaining the standard deviation	76
4.8	Cluster randomized controlled trials	76
4.9	Non-response	77
4.10	Unequal groups	77
4.11	Multiple outcomes/endpoints	79
4.12	Three or more groups	80
4.13	What if we are doing a survey, not a clinical trial?	80
4.13.1	Sample sizes for surveys	80
4.13.2	Confidence intervals for estimating the mean QoL of a population	81
4.13.3	Confidence intervals for a proportion	82
4.14	Sample sizes for reliability and method comparison studies	85
4.15	Post-hoc sample size calculations	86
4.16	Conclusion: Usefulness of sample size calculations	86
4.17	Further reading	86
5	Reliability and method comparison studies for quality of life measurements	91
	Summary	91
5.1	Introduction	91
5.2	Intra-class correlation coefficient	92
5.2.1	Inappropriate method	94
5.3	Agreement between individual items on a quality of life questionnaire	95
5.3.1	Binary data: Proportion of agreement	95
5.3.2	Binary data: Kappa	95
5.3.3	Ordered categorical data: Weighted kappa	96
5.4	Internal consistency and Cronbach's alpha	98
5.5	Graphical methods for assessing reliability or agreement between two quality of life measures or assessments	99
5.6	Further reading	102
5.7	Technical details	102
5.7.1	Calculation of ICC	102
5.7.2	Calculation of kappa	103
5.7.3	Calculation of weighted kappa	104
5.7.4	Calculation of Cronbach's alpha	104
6	Summarizing, tabulating and graphically displaying quality of life outcomes	109
	Summary	109
6.1	Introduction	109
6.2	Graphs	110
6.2.1	Dot plots	112
6.2.2	Histograms	112
6.2.3	Box-and-whisker plot	114
6.2.4	Scatter plots	116
6.3	Describing and summarizing quality of life data	116
6.3.1	Measures of location	117
6.3.2	Measures of spread	119

6.4	Presenting quality of life data and results in tables and graphs	122
6.4.1	Tables for summarizing QoL outcomes	122
6.4.2	Tables for multiple outcome measures	124
6.4.3	Tables and graphs for comparing two groups	126
6.4.4	Profile graphs	129
7	Cross-sectional analysis of quality of life outcomes	133
	Summary	133
7.1	Introduction	133
7.2	Hypothesis testing (using <i>P</i> -values)	134
7.3	Estimation (using confidence intervals)	137
7.4	Choosing the statistical method	138
7.5	Comparison of two independent groups	138
7.5.1	Independent samples <i>t</i> -test for continuous outcome data	140
7.5.2	Mann–Whitney <i>U</i> -test	144
7.6	Comparing more than two groups	146
7.6.1	One-way analysis of variance	147
7.6.2	The Kruskal–Wallis test	150
7.7	Two groups of paired observations	150
7.7.1	Paired <i>t</i> -test	153
7.7.2	Wilcoxon test	157
7.8	The relationship between two continuous variables	157
7.9	Correlation	160
7.10	Regression	165
7.11	Multiple regression	168
7.12	Regression or correlation?	171
7.13	Parametric versus non-parametric methods	171
7.14	Technical details: Checking the assumptions for a linear regression analysis	173
8	Randomized controlled trials	181
	Summary	181
8.1	Introduction	181
8.2	Randomized controlled trials	182
8.3	Protocols	182
8.4	Pragmatic and explanatory trials	182
8.5	Intention-to-treat and per-protocol analyses	183
8.6	Patient flow diagram	186
8.7	Comparison of entry characteristics	186
8.8	Incomplete data	189
8.9	Main analysis	191
8.10	Interpretation of changes/differences in quality of life scores	196
8.11	Superiority and equivalence trials	197
8.12	Adjusting for other variables	199
8.13	Three methods of analysis for pre-test/post-test control group designs	202
8.14	Cross-over trials	203
8.15	Factorial trials	206

8.16	Cluster randomized controlled trials	209
8.17	Further reading	210
9	Exploring and modelling longitudinal quality of life data	217
	Summary	217
9.1	Introduction	217
9.2	Summarizing, tabulating and graphically displaying repeated QoL assessments	218
9.3	Time-by-time analysis	222
9.4	Response feature analysis – the use of summary measures	223
	9.4.1 Area under the curve	223
	9.4.2 Acupuncture study – analysis of covariance	227
9.5	Modelling of longitudinal data	229
	9.5.1 Autocorrelation	231
	9.5.2 Repeated measures analysis of variance	232
	9.5.3 Marginal general linear models – generalized estimating equations	232
	9.5.4 Random effects models	237
	9.5.5 Random effects versus marginal modelling	239
	9.5.6 Use of marginal and random effects models to analyse data from a cluster RCT	241
9.6	Conclusions	243
10	Advanced methods for analysing quality of life outcomes	249
	Summary	249
10.1	Introduction	249
10.2	Bootstrap methods	251
10.3	Bootstrap methods for confidence interval estimation	251
10.4	Ordinal regression	255
10.5	Comparing two independent groups: Ordinal quality of life measures (with less than 7 categories)	257
10.6	Proportional odds or cumulative logit model	258
10.7	Continuation ratio model	259
10.8	Stereotype logistic model	260
10.9	Conclusions and further reading	264
11	Economic evaluations	265
	Summary	265
11.1	Introduction	265
11.2	Economic evaluations	266
11.3	Utilities and QALYs	266
11.4	Economic evaluations alongside a controlled trial	267
11.5	Cost-effectiveness analysis	267
11.6	Cost-effectiveness ratios	268
11.7	Cost-utility analysis and cost-utility ratios	269
11.8	Incremental cost per QALY	270

11.9	The problem of negative (and positive) incremental cost–effectiveness ratios	272
11.10	Cost-effectiveness acceptability curves	273
11.11	Further reading	275
12	Meta-analysis	277
	Summary	277
12.1	Introduction	277
12.2	Planning a meta-analysis	278
12.2.1	Is a meta-analysis appropriate?	279
12.2.2	Combining the results of different studies	279
12.2.3	Choosing the appropriate statistical method	280
12.3	Statistical methods in meta-analysis	282
12.3.1	The choice of effect measure: What outcome measures am I combining?	282
12.3.2	Model choice: fixed or random?	283
12.3.3	Homogeneity	285
12.3.4	Fixed effects model	285
12.3.5	Forest plots	287
12.3.6	Random effects	289
12.3.7	Funnel plots	289
12.4	Presentation of results	293
12.5	Conclusion	294
12.6	Further reading	295
13	Practical issues	297
	Summary	297
13.1	Missing data	297
13.1.1	Why do missing data matter?	297
13.1.2	Methods for missing items within a form	298
13.1.3	Methods for missing forms	300
13.1.4	The regulator’s view on statistical considerations for patient-level missing data	308
13.1.5	Conclusions and further reading on missing QoL data	309
13.2	Multiplicity, multi-dimensionality and multiple quality of life outcomes	310
13.2.1	Which multiple comparison procedure to use?	312
13.3	Guidelines for reporting quality of life studies	314
	Solutions to exercises	319
	Appendix A: Examples of questionnaires	335
	Appendix B: Statistical tables	345
	References	351
	Index	361

Preface

Quality of life (QoL) outcomes or person/patient reported outcome measures (PROMs) are now frequently being used in randomized controlled trials (RCTs) and observational studies. This book aims to be a practical guide to the design, analysis and interpretation of studies that use such outcomes. Since there are numerous QoL instruments now available, it emphasizes that, for busy and time-constrained researchers, it is easier to use an ‘off-the-shelf’ QoL instrument than to design your own. This book gives practical guidance on how to choose between the various instruments.

QoL outcomes tend to generate data with discrete, bounded and skewed distributions. Hence, many investigators are concerned about the appropriateness of using standard statistical methods to analyse QoL data and want guidance on what methods to use. This book provides such practical guidance, based on the author’s extensive experience. Other texts, on the analysis of QoL outcomes, concentrate mainly on clinical trials and ignore other frequently used study designs such as cross-sectional surveys and non-randomized health-care evaluations. Again this book rectifies this and provides practical guidance on the analysis of QoL outcomes from such observational designs. It presents simple conventional methods to tackle these problems (such as linear regression), before addressing more advanced approaches, including ordinal regression and computer-intensive methods (such as the bootstrap).

The book is illustrated throughout with real-life case studies and worked examples from RCTs and other observational studies, taken from the author’s own experience of designing and analysing studies with QoL outcomes. Each analysis technique is carefully explained and the mathematics, as far as possible, is kept to a minimum. Hopefully, it is written in a style suitable for statisticians and clinicians alike!

The practical guidance provided by this book will be of use to professionals working in and/or managing clinical trials, in academic, government and industrial settings, particularly medical statisticians, clinicians and trial co-ordinators. Its practical approach will appeal to applied statisticians and biomedical researchers, in particular those in the biopharmaceutical industry, medical and public health organizations. Graduate students of medical statistics will also find much of benefit, as will graduate students of the medical and health sciences who have to analyse QoL data for their dissertations and projects.

Most of the book is written at an intermediate level for readers who are going to collect and analyse their own QoL data. It is expected that readers will be familiar with basic statistical concepts such as hypothesis testing (P -values), confidence intervals, simple

statistical tests (e.g. the t -test and chi-square test) and simple linear regression. The more advanced topics, in the later chapters, such as marginal generalized linear models for longitudinal data, will require a more thorough statistical knowledge, but are explained in as simple a way as possible with examples.

Stephen J. Walters
Sheffield, UK

1

Introduction

Summary

Quality of life (QoL) is a complex concept with multiple dimensions. This book will assume a wide definition for this concept. It will describe the design, assessment, analysis and interpretation of single- and multi-item, subjective measurement scales. These measurement scales all have the common feature of using a standardized approach to assessing a person's perception of their own health by using numerical scoring systems, and may include one or several dimensions of QoL. This chapter will provide a brief history of QoL assessment; describe the different types of QoL assessment tools available and give reasons why it is important to measure QoL.

1.1 What is quality of life?

Quality of life (QoL) is a complex concept with multiple aspects. These aspects (usually referred to as domains or dimensions) can include: cognitive functioning; emotional functioning; psychological well-being; general health; physical functioning; physical symptoms and toxicity; role functioning; sexual functioning; social well-being and functioning; and spiritual/existential issues (see Figure 1.1). This book will assume a wide definition for this concept. It will describe the design, assessment, analysis and interpretation of single- and multi-item, subjective measurement scales. This broad definition will include scales or instruments that ask general questions, such as 'In general, how would you rate your health now?', and more specific questions on particular symptoms and side effects, such as 'During the past week have you felt nauseated?'. These measurement scales all have the common feature of using a standardized approach to assessing a person's perception of their own health by using numerical scoring systems, and may include one or several dimensions of QoL.

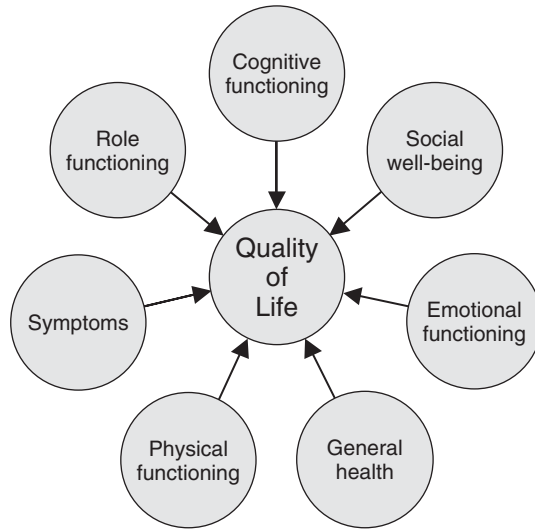


Figure 1.1 Examples of QoL domains.

1.2 Terminology

Researchers have used a variety of names to describe QoL measurement scales. Some prefer to use the term *health-related quality of life* (HRQoL or HRQL), to stress that we are only concerned with health aspects. Others have used the terms *health status* and *self-reported health*. The United States Food and Drug Administration (FDA) has adopted the term *patient-reported outcome* (PRO) in its guidance to the pharmaceutical industry for supporting labelling claims for medical product development (FDA, 2006). However, not all people who complete such outcomes are ill and patients, and hence PRO could legitimately stand for *person-reported outcome*. Mostly, we shall assume that the QoL instrument or outcome is self-reported, by the person whose experience we are interested in, but it could be completed by another person or proxy. The term *health outcome assessment* has been put forward as an alternative which avoids specifying the respondent. This book will follow convention and use the now well-established term *quality of life*.

1.3 History

The World Health Organisation (WHO, 1948) declared health to be ‘A state of complete physical and mental social well-being, and not merely the absence of disease and infirmity’. This definition was one of the first to emphasize other facets of health, such as physical, mental and social, in connection with disease and infirmity.

The Karnofsky Performance Scale (Karnofsky and Burchenal, 1949) was one of the first instruments to undertake a wider assessment of patients’ functional impairment apart from clinical and physiological examination. It involves health-care staff assessing patients, using a simple single-item 11-point scale ranging from 0 for ‘dead’ to 100

Table 1.1 The Karnofsky Performance Scale.

Description	Score
Normal; no complaints; no evidence of disease	100
Able to carry on normal activity; minor signs and symptoms of disease	90
Normal activity with effort; some signs and symptoms of disease	80
Cares for self; unable to carry on normal activity or do work	70
Requires occasional assistance, but is able to care for most personal needs	60
Requires considerable assistance and frequent medical care	50
Disabled; requires special care and assistance	40
Severely disabled; hospitalization indicated although death not imminent	30
Very sick; hospitalization necessary; requires active support treatment	20
Moribund; fatal processes progressing rapidly	10
Dead	0

for 'Normal' (see Table 1.1). It can be used to compare effectiveness of different therapies and to assess the prognosis in individual patients.

This led to the development of the next generation of questionnaires which focused on broader aspects of QoL, such as emotional well-being, social functioning, impact of illness, perceived distress and life satisfaction. These included the Nottingham Health Profile (NHP, Hunt *et al.*, 1980, 1981) and the Sickness Impact Profile (SIP, Deyo *et al.*, 1982). Again, I shall describe the NHP and SIP as QoL scales although their developers neither designed them nor claimed them as QoL scales.

Newer instruments such as the Medical Outcomes Study (MOS) Short Form (SF)-36 (Ware and Sherbourne, 1992) now place more emphasis on the subjective aspects of QoL, such as emotional, role, social and cognitive functioning. The SF-36 is the most commonly used QoL measure in the world today. It contains 36 questions measuring health across eight dimensions: Physical Functioning (PF); Role-Physical (role limitations due to physical health, RP); Social Functioning (SF); Vitality (VT); Bodily Pain (BP); Mental Health (MH); Role-Emotional (role limitations due to emotional problems, RE); and General Health (GH).

Quality of life was introduced by the MEDLINE (Medical Literature Analysis and Retrieval System Online) international literature database of life sciences and biomedical information as a heading in 1975, and accepted as a concept by Index Medicus in 1977. Since then there has been a rapid expansion of interest in the topic, with an exponential increase in the number of citations of QoL in the medical literature (see Figure 1.2).

In 1991, the first edition of a new international, multidisciplinary journal devoted to the rapid communication of original research, theoretical articles and methodological reports related to the field of QoL in all the health sciences was published, entitled *Quality of Life Research*. The February 2004 issue was largely devoted to the publication of abstracts from the first meeting of the International Society for Quality of Life Research (ISOQOL), held in Brussels. ISOQOL's mission is the scientific study of QoL relevant to health and health care. The Society promotes the rigorous investigation of health-related QoL measurement from conceptualization to application and practice. ISOQOL fosters the worldwide exchange of information through scientific publications, international conferences, educational outreach, and collaborative support for QoL initiatives.

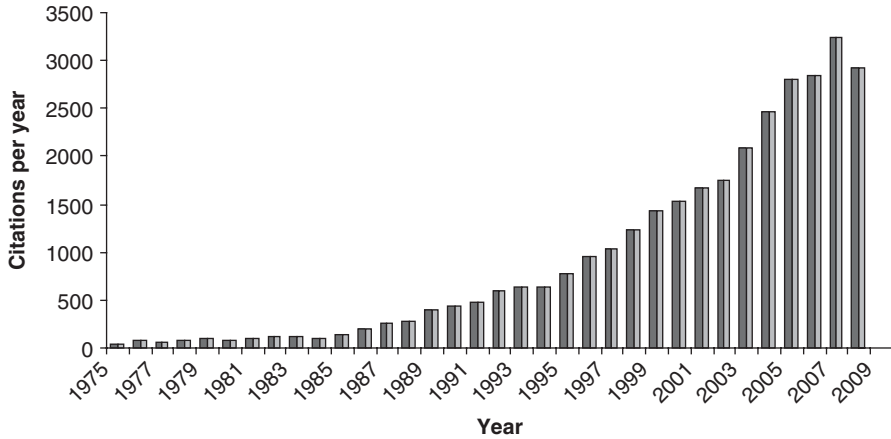


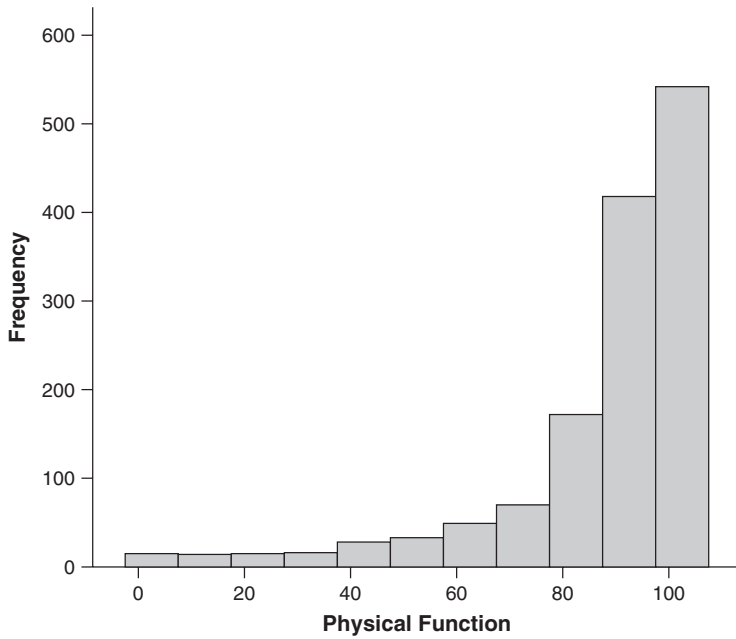
Figure 1.2 MEDLINE database ‘quality of life’ focused subject heading citations, 1975–2008.

1.4 Types of quality of life measures

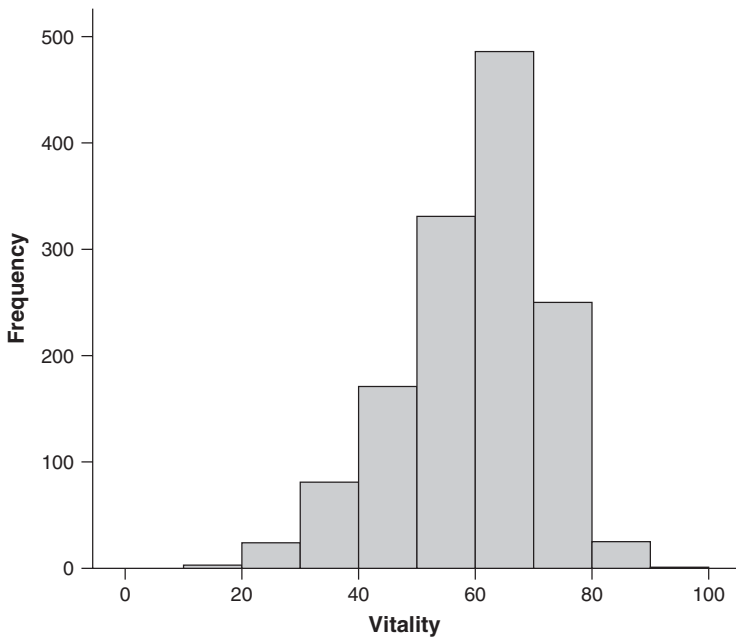
The SF-36 is an example of a QoL instrument that is intended for general use, irrespective of the illness or condition of the patient. Such instruments are often termed *generic* measures and may often be applicable to healthy people too and hence used in population surveys. Figure 1.3 shows the distribution of the eight main dimensions of the SF-36 from a general population survey of United Kingdom residents (Brazier *et al.*, 1992). The SF-36 dimensions are scored on a 0 to 100 (‘good health’) scale. Figure 1.3 shows that the SF-36 outcome, in common with many other QoL scales, generates data with a discrete, bounded and skewed distribution. Figure 1.4 shows how physical functioning in the general population (Walters *et al.*, 2001a) declines rapidly with increasing age.

The SF-36 is also an example of a *profile* QoL measure since it generates eight separate scores for each dimension of health (Figure 1.3). Other generic profile instruments include the SIP and NHP (see Section 1.3). Conversely, some other QoL measures generate a single summary score or *single index*, which combines the different dimensions of health into a single number. An example of a single index QoL outcome is the EuroQol or EQ-5D as it is now named (EuroQol Group, 1990).

Generic instruments are intended to cover a wide range of conditions and have the advantage that the scores from patients with various diseases may be compared against each other and against the general population. For example, Figure 1.5 compares the mean SF-36 dimension scores of a group of patients six months after acute myocardial infarction (AMI) with an age and sex matched general population sample (Lacey and Walters, 2003). The AMI sample has lower QoL on all eight dimensions of the SF-36 than the general population sample. On the other hand, generic instruments may fail to focus on the issues of particular concern to patients with disease, and may often lack the sensitivity to detect differences that arise as a consequence of treatments that are compared in clinical trials. This has led to the development of *condition-* or *disease-specific* questionnaires. Disease-specific QoL measurement scales are comprehensively reviewed by Bowling (2001, 2004). Examples of disease-specific QoL questionnaires include the



(a)



(b)

Figure 1.3 Distribution of the eight SF-36 dimensions from a general population survey ($n = 1372$); a score of 100 indicates ‘good health’ (data from Brazier *et al.*, 1992).

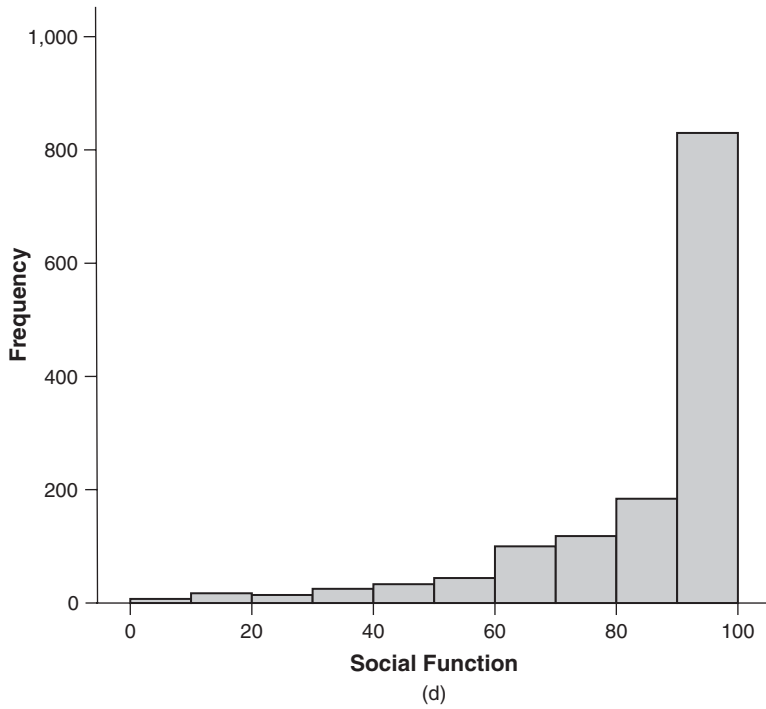
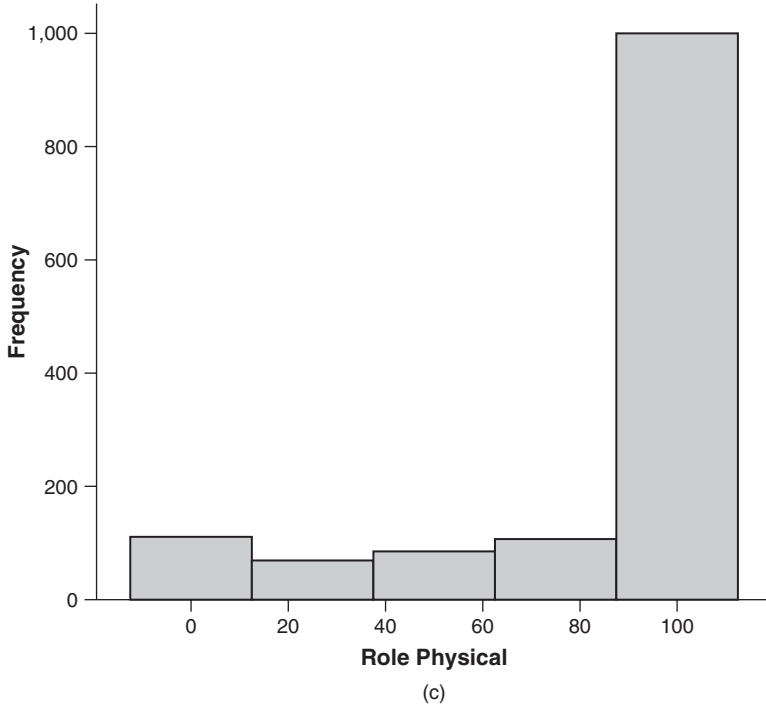


Figure 1.3 (Continued)

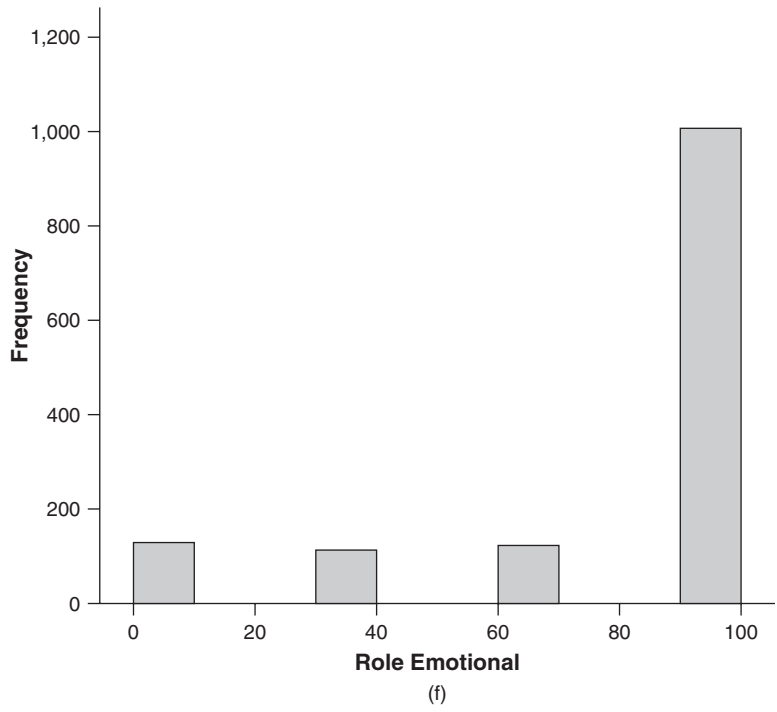
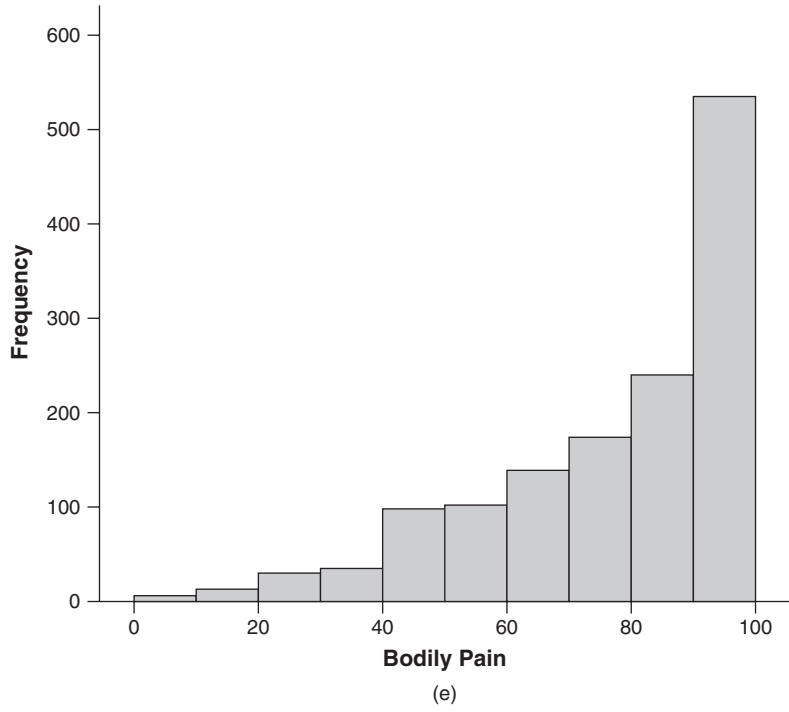


Figure 1.3 (Continued)

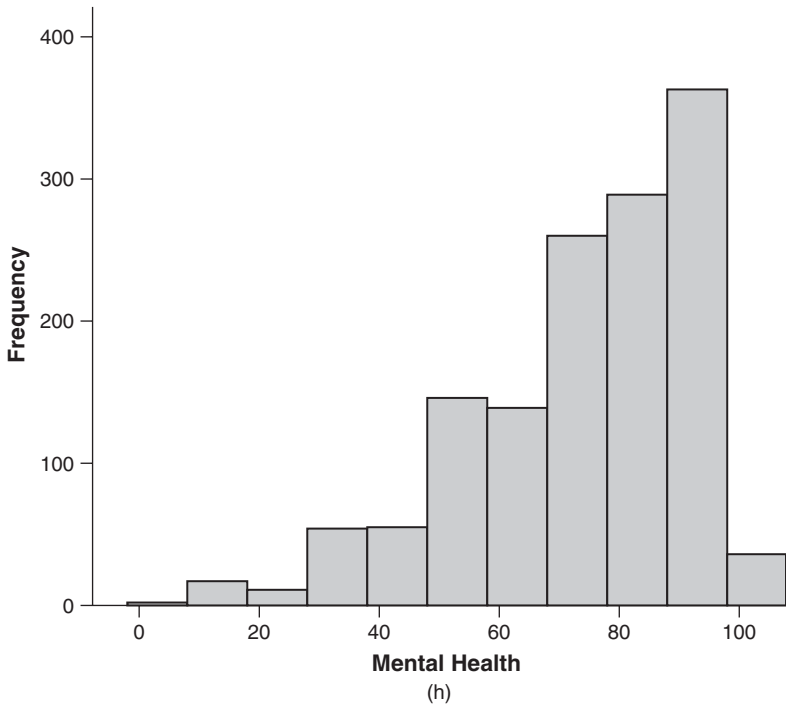
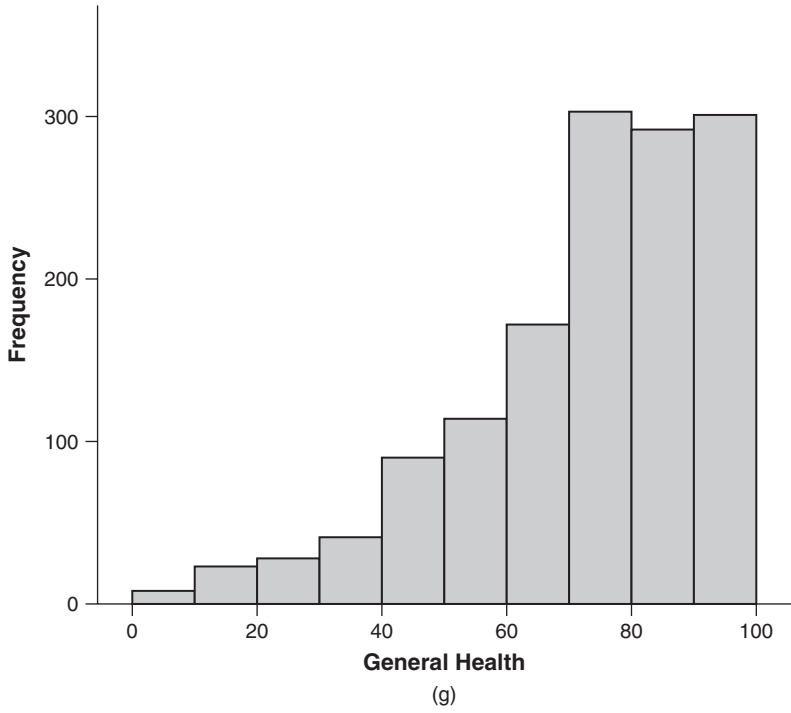


Figure 1.3 (Continued)

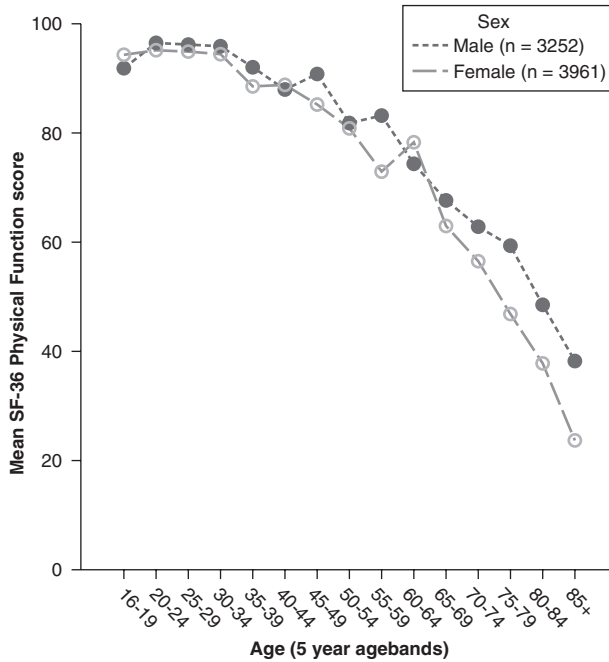


Figure 1.4 Mean SF-36 Physical Functioning age profile by sex (data from Walters *et al.*, 2001a).

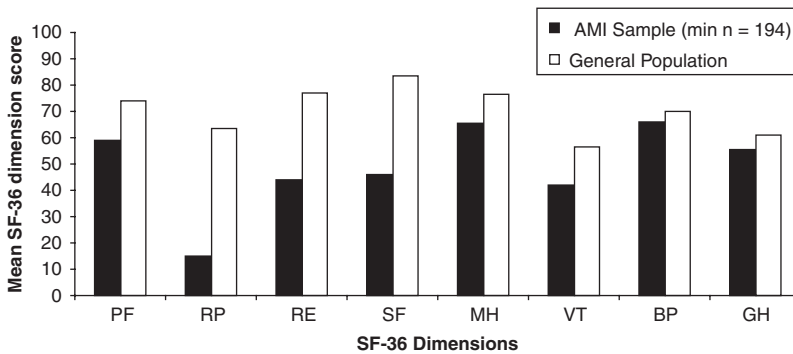


Figure 1.5 Profile of mean SF-36 scores for an acute myocardial infarction sample (six weeks after infarction) compared with an age and sex matched general population sample (data from Lacey and Walters, 2003).

cancer-specific 30-item European Organisation for Research and Treatment of Cancer (EORTC) QLC-30 questionnaire (Aaronson *et al.*, 1993) and the cancer-specific 30-item Rotterdam Symptom Checklist (RSCL, de Haes *et al.*, 1990).

The instruments described above claim to measure general QoL, and usually include at least one question about overall QoL or health. Sometimes investigators may wish to explore particular aspects or concepts in greater depth. There are also instruments

for specific aspects of QoL. These specific aspects may include anxiety and depression, physical functioning, pain and fatigue. Examples of instruments which evaluate specific aspects of QoL are: the Hospital Anxiety and Depression Scale (HADS, Zigmond and Snaith, 1983) and the Beck Depression Inventory (BDI, Beck *et al.*, 1961) instruments for measuring anxiety and depression; the McGill Pain Questionnaire (MPQ, Melzack, 1975) for the measurement of pain; the Multidimensional Fatigue Inventory (MFI, Smets *et al.*, 1995) for assessing fatigue and the Barthel Index (Mahoney and Barthel, 1965) for assessing disability and functioning.

1.5 Why measure quality of life?

There are several reasons why we should measure quality of life in both a research setting and in routine clinical practice. The use of QoL assessment in routine clinical practice may make communication with patients easier and help find out information about the range of problems that affect patients. Medicine and health care have traditionally tended to focus on symptom relief as the main outcome measure. QoL assessment may help improve symptom relief, care or rehabilitation for an individual patient. Using QoL instruments may reveal other issues that are equally or more important to patients than just symptom relief. The patient's self-assessment of their own QoL may differ substantially from the judgement of other health-care staff. Individual patient preferences may also differ from those of other patients. Therefore it is important to measure QoL from the patient's perspective, using a self-completed questionnaire to establish their views and preferences. Cured patients and long-term survivors may have ongoing problems long after their treatment is successfully completed. These ongoing problems may be overlooked, so again it is important to measure QoL long term and to look for late problems of psychosocial adaptation.

QoL assessments may be included in research studies such as randomized controlled trials (RCTs). The main reason is to compare the study treatments with respect to those aspects of QoL that may be affected by the treatment. These treatment comparisons will include both the positive benefits from trials that are expected to improve QoL, and any negative changes, from toxicity and side effects of treatment.

QoL can be a predictor of treatment success, and hence pre-treatment assessment of QoL may have prognostic value. Fayers and Machin (2007) suggest that the direction of the association between QoL scores and treatment outcome is not clear. Do QoL scores reflect an early perception by the patient of the disease progression? Alternatively, does QoL status in some way influence the course of the disease? Whatever the nature of the association, it is important to assess QoL and use it when making medical decisions for individual patients.

QoL assessment can also be used to make decisions on treatments at a population level, rather than an individual patient level. QoL outcomes can be used in economic evaluations alongside clinical trials to assess the clinical and cost-effectiveness of new health technologies.

There is an ongoing thoughtful discussion about the meaning of QoL, and about what should be measured. In the face of this debate, it is still important to measure quality of life as well as clinical and process-based outcomes. This is because 'All of the these [QoL] concepts reflect issues that are of fundamental importance to patients' well-being. They are all worth investigating and quantifying' (Fayers and Machin, 2007).

1.6 Further reading

The two books by Bowling extensively describe the numerous QoL instruments now available (Bowling 2001, 2004). The book by Fayers and Machin (2007) covers all aspects of QoL assessment, analysis and interpretation. Fairclough (2002) goes into more detail about the statistical analysis of QoL data in RCTs with a strong emphasis on imputation methods for missing data and the modelling of longitudinal data. The book edited by Fayers and Hays (2005) covers a variety of topics in its 27 chapters with contributions from 31 authors and provides an overview of QoL assessment, analysis and interpretation.

Measuring quality of life

Summary

This chapter describes the principles of measurement scales and introduces the methods for developing and validating new questionnaires. Psychometric methods lead to scales that are based upon items reflecting a patient's level of QoL. The clinimetric approach makes use of composite scales that may include symptoms and side effects. The remainder of the chapter provides an overview of the stages of developing and validating new questionnaires and the principles that are involved.

2.1 Introduction

Questionnaires for assessing QoL usually contain multiple questions or items, although rarely a few may attempt to rely upon a single global question to assess overall QoL. For example, 'Overall, what has your quality of life been like over the last week?'. Some QoL questionnaires are designed so that all items are combined together to produce an overall score. Most instruments attempt to group the items into separate 'scales' corresponding to different dimensions of QoL. This chapter explores the relationship between items and scales and introduces the concepts underlying QoL scales and their measurement.

2.2 Principles of measurement scales

2.2.1 Scales and items

Most QoL instruments consist of many questions or items. These items are usually combined to generate a dimension or domain score. Figure 2.1 shows this process graphically. Some of these items may aim to measure a relatively simple aspect of QoL, such as physical symptoms like nausea, vomiting or constipation. For these relatively simple aspects of QoL a single question or item may be sufficient to measure the underlying dimension.

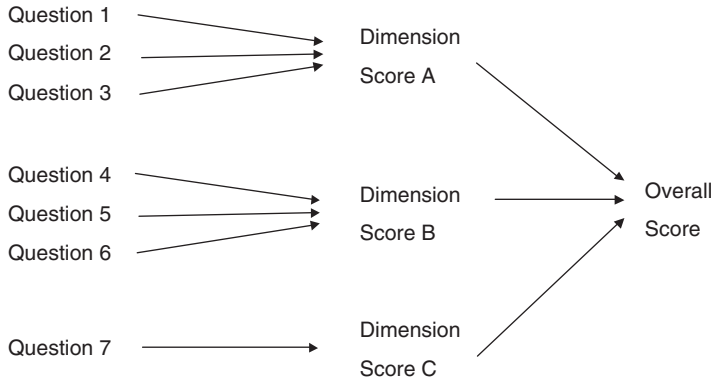


Figure 2.1 Items and scales.

For example, the EORTC QLQ-C30 questionnaire (Aaronson *et al.*, 1993), measures the symptom of constipation with the single question, ‘During the past week, have you been constipated?’. The question has four possible response options: not at all; a little; quite a bit; very much.

The more complex psychological dimensions of QoL such as anxiety and depression are usually more vaguely defined in a subject’s understanding of QoL. These dimensions are typically measured by the use of several questions in multi-item scales. For example, the Hospital Anxiety and Depression scale (HADS) consists of 14 items, with seven items on the ‘anxiety’ aspect and the other seven items assessing ‘depression’ (Zigmond and Snaith, 1983).

2.2.2 Constructs and latent variables

Fayers and Machin (2007) describe QoL as a complex construct that cannot be adequately measured by a single global question. They suggest that QoL has a number of dimensions (see Figure 1.1), each of which should be thought of as an underlying ‘construct’. These constructs are represented or measured by ‘latent variables’, which we measure by asking the subject one or, more typically, a number of separate questions. For this reason QoL instruments commonly contain multiple questions to assess the underlying latent variables.

Example: Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983)

The HADS questionnaire (see Appendix A) is a QoL instrument with a simple theoretical structure (see Figure 2.2). It assumes that there are two different and distinct constructs of ‘anxiety’ and ‘depression’, which are meaningful to patients and can be quantified. It is assumed that anxiety and depression cannot be adequately measured by a single question, such as ‘How anxious are you today?’ (not at all, a little, quite a bit, very much), and that multiple questions must be employed. The HADS consists of 14 items, with seven questions relating to anxiety and seven questions relating to depression. □

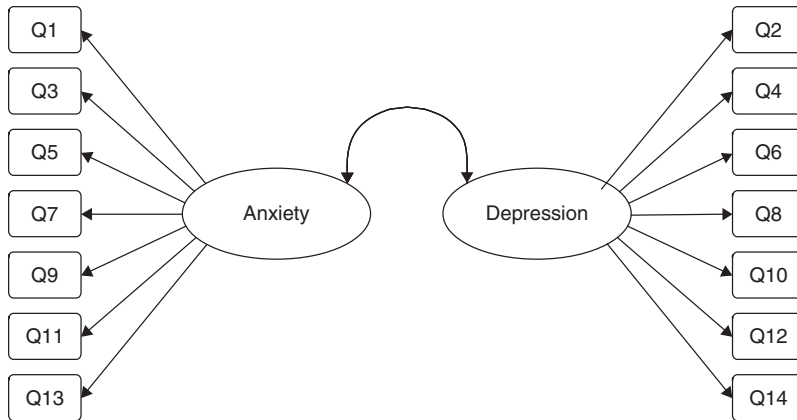


Figure 2.2 The theoretical structure of HADS. Reproduced with permission from *Fayers, P.M., Machin, D., Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes*. 2nd edition. John Wiley & Sons Ltd, Chichester. © 2007 John Wiley & Sons Ltd.

2.3 Indicator and causal variables

2.3.1 Indicator variables

Most items in personality tests, intelligence tests, educational attainment tests and other psychometric assessments reflect a level of ability or a state of mind. Such items do not alter or influence the latent construct that they measure. These items are *indicator variables* (Fayers and Machin, 2007). In common with most questionnaires that assess psychological aspects of QoL, the HADS items (see Appendix A) are mainly indicator variables. For example, ‘During the past week, I feel tense or “wound up”’. The question has four possible response options: most of the time; a lot of the time; from time to time; not at all.

2.3.2 Causal variables

The symptoms (such as nausea, vomiting and constipation) assessed in QoL scales, such as the EORTC QLQ-C30 (Aaronson *et al.*, 1993) may cause a change in QoL. A patient who gets serious symptoms is likely to have their QoL affected by those symptoms. The reason for including symptoms in QoL instruments is principally that symptoms are believed to affect QoL. However, having a poor QoL does not imply that the patient has specific symptoms (such as nausea, vomiting and constipation). Typically, a single causal item may be enough to change the latent QoL variable. It is unnecessary, and usually rare, for each patient to suffer from all the symptoms in order to have a poor QoL. One serious symptom, such as extreme nausea, may be enough to reduce overall QoL.

Fayers and Machin (2007) caution that the above distinction between indicator and causal variables is not entirely clear-cut. Variables may frequently be partly indicator

Box 2.1 Identifying causal items (Fayers and Machin, 2007)

- “Thought test”
 - Consider a typical patient from the target population. For an item called, say, item X:
 - a) If the level of item X changes, is the patient’s quality of life likely to change?
 - b) If the patient’s quality of life improves (or deteriorates), do we expect this to be reflected by a change in item X?
 - If the answer to (a) is “yes” and (b) is “no”, the item is likely to have a causal component.

and partly causal. For example, a patient may experience symptoms such as nausea and vomiting, become anxious and depressed, and then perceive and report the symptoms as being worse than they are. An initial causal variable has acquired indicator properties. So how can we identify causal variables? Fayers and Machin (2007) describe the *thought test* for identifying causal variables (see Box 2.1).

2.3.3 Why do we need to worry about the distinction between indicator and causal items?

Indicator variables assume that the observed responses to the items depend solely upon the level of the underlying latent variable. That is, if QoL is ‘good’, then this should be reflected in good or high levels of response on the various items. Furthermore, if the observed values of the items are correlated, then these correlations arise solely because of the effect of the latent variable. Causal variables are not correlated with each other through the different levels of QoL. They do not have correlations that arise through their parallel nature. Their correlations arise through an underlying variable – such as treatment, or stage or extent of disease. Thus causal variables may exhibit seemingly strange correlations that are nothing to do with changes in QoL. Causal items do not reflect QoL, they affect it. Therefore indicator and causal items behave in fundamentally different ways and this will have a considerable impact upon the design of QoL scales.

2.3.4 Single-item versus multi-item scales

Multi-item scales are commonly used to assess specific aspects of QoL. Responses from multiple items usually have several advantages over a score estimated from the responses to a single item in terms of reliability, precision, validity and scope (see Box 2.2).

2.4 The traditional psychometric model

The most common psychometric model is the *parallel tests* model. In this model each measurement item is a ‘test’ or question that reflects the level of the underlying construct or latent variable. Each item is distinct from the others, but is similar and comparable in all important respects. They differ only as a consequence of random error. These items

Box 2.2 Single-item versus multi-item scales

- *Reliability.* A reliable test is one that measures something in a consistent, repeatable and reproducible manner. Patient variability means that a single-item test is potentially unreliable since we only have one attempt to measure the QoL aspect we are interested in. Reliability is increased by including and averaging a number of ‘parallel’ items.
- *Precision.* Multi-item tests can have greater precision.
- *Validation.* The items of a multi-item scale can be compared against each other.
- *Scope.* QoL is a complex issue and not easily assessed by a single question.

Box 2.3 The traditional psychometric model – parallel tests

1. Each item is a test, which gives an unbiased estimate of the latent variable, with a random error term ε .
2. The error terms of the items are uncorrelated.
3. The error terms are uncorrelated with the latent variable.
4. The amount of influence from the latent variable to each item is assumed to be the same for all items.
5. Each item is assumed to have the same amount of error as any other item. The influence of extraneous factors is assumed to be equal for all items.

are then described as being parallel. The theory of parallel tests underpins the majority of QoL instruments which use simple summated (Likert) scales (see Box 2.3).

2.4.1 Psychometrics and QoL scales

The majority of QoL instruments have been designed on the principles of parallel tests and summated Likert scales. The related psychometric methods to a large extent assume that the scales contain solely indicator variables. The inter-item correlations that exist between causal variables can render these methods inapplicable (Fayers and Machin, 2007).

2.5 Item response theory

So-called modern psychometric theory largely centres on *item response theory* (IRT, Van der Linden and Hambleton, 1997). Items have varying ‘difficulty’. It is assumed that patients will have different probabilities of responding positively to each item, according to their level of ability (that is, the level of the latent variable). Traditional methods focus upon averages; whereas IRT emphasizes the probabilities of responses. The design of scales using IRT methods is markedly different from traditional methods.

2.5.1 Traditional scales versus IRT

Traditional Likert summated scales assume items of broadly similar difficulty, with response categories to reflect severity or degree of response level. In contrast, IRT scales are based upon items of varying difficulty, and frequently each item will have only two response categories, such as ‘yes’ or ‘no’. IRT models assume that the observed variables reflect the value of the latent variable, and that the item correlations arise solely by virtue of this relationship with the latent variable. Thus it is implicit that all items are indicator variables. The IRT model is inappropriate for symptoms and other causal variables (Fayers and Machin, 2007).

Example: A scale with items of varying difficulty – the SF-36 Physical Functioning dimension score

Question 3 of the SF-36 (see Appendix A) has 10 items of varying difficulty about activities that you might do during a typical day. The least difficult or ‘easiest’ item to answer, if the respondent has a good level of physical functioning, is the question on bathing and dressing oneself (question 3j). In general, most people with good physical functioning will not be limited at all in carrying out this daily activity. Conversely, the most difficult or ‘hardest’ item to answer, if the respondent has a poor level of physical functioning, is the question on vigorous activities (question 3a). The other eight items on the questionnaire appear to reflect levels of varying difficulty on the underlying latent physical functioning scale. □

2.6 Clinimetric scales

Many clinical scales possess fundamentally different attributes from psychometric scales. Their development and validation should therefore proceed along separate paths. A ‘good’ and useful *clinimetric* scale may consist of items comprising a variety of symptoms and other clinical indices. It does not necessarily need to satisfy the same requirements that are demanded of other scales. Clinicians try to measure multiple attributes with a single index – for example, the Apgar Score (Apgar, 1953) for assessing the health of newborn babies or the Glasgow Coma Score (Teasdale and Jennett, 1974).

Example: The Glasgow Coma Score (Teasdale and Jennett, 1974)

The Glasgow Coma Scale (GCS) is a neurological scale which aims to give a reliable, objective way of recording the conscious state of a person, for initial as well as continuing assessment. The GCS was initially used to assess the level of consciousness after head injury, and it is now used by doctors as being applicable to all acute medical and trauma patients. In hospital it is also used in chronic patient monitoring in, for instance, intensive care. The scale combines three seemingly disparate symptoms related to the eye, verbal and motor responses. The lowest possible GCS score is 3 (deep coma or death), whilst the highest is 15 (fully awake person). Generally, comas are classified as: severe, with $GCS \leq 8$; moderate, $GCS 9-12$; minor, $GCS \geq 13$. □

2.7 Measuring quality of life: Indicator or causal items

QoL instruments commonly contain both indicator and causal variables. Psychometric methods assume that all items are indicator variables. Clinimetric methods are more relevant for causal items. Fayers and Machin (2007) suggest that QoL instruments serve two different functions:

1. They alert the clinician to problems concerning symptoms and side effects.
2. They assess overall QoL and its aspects.

For the first purpose each symptom is usually reported separately, but if a multi-item scale is needed, then this is best constructed on clinimetric principles. For the second purpose, indicator variables are usually the most effective, chosen and validated using psychometric techniques.

2.8 Developing and testing questionnaires

Development of QoL instruments requires much painstakingly detailed work, patience, time and resources. The validation of a new QoL scale depends upon collecting and analysing data from samples of patients or others. Statistical and psychometric techniques can only confirm that a scale is valid in so far as it performs in the manner that is expected. Thus quantitative techniques presuppose that the scale has been carefully and sensibly designed. Therefore QoL scale development should follow rigorous pre-specified qualitative and quantitative procedures.

QoL instrument development, modification and validation usually occur in a non-linear fashion with a varying sequence of events, simultaneous processes or iterations. This iterative process is shown in Figure 2.3 and is discussed in detail below. One or more parts of the original process may be repeated in a new QoL instrument development, modification, or change in application of an existing instrument. This section describes the steps usually taken in the development of a QoL instrument.

2.8.1 Specify the research question and define the target population

The first stage when developing a QoL instrument is to clearly specify the research question. This should include specification of the objectives in measuring QoL, a working definition of what is meant by 'quality of life', the identification of the intended groups of respondents, and suggestions as to the concepts or dimensions of QoL that are to be assessed (Fayers and Machin, 2007).

Examples of the objectives are whether the new instrument is intended for comparison of treatment groups in randomized clinical trials, or for use in routine clinical practice for individual patient treatment and management. Possible definitions of QoL might place greater or lesser importance upon symptoms, psychological, spiritual or other aspects. Depending on the definition of the target population of respondents, there may be more or less prominence given to disease and treatment-related issues. All these factors will

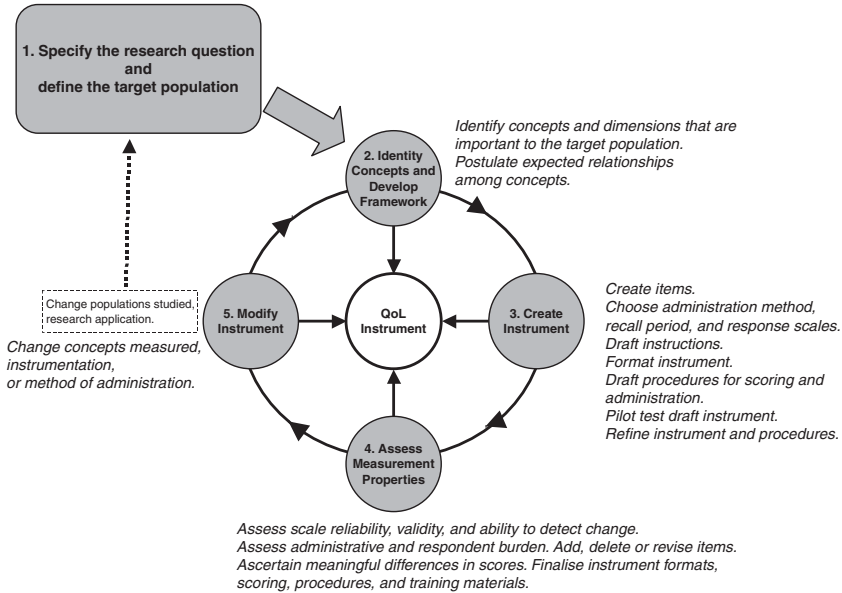


Figure 2.3 The QoL instrument development and modification process (adapted from FDA, 2006).

affect decisions about the dimensions of QoL to be assessed, the number of questions, length of the questionnaire and the scope and content of the questions.

Before identifying the concepts to be measured by the new QoL instrument, it is essential to define the target population. What is the range of diseases to be investigated? Are the symptoms and QoL issues the same for all disease subgroups? What is the range of treatments? For example, in cancer there can be a wide range of treatments, including chemotherapy, radiotherapy, hormone therapy and surgery. What is the severity range (advanced or early disease)? A QoL instrument should ensure that it is appropriate for the range of diseases and treatments to be investigated. As Fayers and Machin (2007) comment, the detailed specification of the intended target population is second in importance only to the specification of the research questions and the definition of QoL or of the aspects of QoL that are to be investigated.

2.8.2 Identify concepts

The second phase of developing a QoL instrument is to identify the concepts and dimensions that are important to the subjects (the intended target population) and the research application (both of which were defined in the previous step). This phase involves generating an extensive and exhaustive list of all QoL issues that are relevant to the dimensions of interest, by searching the literature, interviews with health-care professionals, discussions with patients and expert opinion.

Once the concepts have been identified, it is helpful to hypothesize the expected relationships among these concepts. This should include how individual items are associated with each other, how items are associated with each dimension, and how dimensions

are associated with each other and the general concept of interest. A diagram of the expected relationships among the items and dimensions can help show these relationships. Figure 2.1 shows a general example of a conceptual framework where dimensions A, B and C each represent related but separate concepts. Items in this diagram are aggregated into dimensions. For dimension C a single item is sufficient to measure this aspect of QoL. In some measures, dimensions can be aggregated into an overall score.

Example: EORTC head and neck cancer-specific module (Bjordal *et al.*, 1994)

Literature searches:

46 relevant references were found. Hence 57 potential issues were identified. These were divided into five areas: pain-related, nutrition, dental status, other symptoms, and functional aspects.

Specialist interviews:

21 specialist nurses, oncologists and surgeons.

17 of the 57 issues identified in the literature search were regarded as being irrelevant, too rare, or too broad in scope.

59 new issues were also proposed; 11 were added, resulting in a provisional list of 43 issues.

Patient interviews:

6 of the issues that were felt by patients to be of low relevance or unimportant were deleted.

21 new symptom or problem issues and 5 new function issues were identified. The revised list covered 37 issues. □

2.8.3 Create instrument

2.8.3.1 Item generation

Having identified all the relevant dimensions, items can be generated to reflect these dimensions. The first stage of item generation usually involves searches of relevant journals and bibliographic databases, to ensure that all the dimensions previously thought to be relevant are included. Any existing instruments that address the same or related areas of QoL assessment should be identified and reviewed. From these sources, a list of potential QoL items for inclusion in the questionnaire can be identified.

As before, there are several approaches to generating items: reviewing the literature (i.e. other questionnaires – this is a very incestuous business!); interviews with health-care professionals; interviews with patients; and expert opinion. There is always a certain amount of editing by instrument designers in order to limit the size of the questionnaire, remove ambiguity (see Box 2.4 on the wording of the questions) and to fit into a standard format. There is a trade-off between reliability, requiring more than one item per concept, and practicality, requiring the minimum number of items.

Box 2.4 Wording of questions

- Make questions and instructions brief and simple.
For example, ill patients and the elderly may be confused by long, complicated sentences.
- Avoid small, unclear typefaces.
Elderly patients may not have good eyesight.
- ‘Not applicable’ questions may result in missing or ambiguous answers.
For example, ‘Do you have difficulty going up stairs?’ is not applicable to someone bedridden.
- Potentially embarrassing or offending questions should be avoided, put at the end, or made optional.
For example, before a question about sex life, the FACT-G writes: ‘If you prefer not to answer it, please check this box and go to the next section.’
- Avoid double negatives.
For example, ‘I don’t feel less interest in sex (Yes/No)’.
- If two questions have similar wording, emphasize the differences, using underlining, bold, or italics.
For example, questions 4 and 5 of the SF-36 are very similar apart from the underlined phrases ‘as a result of your physical health’ and ‘as a result of any emotional problems’.
- Underlining and similar methods also draw attention to key words or phrases.
For example, emphasize the time frame of questions, such as ‘during the past 7 days’.
- Consider including both positively phrased and negatively phrased items.
For example, the HADS includes ‘I feel tense or “wound up”’ and ‘I can sit at ease and feel relaxed’.

2.8.3.2 Choice of administration or data collection method

Developers of QoL instruments should consider how the new QoL questionnaire is to be administered to the subjects. Possible modes of administration include: interview, paper-based self-administration, electronic, web-based, and interactive voice response formats. The majority of QoL instruments are designed to be self-completed paper-based questionnaires.

2.8.3.3 Choice of recall period

The development of the items for the QoL instrument should also consider the choice of recall period for the questions. The choice of recall period that is most suitable will depend on the purpose and intended use of the instrument, the characteristics of the disease or condition and the treatment to be tested. However, QoL instruments with items that require respondents to rely on memory, particularly if they must recall their

QoL over a long period of time or to average their response over a period of time, may have reduced accuracy and should be avoided. It is usually best to construct items that ask subjects to describe their current QoL status, rather than to ask them to compare their current state with an earlier period or attempt to average their experiences of over time. For example, the question ‘How would you rate your overall quality of life today?’ is to be preferred to ‘Compared to one year ago, how would you rate your overall quality of life now?’.

2.8.3.4 Choice of response options

It is also important to make sure that the response options to the items are consistent with the purpose and intended use of the QoL instrument. Table 2.1 describes the types of response options that are typically used in QoL instruments.

Response choices are usually regarded as being suitable when:

- The wording used in the responses is clear and appropriate.
- Responses are appropriate for the intended target population.
- Responses offer a clear distinction between choices.
- Instructions to subjects for completing the questionnaire and selecting response options are adequate.
- Response options are appropriately ordered and appear to represent equal intervals.
- Response options avoid potential floor or ceiling effects.
- Response options do not bias the direction of responses.

2.8.3.5 Draft procedures for scoring of items and dimensions




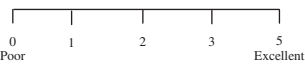
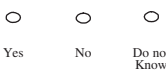
For each item, numerical scores are generally assigned to each response category based on the most appropriate scale of measurement for the item (e.g. nominal, ordinal, interval or ratio scales – see Box 2.5).

The scoring systems in most questionnaires are often arbitrary and chosen for their simplicity. A common method involves ‘adding the ticks’, or for Likert scales, with n response options, simply assigning scores between 1 and n or between 0 and $n - 1$, for each response, and then adding the item response numbers to derive an overall score for the domain (e.g. the SF-36 and the HADs).

Example: Scoring the HADS (Zigmond and Snaith, 1983)

- The HAD scale consists of 14 items on two subscales (seven for anxiety and seven for depression).
- Ratings by subjects are made on four-point ordinal scales, which represent the degree of distress: 0 = not at all; 1 = occasionally; 2 = a lot of the time; 3 = most of the time.
- Items are summed on each of the seven-item anxiety and depression subscales to generate a score ranging from 0 to 21.

Table 2.1 Types of response options.

Type	Description	Example
Visual analogue scale (VAS)	A horizontal or vertical line of fixed length (usually 100 mm) with words that anchor the scale at the extreme ends and no words describing intermediate positions. Subjects are instructed to place a mark on the line corresponding to their perceived state.	<ul style="list-style-type: none"> How would you rate your overall quality of life, today? 
Anchored or categorized VAS	A VAS that has the horizontal or vertical line of fixed length (usually 100 mm) with words that anchor the scale at the extreme ends and words describing intermediate positions.	<ul style="list-style-type: none"> How would you rate your overall quality of life, today? 
Likert scale	An ordered set of discrete terms or statements from which subjects are asked to choose the response that best describes their state or experience.	<ul style="list-style-type: none"> How would you rate your overall quality of life, today? 
Rating scale	A set of numerical categories from which subjects are asked to choose a category that best describes their state or experience. The ends of the rating scales are anchored with words but the intermediate categories do not have descriptive labels.	<ul style="list-style-type: none"> How would you rate your overall quality of life, today? 
Checklist	Checklists provide a simple choice between a limited set of response options such as Yes, No, and Don't know.	<ul style="list-style-type: none"> Today would you rate your overall quality of life <u>as good</u>? 
Binary format	The simplest checklist with only two responses options such as yes or no.	<ul style="list-style-type: none"> Today would you rate your overall quality of life <u>as good</u>? 