WHO Long Form Scoring, Reliability, Validity and Norms for New Zealand

A thesis presented in fulfilment of the requirements for the degree of Master of Public Health

at Massey University, Wellington Campus, New Zealand

Karen Sarah Blakey

July 2006

Abstract

Background

Self-reported health measures provide information about a wider range of health outcomes than objective measures of health status, such as mortality and hospitalisation rates. National health surveys play a role in monitoring population health. The New Zealand Health Monitor (NZHM) is the organised, co-ordinated and integrated survey programme of the Ministry of Health in New Zealand. The New Zealand Health Survey (NZHS) is one of the chief surveys of the NZHM. One of the categories of information collected in the NZHM is health outcomes, and within this there is the subcategory of health status. The International Classification of Functioning and Disability (ICF) provides the framework to describe the critical elements of non-fatal health outcomes captured by health status instruments. NZHM is to collect data on most if not all of these 21 ICF dimensions.

The WHO Long Form was developed as the health module in the WHO Multi-country Survey Study. The WHO Long Form is made up of 20 health domains, some overlapping with the eight SF-36 domains. The WHO Long Form did not have a set scoring system for scales, unlike the SF-36 instrument. The SF-36 has been previously tested and validated in New Zealand in the 1996/97 NZHS.

Methods

The 2002/03 NZHS used a complex sample design. A total of 12, 929 people responded to the survey, with 12,529 respondents being included in the CURF dataset available for research. The health status section of the 2002/03 NZHS measures health-related quality of life (HRQL) covered 16 health and health-related domains. The questions were derived from the SF-36 and the WHO Long Form questionnaire on health status. The health domains covered in the 2002/03 NZHS were general health, vision, hearing, digestion, breathing, pain, sleep, energy and vitality, understanding, communication, physical functioning, self-care. The health-related domains covered in the 2002/03 NZHS were mental health, role-physical and role-emotional (usual activities), and social functioning.

There were five key aims specific to the current thesis. First, to group the WHO Long Form items in the 2002/03 NZHS into scales for each health domain and develop standard scoring protocols for each scale. Second, to test the reliability of the scales using standard

psychometric tests for the total NZ population and for major population subgroups. Third, to test the validity of the scales using the standard psychometric tests for the total NZ population and for major population subgroups. Fourth, to construct norms for the WHO Long Form scales for the NZ population. And finally, to provide recommendations for the health status component of future NZ health surveys.

Results

In summary, this thesis developed a method for producing scale scores for domains of health not previously measured in New Zealand Health Surveys, providing greater coverage of domains from the ICF. There were virtually no missing data for all items and subgroups within the questions used to develop the scales. The scaling approach was consistent with that for the SF-36, allowing the new scales to be presented alongside the SF-36 scales. All scales for the total population and major population subgroups met the required criterion for satisfactory psychometric properties, with the exception of digestion and bodily excretions scale. For the digestion and bodily excretions scale, the Cronbach's alpha was lower than that required for between group comparisons. The composite physical functioning and social functioning scales performed no better than the existing SF-36 scales and were highly correlated with these scales.

Conclusion

Notwithstanding the limitations of this study, key findings of interest are that the new WHO Long Form questions can be used to form scales that cover physical functioning, social functioning, vision, hearing, digestion and bodily excretions, breathing, self-care, understanding, communication and sleep. The majority of the questions and scales work for the NZ population and subgroups. All but one of the scales, digestion and bodily excretions, have satisfactory psychometric properties for the total population and major subpopulation groups of interest. The respondent burden is an important consideration for the NZHS, thus it cannot be argued that enough is gained from adding questions to the physical functioning and Social Functioning domains, thus it would be recommended that the SF-36 scales are used to measure there two domains of health. The new WHO Long Form scales can now be presented alongside the SF-36 scales and used in future analyses looking at interrelationships between factors such as health risk and health status.

Acknowledgements

Firstly I would like to thank the participants of the 2002/03 NZHS who gave freely of their time to take part in the survey. Without them this research would never have been possible!

A large number of people have provided me with support and encouragement during the course of this research. In particular I am indebted to Professor Neil Pearce, Director Centre for Public Health Research (CPHR) at Massey University in Wellington and Dr Barry Borman, Manager of Public Health Intelligence (PHI) in the Ministry of Health. Neil and Barry were both my two bosses and thesis supervisors during my time as Training Fellow with CPHR and PHI. I thank you both for believing that I would finish the thesis (despite there being times when I thought I would never get to the end of the tunnel!) and providing the environment for me to develop both professionally and personally. I am also especially grateful to Martin Tobias, Principal Technical Specialist (Epidemiology) at PHI for providing the initial motivation for this research. I am indebted to Matthew Cronin, previously employed as a statistician with PHI for his assistance with the statistical analyses. I would also like to thank Maria Turley, Senior Advisor (Nutrition/Epidemiology) for ongoing support as I carried out this research.

Finally I wish to acknowledge my wonderful partner Ralf who cooked me the most delicious meal ever to celebrate the completion of my first full thesis draft. Ralf, you have greatly enriched my life and I thank you for all the love and support that you give me.

I thought I would end with a quote from Albert Einstein, "Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning."

Table of contents

List of Tables

List of Figures

Chapter 1 - Introduction

Background

Due to the rising burden of chronic disease and a decrease in mortality from infectious diseases, the use of measures of health-related quality of life (HRQL) has increased. Self-reported health measures provide information about a wider spectrum of health outcomes than objective measures of health status, such as mortality and hospitalisation rates.(Scott, Tobias, & Sarfati, 1999)

A common way to measure health status or health-related quality of life is through scoring standardised responses to standardised questions (Lohr, 1989). Measures of positive health are important when measuring the health of the general population (Bowling, 1997). Self-reports introduce a consumer perspective to population health monitoring. The emphasis of such measures is on quality of life and wellbeing (Ministry of Health, 1999a, 1999b). Numerous questionnaires have been developed to measure HRQL, but the Medical Outcomes Study Short Form 36 (SF-36), which was developed in the United States, is the most widely used (Bowden & Fox-Rushby, 2003)) & (Scott et al., 1999).

National health surveys play a role in monitoring population health. The New Zealand Health Monitor (NZHM) is the organised, co-ordinated and integrated survey programme of the Ministry of Health in New Zealand (NZ). The aim of the New Zealand Health Monitor (NZHM) is to collect data that cannot be collected more effectively and efficiently through other means (e.g. administrative data collection or epidemiologic studies), are needed to inform decision making of the Ministry of Health or district health boards, and are population based (Ministry of Health, 2005).

One of the information domains of the NZHM is "health outcomes", and within this there is "health status". The two portions of health status relevant to this thesis are subjective (self-rated) health and functional limitation. The International Classification of Functioning and Disability (ICF) provides the framework to describe the essential elements of non-fatal health outcomes measured by health status instruments. The ICF was approved by the World Health Organisation (WHO) in May 2001 and identifies 21 key domains of health (World Health Organization, 2001), and the NZHM aims to collect data on most if not all of these dimensions.

The New Zealand Health Survey (NZHS) is one of the main surveys of the NZHM. The 1996/97 NZHS was a nationally representative survey of the general population that measured health status, health-related behaviour and health service utilisation. The major measure of self-reported health status (HRQL) was the SF-36 (Australia/New Zealand adaptation). The SF-36 measures eight health domains using eight scales: physical functioning, role-limitation, bodily pain, general health, vitality, social functioning, role-emotional and mental health. From these eight scales, two summary measures are calculated, the physical component score and the mental component score (Scott et al., 1999)). The SF-36 performed reasonably well nationally meeting standard criteria fort psychometric assessment, but not for older Māori or Pacific peoples (Scott, Sarfati, Tobias, & Haslett, 2000).

An alternative survey instrument for measuring health status is the WHO Long Form which was developed as the health module in the WHO Multi-country Survey Study. The aim of the health module was to develop valid, reliable and comparable instruments to describe a core set of health domains (Ustin et al., 2001). An extensive review of existing instruments was carried out using the International Classification of Functioning, Disability and Health (ICF) Framework (Ustin et al., 2001). The WHO Long Form consists of 20 health domains, some overlapping with the eight SF-36 domains (Ustin et al., 2001). The WHO Long Form does not have a standard scoring system for scales, unlike the SF-36 instrument.

One of the aims of the 2002/03 NZHS was "to measure the health status of Hew Zealand adults, including their self-reported physical and mental health status, and the prevalence of selected health conditions" (Ministry of Health, 2004 p1). The 2002/03 New Zealand Health Survey self-reported health status (HRQL) module included a combination of the SF-36 and the WHO Long Form (NZ Version.

The health status component of the health survey was broadened to cover a greater proportion of the health and health-related domains from the ICF - fifteen health domains were covered in the 2002/03 NZHS, compared to the eight covered by the SF-36 alone in the previous health survey completed in 1996-1997. These domains were the SF-36 domains outlined previously, plus physical functioning, social functioning, vision, hearing, digestion and bodily excretions, breathing, self-care, understanding, communication and sleep.

Questionnaires can have different meanings in different cultures and countries, and also within countries between population subgroups. Also, one aim of the 2002/03 NZHS was to "examine differences between population subgroups (as defined by sex, ethnicity, age and the

New Zealand Deprivation Index 2001 (NZDep01)" (Ministry of Health, 2004 p1) Thus it is important to assess the psychometric performance of instruments for the NZ population and subgroups within it, before national norms can be used in practice (Scott et al., 1999, Gandek, 1998). Thus it is necessary to validate instruments for population subgroups as well as the total NZ population. The SF-36 was validated in the previous 1996/97 health survey, but the WHO Long Form questions had not been previously validated. The focus of this thesis is on the validation of the WHO Long Form questions and scales for the NZ population and important population subgroups.

The primary objectives of this project were:

i. to group the WHO Long Form items in the 2002/03 NZHS into scales for each health domain and develop standard scoring protocols for each scale

ii. to test the reliability of the scales using standard psychometric tests for the total NZ population and for major population subgroups

iii. to test the validity of the scales using the standard psychometric tests for the total NZ population and for major population subgroups

iv. to construct norms for the WHO Long Form scales for the NZ population

v. to provide recommendations for the health status component of future NZ health surveys

Thesis outline

The thesis, which describes the rationale, data collection, scoring development and psychometric testing of the WHO Long form in the New Zealand Health Survey 2002/03, of is organised as follows.

Chapter 2 discusses the background to the measurement of health-related quality of life (HRQL), both internationally and in the New Zealand context. The chapter contains an introduction to the measurement of health status and its decomposition into different health and health-related domains. I review the development and content of the major instruments for measuring HRQL, i.e. the SF-20, SF-36 and WHO Long Form. This is followed by a discussion of their adaptation and application in the New Zealand context.

Chapter 3 is divided into two main sections. First the methodology of the 2002/03 New Zealand Health Survey (2002/03 NZHS) is described followed by a discussion of the use of

SF-36 and the WHO Long Form in this survey. The second half of the chapter covers the specific methods used to apply, and to test, the WHO Long Form in the 2002/03 NZHS. The methods for developing scoring guidelines for the scales is outlined. This is followed by a description of the methods for testing the reliability and validity of the scales and for producing population norms.

Chapter 4 has four main sections. First, the process for creating the WHO Long Form scales in the 2002/03 NZHS is described. Second, the results of the reliability tests are discussed for items and scales. Third, the validity analysis for WHO Long Form scales is presented. Finally, population norms for the scales are presented. The analyses in this chapter were performed for the whole population and major subgroups of interest within the population, separating the population by ethnic group, age group and deprivation.

The thesis concludes with a summary in Chapter 5 of the major findings and a discussion of the implications of this work for future health surveys in New Zealand.

Chapter 2: Background - Measuring health-related quality of life

Measuring health status

In this chapter I review methods of measuring health-related quality of life, including the SF-36 and the WHO Long Form, and their adaptation to, and application in, the New Zealand context. In the following chapter I then describe the methods used to employ these instruments in the 2002/03 New Zealand Health Survey (2002/03 NZHS).

Concepts of health

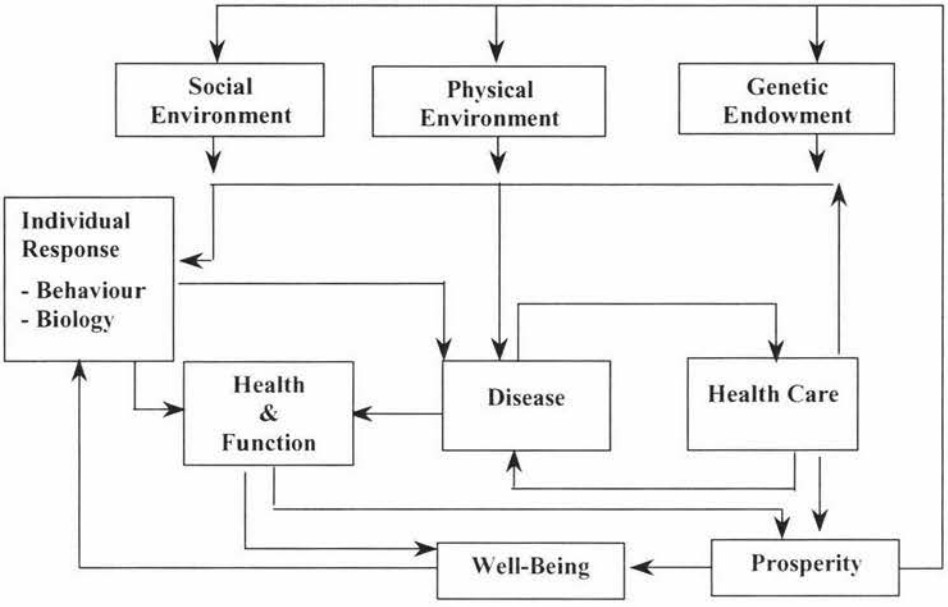
Health is generally referred to as a negative concept, relating to the absence of disease or illness. However, measuring departures from health is usually easier than measuring health itself. Within a general population in a Western society, the percentages of people with a chronic physical or psychiatric impairment are usually about 15 and 10-20 percent respectively. Thus negative health indicators provide scarce information about the health of the remaining 80-90 percent of the population. Measures of positive health are thus important when measuring the health of the general population (Bowling, 1997).

In 1958 the World Health Organisation defined health as "a state of complete physical, mental and social wellbeing and not merely the absence of disease or infirmity." (World Health Organisation, 1958). While this model of health generated some controversies, there is now broad agreement that the focus should be on positive concepts of health, which are not merely the absence of disease or disability. There is however no one accepted definition of health. Positive health is made up of different components which must be measured and interpreted separately (Bowling, 1997).

A model of health widely used within the New Zealand health sector was developed by Evans and Stoddard (Evans & Stoddart, 1994), as reproduced in Figure 1. This model is one of the two that form the theoretical basis for the plan of population health surveys in New Zealand (Ministry of Health, 2002). An individual's health and function are impacted upon a variety of factors, which can be categorised into a number of types. The social environment, physical environment and genetic makeup of an individual, through the behavioural and biological response at an individual level produce a certain level of health and functioning. Diseases impact upon the health and functioning of an individual. The health care system has impacts

upon disease, but can in turn be impacted upon by disease. The level of health and functioning of an individual impacts upon that individual's wellbeing. Prosperity in turn affects an individual's state of wellbeing (Evans & Stoddart, 1994).

Figure 1. Evan and Stoddard's model of health



Note: Adapted from (Evans & Stoddart, 1994 p53)

International Classification of Functioning, Disability and Health (ICF)

The overall aim of the International Classification of Functioning, Disability and Health (ICF) is to provide a common framework and language for the description of health and health-related states. The ICF describes domains of health, which can be divided into health and health-related domains. The ICF consists of two major components that can be used to describe health states, the level of the body and the level of the person. At the level of the person there are various domains of activity and participation. Health status instruments are developed to measure health and health-related domains at the level of the person. The ICF provides the framework to describe the essential elements of non-fatal health outcomes measured by health status instruments (World Health Organization, 2001)

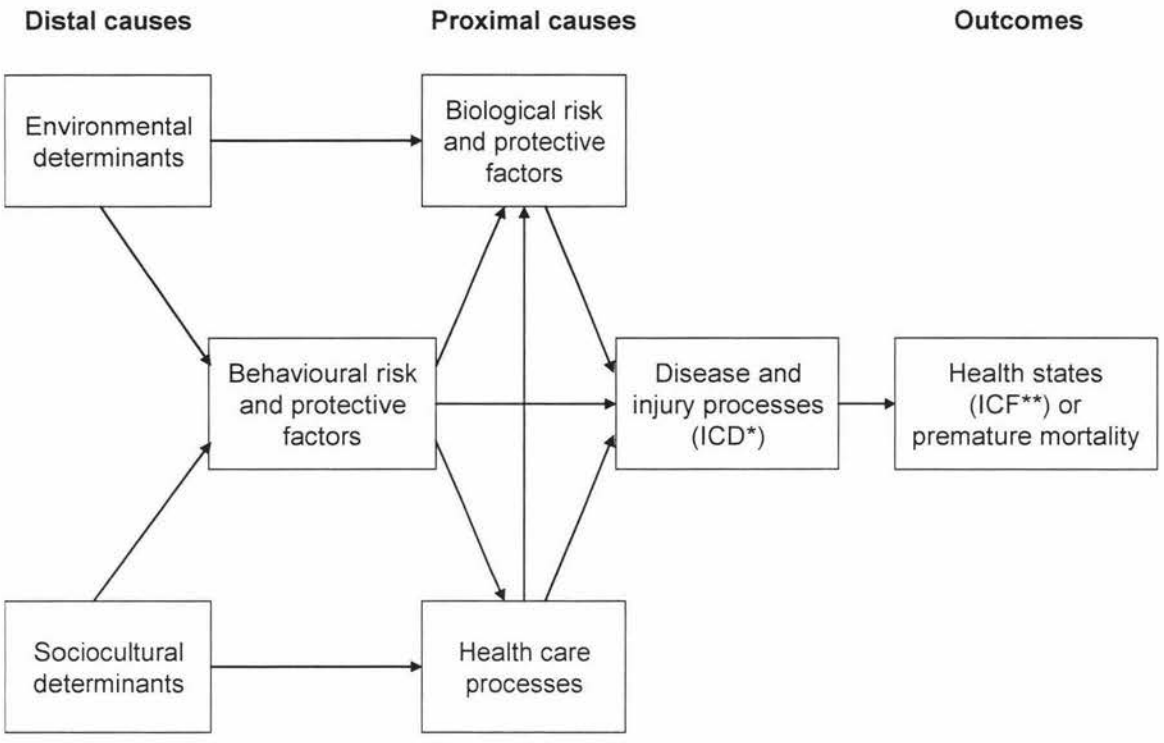
The ICF recognises 21 key dimensions of health, split into health and health-related domains. These are presented in Table 1.

Table 1. Key dimensions of health recognised by the ICF

Health domains	Health-related domains
Vision	Self-care
Sleep	Interpersonal relations
Communication	Usual activities
Skin and disfigurement	Social functioning
Pain	
Excretion	
Energy/vitality	
Speech	
Fertility	
Affect	
Digestion	
Dexterity	
Hearing	
Sexual functioning	
Cognition	
Breathing	
Mobility	

Further to the model of health described above (Evans & Stoddart, 1994), the relationship between the ICF and ICD is described in the updated New Zealand Health Monitor (NZHM) (Ministry of Health, 2005). This conceptual model of health, as reproduced in Figure 2, further informs the theoretical basis of the plan of population health surveys (NZHM).

Figure 2. Model of health and its causes



Notes:
* The World Health Organization's International Classification of Disease (ICD)
** The World Health Organization's International Classification of Functioning, Disability and Health (ICF)

Note: Adapted from (Ministry of Health, 2005), p. 9.

Measurement instruments

Health status or health-related quality of life can be measured objectively through measures such as mortality rates and hospitalisation records. It is also measured subjectively through surveying an individual's perceptions of his or her state of health. Subjective or self-reported health indicators complement objective indicators. Self-reports introduce a consumer perspective to population health monitoring. The emphasis of such measures is on quality of life and wellbeing (Ministry of Health, 1999a, 1999b).

Health status and health-related quality of life are not necessarily interchangeable terms. One useful definition of health-related quality of life is "the value assigned to duration of life as modified by the impairments, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy." (Patrick & Erickson, 1993), p 22. Not all self-reported measures of health emphasise the quality of life or wellbeing components of the

WHO definition of health above. For the purposes of this thesis, the terms 'health status' and 'health-related quality of life' will be used interchangeably[1].

In the 1980s subjective health assessment was limited to the single global self rated health question, in which survey participants were asked to rate their health on a scale ranging from "excellent" to "poor" (Ministry of Health, 1999b).

An efficient way to measure health status or health-related quality of life is through scoring standardised responses to standardised questions. Sets of survey questions that are carefully constructed can greatly assist research (Lohr, 1989). Surveys that are most useful for diverse groups of people are those which cover general health concepts, such as wellbeing (Ware & Sherbourne, 1992).

Interest in short-form health surveys became essential during the Health Insurance Experiment when some study participants declined to complete a lengthy health survey. By developing a short survey, able to be administered in five minutes over the phone, cooperation from participants was gained and the beginning of interest in such scales began. Following on from this development, several of the short-form scales have been used in various studies successfully. Studies demonstrated that well constructed short multi-item scales achieved better validity than single item questions. The same studies also found that longer scales and more comprehensive questionnaires had higher levels of validity then short multi-item scales. This demonstrates that there is a trade-off involved in choosing short- over long-form scales (Ware, Snow, Kosinski, & Gandek, 1993).

It is possible to shorten a survey by not including some health concepts. There are however minimum standards in terms of comprehensiveness requiring numerous health concepts to be represented ie, content validity in relation to conventional definitions of health (Ware & Sherbourne, 1992). Both physical and mental health concepts must be represented, with four operational definitions with each, behavioural functioning, perceived wellbeing, social and role disability and personal evaluations. Self-reports of behavioural functioning are used to measure limitations due to poor health, which are observable and tangible. Perceived wellbeing is more subjective and relates to how an individual feels. Personal evaluations are

[1] Although the complexities of the definition and measurements of health-related quality of life and related concepts are not considered by this thesis, they are acknowledged. For a good overview of such discussions, see, (Patrick & Erickson, 1993).

included due to the importance of capturing the values or preferences of the individual and summaries of health status (Ware et al., 1993).

SF-20

I will now review commonly used instruments, starting with the SF-20, which is a 20 item instrument covering six dimensions of health. Prior to the development of the SF-36 it was one of the most commonly used general health instruments.

SF-20 Development

The comprehensive short-form precursor to the Short-Form 36 (SF-36) was an 18-item form measuring physical functioning, role limitations, general mental health, and perceptions of current health. In 1986, two further items were added to form the SF-20. The SF-20 was used in the Health Insurance Study Experiment/Medical Outcomes Study (HIS/MOS), which included 11,336 participants sampled from 523 practices in Boston, Chicago and Los Angeles (Ware et al., 1993).

SF-20 Content

The SF-20 consists of one multi-item scale per health concept. There are 20 items in the SF-20 divided into six health domains or concepts. The six health concepts or domains covered are physical functioning, role functioning, bodily pain, current health, social functioning and mental health. The scales are comprised of a range of items, which sum to provide a range of levels on the scale (Ware et al., 1993). The items and levels for each scale are summarised in Table 2.

Table 2. Number of items and scale levels in the SF-20

Domain	No. of Items	No. of Levels
Physical functioning	6	7
Role functioning	2	3
Bodily pain	1	6
Current health	5	21
Social functioning	1	6
Mental health	5	26

Note: Adapted from (Ware et al., 1993) p. 3:6

Physical functioning

The Physical Functioning scale has 10 items. The response choices measure duration of physical limitation (Ware et al., 1993).

Role functioning

The SF-20 has two items in a scale called *Role Functioning*, which focus on health-related limitations in the kind and amount of work, producing a three-level scale (Ware et al., 1993).

Bodily pain

The SF-20 scale consists of one item about the intensity of bodily pain or discomfort, resulting in a 6-level scale (Ware et al., 1993).

Current health

The SF-20 has five questions in this scale, the commonly used single item rating of health, from excellent to poor, and four items from the Current Health scale of the Health Perceptions Questionnaire (HPQ) (Ware et al., 1993).

Social functioning

The social functioning scale extends measurement outside the individual to summarize the amount and quality of an individual's social interactions with others. The SF-20 included one item for this concept (Ware et al., 1993).

Mental health

The SF-20 mental health scale (MHI-5) has five items and is a bipolar scale. The MHI-5 was formed from the 5 items from the MHI-38 (38 items) that best predicted the MHI total score. There are one or more items from the four chief mental health dimensions, anxiety, depression, loss of behavioural or emotional control and psychological wellbeing. The sum of the five items had a correlation of 0.95 in the MOS and 0.93 in the HIS/MOS, with the total score from the MHI-38 in. MHI-38 was used as the "gold-standard" for constructing the MHI-5 (Ware et al., 1993).

SF-36

Development of SF-36

The SF-36 includes and expands upon the SF-20. It was developed in 1986 at the Rand Corporation in the USA for use in the HIS/MOS study. It's use across the world has grown since 1990. The aim of the authors was to develop a short, generic measure of subjective health status that could be used in a wide range of settings and was psychometrically sound (Bowling, 1997). Many of the items contained in the SF-36 have been selected from pre-existing instruments. The SF-36 consists of the items which had the best factor loadings from the 149 items from full-length scales which were based on results from the more than 22,00 patients in the Rand HIS/MOS studies (Bowling, 1997). The factor loading is the correlation between a variable or scale and the factor (underlying theoretical construct) (Ware et al., 1993). The SF-36 covers subset of 8 health concepts or domains, from the 40 concepts or domains that the full-length scales measured. There were four concepts seriously considered, but not included. These were health distress, sexual functioning, family functioning and sleep adequacy (Ware et al., 1993). The SF-36 takes five minutes to complete and is self-administered (Bowling, 1997).

SF-36 Content

The SF-36 consists of one multi-item scale per health concept. There are 36 items in the SF-36 divided into eight health domains or concepts. The eight health concepts or domains covered are physical functioning, role-physical, role-emotional, bodily pain, general health, vitality, social functioning and mental health. The scales are comprised of a range of items, which sum to provide a range of levels on the scale. A difference in five points between groups or over time on any scale is deemed to be clinically and socially relevant (Ware et al., 1993). Table 3 summarises the items and levels for each scale.

Table 3. Number of items and scale levels in the SF-36

Domain	No. of Items	No. of Levels
Physical functioning	10	21
Role-physical	4	5
Role-emotional	3	4
Bodily pain	2	11
General health	5	21
Vitality	4	21
Social functioning	2	9
Mental health	5	26

Note: Adapted from (Ware et al., 1993) p. 3:6

There is also a single item measuring the respondents perceptions of the amount of change in their health status over the previous 12 months (Ware et al., 1993). The abbreviated item content for the SF-36 is given in Table 4. A full copy of the SF-36 questionnaire used in the 2002/03 NZHS is included as Appendix II (as part of the relevant portion of the 2002/03 NZHS given).

Table 4. SF-36 Domains and abbreviated item content

Domain	Abbreviated item content
Physical Functioning	Vigorous activities, such as running, lifting heavy objects, strenuous sports Moderate activities, such as moving a table, vacuuming, bowling Lifting or carrying groceries Climbing several flights of stairs Climbing one flight of stairs Bending, kneeling or stooping Walking more than one mile Walking several blocks Walking one block Bathing or dressing
Role Physical	Were limited in the kind of work or other activities Cut down the amount of time you spent on work or other activities Accomplished less than you would like Had difficulty performing the work or other activities
Bodily pain	Intensity of bodily pain Extent pain interfered with normal work
General health	Is your health: excellent, very good, good, fair, poor My health is excellent I am as healthy as anybody I know I seem to get sick a little easier than other people I expect my health to get worse
Vitality	Feel full of pep Have a lot of energy Feel worn out Feel tired
Social functioning	Extent health problems interfered with normal social activities Frequency health problems interfered with social activities
Role Emotional	Cut down the amount of time you spent on work or other activities Accomplished less than you would like Didn't do work or other activities as carefully as usual
Mental health	Been a very nervous person Felt so down in the dumps nothing could chair you up Felt calm and peaceful Felt down Been a happy person
Health Transition	General Health compared to one year ago

Note: Adapted from (Ware et al., 1993) p. 3:6

Physical functioning

The whole HIS/MOS physical functioning scale (PF), a unipolar scale with 10 items was included in the SF-36 without any changes. This unipolar scale has 10 items. The SF-36 items cover both the presence and level of physical limitation by using three-level response categories. These three categories are "Yes, limited a lot", "Yes, limited a little" and "No, not limited at all". By using the three level of responses, the number of scale levels achieved by the 10-items was doubled relative to that achieved by using dichotomous responses. Thus, without increasing respondent burden the statistical precision of the analyses is increased (Ware et al., 1993).

Role functioning – Physical and Emotional

The SF-36 has seven role functioning items, spilt into two unipolar scales, r*ole-physical* and r*ole-emotional*. The role-physical scale has four items and the role-emotional scale has three items (Ware et al., 1993).

Bodily pain

The SF-36 bodily pain (BP) unipolar scale consists of the SF-20 item about the intensity of bodily pain or discomfort and a second item that measures the level of interference of the pain with normal activities. This results in an 11-level unipolar scale (Ware et al., 1993).

General health

The General Health Rating Index (GHRI) summary score is used to capture general health, instead of the Current Health scale. There is a sixth general health question included in the SF-36. This question asks respondents how much their general health has changed over the previous year. This item is analysed on its own, without being added to any of the eight multi-item scales. This question can be analysed as either an ordinal- or interval-level scale. This scale is a bipolar scale (Ware et al., 1993).

Vitality

The Vitality scale has four items and is bipolar. The items were adapted from the Mental Health Inventory (MHI) used in the HIS/MOS, which was sourced from the 1976 Health and Nutrition Examination Survey (HANES) (Ware et al., 1993).

Social functioning

The SF-36 has two items in the social functioning scale. This two-item unipolar scale assesses the relationship between health and social interactions. The SF-36 focuses specifically on the impact of either physical or emotional health on social activities (Ware et al., 1993).

Mental health

The 5-item mental health scale (MHI-5) from the SF-20 was retained in the SF-36, with changes in the format (Ware et al., 1993).

Comparison of SF-20 and SF-36 scales

Physical functioning

The SF-36 scale has two major improvements compared to the SF-20.

Firstly, items were added to SF-36 to correspond to the types and levels of limitations along the full range of physical limitations. Some examples are carrying the groceries, climbing stairs, kneeling and walking moderate distances. Only one self-care item was included to capture limitations in self-care activities. The rationale for this was that although such limitations are important, they are rare in both patient and population groups. For example, only 7.4% of the 11,336 patients in the MOS had self-care limitations (Ware et al., 1993). Thus in a general health survey it is inefficient to administer a number of items on self-care limitation.

Secondly, the standard response options were adapted to estimate the severity of each physical limitation. In the SF-20 the response choices measured duration of physical limitation. However, due to the fact that most physical limitations are chronic, measures of duration are not useful for data analysis. Comparisons between those with and without difficulties in performing such activities and increase the precision of the score on the scale (Ware & Sherbourne, 1992; Ware et al., 1993)

Role functioning – Physical and Emotional

The SF-36 r*ole-physical* and r*ole-emotional* scales differ from the SF-20 and other health surveys in two major ways. First, they cover a wider range of role limitations in three major categories and are thus applicable to a wider range of people. Second, the two scales differentiate between role limitations, which are caused by physical or mental issues. By doing so, the validity and precision of this scale of the SF-36 is improved from the SF-20. (Ware et al., 1993)

Bodily pain

The SF-36 scale has 11-levels compared to the 6-levels of the SF-20 scale. This leads to gains in content validity, scale reliability and precision (Ware et al., 1993)

General health

Improvements were made in the SF-36 by choosing the General Health Rating Index (GHRI) summary score to capture general health, instead of the Current Health scale. This has several advantages. First, it achieves a more complete coverage of the content of the HPQ, including current health, resistance to illness and health outlook. Second, it correlates highly with the GHRI. Third, respondents find it more acceptable, with a balance between favourably and unfavourably worded items, which controls for effects of response set (Ware et al., 1993)

There is a sixth general health question included in the SF-36. This item has been found to provide useful information regarding the actual change in the respondents health in the previous year (Ware et al., 1993)

Vitality

The Vitality scale was not included in the SF-20, but was added to the SF-36 to better portray difference in subjective-wellbeing. These items achieve a balance between favourably and unfavourably worded items to control for response set effects. The item-discriminant validity and reliability of this scale have been demonstrated (Ware et al., 1993)

Social functioning

The SF-20 included only one item for this concept. The SF-36 improved this item and added a second question. The SF-36 scales covers a wider range of levels of social functioning with a higher degree of precision (Ware et al., 1993)

Mental health

The scales in the SF-20 and SF-36 mental health domain are identical. The validity of this scale has been demonstrated with numerous studies involving the general population and mental health consumers (Ware et al., 1993).

Scoring

Items are summed to and normalised to a score of 0-100 for the scales, where higher scores represent better health (Ware et al., 1993).

Standardisation

By standardising the content and scoring of the SF-36, interpretation of the SF-36 scales is made possible. Both the content and the scoring algorithms of the SF-36 were chosen after thoughtful consideration of a range of alternative options by Ware et al. (1993). In selecting the scoring algorithms, a balance was sort between keeping things simple and satisfying the assumptions of the scale construction methods. There are two main reasons for adhering to the SF-36 content and scoring standards. Firstly, they are most likely so achieve the same reliability and validity as reported in the MOS publications. Secondly, this allows comparisons amongst studies using the SF-36, benefiting all those adhering to the standardisation (Ware et al., 1993).

Item recoding

Responses to the SF-36 are recoded to derive the item values that will be used to calculate scale scores. The three parts included in this stage of scoring are to change out-of-range values to missing values, recode values on 10 items and to replace missing estimates with person-specific estimates. SF-36 items for which a higher score represents poorer health needed to be reverse scored. Seven items are reverse scored, to ensure that a higher value indicates better health on all SF-36 items and scales (Ware et al., 1993). For respondents who have answered at least half the questions of a scale, person specific estimates are imputed for missing values, based on the average of the scores across other items on the scale.

Item recalibration

There are two items in the SF-36, which are recalibrated to satisfy the assumption of a linear relationship between the item scores and the underlying health dimension of the scale. There is one general health (GH) item and one Bodily pain (BP) item. For the GH question (Item 1) "In general would you say your health is...Excellent, Very good, Good, Fair, Poor", the "Very Good" and "Good" responses are recalibrated to achieve a better linear fit with the general health concept measured by the GH scale. Studies have shown that the distance between these "Excellent" and "Very Good" is half the size of that between "Fair" and "Good" responses. In the MOS studies, the mean current health scores for respondents who chose the same option for GH1 were very similar in the screening and longitudinal samples. The intervals between the response categories were unequal. After the transformation, this item had a correlation of 0.7 with the sum of the other four items in the GH scale (Ware et al., 1993).

The BP item that is recalibrated is the question, "How much bodily pain have you had during the past four weeks?" and the response categories are "None", "Very mild", "Mild", "Moderate", "Severe" and "Very severe". The two BP items have a different number of responses, which means that they do not satisfy the assumption of equal various for summated rating scales. The recalibration of the above item gives it an equal variance to the other BP item by converting it to a six-level item. The recalibration is also necessary to satisfy the linear association between the BP items and the underlying concept of bodily pain in terms of criterion validity (Ware et al., 1993).

Computing scale scores

After item recoding is completed, a raw score is computed for each scale. This score is the algebraic sum of all items in the scale. This simple method of summation is possible due to the fact that each of the items in a scale have an approximately equal relationship with the underlying health dimension which is measured. This means that weighting of items is not necessary. After the raw score is calculated, the scores for the scales are transformed to a 0-100 scale using the following formula:

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score - lowest } \textit{possible} \text{ raw score})}{\textit{Possible} \text{ raw score range}} \right] \times 100$$

Scores between 0 and 100 represent the percentage of the total score possible which is achieved. These transformed scores can be compared with published norms. Raw and transformed scale scores are not calculated for the Reported Health Transition item (Ware et al., 1993)

Summary measures

There are two summary indexes covering two major dimensions of health, physical and mental health, which are derived from the scale scores for the SF-36. These indexes are based on the physical and mental health components or factors, which are shown to explain 82% of the variance in the scale scores. These two summary indexes have several advantages including the following. First, they reduce the number of statistical comparisons required to capture differences in health outcome. Second, they adjust for extensive correlations amongst the eight scales. Third, the interpretation of differences in physical and mental health scores is made easier. Fourth, there is greater precision for general physical and mental health outcomes (Ware et al., 1993).

Psychometric properties

Reliability

Estimates of score reliability for the scales within the SF-36 have been published for population samples from 11 countries within the International Quality of Life Assessment (IQOLA) Project. Countries included in this project were Denmark, France, Germany, Italy, Japan, the Netherlands, Norway, Spain, the United Kingdom, the United States and Sweden. For all scales except Social Functioning (SF) reliability estimates were greater then 0.70, the desired minimum for group comparisons (Gandek et al., 1998).

Validity

The construct validity of the SF-36 was assessed in the IQOLA Project by examining the factor structure using principal components in population samples within 10 countries. Ware et al (Ware et al., 1998) used five criteria to assess the support for the two underlying dimensions of physical and mental health of the SF-36. First, the eigenvalues for the first two components should be greater than one. Second, more than 60% of the total variance in scale scores should be explained by the two components. Third, the physical functioning (PF) scale should correlate highest with the physical component, followed by the role-physical (RP); and bodily pain (BP) scales, all of which should correlate lowest with the mental component. Fourth, the mental health (MH) scale should correlate highest with the mental component, followed by the role-emotional (RE) and social functioning (SF) scales, with all three scales correlating the least with the physical component. Finally, the general health (GH) and vitality (VT) scales should correlate moderately with both the physical and mental components, with the GH scale correlating higher with the physical component and the VT higher with the mental component (Ware et al., 1998).

The IQOLA analyses supported the existence of two dimensions, physical and mental health, for all countries and across age and gender subgroups within the countries (Ware et al., 1998). However although the factor structure for the SF-36 was found to be supported in the Western European and United States populations, the results for the same analysis in Japan were more uncertain. The two distinct factors of physical and mental health rely on the assumption of mind-body dualism which may not fit with the concepts of health in all populations or ethnic groups (Scott et al., 2000).

The Australian adaptation of the SF-36 has been validated. The principal components analysis supported the two underlying dimensions, physical and mental health for the SF-36. The clinical construct validity was also demonstrated, with decreasing SF-36 scale score with increasing level of medical and/or depressive condition (Sanson-Fisher & Perkins, 1998).

Data quality

The quality of SF-36 data can be assessed through the results of tests of score reliability, tests of validity, completeness of data, response consistency and scaling assumptions. The first two categories have already been discussed, but the latter three will now be considered. The percentage of missing data tends to vary by population subgroup, with the highest being observed in low socioeconomic status and older age groups (Ware et al., 1993). In the IQOLA project the percent of missing item-level data was low for most countries, but tended to be higher in the Scandinavian countries. Missing data percentages were usually lowest in countries which had used interview or telephone administration of the questionnaire (Gandek et al., 1998).

Response consistency can be assessed by analysing individual responses. There are fifteen pairs of SF-36 items that can be used to check internal consistency, producing the SF-36 Response Consistency Index (RCI). For the general population in the United States, 90.3% of people had no inconsistent responses (Ware et al., 1993).

Scaling assumptions

The IQOLA Project also evaluated the scaling assumptions of the SF-36. This was measured using the scaling success rate, the percentage of the item-scale correlations significantly greater than the item-competing scale correlations. Items discriminated across all scales for most items in all of the participating countries, with scaling success rates ranging from 81.3 to 100%. This provides strong support for the method of summated rating scores for the SF-36 scales (Gandek et al., 1998).

Scales can be described in terms of their floor and ceiling effects. The floor is the lowest score on a scale and the ceiling the highest (Streiner & Norman, 2003). The proportion of respondents that score either at the floor or the ceiling is calculated to test for skewness in scale scores. If either value is high, the potential of the SF-36 to detect change in the population over time is limited. The percent of respondents scoring at the floor was less than 1.5% for six scales, but between 6.1% and 13.2% for the role-physical and role-emotional

scales. The Role Functioning scales and the Social Functioning scale had high ceiling effects (Gandek et al., 1998).

WHO Long Form

Development of the WHO Long Form

The WHO Multi-country Survey Study was a research project set up to develop instruments that would enable the measurement of health, responsiveness and other health-related dimensions in a way which allowed valid comparisons across countries. The WHO Long Form was developed as the health module in the WHO Multi-country Survey Study. The aim of the health module was to develop valid, reliable and comparable instruments to describe a core set of health domains (Ustin et al., 2001).

An extensive review of existing instruments for measuring health status was carried out. This review was synchronised with the revision of the ICF. A pool of items was developed, with psychometric properties of the items documented. Qualitative research was conducted to develop a list of core constructs in the different countries. The pool of items and outcomes of the qualitative research were presented at the UN/OECD meeting in Ottawa in 2000. From the item pool the health domain items were selected for the WHO Long Form using the following six criteria. First, they should be linked the ICF. Second, they should have face and construct validity. Third, they should be able to be measured by self-report. Fourth, they should be drawn from existing common questionnaires. Fifth, they should be comparable across populations. Finally, it should be possible for some of the domains to be linked to a calibration test (Ustin et al., 2001).

WHO Long Form Content

There were 20 health and health-related domains included in the WHO Long Form (Table 5).

Table 5. WHO Long Form Health and Health-Related Domains

Health Domains	Health-Related Domains
Vision	Self-care
Hearing	Usual activities
Speech	Social functioning
Digestion	Participation
Bodily excretion	
Fertility	
Sexual activity	
Skin & disfigurement	
Breathing	
Pain	
Affect	
Sleep	
Energy/vitality	
Cognition	
Communication	
Mobility and Dexterity	

Based on the list of domains chosen, the questions were selected from existing survey instruments (Ustin et al., 2001).

Application of SF-36 to New Zealand

Australian/New Zealand adaptation of SF-36

The validity of the SF-36 for the Australian population was examined in 1992, as part of the International Quality of Life Assessment (IQOLA) Project. Overall the SF-36 was found to be a valid measure of general health status in the Australian population, with desirable psychometric properties (Sanson-Fisher & Perkins, 1998), which had been demonstrated in an earlier study by McCallum (1995).

Adaptation

All questions from the developmental version of the SF-36 were pretested twice. In the vitality and mental health scales, two "Americanisms", "did you feel full of pep?" and "have you felt downhearted and blue?" were changed following the first pre-test to "did you feel full of enthusiasm?" and "have you felt sad?". They were subsequently further modified and now

are "did you feel full of life?" and "have you felt down?" in the standard Australian version. Changes were made to questions which used to the imperial system, to refer to the metric system instead. Three items in the physical functioning scale were adapted in the following ways: "walking more than one mile" was changed to "walking more than one kilometre"; "walking several blocks" was changed to "walking half a kilometre"; and "walking one block" was changed to "walking 100 metres" (Sanson-Fisher & Perkins, 1998)

Internal consistency

All scales had high internal consistency measures with Cronbach alphas of greater than or equal to 0.8. All items had higher correlations with their own scale totals than with any other scales (Sanson-Fisher & Perkins, 1998).

Construct validity

The eight different health domains and two general health dimensions were associated with the health status of respondents as the United States results predicted. Correlations between the scale scores and the physical and mental health components were consistent with previous research. Thus the eight health domains behaved in a similar way to that reported in United States and United Kingdom validation studies (Sanson-Fisher & Perkins, 1998).

Items and domains

The abbreviated item content of the scales of the Australia/New Zealand adaptation of the SF-36 used in the 1996/97 NZHS and 2002/03 NZHS is presented in Table 6.

Table 6. SF-36 scale and abbreviated item content in Health status module of NZHS 1996/97 and 2002/03

Scale	Abbreviated item content
Physical functioning	Vigorous activities, such as running, lifting heavy objects, strenuous sports
	Moderate activities, such as moving a table, vacuuming, bowling
	Lifting or carrying groceries
	Climbing several flights of stairs
	Climbing one flight of stairs
	Bending, kneeling or stooping
	Walking more than a kilometre
	Walking half a kilometre
	Walking 100 metres
	Bathing or dressing yourself
Role Physical	Were limited in the kind of work or other activities
	Cut down the amount of time you spent on work or other activities
	Accomplished less than you would like
	Had difficulty performing the work or other activities
Bodily pain	Intensity of bodily pain
	Extent pain interfered with normal work
General health	Is your health: excellent, very good, good, fair, poor
	My health is excellent
	I am as healthy as anybody I know
	I seem to get sick a little easier than other people
	I expect my health to get worse
Vitality	Feel full of life
	Have a lot of energy
	Feel worn out
	Feel tired
Social functioning	Extent health problems interfered with normal social activities
	Frequency health problems interfered with social activities
Role Emotional	Cut down the amount of time you spent on work or other activities
	Accomplished less than you would like
	Didn't do work or other activities as carefully as usual
Mental health	Been a very nervous person
	Felt so down in the dumps nothing could chair you up
	Felt calm and peaceful
	Felt down
	Been a happy person
Health Transition	General health compared to one year ago

Note: Adapted from (Ware et al., 1993) p. 3:6

The SF-36 was self-administered by survey participants in the 1996/97 NZHS and interviewer administered in the 2002/03 NZHS.

Psychometric properties

Data quality

Scott et al performed the validation for the SF-36 in the 1996/97 population health survey in New Zealand (Scott et al., 1999), following the same process conducted in Sweden (Sullivan, Karlsson, & Ware, 1995). Data quality was assessed by calculating the percentage of missing data per item, and how this varied across population subgroups. The percentage of missing data ranged from 1.5-3.2%. When analysed by population subgroup, the proportion of missing data was found to be higher for respondents with lower socioeconomic status and in the older age groups. The proportion of missing data also varied by ethnicity, but not by education level (Scott et al., 1999).

Scaling assumptions

There was good item internal consistency, with correlations for the scales ranging from 0.62-0.9, all more than 0.4 (Scott et al., 1999). One hundred percent scaling success rates were observed for all eight scales. Negative skewness was marked for scale scores, with respondents more likely to endorse higher response choices, indicating better health. The floor effects were low for all but two scales - role-physical and role-emotional. The ceiling effects were highest for the both Role- scales and Social Functioning (Scott et al., 1999).

Reliability

There was good internal consistency reliability as measured by Cronbach's alpha. For the eight SF-36 scales these coefficients ranged between 0.78 and 0.93, all being higher than the recommended level of 0.7 for between group comparisons (Scott et al., 1999).

Validity

Construct validity was assessed using item discriminant validity and factorial validity. The former was discussed with respect to scaling assumptions section. Factorial validity was tested using confirmatory factor analysis to test for the existence of the two underlying dimensions, mental and physical health. The two factor orthogonal solution explained 67% of the variance in the data, consistent with results from previous overseas studies. Scales correlated with these two components in the expected pattern, with the physical functioning scale correlating most strongly with the physical health component and the mental health scale correlating most strongly with the mental health component. However, the factor

structure differed by ethnic group, which will be discussed later in this chapter (Scott et al., 1999).

Norms

Males scored significantly higher than females on all of the SF-36 scales, except the general health scale. Scores generally decreased with increasing age, with the most marked decline on the physical health-related scales. Comparisons by ethnic group were performed for the three major ethnic groups, Māori, Pacific and New Zealand (NZ) European. The New Zealand European group had significantly higher scores than Māori on all scales except bodily pain and vitality, and than Pacific for all scales except the latter plus the mental health scale (Scott et al., 1999).

Issues for use with Māori and Pacific peoples

To understand whether a health status instrument is interpreted in the same way by different subgroups of the population, the structural model of the questionnaire should be assessed for these different groups. This involves assessing the pattern of correlations between the scales within the instrument using principal components analysis to derive components, which represent the underlying constructs the instrument purports to measure (Scott et al., 2000).

Problems with principal components

Ware et al's (1998) five criteria for assessing the factor structure of the SF-36, outlined earlier, were assessed for the three major ethnic groups in New Zealand, NZ European, Māori and Pacific peoples. The first criterion was met for NZ European and Māori, but not Pacific with the eigenvalue of 0.91 for the second component being less than one. For the NZ European population, the factor structure was found to be very similar to that found in the IQOLA analyses, meeting almost all of the criteria. For the Māori population, the factor loadings were similar to that of NZ Europeans and the structure of the physical component was supported (third criterion). However, the fourth criterion was not met, as the social functioning and role-emotional scales correlated less highly with the mental component than did the vitality scale, but the social functioning and role-emotional scales did correlate higher with the mental than the physical component (Scott et al., 2000).

For Pacific peoples, criteria three to five were not met. The scales which correlated highest with the physical and mental components respectively were role-physical and vitality scales.

Many of the scales correlated higher with the opposite component to what was observed in the NZ European population and in the IQOLA study countries. For Pacific peoples evidence of clusters of correlations of the scales are evident, especially between the role-emotional, role-physical and social functioning scales. Due to the factor loadings and features of the clusters of correlations between scales, the interpretation of the two components as representing physical and mental dimensions is highly disputed (Scott et al., 2000).

The same principal components analysis was carried out to evaluate the stability of the observed factor structure across genders, age groups and socioeconomic groups, by gender, age group and socioeconomic status. There was no major difference across age groups in the NZ European and Pacific groups (same as people(s). For Māori, large differences were observed when comparing younger (<45 years) and older (>44 years) age groups. The factor structure for younger Māori was similar to that of NZ Europeans, but for older Māori the extraction of two factors was not supported. Instead a one-factor structure of the questionnaire is suggested for older Māori. There were no differences in the factor structure of the socioeconomic groups with higher and lower household incomes (age and sex standardised). Thus it appears that the differences in factor structure across ethnic groups are not due to differences in factor structure by income group (Scott et al., 2000).

Māori/Pacific models of health

Overall the construct validity of the SF-36 is supported for the NZ European ethnic group, but is more questionable for the Māori and Pacific ethnic groups. There are a variety of explanations for this finding. The results for Māori were related to age, with the two factor structure being supported for younger Māori (<45 years), but a one-factor structure being found for older Māori. This could possibly be due to the fact that many younger Māori are urbanised with weaker cultural ties, whereas older Māori may more closely identify with Māori constructs of health, which do not dichotomise health into physical and mental dimensions. Interpretation of the two-factors that appeared for Pacific peoples was extremely challenging, due to the correlations of scales being different to that which is hypothesised for the proposed two-dimensional model (Scott et al., 2000).

There are also issues of differences between respondents' own understandings of the questions and the intended meanings of the questions. For older Māori it appears that their traditional views of health have guided their responses, which do not separate physical and mental dimensions of health. For Pacific peoples the patterns of answering questions are

harder to interpret. Thus the application of the SF-36 to the NZ population has questioned the cross-cultural validity of the instrument. Scott et al (Scott et al., 2000) draw two conclusions. First, that quantitative evidence is provided for the lack of simple mind-body dualism in traditional Māori and Pacific models of health. Second, the model of health which will determine a respondents pattern of answers to questions is their own one.

Application of WHO Long Form to New Zealand

Items and domains

Including questions and scales from the WHO Long Form in the self-reported health status section increases the number of domains of health covered from eight to fifteen (Table 7). The number of new domains included as a result of this addition is eight. A greater coverage of the health and health-related domains identified by the ICF (World Health Organization, 2001) is achieved with the extension of this section.

WHO Long Form Validation

The SF-36 has been tested and validated in New Zealand in the 1996/97 NZHS (Scott et al., 1999). The WHO Long form questions had not previously were used for the first time in New Zealand NZ in the 2002/03 NZHS. There was also a need to develop an approach to scoring the questions and forming scales. The validation of these questions was the focus of this thesis. The questions required validation for the total NZ population and subgroups within the population.

Table 7. WHO Long Form scales and abbreviated item content in Health status module of 2002/03 NZHS

Scale	Abbreviated item content
Self care	Bathing yourself Dressing yourself Grooming yourself Eating Using the toilet Staying by yourself for a few days
Vision	Distance of about 20 metres Distance or about 5 metres Reading a book or newspaper
Hearing	Conversation with one other person in a quiet room Someone talking on other side of room Group conversation with at least three other people
Digestion & bodily excretions	Indigestion Constipation Passing urine Controlling urine
Breathing	Short of breath with mild exercise Short of breath at rest Coughing or wheezing for 10 minutes or more
Sleep	Problem falling asleep Waking up frequently at night Waking up too early
Understanding	Concentrating on doing something for 10 minutes Remembering to do important things Analysing and solving problems in day-to-day life Learning a new task
Communication	Understanding what people say Starting and maintaining a conversation Speaking clearly
Social Functioning	Dealing with people you do not know Maintaining a friendship Getting along with people who are close to you

Chapter 3: Methods

In this chapter I will describe the methodology of the use of SF-36 and the WHO Long Form in the 2002/03 New Zealand Health Survey (2002/03 NZHS). I will start by describing the general methodology of the 2002/03 NZHS, and then in the second half of the chapter I will describe the specific methods used to apply, and to test, the WHO Long Form.

The first section is provided as background to the analysis presented in this thesis, and the author was not involved with this part. The author was given access to the final dataset of the 2002/03 NZHS, which was developed by collaborators at the Ministry of Health. The second half of this chapter describing the specific methods used to test the WHO Long Form were the authors own work.

New Zealand Health Monitor

The New Zealand Health Monitor (NZHM) (Ministry of Health, 2002) outlines a 10 year strategic plan for the design, funding and management of the national population-based survey programme carried out by the Ministry of Health. The Ministry of Health has a clear mandate under Section 3(c) of the Health Act 1956 to carry out health-related surveys. The purpose of such surveys is to provide information to support decision making within the health sector. The following surveys are included within the NZHM; general health survey, health behaviour survey, children's nutrition survey, adults' nutrition survey, tobacco use survey, alcohol and drug survey, sexual and reproductive health survey, mental health and wellbeing survey, NZ Census Mortality study and the NZ Birth Linkage study.(Ministry of Health, 2002, 2005).

The purpose of the NZHM is to provide essential information for the Ministry of Health and District Health Boards to develop and evaluate policies, plan and allocate resources to services or programmes and to make strategic decisions in the health sector. The health survey programme provides information for evidence-based policy making and decision making (Ministry of Health, 2002).

The NZHM is an organised, co-ordinated and integrated survey programme, which operates over 10-year cycles. It offers several benefits which include enhancing public understanding and acceptance; providing a continuous stream of relevant time series data; meeting the diverse information needs of all users; providing opportunities for systematic record linkage;

and proving efficiency gains due to reducing duplications in contents across surveys (Ministry of Health, 2002).

The two key questions the NZHM is intended to answer are "How healthy are we?" and "How healthy is the health system?" The first of these two questions applies to the current thesis particularly (Ministry of Health, 2002).

The data collected by the NZHM should meet the following criteria: policy relevance, logical data structure, survey appropriateness, local specificity and modifiability and impact. Variables should be related to an underlying conceptual model of health and health systems. These models should be culturally appropriate for use in New Zealand (Ministry of Health, 2002).

There are three information domains included within the NZHM: health outcomes, health causes and health services. Standard classifications for health outcomes are developed by the WHO, to which New Zealand is a signatory. These standard classifications include the International Classification of Functioning, Disability and Health (ICF) and the International Classification of Diseases, now in its tenth revision (ICD10).

The overall aim of the ICF is to provide a standard language and framework for the description of health and health-related states. The domains included in the ICF can be broken down into *health domains* and *health-related domains*. Functioning is the term used to describe all body functions and disability is used to refer to impairments (World Health Organization, 2001).

Using these classifications, four sub-domains of health outcomes can be distinguished: subjective health and wellbeing, functional limitation, chronic conditions and injuries (World Health Organization, 2001). This thesis is concerned with the measurement of subjective health and wellbeing the first of these four sub-domains.

There are a variety of scales available to measure the construct of subjective health. The minimum is the global single self-rated health assessment. The ICF identifies 21 key dimensions of health, split into health domains and health-related domains. The health domains are vision, hearing, speech, communication, cognition, affect, pain, mobility, dexterity, energy/vitality, sleep, sexual functioning, fertility, skin and disfigurement, breathing, digestion and excretion. The health-related domains are self-care, interpersonal relations, usual activities and social functioning (World Health Organization, 2001). The NZHM aims to collect data on all of these key dimensions (Ministry of Health, 2002).

New Zealand Health Survey

The General Health Survey, known as the New Zealand Health Survey (NZHS) is the core of the NZHM. The NZHS collects data from all age groups on health outcomes, health causes and health risks. Unlike most other developed countries, New Zealand lacked national population health surveys until fairly recently. The first nutrition survey was the National Dietary Survey, carried out by the Heart Foundation in 1978. Similar nutrition surveys were carried out in 1987 by the Hillary Commission and 1997 by the Ministry of Health. The first general health survey was completed in 1993/94, followed by a second survey in 1996/97 and a third in 2002/03 (Ministry of Health, 2002).

The 2002/03 NZHS was similar to the two previous national health surveys, particularly the 1996/97 survey. The new components of the 2002/03 survey were the inclusion questions on of a broader range of chronic conditions, risk and protective factors, additional measures of self-reported physical and mental health status and a more detailed range of health services. Another new dimension was the collection of height, weight and circumference measurements. The 1992/93 survey was conducted via telephone interviews, whereas the 1996/97 and 2002/03 surveys were carried out face-to-face. However, in the 1996/97 survey the General Health Questionnaire (including the SF-36) was completed by respondents themselves at the end of the interview (Ministry of Health, 2004)

Methodology of the 2002/03 NZHS

Aims

The 2002/03 NZHS had the following five aims:

i. to measure the health status of New Zealand adults, including self- reported physical and mental health status and the prevalence of specific health conditions;

ii. to measure the prevalence of risk and protective factors, which are associated with the specific health conditions; to measure the use of health services, including satisfaction with health services and barriers to the use of such services;

iii. to examine the differences between subgroups of the population, such as gender, ethnicity and socio-economic status, in the survey questions;

iv. and to examine changes over time.

(Ministry of Health, 2004).

Sample design

The target population was the New Zealand adult population aged 15 years and over living in permanent private dwellings, approximately 2.6 million people, according to the 2001 Census of Populations and Dwelling (2001 Census). All adults aged 15 years and older who were usually resident within permanent private dwellings were eligible for selection as survey respondents. The sample frame was area-based. The frame was a list of small geographic areas, called *meshblocks*, that fell within the geographic coverage of the survey. The meshblock is defined by Statistics New Zealand as a defined geographic area, which varies in size from a block within a city to a large area of rural land. It is the smallest geographic unit for which data is collected (Statistics New Zealand, 2005). The meshblocks were used as the primary sampling units (PSUs). All New Zealand households were clustered to avoid having to develop a list of all households in New Zealand. Lists of households were only maintained for the PSUs that were selected (Ministry of Health, 2004).

The 2002/03 NZHS used a complex sample design, to achieve quality estimates, minimal cost and satisfactory respondent burden. A stratified design was used with strata defined according to the ethnicity variable defined by Question 11 on the Individual Form of the 2001 Census. A constant sampling fraction was taken from each PSU, to give equal probability of selection for all dwellings within a particular stratum. Due to a change in the survey objectives early in the fieldwork stage, the survey had two sample designs (Ministry of Health, 2004).

The initial sample design consisted of four strata: Māori, Asian, Pacific peoples and Other. The Māori stratum comprised all PSUs containing 60 percent or more eligible persons who identified themselves as Māori, according to the 2001 Census. For each PSU, 50 percent of dwellings were chosen. There were 696 out of 861 PSUs selected, but only 32 were surveyed before the sample design changed. The Asian stratum comprised all PSUs containing 40 percent or more eligible people who identified as Asian. For each PSU chosen, two out of nine dwellings were selected. There were 189 out of 340 PSUs chosen, but only two were surveyed before the change in sample design. The Pacific peoples stratum comprised all PSUs with 55 percent or more eligible people who identified as being Pacific peoples. For each selected PSU, a constant sample size of 12 dwellings were chosen. There were 125 out of 439 PSUs selected, with only three being surveyed before the sample design changed. The Other stratum was made up of all remaining PSUs. A constant sample of 12 dwellings were

chosen from each PSU. There were 490 out of the 36, 712 PSUs selected, but only 77 were surveyed before the change in sample design (Ministry of Health, 2004).

The latter sample design contained only two strata: Māori and Other. The Māori stratum comprised all PSUs containing 70 percent or more eligible persons who identified themselves as Māori, according to the 2001 Census. For each PSU, two out of three dwellings were selected. This sampling fraction was reduced to one in seven dwellings in the later stages of the survey. The Other stratum PSUs were selected with probability proportional to size. Twelve dwellings were chosen from each selected PSU (Ministry of Health, 2004).

In addition to the two strata in the latter sample design, there were three ethnic over-samples within the Other stratum for Māori, Pacific peoples and Asian. The purpose of the ethnic over-samples was to increase the number of respondents in the sample identifying themselves as Māori, Pacific peoples or Asian.

For the Māori over-sample, only respondents identifying themselves as Māori, Pacific or Asian were eligible for selection. This over-sample consisted of all PSUs in the Other stratum which contained 10 percent or more eligible respondents who identified themselves as Māori, according to the 2001 Census.

For the Asian over-sample only respondents identifying themselves as Pacific or Asian were eligible for interview. The Asian over-sample consisted of all PSUs in the Other stratum not already oversampled and containing at least 30 percent of eligible respondents who identified themselves as Asian according to the 2001 Census.

For the Pacific peoples oversample only respondents identifying themselves as Pacific were eligible for interview. This oversample consisted of all PSUs in the Other stratum not already oversampled and containing 25 percent or more eligible respondents identifying themselves as Pacific according to the 2001 Census (Ministry of Health, 2004).

Sample selection

The first stage of sampling took place at the meshblock level. A systematic sample was taken from each stratum starting from a random point. For some strata, meshblocks were selected with a probability proportional to their size ie. number of eligible respondents, whereas for other strata, meshblocks were selected with equal probability. The secondary sampling unit is the dwelling. Each PSU is described in relation to the streets, side and section of the street within the PSU. The sampling fraction for each PSU was expressed as an integer, which

specified the step between successive dwellings that were chosen for inclusion in the sample (Ministry of Health, 2004).

The final level of sample selection was the respondent. Within each selected dwelling, the number of adults ages 15 years and over were identified. One person was selected for interview. The Kish grid was used to choose the single eligible person from the individuals in the dwelling. The names of the individuals in a household were listed in order of descending age on the sample grid. The selected respondent was the individual's whose name fell alongside a prechosen indicator (Ministry of Health, 2004).

Data collection

The mode of collection was face-to-face interviewing by trained interviewers. This method was chosen for the following reasons. First, the 2001 Census showed that 3.7 percent of households in private occupied dwellings had no access to a telephone. Secondly, it has become accepted knowledge in the telecommunications industry that a proportion of homes no longer use landline phones, although this is not well documented. The third and final reason was that from an analysis that the National Research Bureau (NRB) commissioned on the Statistics New Zealand Household Income Survey, non-ownership of phones is markedly skewed to those Māori and Pacific peoples in the lower income groups. Thus a phone sample frame would prejudice those who most need to be represented in the NZHS sample. Another advantage of face-to-face interviewing is that is allows for comparability with the previous health survey, the 1996/97 NZHS (Ministry of Health, 2004)).

Selection of competent interviewers was important step towards achieving a good response rate. Interviewers were trained and received in-field support. Their performance was regularly monitored. The call pattern was also an important part of achieving a good response rate. The 'call' is defined as one visit on one day during a particular time band, eg, 5-8pm. NRB carried out 10 calls at each sampled dwelling before accepting that no contact was possible for that dwelling. The survey was voluntary. Adults that were selected as respondents were told about the survey and given an information brochure. If they did agree to participate in the survey they were asked to sign a consent form.

The pilot test for the survey, with a sample size of 114 people was conducted between August and September 2002. The main survey was carried out from September 2002 to January 2004. The pilot test was a test of the performance of the questionnaire and sample design (Ministry of Health, 2004).

Sample allocation

More respondents were interviewed in the second half of the interview period, in particular more Māori, Asian and Pacific peoples. However, subsequent analyses of the data from the survey found that the effects of seasonality on the estimates were found to be insignificant. However, it is advised that care be taken whenever analysing health data where seasonality may be important (Ministry of Health, 2004).

Questionnaire

Most of the questions in the 2002/03 NZHS were sourced from international or local health surveys, or were developed by researchers with expertise in the topic area. The survey had four health-related and one demographic module. The four health-related sections in the questionnaire were chronic diseases, health service utilisation, risk and protective factors and self-reported health status (Ministry of Health, 2004).

The topics covered in the demographics module were the standard socio-demographic variables such as age, gender, ethnicity, country of birth, household details, education, income support, employment, personal and household income, medical insurance and geographic area unit (for use to determine rural/urban and NZDep score). The demographic and socioeconomic questions were sourced from the 2001 Census and/or Household Labour Force Survey to ensure comparability with other national surveys. Minor changes were needed in some questions to change from self-administered to interviewer administered format (Ministry of Health, 2004). Details about the self-reported health status module are given later in this chapter. The relevant questions from the 2002/03 NZHS questionnaire are given in Appendix I.

Response rate

The main performance indicator for the survey was a true response rate of at least 70 percent. The survey was well received by the New Zealand public. Seventy two percent of eligible people approached agreed to complete and interview. There were four parts to the response rate calculation, ineligibles, eligible responding, eligible non-responding and unknown eligibility. The response rate was calculated as shown in the following formula:

$$\text{Response rate} = \frac{\text{number of eligible responding}}{\left[\begin{array}{c}\text{number of eligible}\\ \text{responding}\end{array}\right] + \left[\begin{array}{c}\text{number of eligible}\\ \text{non - responding}\end{array}\right] + \left[\begin{array}{c}\text{estimated number of eligibles}\\ \text{from the unknowns}\end{array}\right]} \times 100$$

(Ministry of Health, 2004).

Quality control

Quality control was implemented through comprehensive testing, ongoing performance monitoring of interviewers, peer review, using standard classifications and concept and by using specialist staff (Ministry of Health, 2004).

Weighting estimation

The weighting estimation was carried out by the Ministry of Health once the quality checks were conducted. As the survey was a carried out on a sample of respondents, each individual represented a number of people in the population. Thus each person was assigned a weight, which was based on how many population units they represented, and weighted estimates were calculated for the whole population. The resulting weight is the sample design weight, unadjusted for non-response or post-stratification (Ministry of Health, 2004).

The methodology for calculating the selection weight was complicated by having a 'main' sample and an over-sample. For PSUs in the 'main' sample, where no over-sampling was carried out, the probability of selecting a person is the product of the following three probabilities: the probability of selecting a PSU, the probability of selecting a dwelling within a PSU and the probability of selecting an eligible person in a selected dwelling (Ministry of Health, 2004).

For PSUs where over-sampling occurred, the calculation is more complicated as probability of selection now equals the sum of the probability of being selected for the 'main' sample and the screened dwellings (over-sample). The inverse of the probability of selection is equal to the survey weight ie. survey weight = 1/probability of selection. Weights were calculated taking into account the early and latter sample designs (Ministry of Health, 2004).

To ensure that the final weights were consistent with the 2001 Census population, generalised linear weighting was carried out. This method allows more flexibility to include several population benchmarks. The benchmarks used were the number of people aged 15 years and over living in permanent private dwellings by 10-year age group, by prioritised ethnicity and lifecycle age groups and by sex and NZDep2001 deciles (Ministry of Health, 2004).

Age is an important determinant of health status. Thus age-standardisation was conducted by the direct method using the WHO World population age distributions applied to the 2001 Census population counts. Thus there are two weights, the New Zealand population survey weight and the age standardised weight. By using age standardised weights any difference in ethnic groups for either sex cannot be attributed to difference in the age structure of the population groups. A third set of weights included is replicate weights to allow confidence intervals to be calculated (Ministry of Health, 2004).

CURF

The full 2002/03 NZHS dataset was confidentialised to form a National Confidentialised Unit Record File (CURF). The following steps were undertaken to confidentialise the full dataset; all respondent contact details were removed; all geographic identifiers were removed; Chatham and Pitt Island records were removed and the data reweighted so that these individuals were represented by the remaining respondents; for some questions the responses were recoded due to small sample sizes; continuous variables were ranged. The full dataset had 12,929 records and the CURF 12,529 records, due to the removal of the 400 Chatham and Pitt Island records (Ministry of Health, 2004b)

Data reliability

Survey estimates inevitably involve sampling errors due to the fact they are based on a sample, not a full census of the population. The Ministry of Health calculated sampling errors for the survey estimated using a replicated method termed the Delete-a-Group (DAG) jackknife method (Kott, 1998). This approach involves dividing the sample into G random groups, then estimating the variance of the complete sample survey estimate. G subsamples were produced by deleting groups from the full sample one at a time. Weighting estimation is carried out for each subsample, generating G 'replicate weights' for each unit record in the dataset. To calculate the variance of an estimate the DAG jackknife formula was used. The DAG jackknife requires the number of first-phase samples in each stratum to be larger the 5 (Ministry of Health, 2004).

Health Status module of 2002/03 NZHS

The self-reported health status section of the 2002/03 NZHS was expanded from the section used in the 1996/97 survey. In the 1996/97 NZHS only the SF-36 questionnaire was used. The reason for the expansion of this section was to give a better overall measure of self-

reported health status. The WHO Long Form (WHOLF) provides coverage of more of the International Classification of Functioning, Disability and Health (ICF) health domains (Ministry of Health, 2004). This allows for comparisons to be drawn from the data collected. Two WHOLF domains were not included in the 2002/03 NZHS: fertility as data is to be collected in the NZ Sexual and Reproductive Health Survey, due to go in to the field in 2006; skin and disfigurement because this domain not considered to be important from a policy perspective and was thus omitted (Ministry of Health, Unpublished report).

Table 8. WHO Long Form and Composite Domains and abbreviated item content

Domain	Abbreviated item content
Social functioning	Extent health problems interfered with normal social activities Frequency health problems interfered with social activities Dealing with people you do not know Maintaining a friendship Getting along with people who are close to you
Physical functioning	Vigorous activities, such as running, lifting heavy objects, strenuous sports Moderate activities, such as moving a table, vacuuming, bowling Lifting or carrying groceries Climbing several flights of stairs Climbing one flight of stairs Bending, kneeling or stooping Walking more than one mile Walking several blocks Walking one block Standing up from sitting down Placing your hands behind your head Using your hands and fingers
Vision	Seeing and recognising a person from 20 metres Seeing and recognising a person from across the room Reading a book or newspaper
Hearing	Hearing a conversation with a person in a quiet room Hearing someone talking on the other side of the room Hearing a group conversation with at least 3 people
Digestion & bodily excretions	Had indigestion Had constipation Had difficulty passing urine Had difficulty controlling urine
Breathing	Short of breath with mild exercise Short of breath at rest Coughing or wheezing for ten minutes or more
Self-care	Bathing Dressing Grooming Eating Using the toilet Staying by yourself for a few days
Understanding	Concentrating on doing something for at least 10 minutes Remembering to do important things Analysing and solving problems in day-to-day life Learning a new task
Communication	Understanding what people say Starting and maintaining a conversation Speaking clearly
Sleep	Problem falling asleep Waking up frequently during night Waking up too early

The health status section of the 2002/03 NZHS measures health-related quality of life (HRQL). The questions were derived from the SF-36 and the WHO Long Form questionnaire on health status. There were 16 health and health-related domains covered. The health domains covered in the 2002/03 NZHS were general health, vision, hearing, digestion, breathing, pain, sleep, energy and vitality, understanding, communication, physical functioning, self-care. The health-related domains covered in the 2002/03 NZHS were mental health, role-physical and role-emotional (usual activities), and social functioning, as shown in Table 9.

The total number of questions in this module was 74, of which 71 are used to derive the scale scores. The remaining questions are related to specific health issues eg, location of pain or use of hearing equipment or services. The SF-36 is embedded within this instrument, covering 8 domains and there are 8 new domains from the WHOLF-NZ. There were two composite domains, physical functioning and social functioning, which included as the base the SF-36 items plus additional items from the WHOLF.

The analysis in this section focussed on the new eight WHOLF-NZ and the two composite domains in the Health Status module.

Table 9. Number of domains covered by SF-36, WHO Long Form (NZ Version) and composite instrument used in Health status module of 2002/03 NZHS

Domain	SF-36		WHO Long Form NZ Version (WHOLF-NZ)		Composite instrument in Health Status module of 2002/03 NZHS	
	Included	No. of items	Included	No. of items	Instrument source	No. of items
Role-physical	✓	4			SF-36	4
Bodily pain	✓	2	✓	2	SF-36	2
General health	✓	5	✓	1	SF-36	5
Energy and vitality	✓	4	✓	2	SF-36	4
Role-emotional	✓	3			SF-36	3
Mental health	✓	5	✓	1	SF-36	5
Physical functioning	✓	10	✓	2	SF-36 + WHOLF-NZ	12
Social functioning	✓	2	✓	3	SF-36 + WHOLF-NZ	5
Vision			✓	3	WHOLF-NZ	3
Hearing			✓	5	WHOLF-NZ	3
Digestion & bodily excretions			✓	4	WHOLF-NZ	4
Breathing			✓	3	WHOLF-NZ	3
Self care			✓	6	WHOLF-NZ	6
Understanding			✓	4	WHOLF-NZ	4
Communication			✓	3	WHOLF-NZ	3
Sleep			✓	3	WHOLF-NZ	3

As shown in Table 9, for the social functioning (SF) domain, the scale is comprised of the original SF-36 scale, plus 3 additional questions from the WHOLF-NZ. For the physical functioning domain the scale was comprised of the original SF-36 scale, plus two additional questions from the WHOLF-NZ. For all remaining scales they are either sourced from the SF-36 or WHOLF-NZ exclusively.

The names of the ten new WHO derived scales derived were physical functioning (WPF), social functioning (WSF), vision (V), hearing (H), digestion and bodily excretions (DB), breathing (BR), self-care (SC), understanding (US), communication (CM) and sleep (S).

In the second half of this chapter I will describe the specific methods used to apply the SF-36 and WHO Long Form in the 2002/03 NZHS, and to test their reliability and validity.

Embedding of SF-36 within WHO Long Form

Most new questions on health status were either taken exactly or slightly adapted from the WHO Long Form instrument. This instrument was developed in association with the World Health Survey and is based on the ICF (World Health Organization, 2001). The SF-36 items (version 1, Australia and New Zealand adaptation) were embedded in the WHO Long Form by slightly modifying the wording or response categories of the WHO Long Form questions. This was done to enable the SF-36 analysis to be carried out as a subset of the self-reported health module and also to enable comparisons with using the full set of scales from the WHO Long Form (Ministry of Health, 2004).

As a result of the embedding of the SF-36 within the WHO Long Form, one of the items from the physical functioning scale was asked as two separate questions, forming part of the self-care scale. The original SF-36 question was "Does **your health now limit you** in these activities? If so, how much?....Bathing or dressing yourself". In the 2002/03 NZHS this question was separated into "Bathing yourself" and "Dressing yourself". These items were adapted from the Self care domain of the WHO Long Form, a domain not included in the SF-36. These two items were then combined to produce a single item for use in creating the physical functioning scale score.

Statistical analysis

Aaronson et al's (Aaronson et al., 2002) framework and the methods used in the IQOLA project (Ware & Gandek, 1998) were used to guide the analysis of the Health Status module of the 2002/3 NZHS. The analysis carried out by Scott et al for the SF-36 (Scott et al., 1999) was repeated for the WHO Long form (NZ version). In addition effect sizes and two sided Mann-Whitney U tests of differences between the subgroup with chronic disease(s) compared with the subgroup with no chronic disease to assess the discriminative validity of the short child health questionnaire parent form (CHG-PF28) (Ratt, Botterweck, Landgraf, Hoogeveen, & Essink-Bot, 2005).

Scoring

First, scaling recommendations were made for all new scales from the WHO Long Form (NZ version). The methods used for the SF-36 scale construction were the preferred methods, as they are used widely throughout the world. As the scales from the WHO Long Form (NZ version) are presented alongside the SF-36 scales, using the same process of scoring and

summing items to create scales was the most appropriate. Where possible, items were summed and normalised to a score of 0-100, where higher scores represent better health, using the following formula:

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score - lowest } \textit{possible} \text{ raw score})}{\textit{Possible} \text{ raw score range}} \right] \times 100$$

Once scales were constructed, the psychometric properties of the items and scales were assessed for the total population and for major subgroups in the population.

Subgroups

The psychometric performance of the scales was tested for the major subgroups in the population that the NZHS 2002/03 was designed to focus on. The variables used to break the sample up into population subgroups were sex, ethnic group, age group and socio-economic status. The variable used as a measure of socio-economic status was the NZ Deprivation Index 2001 (NZDep2001) (Salmond & Crampton, 2002).

The ethnic group categories were: Māori, Pacific peoples, Asian and European/Other. The assignment of ethnicity to individuals was carried out using the standard prioritisation of ethnic groups chosen by individuals in the following order: Māori, Pacific peoples, Asian, Other, European (Ministry of Health, 2004). The age group categories used were those defined as 'life cycle' age groups: 15-24 years, 25-44 years, 45-64 years, 65-74 years and 75+ years. The NZDep2001 index categories used were quintiles rather than deciles to limit the number of response categories. Quintile 1 is the least deprived quintile and quintile 5 the most deprived quintile. In total there were 16 subgroups for which results were assessed.

Reliability analysis

Items

The first level of analysis was the item. The following item level characteristics were examined, extent of missing data, frequency distribution, mean and standard deviation.

Completeness of data

For those scales for which 50% or more of the items were answered, values were imputed for the remaining missing items with an average of the complete items. If less than 50% of the

scale was complete, values were not imputed. The percent of missing data before and after imputation was assessed. Acceptability of the questionnaire was assessed via the completeness of the data ie. examining the percentage of missing data. The score for a scale cannot be estimated with the same degree of confidence if there is a large amount of missing data for a particular item.

Descriptive statistics

The frequency distribution for each scale was examined to see if all response choices were used by respondents and whether the items had a symmetrical distribution. The best items should have a high degree of variability and a symmetrical, roughly normal distribution.

Under traditional Likert scaling criteria, items should contribute roughly the same amount of information to the total scale score. Thus item means were examined with regards to their equivalence within a scale. Item standard deviations were also examined as these should also be roughly equivalent. For five choice response options the standard deviations were checked to see if they were around the desired value of unity. If the variances of the items differed greatly, standardisation was required (Ware & Gandek, 1998).

Scales

The next level of analysis was the scale. The following scale characteristics were examined, percent of missing data, descriptive statistics, summated rated scaling assumptions and reliability and scaling tests.

Completeness of data

The percent of missing data for the scales was assessed.

Descriptive statistics

The mean, standard deviation and percentiles were calculated for the scales. The percentage of scale scores at the floor (minimum) and ceiling (maximum) was calculated. If more than 25% of respondents score at the minimum or maximum of the scale a floor or ceiling effect was said to exist for the scale. This standard was deemed to be met if all items contributed to the scale score, even if the inter-scale correlations varied eg. From 0.4 to 0.7 within a scale.

Summated rating scale assumptions

Three assumptions were examined. First, the assumption that the items were substantially linearly related to the scale score was tested using item internal consistency correlations.

Item-scale correlations were computed after correcting for the item-scale overlap. The item-scale overlap is corrected for estimating the correlation between and item and its hypothesised scale as if the item was not in the total scale score. This avoids the inflating of the item-scale correlation coefficient (Ware & Gandek, 1998). The assumption of item internal consistency is met if item-total correlations (after correcting for overlap) are all greater than 0.4 (Streiner & Norman, 2003).

The second assumption that was tested was that the items explained about the same amount of information about the construct being measured. The test for this assumption is the equality of the item-scale correlations for each scale. The item-scale correlations should be roughly equivalent, but no formal statistical test is carried out.

The third assumption that was tested is that the items are stronger measures of their own scale than all other scales. The correlations of the items with each of the nine other scales were calculated. Lower correlations with other scales indicate good discrimination between scales. Hence these are referred to as item discriminant validity correlations. The number of item discriminant validity tests for each scale equals the number of items in a scale, multiplied by the number of scales, minus one.

The measure of this assumption was the scaling success rate. The scaling success rate is equal to the number of significantly higher correlations with item's own scale divided by the total correlations with all other nine scales. This assumption was met if the correlation between an item and its own scale was significantly higher than the other correlations between the item and all the nine other scales, equating to a scaling success rate of 100%. The significance level used to test this assumption was equal to two standard errors, an approximation for the 95% confidence interval. The following formula was used for calculating the standard error.

Standard error = $1 / \sqrt{\text{sample size}}$

(Ware & Gandek, 1998).

Reliability

The reliability coefficient represents the extent to which the measured variance in a score reflects the true score as opposed to random error. For example, a reliability coefficient of 0.70 means that 70% of the measured variance is reliable. There are a number of different methods for estimating reliability, such as test-retest, alternate forms and internal consistency (Ware & Gandek, 1998).

For this analysis, reliability was assessed using the internal consistency method. The reliability of the scales was estimated by calculating Cronbach's alpha coefficient for each scale. Items in a scale should all be measuring the same thing. Cronbach's alpha is based on the number of items in a scale and the average correlation among items. Item homogeneity is defined as the average of all inter-item correlations within a scale and reflects the degree to which items share common variance. The following formula was used to calculate Cronbach's alpha (α):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right)$$

Where k = number of items in scale

s_i^2 = variance of the *i*th item

s_T^2 = variance of the total score

The minimum reliability level needed for between group comparisons is Cronbach's alpha equal to 0.7 and for individual level comparisons is 0.9. The reliability coefficient represents the extent to which the measured variance in a score reflects the true score as opposed to random error. For example, an α of 0.70 means that 70% of the measured variance is reliable (Bland & Altman, 1997).

Validity analysis

Construct validity

Relationships between scales

To assess the construct validity of the scales, the correlations of the scales with each other were calculated. First, they were compared to the reliability coefficient for each scale which captures the correlation of a scale with itself. For the scales to be measuring unique constructs, all correlations with other scales should be less than the reliability coefficient. The patterns of correlations between scales were also examined as a further means of assessing construct validity to understand the inter-relationships between the scales and their underlying constructs.

Comparative analysis for SF-36 and Composite scales

Two of the new scales, the physical functioning and social functioning scales, were composite measures of SF-36 and WHO items. The performance of these two new scales was compared to the SF-36 scales. The number of items, mean, standard deviation, percentiles, percent at scoring at the ceiling and Cronbach's alpha for the corresponding scales were compared. The correlation between the SF-36 and the Composite scale was calculated.

Discriminative validity

Discriminative validity is construct validation with extreme groups. The discriminative validity of the scales was assessed by comparing scale scores for those in the sample with and without chronic diseases, to see if the scale scores discriminated between these two groups. Discriminative validity was assessed with effects sizes and Mann Whitney U tests, as done by Raat et al (Ratt et al., 2005).

The chronic conditions group was defined as all respondents who had answered 'yes' to any of the chronic conditions questions included in the 2002/03 NZHS. The conditions included were the following nine conditions, plus a question asking about any other long-term illnesses. The specified conditions were: heart disease, stroke, diabetes, asthma, COPD, arthritis, spinal disorders, osteoporosis and cancer. Respondents that answered "Don't know", "Not specified/refused" to all of these questions were excluded from this analysis.

The effect size was defined as:

$$\text{Effect size (d)} = \frac{[\text{mean(no conditions subgroup)} - \text{mean(chronic conditions subgroup)}]}{\text{Standard deviation in the chronic condition subgroup}}$$

Effect sizes were categorised as being small, medium or large, using the following criteria. For a small effect size, $0.2 \leq d < 0.5$; for a medium effect size, $0.5 \leq d < 0.8$ and for a large effect, $d \geq 0.8$. As the standard deviations in the chronic condition group tended to be higher than in the "no condition" subgroup this produced conservative effect sizes (Cohen, 1977).

Two sided Mann-Whitney U tests were carried out to test the differences between the chronic conditions subgroup and the no conditions subgroup. This non parametric test was used due to the data not being normally distributed.

Norms

Finally, norms for the whole population are presented. The value presented is the scale mean and standard error. Norms are also presented by sex, age groups, ethnic groups and deprivation quintiles. As the scales are independent of each other, scale scores cannot be compared. However, within one scale, population subgroup means can be compared.

Data

The CURF data file for the 2002/03 NZHS was used, as this was the file available for use by researchers.

Software

The data were analysed using SAS (SAS Institute Inc.) statistical package. All analyses applied a unique survey weight for all respondents. For further information on how these weights were estimated, see the following section on the methodology of the 2002/03 NZHS. The SAS programmes for conducting the analysis are available from the author.

Chapter 4: Results - Scoring, reliability, validity and norms

In this chapter I describe the process for creating the WHO Long Form scales in the 2002/03 NZHS, followed by the results of the reliability and validity testing of this instrument. I then present population norms for the scales. The testing was performed for the whole population and major subgroups of interest within the population. Tables and figures for the total population and population subgroups are given within this chapter.

I start by describing the sample characteristics of the 2002/03 NZHS and the explanation of how scales were calculated. Then I present the reliability analysis first for individual items and then scales.

Sample characteristics

Response rate

The overall response rate for the 2002/03 NZHS was 72%. The response rates for the Māori, Pacific peoples, Asian ethnic and European/Other groups were respectively 70%, 60%, 62% and 77%.[2] The response rates for the NZDep01 quintiles were respectively 73%, 73%, 72%, 69% and 73%. The response rates were not able to be accurately estimated for the different sexes and age groups due to the sampling process.[3]

The sample characteristics are presented for the population subgroups used in the analysis. The variables used were sex, ethnic group, age group and socioeconomic status. Survey weights were used to produce estimates for the NZ population. Comparisons are provided to the target population for the survey, which was based on the 2001 New Zealand Census of Population and Dwellings.

Sex

There were more females (61.1%) than males (38.9%) included in the sample. There were more females in the sample than in the target population (51.9%).

---

[2] Heineman, A. 2005. Personal communication. NRB: Auckland. 3 August 2005.

[3] Hill, L. 2006. Personal communication. NRB: Auckland 7 February 2006.

Age Group

The age distribution of the sample is shown in Table 10. The largest age group in the sample were those aged 25-44 years (40.2%), followed by those aged 45-64 years (29.7%). The mean age for the sample was 45 years, with a range of 15-97 years.

Table 10. Total sample of 2002/03 NZHS and target population by age group

Age Group	Number	Percent of sample (%)	Percent of target population (%)
15-24 years	1566	12.5	17.3
25-44 years	5039	40.2	38.9
45-64 years	3718	29.7	29.0
65-74 years	1236	9.9	8.0
75+ years	970	7.7	6.8
TOTAL	12529	100.0	100.0

The proportion of participants in the 25-44 years age group (38.9%) was slightly higher in the sample then in the target population, and this was the reverse for the 15-24 years age group (17.3%).

Ethnic Group

The ethnic distribution of the sample for the 2002/03 NZHS is shown in Table 11. When categorised by prioritised ethnicity, 32.9% of the sample were Māori, 7.3% Pacific peoples, 9.4% Asian and the remaining 50.5% were Other, with the major ethnic group being European. As previously discussed in the Chapter 3, the study included oversampling of Pacific and Asian peoples. There were no participants for which ethnic group was missing.

Table 11. Total sample of 2002/03 NZHS and target population by ethnic group

Ethnicity	Number	Percent of sample (%)	Percent of target population (%)
Māori	4120	32.9	10.9
Pacific	908	7.3	4.4
Asian	1172	9.4	6.0
European/Other	6329	50.5	78.7
TOTAL	12529	100.0	100.0

Due to the oversampling in the survey methodology, the proportion of Māori (10.9%) and Pacific (4.4%) participants in the sample was higher than in the target population. Conversely this meant that the proportion of participants in the "Other" ethnic group (78.7%) was lower in the sample than in the target population.

Deprivation

The distribution of the sample by NZDep01 quintile is shown in Table 12. Quintile 5 is the most deprived quintile and Quintile 1 the least deprived. The largest proportion of the sample were from Quintile 5 (39.7%), followed by Quintile 4 (19.1%), with similar numbers in the remaining three quintiles. There were a small proportion of participants (0.23%) of the target population (0.1%) for which deprivation information was missing.

Table 12. Total sample of 2002/03 NZHS and target population by deprivation quintile

Quintile	Number	Percent of sample (%)	Percent of target population (%)
Quintile 1 (=least deprived)	1705	13.6	18.5
Quintile 2	1573	12.6	19.4
Quintile 3	1854	14.8	20.5
Quintile 4	2389	19.1	21.6
Quintile 5 (=most deprived)	4979	39.7	19.9
TOTAL	12500	100.0	99.9

The proportions in the more deprived quintiles were higher in the sample than in the target population, due to the fact that Māori and Pacific peoples are overrepresented within these quintiles.

When sample estimates were generated for the total population, weighted estimates for the total population. Due to the sampling methodology as described in Chapter 2, the sample and target population characteristics could be expected to differ. This is largely due to the oversamples included in the total sample.

The following analysis was carried out for the total population, and for the 16 subgroups described above, which divide the sample by sex, life-cycle age group, ethnic group and deprivation quintile. This allowed the psychometric performance of the WHO scales to be

tested for the whole NZ population and for population subgroups. All results are weighted unless otherwise specified.

Scoring

SF-36 Domains

The SF-36 items in the 2002/03 NZHS were scored in accordance with the guidelines provided in SF-36 Manual (Ware et al., 1993) The scoring process is discussed in Chapter 2. The only deviation from this procedure was in the scoring for the tenth item of the physical functioning scale, due to the splitting of this item into two items as used in the WHO Long Form. The process for combining the two items to produce one score was the following. If the answer to either of the questions was "Yes, limited a lot", then the answer to the combined item was "Yes, limited a lot". If the answers to both the items were "No, not limited at all", then the answer for the combined item was "No, not limited at all". The scoring procedure was applied uniformly to the whole sample.

WHO Long Form Domains

Process for summing scales

There were eight new WHO and two new Composite scales derived from 46 items in the Health Status module of the 2002/03 NZHS. The same process used for deriving SF-36 scales was used to derive the new WHO scales and the Composite scales. Items were recoded where necessary to reverse the order of item values, such that high scores represented good health and low scores poor health. Imputation was carried out for missing items within scales using the same process as for the SF-36 scales. For scales where at least half the items were completed, the values for the missing items were imputed by averaging the scores of the completed items.

All scales had items with the same number of response options, except for one of the two Composite scales, the social functioning scale. This scale had five items. Four items had five response options and one item had six response options. To standardise the number of response options in order to equalise the variance, the item with six response options was recoded to having five response options, with equal distance between the categories. The new values for the six categories were respectively, 1.0, 1.8, 2.6, 3.4, 4.2 and 5.0.

After imputation and recoding items were summed. Items were all given an equal weighting within all scales. Unequal weighting of items rarely improves the performance of the scale enough to justify the added complexity associated with differential item weighting. After summing, the scale totals were then transformed to give scores on all scales from 0-100, where 100 represented the best possible health. The formula used to transform the raw score to a standardised score was the following.

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score - lowest } \textit{possible} \text{ raw score})}{\textit{Possible} \text{ raw score range}}\right] \times 100$$

The names of the ten new WHO derived scales derived were physical functioning (WPF), social functioning (WSF), vision (V), hearing (H), digestion and bodily excretions (DB), breathing (BR), self-care (SC), understanding (US), communication (CM) and sleep (S).

Reliability analysis

Items

Completeness of data

For 20 items there was no missing data before or after imputation. For 26 items there was missing data before imputation (maximum count missing was 8 for any item). For those scales for which 50% or more of the items were answered, values were imputed for the remaining missing items with an average of the complete items. If less than 50% of the scale was complete, values were not imputed. There were 15 items for which there was missing data after imputation (maximum count missing was 2 for any item). For those items for which there was missing data, the count was so small that it equated to zero percent missing data when the percentage was rounded. There was no difference in the percentage of missing data by item across population subgroups, so information regarding missing data by item by subgroup is not presented.

Descriptive statistics

The item frequency distributions and the mean and standard deviations for each of the 46 items are presented in Table 13 for the total population.

Table 13. Item frequency distributions in percent for total population

		Mean	SD	Item frequency distribution % in each category					
Scale	Item			1	2	3	4	5	6
Physical functioning	WPF1	2.27	0.81	26.1	26.7	47.1	0	0	0
	WPF2	2.76	0.56	7.4	12.6	80	0	0	0
(WPF)	WPF3	2.84	0.47	4.8	9.5	85.7	0	0	0
	WPF4	2.68	0.63	11.4	15.8	72.9	0	0	0
	WPF5	2.86	0.44	4.9	8	87.2	0	0	0
	WPF6	2.79	0.55	8.1	9.2	82.7	0	0	0
	WPF7	2.86	0.45	5.3	6.4	88.3	0	0	0
	WPF8	2.92	0.33	2.6	4.4	93.1	0	0	0
	WPF9	2.69	0.6	8.5	16.6	74.9	0	0	0
	WPF10	2.83	0.44	3.1	12.9	84	0	0	0
	WPF11	2.95	0.26	1.4	3.2	95.4	0	0	0
	WPF12	2.91	0.33	1.9	5.7	92.4	0	0	0
Social functioning	WSF1	4.68	0.76	1.1	3.7	3.6	10.9	80.8	0
	WSF2	5.44	1.28	4.6	2.1	2.6	3.6	9.2	77.9
(WSF)	WSF3	4.84	0.53	0.4	1.4	2.2	7.7	88.3	0
	WSF4	4.91	0.41	0.2	0.8	1.3	4.1	93.6	0
	WSF5	4.88	0.46	0.3	0.8	1.2	6	91.7	0
Vision	V1	4.84	0.57	0.7	1.9	2.2	6	89.2	0
	V2	4.95	0.34	0.3	0.5	0.8	2.1	96.3	0
(V)	V3	4.79	0.61	0.8	1.6	2.8	9.5	85.4	0
Hearing	H1	4.88	0.43	0.2	0.9	1.5	6.8	90.5	0
	H2	4.77	0.64	0.8	1.8	2.8	10.4	84.2	0
(H)	H3	4.57	0.87	1.7	3.9	5.3	14.3	74.7	0
Digestion & bodily	DB1	5.57	0.88	0.7	1.3	2.1	7.2	14.5	74.4
excretions	DB2	5.77	0.72	0.6	1	1.3	4.4	6.9	85.8
	DB3	5.93	0.42	0.2	0.4	0.3	1.4	1.6	96.1
(DB)	DB4	5.82	0.62	0.5	0.8	0.7	3.6	6.5	87.9
Breathing	BR1	5.63	0.92	1.5	2.1	2.5	7.2	9.2	77.5
	BR2	5.89	0.52	0.2	0.7	0.8	2.7	3.4	92.1
(BR)	BR3	5.81	0.64	0.5	1.1	1.5	5	5.5	86.4
Self-care	SC1	2.97	0.22	1.1	2	96.9	0	0	0
	SC2	2.97	0.2	0.5	2.3	97.2	0	0	0
(SC)	SC3	2.98	0.15	0.3	1.2	98.4	0	0	0
	SC4	2.99	0.14	0.4	1.0	98.6	0	0	0
	SC5	2.99	0.13	0.3	1.0	98.8	0	0	0
	SC6	2.97	0.23	0.9	1.4	97.7	0	0	0
Understanding	U1	4.73	0.65	0.4	2.4	3.9	13.2	80.2	0
	U2	4.64	0.69	0.5	2.3	4.3	20.8	72	0
(US)	U3	4.74	0.61	0.5	1.6	3	15	79.9	0
	U4	4.72	0.64	0.6	1.7	3.5	15.3	79	0
Communication	CM1	4.83	0.48	0.2	1.1	2.1	11.3	85.2	0
	CM2	4.82	0.53	0.3	1.1	2.4	9.8	86.4	0
(CM)	CM3	4.86	0.46	0.2	0.7	1.8	8	89.3	0
Sleep	S1	5.13	1.31	3	5	4.4	12	14.5	61.1
	S2	4.8	1.43	4.1	7	6.7	16.1	19	47.2
(S)	S3	5.14	1.41	4.8	6.7	3.7	11.4	11.4	62

Note: A high value indicates better health for all items

Physical functioning

All response options are used for all items in the physical functioning (PF) scale. The frequency distribution was non-symmetrical and skewed towards response option three, the end representing better health, for all items in the scale.

The first PF item had a lower mean and higher standard deviation then other items in the scale.

Social functioning

All response options are used for all items in the social functioning (SF) scale. The frequency distribution was skewed towards response option five or six, the end representing better health, for all items in the scale.

To equalise the variance the second item was transformed from six to five response options, changing the mean from 5.44 to 4.55 and the standard deviation from 1.28.

Vision

All response options are used for all items in the vision (V) scale. The frequency distribution was skewed towards response option five, the end representing better health, for all items in the scale, especially for the second item.

Hearing

All response options are used for all items in the hearing (H) scale. The frequency distribution was skewed towards response option five, the end representing better health, for all items in the scale.

Digestion and bodily excretions

All response options are used for all items in the digestion & bodily excretions (DB) scale. The frequency distribution was skewed towards response option six, the end representing better health, for all items in the scale.

Breathing

All response options are used for all items in the breathing (BR) scale. The frequency distribution was skewed towards response option five, the end representing better health, for all items in the scale.

Self-care

All response options are used for all items in the self-care (SC) scale. The frequency distribution was skewed towards response option three, the end representing better health, for all items in the scale.

Understanding

All response options are used for all items in the understanding (U) scale. The frequency distribution was skewed heavily towards response option five, the end representing better health, for all items in the scale.

Communication

All response options are used for all items in the communication (CM) scale. The frequency distribution was skewed heavily towards response option five, the end representing better health, for all items in the scale.

Sleep

All response options are used for all items in the sleep (S) scale. The frequency distribution was skewed heavily towards response option six, the end representing better health, for all items in the scale.

Summary

For each of the scales, all response options were observed for each item but response distributions were skewed towards the healthier end of the distribution, as is expected for general population samples. A less positively skewed distribution was observed for the first item in the physical functioning scale (which represents high physical capacity) and all items is the sleep scale.

Scales

Descriptive statistics

Total population

The scale descriptive statistics for the total population are given in Table 14. The full range of possible scores from 0-100 was observed for all scales. The missing data for all scales was negligible, when rounded producing zero percent missing data for all scales. The median exceeded the mean for all scales, as expected in a general population sample where most people are reasonably well. Negative skewness was pronounced for all scales, with most respondents scoring towards the positive end of the scales. For eight of the ten scales the median was equal to 100, the maximum possible scale score.

For the total population, the highest mean scores were observed for the SC scale (98.8) followed by the V scale (96.5), with the lowest scale means being observed for the S scale (80.5) and WPF scale (89.0). This pattern was reversed for the standard deviations of these scales, with the SC scale having the smallest standard deviation (6.6) and the S (22.1) and WPF (18.2) scales the largest standard deviations.

The number of respondents scoring at the floor of the scale was close to zero for all scales, as would be expected in a general population sample. All scales exhibited a ceiling effect with greater than 25% of the sample scoring at the 100 for all scales. The heaviest floor effect was observed for the SC scale with 94.9% of respondents scoring at the ceiling. The weakest ceiling effect was observed for the S scale with 30.4% of respondents scoring at the ceiling.

Table 14. Scale descriptive statistics for total population

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	89	0.2	87.5	95.8	100	100	18.2	0	0.2	44.7
Social Functioning (WSF)	94.3	0.2	95	100	100	100	11.6	0	0	68.8
Vision (V)	96.5	0.1	100	100	100	100	10.3	0	0.3	80.6
Hearing (H)	93.5	0.2	91.7	100	100	100	14.1	0	0.1	71.0
Digestion & bodily excretions (DB)	95.5	0.1	95	100	100	100	8.3	0	0	61.6
Breathing (BR)	95.5	0.1	100	100	100	100	11.1	0	0.1	76.1
Self-care (SC)	98.8	0.1	100	100	100	100	6.6	0	0.1	94.9
Understanding (US)	92.7	0.2	87.5	100	100	100	12.4	0	0.1	56.6
Communication (CM)	95.9	0.2	100	100	100	100	9.9	0	0	77.4
Sleep (S)	80.5	0.3	73.3	86.7	100	100	22.1	0	0.9	30.4

Sex

The scale descriptive statistics for the males and female are given in Table 15 and Table 16. The full range of scale scores from 0-100 was observed for all scales for females and all except DB for males. For all scales, there were no missing data for gender. For both sexes the median exceeded the mean for all scales. Negative skewness (with the tail of low values higher than the tail of high values on the scale) was strong for all scales for males and

females, with the median equal to 100, the maximum possible scale score for eight of the ten scales.

For males, the highest mean scores were observed for the SC scale (98.9) followed by the V scale (96.9), with the lowest scale means being observed for the S scale (82.4) followed by the WPF scale (90.8). For females, the highest mean scores were observed for the SC scale (98.8) followed by the CM scale (96.2), with the lowest scale means being observed for the S scale (78.6) followed by the PF scale (87.4). This pattern was reversed for the standard deviations of these scales, with the SC scale having the smallest standard deviation (6.5 and 6.6 for males and females respectively) and the S (21.1 and 16.5 for males and females respectively) and WPF (16.5 and 19.5 for males and females respectively) scales the largest standard deviations.

For both sexes, the number of respondents scoring at the floor of the scale was close to zero for all scales. All scales exhibited a ceiling effect. The heaviest ceiling effect was observed for the SC scale with 95.4% of males and 94.5% of females scoring at the ceiling. The weakest ceiling effect was observed for the S scale with 34.0% and 27.1% of males and females respectively scoring at the ceiling.

Table 15. Scale descriptive statistics for males

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	90.8	0.3	91.7	95.8	100	100	16.5	0	0.2	49
Social functioning (WSF)	94.8	0.3	95	100	100	100	11.1	0	0	71
Vision (V)	96.9	0.2	100	100	100	100	10.2	0	0.3	82.5
Hearing (H)	92.4	0.3	91.7	100	100	100	15.1	0	0.1	66.7
Digestion & bodily excretions (DB)	96.4	0.2	95	100	100	85	7.6	0	0	66.9
Breathing (BR)	96.4	0.2	100	100	100	100	9.6	0	0.1	79.1
Self-care (SC)	98.9	0.1	100	100	100	100	6.5	0	0.1	95.4
Understanding (US)	92.9	0.3	87.5	100	100	100	12.1	0	0.1	57.2
Communication (CM)	95.6	0.2	100	100	100	100	10.4	0	0	76.0
Sleep (S)	82.4	0.5	73.3	86.7	100	100	21.1	0	0.7	34.0

Table 16. Scale descriptive statistics for females

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	87.4	0.3	83.3	95.8	100	100	19.5	0	0.1	40.7
Social functioning (WSF)	93.8	0.3	91	100	100	100	12.0	0	0	66.7
Vision (V)	96.1	0.2	100	100	100	100	10.5	0	0.2	78.9
Hearing (H)	94.6	0.2	100	100	100	100	13.0	0	0.1	75.0
Digestion & bodily excretions (DB)	94.7	0.2	90	100	100	100	8.8	0	0	56.7
Breathing (BR)	94.7	0.2	93.3	100	100	100	12.3	0	0.1	73.2
Self-care (SC)	98.8	0.1	100	100	100	100	6.6	0	0	94.5
Understanding (US)	92.4	0.3	87.5	100	100	100	12.7	0	0.1	56.0
Communication (CM)	96.2	0.2	100	100	100	100	9.5	0	0	78.6
Sleep (S)	78.6	0.4	66.7	86.7	100	100	22.9	0	1	27.1

Age Group

The scale descriptive statistics for the life cycle age groups are given in Table 17 to Table 21. The full range of scale scores from 0-100 was not observed for five scales, zero scales, two scales, four scales and three scales, respectively for the 15-24, 25-44, 45-64, 65-75 and 75+ years age groups. There was zero percent missing data for all scales for males and females. For both sexes the median exceeded the mean for all scales. Negative skewness was strong for all scales and age groups.

For all age groups, except those aged 75+ years, the highest mean scores were observed for the SC scale (99.4, 97.9, 99.0 and 98.0 respectively). For those in this age group the highest mean score was observed for the CM scale (95.8). The lowest scale means for all except the oldest two age groups, 65-74 years and 75+ years were observed for the S scale (82.0, 82.0 and 79.1 respectively). For the two other age groups the lowest scale means were observed for the WPF scale (78.0 and 76.5 respectively). For the three youngest age groups the SC scale has the smallest standard deviation (4.3, 5.5 and 6.0 respectively for 15-24, 25-44, and 45-64 year olds respectively). For the two older age groups, the CM scale had the smallest standard deviation (7.9 and 10.5 respectively for the 65-74 and 75+ years age groups).

For all age groups, the number of respondents scoring at the floor of the scale was close to zero for all scales. For the three youngest age groups, all scales exhibited a ceiling effect. The heaviest ceiling effect was observed for the SC scale with 95.4% of 15-24 year olds, 96.9% of 25-44 year olds and 95.2% of 45-64 year olds scoring at the ceiling. The weakest ceiling effect was observed for the S scale (35.5, 33.0 and 28.0 respectively for 15-24, 25-44, and 45-64 year olds respectively scoring at the ceiling). For 65-74 and 75+ year olds the scales that did not exhibit ceiling effects were the WPF and S scales.

Table 17. Scale descriptive statistics for 15-24 years

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	95.5	0.4	95.8	100	100	100	10.1	0	0	66.7
Social functioning (WSF)	93.7	0.5	91	100	100	80	11.5	0	0.1	62.7
Vision (V)	97.5	0.3	100	100	100	83.3	7.8	0	0	85.5
Hearing (H)	96.1	0.4	100	100	100	83.3	9.3	0	0	79.0
Digestion & bodily excretions (DB)	96.8	0.2	95	100	100	75	6.6	0	0	69.3
Breathing (BR)	95.5	0.4	93.3	100	100	80	10.6	0	0.2	73.8
Self-care (SC)	99.4	0.1	100	100	100	100	4.3	0	0	96.8
Understanding (US)	91	0.5	87.5	93.8	100	100	12.9	0	0	49.5
Communication (CM)	94.2	0.4	91.7	100	100	100	11.3	0	0.1	68.2
Sleep (S)	82	0.9	73.3	86.7	100	100	21.8	0	0.4	35.3

Table 18. Scale descriptive statistics for 25-44 years

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	93.6	0.2	91.7	100	100	100	13.2	0	0.2	57.5
Social functioning (WSF)	94.4	0.3	95	100	100	100	11.9	0	0	70.0
Vision (V)	97.9	0.2	100	100	100	100	8.1	0	0.2	87.2
Hearing (H)	95.9	0.3	100	100	100	100	10.9	0	0	79.0
Digestion & bodily excretions (DB)	96.2	0.1	95	100	100	100	7.7	0	0	65.7
Breathing (BR)	96.3	0.2	100	100	100	100	10.1	0	0	79.4
Self-care (SC)	99.3	0.1	100	100	100	100	5.5	0	0.1	96.9
Understanding (US)	93.2	0.3	93.8	100	100	100	11.9	0	0	58.0
Communication (CM)	95.8	0.2	100	100	100	100	10.3	0	0	77.5
Sleep (S)	82.0	0.4	73.3	86.7	100	100	21.5	0	0.9	33.0

Table 19. Scale descriptive statistics for 45-64 years

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	87.9	0.4	83.3	95.8	100	100	17.6	0	0.1	33.3
Social functioning (WSF)	94.8	0.2	95	100	100	100	11	0	0	71.1
Vision (V)	95.7	0.2	91.7	100	100	100	9.5	0	0	72.4
Hearing (H)	93.1	0.3	91.7	100	100	100	13.9	0	0.1	67.6
Digestion & bodily excretions (DB)	95.2	0.2	95	100	100	85	8.4	0	0	59.5
Breathing (BR)	96.0	0.3	100	100	100	100	10.9	0	0.2	77.8
Self-care (SC)	99.0	0.1	100	100	100	100	6.0	0	0	95.2
Understanding (US)	93.5	0.3	93.8	100	100	100	11.7	0	0	60.8
Communication (CM)	96.7	0.2	100	100	100	91.7	8.8	0	0	80.9
Sleep (S)	79.1	0.5	66.7	86.7	100	100	22.7	0	1	28.0

Table 20. Scale descriptive statistics for 65-74 years

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	78.0	0.8	66.7	87.5	95.8	100	22.3	0	0.2	10.8
Social functioning (WSF)	94.8	0.5	95	100	100	80	11.2	0	0	72.3
Vision (V)	95.5	0.6	100	100	100	100	12.2	0	0.6	78.1
Hearing (H)	88.6	0.7	83.3	100	100	100	17.8	0	0.2	54.6
Digestion & bodily excretions (DB)	93.4	0.4	90	95	100	75	10.2	0	0.2	49.2
Breathing (BR)	94.0	0.5	93.3	100	100	100	13	0	0.1	70.8
Self-care (SC)	98.0	0.3	100	100	100	91.7	8.6	0	0.1	91.8
Understanding (US)	93.4	0.5	93.8	100	100	100	11.9	0	0.1	57.1
Communication (CM)	97.3	0.4	100	100	100	91.7	7.9	0	0.1	83.0
Sleep (S)	78.3	0.8	66.7	86.7	93.3	100	22	0	1.3	22.5

Table 21. Scale descriptive statistics for 75+ years

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	63.8	1.1	41.7	70.8	87.5	100	26.7	0	0.5	3.6
Social functioning (WSF)	92.9	0.5	90	100	100	75	12.1	0.1	0.2	63.1
Vision (V)	90.6	0.8	91.7	100	100	100	20.8	0	1.8	68.7
Hearing (H)	80.6	1.0	66.7	91.7	100	100	23.8	0	0.7	39.3
Digestion & bodily excretions (DB)	91.9	0.5	90	95	100	55	10.8	0	0.5	42.3
Breathing (BR)	91.1	0.7	86.7	100	100	100	15	0	0	61.2
Self-care (SC)	95.4	0.6	100	100	100	100	12.5	0	0	81.2
Understanding (US)	89.4	0.7	81.3	93.8	100	100	16.2	0	0.4	47.9
Communication (CM)	95.8	0.4	100	100	100	91.7	10.5	0	0.2	78.5
Sleep (S)	76.5	1.0	66.7	80	93.3	100	23	0	0.9	22.8

Ethnic group

The scale descriptive statistics for the four ethnic groups are given in Table 22 to Table 25. For all ethnic groups, the full range of scale scores from 0-100 were not observed for at least one scale, ranging from one scale for Māori to four scales for Asian peoples. The proportion of missing data was zero for all scales and all ethnic groups except for the S scale for the Asian ethnic group. However this amount of missing data was very low (0.2%).

The median was higher than the mean and negative skewness pronounced for all scales and ethnic groups. The scale with the highest mean and median for all ethnic groups was SC (93.9, 96.1, 97.1 and 94.8 respectively for Māori, Pacific, Asian and European/Other ethnic groups), followed by V for all ethnic groups, except Asian peoples where H is the second highest. The two scales with the highest standard deviations (S and WPF) were the same for all ethnic groups. No floor effects were observed for any scales for any ethnic group. Ceiling

effects were observed for all scales and all ethnic groups. The heaviest ceiling effect was observed for the SC scale for all ethnic groups.

For all ethnic groups the highest mean scores were observed for the SC scale (98.9, 98.9, 99.2 and 98.8 respectively for Māori, Pacific, Asian and European/Other ethnic groups). The lowest scale means for all ethnic groups were observed for the S scale (77.7, 82.8, 87.2 and 80.2 respectively for Māori, Pacific, Asian and European/Other ethnic groups). For all ethnic groups the SC scale had the smallest standard deviation (6.1, 7.9, 5.8 and 6.6 for Māori, Pacific, Asian and European/Other ethnic groups respectively).

For all ethnic groups, the number of respondents scoring at the floor of the scale was close to zero for all scales. For all ethnic groups, all scales exhibited a ceiling effect. The heaviest ceiling effect was observed for the SC scale with 93.9% of Māori, 96.1% of Pacific peoples, 97.1% of Asian and 94.8% of European/Other scoring at the ceiling. The weakest ceiling effects were observed for the S scale (31.5, 43.3, 47.9 and 28.2 for Māori, Pacific, Asian and European/Other ethnic groups respectively scoring at the ceiling). Thus for European/Other the percentage of people scoring at the ceiling approached the cutoff for a ceiling effect of 25%.

Table 22. Scale descriptive statistics for Māori

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	88.9	0.6	83.3	95.8	100	100	17.7	0	0.2	46.3
Social functioning (WSF)	92.3	0.5	90	100	100	100	14.1	0	0.1	62.3
Vision (V)	95.3	0.5	100	100	100	100	12.8	0	0.3	78.9
Hearing (H)	92.6	0.5	91.7	100	100	100	14.7	0	0.1	68.4
Digestion & bodily excretions (DB)	95.1	0.3	95	100	100	85	9.3	0	0	63.3
Breathing (BR)	92.0	0.6	86.7	100	100	100	15	0	0.1	63.7
Self-care (SC)	98.9	0.2	100	100	100	100	6.1	0	0.1	93.9
Understanding (US)	89.6	0.5	81.3	93.8	100	100	15.4	0	0.2	49.0
Communication (CM)	94.0	0.4	91.7	100	100	100	11.8	0	0	69.0
Sleep (S)	77.7	0.9	66.7	86.7	100	100	24.6	0	1.2	31.5

Table 23. Scale descriptive statistics for Pacific peoples

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	90.3	1.1	87.5	100	100	100	18.3	0	0.2	54.1
Social functioning (WSF)	93.3	0.9	95	100	100	80	13.9	0	0.1	70.6
Vision (V)	96.7	0.5	100	100	100	100	9.5	0	0	84.3
Hearing (H)	96.7	0.7	100	100	100	100	11.9	0	0.1	87.9
Digestion & bodily excretions (DB)	94.7	0.6	95	100	100	80	10.9	0	0.1	66.4
Breathing (BR)	92.9	0.7	93.3	100	100	100	15.2	0	0.3	71.8
Self-care (SC)	98.9	0.3	100	100	100	100	7.9	0	0.2	96.1
Understanding (US)	92.6	1.0	93.8	100	100	100	13.5	0	0	66.4
Communication (CM)	94.6	0.8	100	100	100	100	12.9	0	0.2	76.8
Sleep (S)	82.8	1.0	73.3	93.3	100	100	21.1	0.2	0.2	43.3

Table 24. Scale descriptive statistics for Asian peoples

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	93.9	0.7	95.8	100	100	91.7	14.1	0	0.5	60.9
Social functioning (WSF)	96.1	0.5	100	100	100	85	8.7	0.2	0	75.5
Vision (V)	98.0	0.3	100	100	100	58.3	6.2	0	0	85.2
Hearing (H)	98.7	0.3	100	100	100	75	6	0	0	91.7
Digestion & bodily excretions (DB)	96.3	0.4	95	100	100	100	7.8	0	0.1	68.9
Breathing (BR)	97.4	0.5	100	100	100	100	9.9	0	0.4	85.9
Self-care (SC)	99.2	0.4	100	100	100	58.3	5.8	0	0.4	97.1
Understanding (US)	95.1	0.5	93.8	100	100	75	9.9	0	0.1	69.4
Communication (CM)	96.4	0.5	100	100	100	100	9.8	0	0.1	82.2
Sleep (S)	87.2	0.9	80	93.3	100	100	19.6	0	0.6	47.9

Table 25. Scale descriptive statistics for European/Other

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	88.6	0.2	87.5	95.8	100	100	18.5	0	0.2	42.7
Social functioning (WSF)	94.5	0.2	95	100	100	100	11.2	0	0	69.0
Vision (V)	96.5	0.2	100	100	100	100	10.2	0	0.3	80.3
Hearing (H)	93.1	0.2	91.7	100	100	100	14.4	0	0.1	68.9
Digestion & bodily excretions (DB)	95.5	0.1	95	100	100	75	8	0	0	60.6
Breathing (BR)	96.0	0.1	100	100	100	100	10.2	0	0	77.2
Self-care (SC)	98.8	0.1	100	100	100	100	6.6	0	0	94.8
Understanding (US)	92.9	0.2	87.5	100	100	100	12	0	0.1	56.1
Communication (CM)	96.2	0.2	100	100	100	91.7	9.4	0	0	78.2
Sleep (S)	80.2	0.4	66.7	86.7	100	100	21.9	0	0.9	28.2

Deprivation quintile

The scale descriptive statistics for the five deprivation quintiles are given in Table 26 to Table 30. The full range of scale scores from 0-100 was not observed for seven scales, four scales, four scales, one scale and one scale, respectively for quintiles 1 to 5. In all quintiles, there was zero percent missing data for all scales. For all quintiles the median exceeded the mean for all scales. Negative skewness was strong for all scales and quintiles.

For all quintiles, the highest mean scores were observed for the SC scale (99.4, 98.8, 98.9, 98.7 and 98.4 respectively for quintiles 1 to 5). For all quintiles the S scale had the highest standard deviation (19.6, 21.0, 21.4, 23.5 and 24.2 respectively for quintiles 1 to 5).

For all age groups, the number of respondents scoring at the floor of the scale was close to zero for all scales. For all quintiles, all scales exhibited a ceiling effect. The heaviest ceiling effect was observed for the SC scale for all quintiles with 96.9%, 95.1%, 95.6%, 94.4% and 93.0% of those in quintiles 1 to 5 respectively scoring at the ceiling. The weakest ceiling effect was observed for the S scale (33.3%, 29.9%, 27.7%, 28.7% and 32.8% respectively for quintiles 1 to 5 scoring at the ceiling).

Table 26. Scale descriptive statistics for Deprivation Quintile 1

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	92.6	0.3	91.7	100	100	100	13.8	0	0.1	52.4
Social functioning (WSF)	95.6	0.3	95	100	100	80	9.9	0	0.1	73.2
Vision (V)	97.7	0.2	100	100	100	91.7	7	0	0.1	84.0
Hearing (H)	95.1	0.3	91.7	100	100	91.7	11	0	0	73.4
Digestion & bodily excretions (DB)	96.5	0.2	95	100	100	45	6.1	0	0	63.0
Breathing (BR)	97.3	0.2	100	100	100	80	7.7	0	0	82.0
Self-care (SC)	99.4	0.1	100	100	100	75	4.1	0	0	96.9
Understanding (US)	94.1	0.3	93.8	100	100	75	10.2	0	0.2	59.2
Communication (CM)	97.0	0.3	100	100	100	100	8.2	0	0	81.8
Sleep (S)	83.4	0.5	73.3	86.7	100	100	19.6	0	0.5	33.3

Table 27. Scale descriptive statistics for Deprivation Quintile 2

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	90.1	0.6	87.5	95.8	100	100	17.6	0	0.2	47.1
Social functioning (WSF)	94.5	0.3	95	100	100	81	10.9	0	0	69.3
Vision (V)	96.9	0.3	100	100	100	100	9.2	0	0.1	81.3
Hearing (H)	93.8	0.5	91.7	100	100	100	13.5	0	0	70.2
Digestion & bodily excretions (DB)	95.7	0.3	95	100	100	55	7.9	0	0.3	61.8
Breathing (BR)	96.4	0.3	100	100	100	100	9.8	0	0.3	79.6
Self-care (SC)	98.8	0.2	100	100	100	83.3	6.9	0	0.1	95.1
Understanding (US)	93.2	0.4	93.8	100	100	100	12.3	0	0.1	58.2
Communication (CM)	96.2	0.3	100	100	100	91.7	9.2	0	0	78.5
Sleep (S)	81.1	0.6	73.3	86.7	100	100	21	0	0.8	29.9

Table 28. Scale descriptive statistics for Deprivation Quintile 3

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	88.9	0.6	87.5	95.8	100	100	17.8	0	0.1	42.3
Social functioning (WSF)	94.6	0.3	95	100	100	85	11	0	0	68.7
Vision (V)	96.9	0.3	100	100	100	100	9.1	0	0.2	80.8
Hearing (H)	92.7	0.5	91.7	100	100	100	15.3	0	0.1	70.3
Digestion & bodily excretions (DB)	95.7	0.2	95	100	100	70	7.8	0	0.1	63.2
Breathing (BR)	96.2	0.3	100	100	100	93.3	9.7	0	0.2	77.1
Self-care (SC)	98.9	0.2	100	100	100	100	6.5	0	0.1	95.6
Understanding (US)	92.4	0.5	87.5	100	100	100	12.2	0	0.1	54.3
Communication (CM)	95.9	0.3	100	100	100	91.7	9.8	0	0.1	76.9
Sleep (S)	80.6	0.7	66.7	86.7	100	100	21.4	0	0.8	27.7

Table 29. Scale descriptive statistics for Deprivation Quintile 4

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	87.7	0.5	83.3	95.8	100	100	19.2	0	0.2	40.5
Social functioning (WSF)	94.1	0.4	95	100	100	100	11.9	0	0.1	67.6
Vision (V)	95.5	0.4	100	100	100	100	13.1	0	0.8	79.8
Hearing (H)	93.1	0.4	91.7	100	100	100	14	0	0	68.5
Digestion & bodily excretions (DB)	95.0	0.3	95	100	100	100	9	0	0	59.3
Breathing (BR)	94.8	0.3	93.3	100	100	100	12.2	0	0	73.6
Self-care (SC)	98.7	0.2	100	100	100	100	6.8	0	0.1	94.4
Understanding (US)	92.3	0.4	87.5	100	100	100	12.6	0	0	55.2
Communication (CM)	95.5	0.4	100	100	100	91.7	10.6	0	0.1	76.4
Sleep (S)	78.7	0.7	66.7	86.7	100	100	23.5	0	1.1	28.7

Table 30. Scale descriptive statistics for Deprivation Quintile 5

Scale	Mean	Standard error	25th percentile	50th percentile	75th percentile	Range	SD	% Missing Data	% Floor	% Ceiling
Physical functioning (WPF)	86.2	0.7	83.3	95.8	100	100	20.9	0	0.3	42.4
Social functioning (WSF)	92.8	0.5	91	100	100	100	13.5	0	0	65.4
Vision (V)	95.6	0.4	100	100	100	100	11.5	0	0.1	77.6
Hearing (H)	93.3	0.5	91.7	100	100	100	15.5	0	0.3	73.1
Digestion & bodily excretions (DB)	94.6	0.3	90	100	100	85	9.9	0	0	61
Breathing (BR)	93.2	0.4	93.3	100	100	100	14.2	0	0.1	68.6
Self-care (SC)	98.4	0.2	100	100	100	100	7.7	0	0	93
Understanding (US)	91.5	0.6	87.5	100	100	100	14.2	0	0.1	56.3
Communication (CM)	95.0	0.4	91.7	100	100	100	11.3	0	0	73.6
Sleep (S)	78.8	0.9	66.7	86.7	100	100	24.2	0	1.1	32.8

Summated rating scale assumptions

Total population

The correlations between items and scales are presented in Table 31 and the results of scaling tests and reliability estimates in Table 32 for the total population. Internal consistency was high with correlations (after correction for overlap) greater than 0.4 for all scales. The scale with the lowest range correlations between items and scale total was the DB scale. The correlation between item three and the scale total (0.47) approached the lowest desirable correlation value of 0.4.

The internal consistency correlations within scales were similar, meeting the assumption that all items contribute roughly equal proportions of information to the scale score. This supports the equal weighting of items within the scale, as unequal weighting would be unlikely to improve the performance of a scale enough to justify the added scoring complexity.

Item discriminant validity was good for all scales, with low correlations between items within a scale with all other scales. Scaling success rates were 100% for all scales.

Table 31. Correlations between items and scales for total population

		WHO and Composite scales									
Scale	Item	WPF	WSF	WVS	WHR	WDB	WBR	WSC	WUS	WCM	WSL
Physical functioning	WPF1	**0.69**	0.23	0.17	0.24	0.27	0.28	0.24	0.2	0.13	0.25
	WPF2	**0.81**	0.3	0.2	0.19	0.26	0.32	0.41	0.23	0.15	0.2
(WPF)	WPF3	**0.77**	0.3	0.18	0.17	0.24	0.29	0.43	0.22	0.13	0.22
	WPF4	**0.8**	0.27	0.2	0.23	0.26	0.4	0.33	0.22	0.16	0.22
	WPF5	**0.8**	0.28	0.21	0.22	0.25	0.37	0.42	0.22	0.15	0.18
	WPF6	**0.82**	0.28	0.19	0.22	0.27	0.36	0.39	0.23	0.16	0.19
	WPF7	**0.81**	0.27	0.19	0.19	0.25	0.34	0.44	0.19	0.13	0.17
	WPF8	**0.71**	0.26	0.16	0.16	0.22	0.29	0.46	0.18	0.11	0.15
	WPF9	**0.74**	0.25	0.18	0.24	0.26	0.24	0.33	0.2	0.13	0.23
	WPF10	**0.68**	0.27	0.18	0.19	0.26	0.21	0.38	0.2	0.14	0.21
	WPF11	**0.44**	0.22	0.12	0.1	0.16	0.18	0.43	0.18	0.12	0.15
	WPF12	**0.47**	0.19	0.17	0.15	0.17	0.19	0.39	0.2	0.13	0.15
Social functioning	WSF1	0.36	**0.78**	0.15	0.14	0.23	0.26	0.29	0.36	0.33	0.29
	WSF2	0.3	**0.74**	0.11	0.11	0.19	0.21	0.24	0.25	0.21	0.2
(WSF)	WSF3	0.23	**0.62**	0.15	0.18	0.21	0.22	0.21	0.41	0.46	0.21
	WSF4	0.15	**0.63**	0.11	0.09	0.14	0.12	0.14	0.3	0.36	0.18
	WSF5	0.1	**0.6**	0.07	0.07	0.12	0.1	0.08	0.3	0.32	0.19
Vision	V1	0.2	0.15	**0.83**	0.16	0.15	0.15	0.13	0.18	0.14	0.13
	V2	0.18	0.12	**0.76**	0.13	0.16	0.13	0.11	0.12	0.11	0.09
(V)	V3	0.21	0.13	**0.78**	0.17	0.15	0.13	0.13	0.16	0.11	0.13
Hearing	H1	0.18	0.14	0.14	**0.75**	0.12	0.15	0.08	0.18	0.17	0.11
	H2	0.23	0.14	0.18	**0.89**	0.13	0.17	0.12	0.21	0.18	0.12
(H)	H3	0.27	0.14	0.18	**0.89**	0.15	0.14	0.14	0.21	0.19	0.14
Digestion & bodily	DB1	0.18	0.16	0.11	0.08	**0.65**	0.18	0.08	0.17	0.11	0.21
excretions	DB2	0.2	0.17	0.1	0.07	**0.6**	0.19	0.11	0.15	0.09	0.16
	DB3	0.17	0.15	0.11	0.12	**0.47**	0.13	0.14	0.17	0.12	0.15
(DB)	DB4	0.28	0.18	0.14	0.16	**0.52**	0.18	0.17	0.17	0.1	0.18
Breathing	BR1	0.43	0.24	0.16	0.19	0.25	**0.84**	0.21	0.23	0.17	0.23
	BR2	0.27	0.22	0.12	0.12	0.21	**0.75**	0.17	0.2	0.15	0.19
(BR)	BR3	0.2	0.17	0.12	0.1	0.19	**0.71**	0.13	0.17	0.16	0.19
Self-care	SC1	0.47	0.23	0.14	0.12	0.15	0.19	**0.8**	0.18	0.12	0.1
	SC2	0.45	0.24	0.11	0.12	0.16	0.18	**0.82**	0.19	0.14	0.13
(SC)	SC3	0.35	0.21	0.11	0.11	0.14	0.15	**0.74**	0.18	0.11	0.1
	SC4	0.26	0.18	0.1	0.07	0.1	0.1	**0.62**	0.14	0.11	0.12
	SC5	0.31	0.19	0.07	0.07	0.14	0.13	**0.71**	0.16	0.12	0.09
	SC6	0.33	0.22	0.15	0.1	0.15	0.18	**0.65**	0.24	0.21	0.11
Understanding	U1	0.21	0.38	0.14	0.16	0.21	0.24	0.18	**0.74**	0.43	0.25
	U2	0.21	0.29	0.15	0.2	0.2	0.19	0.17	**0.75**	0.4	0.23
(US)	U3	0.19	0.38	0.13	0.17	0.2	0.19	0.19	**0.77**	0.48	0.21
	U4	0.25	0.29	0.18	0.19	0.17	0.15	0.24	**0.73**	0.43	0.19
Communication	CM1	0.18	0.38	0.14	0.18	0.17	0.17	0.16	0.5	**0.79**	0.21
	CM2	0.16	0.37	0.12	0.19	0.11	0.15	0.15	0.46	**0.81**	0.17
(CM)	CM3	0.12	0.31	0.09	0.13	0.12	0.18	0.15	0.4	**0.77**	0.16
Sleep	S1	0.19	0.26	0.11	0.1	0.21	0.21	0.11	0.24	0.2	**0.74**
	S2	0.28	0.24	0.14	0.15	0.27	0.22	0.13	0.22	0.16	**0.81**
(S)	S3	0.18	0.22	0.11	0.1	0.2	0.2	0.11	0.22	0.17	**0.76**

Note: Correlations in **bold** represent item internal consistency (corrected for overlap), remaining correlations represent item discriminant validity.

Table 32. Results of item scaling tests for total population

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.44 – 0.82	0.1-0.46	120/120	100%
Social functioning (WSF)	5	0.6-0.78	0.07-0.46	50/50	100%
Vision (V)	3	0.76-0.83	0.09-0.21	30/30	100%
Hearing (H)	3	0.75-0.89	0.08-0.27	30/30	100%
Digestion & bodily excretions (DB)	4	0.47-0.65	0.07-0.28	40/40	100%
Breathing (BR)	3	0.71-0.84	0.1-0.43	30/30	100%
Self-care (SC)	6	0.62-0.82	0.07-0.47	60/60	100%
Understanding (US)	4	0.73-0.77	0.13-0.48	40/40	100%
Communication (CM)	3	0.77-0.81	0.09-0.5	30/30	100%
Sleep (S)	3	0.74-0.81	0.1-0.28	30/30	100%

Sex

Results of scaling tests and reliability estimates are presented in Table 33 and Table 34 for males and females. Internal consistency was high with correlations (after correction for overlap) greater than 0.4 for all scales for both sexes. For both males and females, the scale with the lowest range correlations between items and scale total was the WPF scale. The internal consistency correlations within scales were similar, meeting the assumption that all items contribute roughly equal proportions of information to the scale score. Item discriminant validity was good for all scales, with low correlations between items within a scale with all other scales. Scaling success rates were 100% for all scales.

Table 33. Results of item scaling tests for males

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.44-0.81	0.09-0.47	120/120	100%
Social functioning (WSF)	5	0.6-0.76	0.05-0.46	50/50	100%
Vision (V)	3	0.78-0.83	0.07-0.21	30/30	100%
Hearing (H)	3	0.75-0.89	0.04-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.54-0.66	0.07-0.27	40/40	100%
Breathing (BR)	3	0.67-0.84	0.07-0.41	30/30	100%
Self-care (SC)	6	0.66-0.82	0.04-0.47	60/60	100%
Understanding (US)	4	0.72-0.77	0.11-0.5	40/40	100%
Communication (CM)	3	0.77-0.82	0.11-0.52	30/30	100%
Sleep (S)	3	0.71-0.81	0.09-0.28	30/30	100%

Table 34. Results of item scaling tests for females

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.44-0.83	0.1-0.5	120/120	100%
Social functioning (WSF)	5	0.59-0.81	0.06-0.47	50/50	100%
Vision (V)	3	0.74-0.83	0.08-0.21	30/30	100%
Hearing (H)	3	0.76-0.9	0.11-0.29	30/30	100%
Digestion & bodily excretions (DB)	4	0.4-0.64	0.06-0.28	40/40	100%
Breathing (BR)	3	0.73-0.84	0.07-0.43	30/30	100%
Self-care (SC)	6	0.59-0.83	0.06-0.47	60/60	100%
Understanding (US)	4	0.72-0.77	0.12-0.45	40/40	100%
Communication (CM)	3	0.77-0.8	0.07-0.49	30/30	100%
Sleep (S)	3	0.76-0.82	0.09-0.28	30/30	100%

Age Group

Results of scaling tests and reliability estimates are presented in Table 35 to

Table 39 for the 15-24 years, 25-44 years, 45-64 years, 65-74 years and 75+ years age groups. Internal consistency was high with correlations (after correction for overlap) greater than 0.4 for almost all scales and age groups. There were three age groups for which there were correlations between an item and the WPF scale less than 0.4: 15-24, 65-74 and 75+ years (also for the WSF scale). For the latter age group this also applied to the WSF scale. Item discriminant validity was good for all scales, with low correlations between items within a scale with all other scales. Scaling success rates were 100% for all scales.

Table 35. Results of item scaling tests for 15-24 years

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.35-0.74	0.01-0.51	120/120	100%
Social functioning (WSF)	5	0.59-0.76	0.07-0.45	50/50	100%
Vision (V)	3	0.71-0.86	0.01-0.25	30/30	100%
Hearing (H)	3	0.78-0.85	-0.01-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.45-0.69	0.01-0.24	40/40	100%
Breathing (BR)	3	0.75-0.82	0.02-0.26	30/30	100%
Self-care (SC)	6	0.52-0.8	-0.02-0.49	60/60	100%
Understanding (US)	4	0.7-0.75	0.09-0.42	40/40	100%
Communication (CM)	3	0.79-0.81	0.11-0.48	30/30	100%
Sleep (S)	3	0.74-0.8	0.03-0.28	30/30	100%

Table 36. Results of item scaling tests for 25-44 years

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.46-0.78	0.03-0.53	120/120	100%
Social functioning (WSF)	5	0.66-0.79	0.08-0.48	50/50	100%
Vision (V)	3	0.76-0.83	0-0.15	30/30	100%
Hearing (H)	3	0.77-0.89	0.04-0.23	30/30	100%
Digestion & bodily excretions (DB)	4	0.48-0.69	0.05-0.28	40/40	100%
Breathing (BR)	3	0.71-0.84	0.07-0.36	30/30	100%
Self-care (SC)	6	0.68-0.82	0.03-0.47	60/60	100%
Understanding (US)	4	0.72-0.79	0.11-0.51	40/40	100%
Communication (CM)	3	0.8-0.82	0.1-0.56	30/30	100%
Sleep (S)	3	0.73-0.82	0.06-0.3	30/30	100%

Table 37. Results of item scaling tests for 45-64 years

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.41-0.79	0.07-0.41	120/120	100%
Social functioning (WSF)	5	0.56-0.79	0.09-0.5	50/50	100%
Vision (V)	3	0.6-0.75	0.05-0.2	30/30	100%
Hearing (H)	3	0.73-0.88	0.03-0.27	30/30	100%
Digestion & bodily excretions (DB)	4	0.45-0.65	0.03-0.23	40/40	100%
Breathing (BR)	3	0.74-0.83	0.11-0.43	30/30	100%
Self-care (SC)	6	0.61-0.82	0.04-0.41	60/60	100%
Understanding (US)	4	0.71-0.76	0.1-0.46	40/40	100%
Communication (CM)	3	0.72-0.8	0.11-0.48	30/30	100%
Sleep (S)	3	0.75-0.82	0.1-0.28	30/30	100%

Table 38. Results of item scaling tests for 65-74 years

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.39-0.82	0.04-0.48	120/120	100%
Social functioning (WSF)	5	0.46-0.79	-0.01-0.49	50/50	100%
Vision (V)	3	0.77-0.85	0.04-0.25	30/30	100%
Hearing (H)	3	0.76-0.9	0.06-0.22	30/30	100%
Digestion & bodily excretions (DB)	4	0.51-0.6	0-0.21	40/40	100%
Breathing (BR)	3	0.72-0.86	0.06-0.48	30/30	100%
Self-care (SC)	6	0.55-0.83	0.05-0.5	60/60	100%
Understanding (US)	4	0.71-0.77	0.07-0.45	40/40	100%
Communication (CM)	3	0.72-0.83	0.06-0.43	30/30	100%
Sleep (S)	3	0.72-0.8	0.04-0.26	30/30	100%

Table 39. Results of item scaling tests for 75+ years

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.37-0.84	0.06-0.45	120/120	100%
Social functioning (WSF)	5	0.33-0.8	0.03-0.48	50/50	100%
Vision (V)	3	0.88-0.9	0.01-0.25	30/30	100%
Hearing (H)	3	0.75-0.92	0.06-0.25	30/30	100%
Digestion & bodily excretions (DB)	4	0.41-0.6	0.02-0.29	40/40	100%
Breathing (BR)	3	0.62-0.88	0.08-0.45	30/30	100%
Self-care (SC)	6	0.56-0.83	-0.02-0.49	60/60	100%
Understanding (US)	4	0.75-0.8	0.09-0.56	40/40	100%
Communication (CM)	3	0.69-0.85	0.02-0.56	30/30	100%
Sleep (S)	3	0.72-0.77	-0.02-0.32	30/30	100%

Ethnic Group

Results of scaling tests and reliability estimates are presented in Table 40 to Table 43 for Māori, Pacific peoples, Asian and European/Other ethnic groups. Internal consistency was high with correlations (after correction for overlap) greater than 0.4 for almost all scales and ethnic groups. There were two ethnic groups, Māori and Asian, for which there were correlations between an item and the WPF scale less than 0.4. Item discriminant validity was good for all scales, with low correlations between items within a scale with all other scales. Scaling success rates were 100% for all scales.

Table 40. Results of item scaling tests for Māori

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.35-0.81	0.01-0.44	120/120	100%
Social functioning (WSF)	5	0.63-0.8	0.04-0.49	50/50	100%
Vision (V)	3	0.73-0.85	0.03-0.19	30/30	100%
Hearing (H)	3	0.77-0.88	0-0.21	30/30	100%
Digestion & bodily excretions (DB)	4	0.53-0.65	0.07-0.33	40/40	100%
Breathing (BR)	3	0.74-0.83	0.1-0.41	30/30	100%
Self-care (SC)	6	0.59-0.8	-0.01-0.38	60/60	100%
Understanding (US)	4	0.73-0.81	0.07-0.47	40/40	100%
Communication (CM)	3	0.75-0.85	0.03-0.51	30/30	100%
Sleep (S)	3	0.76-0.82	0.1-0.34	30/30	100%

Table 41. Results of item scaling tests for Pacific peoples

Scale	No. of items	Range of item correlations		Item scaling tests	
		Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.58-0.85	0.02-0.61	120/120	100%
Social functioning (WSF)	5	0.72-0.81	0.06-0.54	50/50	100%
Vision (V)	3	0.67-0.83	0.01-0.23	30/30	100%
Hearing (H)	3	0.89-0.9	-0.02-0.29	30/30	100%
Digestion & bodily excretions (DB)	4	0.63-0.79	0.03-0.35	40/40	100%
Breathing (BR)	3	0.77-0.87	-0.01-0.53	30/30	100%
Self-care (SC)	6	0.77-0.94	-0.03-0.52	60/60	100%
Understanding (US)	4	0.71-0.78	0.09-0.55	40/40	100%
Communication (CM)	3	0.84-0.89	0.09-0.58	30/30	100%
Sleep (S)	3	0.69-0.82	0-0.33	30/30	100%

Table 42. Results of item scaling tests for Asian peoples

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.35-0.85	0.03-0.86	120/120	100%
Social functioning (WSF)	5	0.51-0.73	0-0.52	50/50	100%
Vision (V)	3	0.53-0.74	0.03-0.31	30/30	100%
Hearing (H)	3	0.78-0.87	0.05-0.34	30/30	100%
Digestion & bodily excretions (DB)	4	0.48-0.7	0.02-0.31	40/40	100%
Breathing (BR)	3	0.78-0.86	0.06-0.49	30/30	100%
Self-care (SC)	6	0.14-0.95	0.02-0.65	60/60	100%
Understanding (US)	4	0.71-0.77	0.09-0.45	40/40	100%
Communication (CM)	3	0.82-0.86	0.04-0.46	30/30	100%
Sleep (S)	3	0.79-0.83	0.01-0.35	30/30	100%

Table 43. Results of item scaling tests for European/Other

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.45-0.82	0.1-0.45	120/120	100%
Social functioning (WSF)	5	0.59-0.78	0.05-0.46	50/50	100%
Vision (V)	3	0.76-0.83	0.07-0.21	30/30	100%
Hearing (H)	3	0.74-0.9	0.09-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.44-0.64	0.07-0.28	40/40	100%
Breathing (BR)	3	0.68-0.84	0.1-0.43	30/30	100%
Self-care (SC)	6	0.62-0.82	0.07-0.46	60/60	100%
Understanding (US)	4	0.72-0.76	0.13-0.48	40/40	100%
Communication (CM)	3	0.76-0.8	0.09-0.5	30/30	100%
Sleep (S)	3	0.73-0.81	0.1-0.27	30/30	100%

Deprivation quintile

Results of scaling tests and reliability estimates are presented in Table 44 to Table 48 for Quintiles 1 to Quintile 5. Internal consistency was high with correlations (after correction for overlap) greater than 0.4 for all except one scale and deprivation quintile. For Quintile 1, there were correlations between an item and the DB scale less than 0.4. Item discriminant validity was good for all scales, with low correlations between items within a scale with all other scales. Scaling success rates were 100% for all scales.

Table 44. Results of item scaling tests for Deprivation Quintile 1

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.41-0.78	0.06-0.51	120/120	100%
Social functioning (WSF)	5	0.6-0.76	0.06-0.47	50/50	100%
Vision (V)	3	0.62-0.8	0.01-0.21	30/30	100%
Hearing (H)	3	0.7-0.89	0.07-0.23	30/30	100%
Digestion & bodily excretions (DB)	4	0.34-0.66	-0.01-0.31	40/40	100%
Breathing (BR)	3	0.64-0.81	0.02-0.37	30/30	100%
Self-care (SC)	6	0.46-0.85	0.03-0.47	60/60	100%
Understanding (US)	4	0.67-0.76	0.04-0.45	40/40	100%
Communication (CM)	3	0.76-0.81	0.1-0.51	30/30	100%
Sleep (S)	3	0.72-0.81	0.06-0.3	30/30	100%

Table 45. Results of item scaling tests for Deprivation Quintile 2

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.46-0.84	0.09-0.52	120/120	100%
Social functioning (WSF)	5	0.57-0.77	0-0.46	50/50	100%
Vision (V)	3	0.69-0.78	0.03-0.25	30/30	100%
Hearing (H)	3	0.71-0.9	0.03-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.47-0.64	0.06-0.31	40/40	100%
Breathing (BR)	3	0.68-0.83	0.11-0.48	30/30	100%
Self-care (SC)	6	0.55-0.83	0.03-0.48	60/60	100%
Understanding (US)	4	0.75-0.82	0.13-0.52	40/40	100%
Communication (CM)	3	0.72-0.82	0.05-0.53	30/30	100%
Sleep (S)	3	0.72-0.81	0.08-0.29	30/30	100%

Table 46. Results of item scaling tests for Deprivation Quintile 3

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.45-0.82	0.09-0.42	120/120	100%
Social functioning (WSF)	5	0.56-0.79	0.05-0.4	50/50	100%
Vision (V)	3	0.73-0.81	0.05-0.2	30/30	100%
Hearing (H)	3	0.76-0.9	0.06-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.4-0.63	0.02-0.23	40/40	100%
Breathing (BR)	3	0.64-0.85	0.02-0.44	30/30	100%
Self-care (SC)	6	0.63-0.82	0.04-0.44	60/60	100%
Understanding (US)	4	0.72-0.75	0.08-0.46	40/40	100%
Communication (CM)	3	0.78-0.79	0.08-0.48	30/30	100%
Sleep (S)	3	0.75-0.8	0.08-0.29	30/30	100%

Table 47. Results of item scaling tests for Deprivation Quintile 4

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.45-0.82	0.06-0.48	120/120	100%
Social functioning (WSF)	5	0.61-0.77	0.02-0.42	50/50	100%
Vision (V)	3	0.82-0.87	0.1-0.26	30/30	100%
Hearing (H)	3	0.72-0.89	0.1-0.28	30/30	100%
Digestion & bodily excretions (DB)	4	0.54-0.62	0.06-0.27	40/40	100%
Breathing (BR)	3	0.73-0.83	0.05-0.37	30/30	100%
Self-care (SC)	6	0.63-0.81	0.01-0.47	60/60	100%
Understanding (US)	4	0.72-0.77	0.11-0.45	40/40	100%
Communication (CM)	3	0.8-0.84	0.06-0.51	30/30	100%
Sleep (S)	3	0.73-0.82	0.08-0.28	30/30	100%

Table 48. Results of item scaling tests for Deprivation Quintile 5

		Range of item correlations		Item scaling tests	
Scale	No. of items	Item internal consistency	Item discriminant validity	Success/ total	Scaling success
Physical functioning (WPF)	12	0.44-0.83	0.06-0.45	120/120	100%
Social functioning (WSF)	5	0.62-0.81	0.06-0.54	50/50	100%
Vision (V)	3	0.76-0.84	0.04-0.19	30/30	100%
Hearing (H)	3	0.82-0.91	0.1-0.27	30/30	100%
Digestion & bodily excretions (DB)	4	0.47-0.67	0.05-0.3	40/40	100%
Breathing (BR)	3	0.72-0.85	0.11-0.43	30/30	100%
Self-care (SC)	6	0.57-0.85	0-0.5	60/60	100%
Understanding (US)	4	0.74-0.76	0.1-0.49	40/40	100%
Communication (CM)	3	0.75-0.82	0.09-0.5	30/30	100%
Sleep (S)	3	0.76-0.82	0.09-0.31	30/30	100%

Reliability of scale scores

The internal consistency reliability of each scale and population subgroup was summarised with Cronbach's alpha coefficient. The recommended level for Cronbach's alpha for between group comparisons is 0.70 and between individual comparisons in 0.90.

Total population

The internal consistency reliability estimates are presented for the total population for all scales in Table 49. All scales except for the DB scale exceeded the 0.70 level recommended for between group comparisons. Cronbach's alpha for the scale was 0.48. The scale with the highest reliability was the WPF, with a Cronbach's alpha of 0.92. This was the only scale that met the recommended standard for individual comparisons of 0.90.

Table 49. Reliability estimates for scales (Cronbach's alpha) for the total population

Scale	Scale reliability
Physical functioning (WPF)	0.92
Social functioning (WSF)	0.77
Vision (V)	0.74
Hearing (H)	0.83
Digestion & bodily excretions (DB)	0.48
Breathing (BR)	0.71
Self-care (SC)	0.83
Understanding (US)	0.77
Communication (CM)	0.74
Sleep (S)	0.71

Sex

Reliability estimates by sex for all scales are presented in Table 50. Cronbach's alpha was higher than 0.7 for all except two scales for males and one for females. The scales under this criterion for males were DB (0.53) and BR (0.66). For females the one scale under the criterion was DB (0.45). One scale, WPF, exceeded the recommended criterion of 0.9 for between individual comparisons for both males and females.

Table 50. Reliability estimates for scales (Cronbach's alpha) by sex

	Scale reliability	
Scale	Males	Females
Physical functioning (WPF)	0.92	0.92
Social functioning (WSF)	0.77	0.77
Vision (V)	0.78	0.71
Hearing (H)	0.83	0.84
Digestion & bodily excretions (DB)	0.53	0.45
Breathing (BR)	0.66	0.74
Self-care (SC)	0.84	0.83
Understanding (US)	0.77	0.77
Communication (CM)	0.74	0.73
Sleep (S)	0.69	0.72

Age Group

Reliability estimates by age group for all scales are presented in Table 51. Cronbach's alpha was higher than 0.7 for all scales for 15-24 and 24-44 year olds. For the other three age groups there were scales for which this recommended level was not satisfied. The scales under this criterion for 45-64 year olds were DB, V and CM; for 65-74 years olds were DB, CM and S and for 75+ year olds were DB, BR, CM and S. One scale, WPF, exceeded the recommended criterion of 0.9 for between individual comparisons for all except the youngest age group, 15-24 year olds.

Table 51. Reliability estimates for scales (Cronbach's alpha) by age group

Scale	Scale reliability				
	15-24 years	25-44 years	45-64 years	65-74 years	75+ years
Physical functioning (WPF)	0.89	0.91	0.91	0.90	0.90
Social functioning (WSF)	0.75	0.8	0.76	0.71	0.68
Vision (V)	0.69	0.73	0.59	0.78	0.88
Hearing (H)	0.79	0.83	0.81	0.84	0.84
Digestion & bodily excretions (DB)	0.51	0.5	0.47	0.47	0.38
Breathing (BR)	0.73	0.72	0.72	0.72	0.66
Self-care (SC)	0.80	0.88	0.83	0.84	0.80
Understanding (US)	0.75	0.79	0.75	0.76	0.81
Communication (CM)	0.75	0.77	0.69	0.69	0.69
Sleep (S)	0.71	0.72	0.73	0.68	0.67

Ethnic Group

Reliability estimates by ethnic group for all scales are presented in Table 52. Cronbach's alpha was not higher than 0.7 for all scales for any ethnic group. The scales with Cronbach's alpha less than this recommended level were the following for Māori, Pacific peoples, Asian and European/Other ethnic groups respectively: DB; V; WSF, V and DB; and DB and BR. The lowest Cronbach's alpha was for the DB scale (0.44) for European/Other ethnic group. One scale, WPF, exceeded the recommended criterion of 0.9 for between individual comparisons for all ethnic groups. One further scale, the SC scale, exceeded this recommended level for Pacific peoples only.

Table 52. Reliability estimates for scales (Cronbach's alpha) by ethnic group

Scale	Scale reliability			
	Māori	Pacific	Asian	European /Other
Physical functioning (WPF)	0.90	0.94	0.92	0.92
Social functioning (WSF)	0.79	0.85	0.68	0.76
Vision (V)	0.73	0.65	0.51	0.75
Hearing (H)	0.82	0.88	0.82	0.83
Digestion & bodily excretions (DB)	0.54	0.71	0.55	0.44
Breathing (BR)	0.74	0.81	0.81	0.68
Self-care (SC)	0.84	0.94	0.8	0.83
Understanding (US)	0.80	0.78	0.76	0.76
Communication (CM)	0.75	0.84	0.82	0.72
Sleep (S)	0.72	0.70	0.78	0.71

Deprivation Quintile

Reliability estimates by deprivation quintile for all scales are presented in Table 53. Cronbach's alpha was not higher than 0.7 for all scales for any deprivation quintile. The scales with Cronbach's alpha less than this recommended level were the following for Quintiles to 5 respectively: V, DB, BR and S; V, DB, BR and S; DB and BR; DB; and DB. The lowest Cronbach's alpha was for the DB scale (0.35) for Quintile 1. One scale, WPF, exceeded the recommended criterion of 0.9 for between individual comparisons for all deprivation quintiles.

Table 53. Reliability estimates for scales (Cronbach's alpha) by deprivation quintile

Scale	Scale reliability				
	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Physical functioning (WPF)	0.90	0.92	0.91	0.92	0.92
Social functioning (WSF)	0.77	0.74	0.73	0.78	0.79
Vision (V)	0.66	0.66	0.70	0.82	0.73
Hearing (H)	0.79	0.81	0.84	0.81	0.87
Digestion & bodily excretions (DB)	0.35	0.46	0.42	0.49	0.54
Breathing (BR)	0.66	0.67	0.64	0.74	0.74
Self-care (SC)	0.79	0.85	0.85	0.82	0.83
Understanding (US)	0.74	0.80	0.75	0.77	0.78
Communication (CM)	0.73	0.71	0.73	0.77	0.74
Sleep (S)	0.69	0.69	0.71	0.72	0.73

Comparative analysis for physical functioning scales

The composite WHO physical functioning scale was composed of the SF-36 physical functioning scale (10 items) plus two additional WHO questions. The scale descriptive statistics, reliability estimates and correlation between the two scales is given in Table 54. The means and standard deviations for the two scales were very similar. Both scales had negatively skewed distributions, with the medians being higher than the means, and exhibited ceiling effects, with similar percentages of respondents scoring at the maximum possible scale score. The internal consistency reliability estimate for both scales was equal and exceeded the required level for between group and individual comparisons. The correlation (0.99) between the two scales was very high.

Table 54. Comparison of SF-36 and Composite physical functioning scales in 2002/03 NZHS for total population

	SF-36 Physical functioning scale	Composite WHO Physical functioning scale
No. of items	10	12
Mean	88.2	89.0
SD	19.7	18.2
25th percentile	85	87.5
50th percentile	95	95.8
75th percentile	100	100
Cronbach's Alpha	0.92	0.92
% Ceiling	45.7	44.7
Correlation with SF-36 Physical Functioning scale	-	0.99

Comparative analysis for social functioning scales

The composite WHO social functioning scale was composed of the SF-36 social functioning scale (2 items) plus three additional WHO questions. The scale descriptive statistics, reliability estimates and correlation between the two scales is given in Table 55. The mean of the Composite WHO scale is higher and the standard deviation lower than the corresponding values for the SF-36 scale. Both scales had negatively skewed distributions, with the medians being higher than the means and medians equal to 100, the maximum possible value. Both scales exhibited ceiling effects, with the SF-36 scale (73.8%) having a greater proportion of respondents scoring at the maximum scale score. The internal consistency reliability of the Composite WHO scale was higher than that of the SF-36 scale. The correlation between the two scales was 0.91.

Table 55. Comparison of SF-36 and Composite social functioning scales in 2002/03 NZHS for total population

	SF-36 social functioning scale	Composite social functioning scale
No. of items	2	5
Mean	90.5	94.3
SD	19.5	11.6
25th percentile	90	95
50th percentile	100	100
75th percentile	100	100
Cronbach's Alpha	0.68	0.77
% Ceiling	73.8	68.8
Correlation with SF-36 Social Functioning Scale	-	0.91

Validity analysis

In the next section of this chapter I present the results of the validity analysis for the scales. First, construct validity is examined, using scaling success rates and the interrelationships between scales. This is followed by discussion of the discriminative validity of the scales.

Construct validity

Scaling success rates

All items had significantly stronger associations with their own scale than any scale, measured by the scaling success rate. The scaling success rates, which directly reflect the construct validity of the WHO Long Form scales, were consistently 100% for all scales for the total population and all population subgroups.

Relationships between scales

Total population

For the total population, the reliability coefficients and inter-scale correlations for all scales are presented in Table 56. The Cronbach's alpha reliability coefficient for all scales was greater than all correlations between each scale and all other scales. Thus, there is evidence

that each of the scales has measured unique reliable variance, representing distinct concepts. The inter-scale correlations ranged from 0.14 to 0.56.

The highest inter-scale correlation was 0.56 between the US and CM scales. The next highest inter-scale correlation was 0.50 between the WPF and SC. Physical functioning is a health domain and self-care is a health-related domain. Both of the pairs of scales measure related concepts which impact on each other, and it would be expected that the correlation between them would be moderately high.

Table 56. Reliability coefficients and inter-scale correlations for total population

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.35	0.25	0.27	0.33	0.4	0.5	0.28	0.19	0.27
WSF		(0.77)	0.17	0.16	0.26	0.27	0.29	0.43	0.44	0.3
V			(0.74)	0.2	0.19	0.17	0.16	0.2	0.15	0.15
H				(0.83)	0.16	0.18	0.14	0.24	0.21	0.15
DB					(0.48)	0.28	0.19	0.26	0.17	0.29
BR						(0.71)	0.22	0.25	0.2	0.26
SC							(0.83)	0.25	0.19	0.14
US								(0.77)	0.56	0.28
CM									(0.74)	0.22
S										(0.71)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Sex

The reliability coefficients and inter-scale correlations for all scales by sex are presented in Table 57 and Table 58. The Cronbach's alpha reliability coefficient for both sexes on all scales was greater than all correlations between each scale and all other scales. For both sexes, the inter-scale correlations ranged from 0.09 to 0.58.

Similar patterns were observed as for the total population. For males and females, the highest inter-scale correlation was between the US and CM scales (0.58 and 0.55 respectively for males and females). The next highest inter-scale correlations were between the WPF and SC (0.51 and 0.50 respectively for males and females).

Table 57. Reliability coefficients and inter-scale correlations for males

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.35	0.23	0.28	0.31	0.38	0.51	0.29	0.2	0.26
WSF		(0.77)	0.16	0.16	0.27	0.25	0.26	0.42	0.41	0.29
V			(0.78)	0.17	0.22	0.16	0.12	0.19	0.16	0.17
H				(0.83)	0.19	0.23	0.09	0.22	0.18	0.15
DB					(0.53)	0.22	0.16	0.27	0.19	0.27
BR						(0.66)	0.18	0.25	0.2	0.26
SC							(0.84)	0.27	0.22	0.15
US								(0.77)	0.58	0.29
CM									(0.74)	0.21
S										(0.69)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 58. Reliability coefficients and inter-scale correlations for females

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.35	0.25	0.29	0.33	0.4	0.5	0.27	0.19	0.27
WSF		(0.77)	0.17	0.18	0.25	0.28	0.31	0.45	0.47	0.31
V			(0.71)	0.23	0.16	0.18	0.19	0.2	0.13	0.13
H				(0.84)	0.15	0.16	0.19	0.25	0.23	0.16
DB					(0.45)	0.3	0.22	0.25	0.16	0.29
BR						(0.74)	0.25	0.25	0.21	0.26
SC							(0.83)	0.24	0.16	0.14
US								(0.77)	0.55	0.27
CM									(0.73)	0.23
S										(0.72)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Age Group

The reliability coefficients and inter-scale correlations for all scales by age group are presented in Table 59 to Table 60. For all age groups on all scales the Cronbach's alpha reliability coefficient for all scales was greater than all correlations between each scale and all other scales. Across all age groups, the inter-scale correlations ranged from 0.05-0.58.

For all age groups, the highest inter-scale correlation was between the US and CM scales (0.53, 0.59, 0.57, 0.51 and 0.58 respectively for those aged 15-24, 25-44, 45-64, 65-74, 75+

years). For all age groups, there were moderately high inter-scale correlations was between the WPF and SC scales (0.50, 0.52, 0.45, 0.51 and 0.52 respectively for those aged 15-24, 25-44, 45-64, 65-74, 75+ years)

There were patterns in the inter-scale correlations that were moderately high by age group. For the youngest two age groups, there were moderate associations between the WSF and CM scales, of 0.51 and 0.46 respectively, yet these associations were lower for the older age groups. In the oldest three age groups there were moderately high inter-scale correlations between WPF and BR of 0.43, 0.46 and 0.47 respectively, with this association not present for the two younger age groups.

Table 59. Reliability coefficients and inter-scale correlations for 15-24 years

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.89)	0.35	0.22	0.18	0.22	0.27	0.5	0.31	0.32	0.2
WSF		(0.75)	0.16	0.18	0.27	0.18	0.17	0.37	0.51	0.32
V			(0.69)	0.19	0.15	0.14	0.07	0.24	0.17	0.15
H				(0.79)	0.16	0.12	0.03	0.31	0.23	0.19
DB					(0.51)	0.27	0.05	0.26	0.18	0.21
BR						(0.73)	0.06	0.21	0.17	0.29
SC							(0.8)	0.17	0.23	0.05
US								(0.75)	0.53	0.28
CM									(0.75)	0.29
S										(0.71)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 60. Reliability coefficients and inter-scale correlations for 25-44 years

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.91)	0.39	0.14	0.14	0.32	0.36	0.52	0.32	0.22	0.29
WSF		(0.8)	0.15	0.18	0.28	0.28	0.3	0.48	0.46	0.33
V			(0.73)	0.11	0.16	0.14	0.08	0.17	0.16	0.16
H				(0.83)	0.09	0.14	0.08	0.2	0.25	0.11
DB					(0.5)	0.29	0.23	0.29	0.19	0.3
BR						(0.72)	0.15	0.28	0.25	0.24
SC							(0.88)	0.26	0.14	0.18
US								(0.79)	0.59	0.33
CM									(0.77)	0.26
S										(0.72)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 61. Reliability coefficients and inter-scale correlations for 45-64 years

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.91)	0.39	0.21	0.18	0.31	0.43	0.45	0.28	0.25	0.31
WSF		(0.76)	0.22	0.17	0.24	0.27	0.24	0.44	0.38	0.29
V			(0.59)	0.19	0.15	0.19	0.1	0.24	0.19	0.18
H				(0.81)	0.12	0.18	0.06	0.28	0.26	0.14
DB					(0.47)	0.28	0.14	0.25	0.21	0.29
BR						(0.72)	0.22	0.23	0.21	0.3
SC							(0.83)	0.15	0.21	0.14
US								(0.75)	0.57	0.28
CM									(0.69)	0.19
S										(0.73)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 62. Reliability coefficients and inter-scale correlations for 65-74 years

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.9)	0.47	0.19	0.13	0.28	0.46	0.51	0.3	0.22	0.29
WSF		(0.71)	0.17	0.14	0.21	0.32	0.41	0.32	0.3	0.23
V			(0.78)	0.1	0.17	0.19	0.17	0.16	0.25	0.09
H				(0.84)	0.12	0.11	0.15	0.17	0.23	0.08
DB					(0.47)	0.23	0.23	0.2	0.14	0.25
BR						(0.72)	0.33	0.26	0.13	0.23
SC							(0.84)	0.36	0.29	0.18
US								(0.76)	0.51	0.2
CM									(0.69)	0.12
S										(0.68)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 63. Reliability coefficients and inter-scale correlations for 75+ years

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.9)	0.48	0.23	0.2	0.29	0.47	0.52	0.34	0.27	0.2
WSF		(0.68)	0.17	0.21	0.33	0.32	0.48	0.45	0.37	0.25
V			(0.88)	0.18	0.22	0.14	0.24	0.18	0.09	0.08
H				(0.84)	0.18	0.19	0.16	0.23	0.23	0.2
DB					(0.38)	0.21	0.16	0.23	0.1	0.31
BR						(0.66)	0.3	0.24	0.2	0.18
SC							(0.8)	0.4	0.3	0.12
US								(0.81)	0.58	0.19
CM									(0.69)	0.1
S										(0.67)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Ethnic Group

The reliability coefficients and inter-scale correlations for all scales by ethnic group are presented in Table 64 to Table 67. For all ethnic groups and all scales the Cronbach's alpha reliability coefficient was greater than all correlations between each scale and all other scales. Thus there is evidence that each of the scales has measured unique reliable variance, representing distinct concepts. Across all age groups, the inter-scale correlations ranged from 0.01-0.72.

For Māori, Pacific peoples and European/Other, the highest inter-scale correlation was between the US and CM scales (0.57, 0.61, and 0.56 respectively). For the Asian ethnic

group, the highest inter-scale correlation (0.72) was between the WPF and SC scales. For the remaining three ethnic groups there was moderately high inter-scale correlations was between the WPF and SC scales (0.43, 0.46 and 0.50 respectively for Māori, Pacific peoples and European/Other).

There were patterns in the inter-scale correlations that were moderately high by ethnic group. For Māori, Pacific peoples and Asian, there were moderate associations between the WPF and BR scales, of 0.42, 0.50 and 0.51 respectively, but this association was lower for European/Other.

Table 64. Reliability coefficients and inter-scale correlations for Māori

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.9)	0.34	0.2	0.17	0.37	0.42	0.43	0.21	0.16	0.27
WSF		(0.79)	0.17	0.2	0.3	0.26	0.19	0.46	0.47	0.38
V			(0.73)	0.17	0.14	0.21	0.1	0.18	0.1	0.18
H				(0.82)	0.16	0.19	0.01	0.23	0.22	0.16
DB					(0.54)	0.33	0.27	0.23	0.15	0.34
BR						(0.74)	0.15	0.21	0.17	0.31
SC							(0.84)	0.16	0.08	0.13
US								(0.8)	0.57	0.28
CM									(0.75)	0.25
S										(0.72)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 65. Reliability coefficients and inter-scale correlations for Pacific peoples

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.94)	0.47	0.24	0.08	0.41	0.5	0.46	0.31	0.27	0.25
WSF		(0.85)	0.2	0.2	0.33	0.29	0.2	0.51	0.47	0.33
V			(0.65)	0.15	0.26	0.18	0.05	0.2	0.17	0.17
H				(0.88)	0.11	-0.01	-0.01	0.18	0.24	0.02
DB					(0.71)	0.31	0.21	0.36	0.19	0.35
BR						(0.81)	0.22	0.27	0.16	0.34
SC							(0.94)	0.21	0.23	0.18
US								(0.78)	0.61	0.3
CM									(0.84)	0.2
S										(0.7)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 66. Reliability coefficients and inter-scale correlations for Asian peoples

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.42	0.3	0.34	0.31	0.51	0.72	0.28	0.18	0.31
WSF		(0.68)	0.23	0.26	0.18	0.33	0.42	0.44	0.41	0.3
V			(0.51)	0.1	0.19	0.24	0.21	0.26	0.2	0.17
H				(0.82)	0.1	0.18	0.24	0.18	0.16	0.08
DB					(0.55)	0.28	0.19	0.24	0.11	0.28
BR						(0.81)	0.52	0.31	0.14	0.39
SC							(0.8)	0.33	0.18	0.23
US								(0.76)	0.47	0.36
CM									(0.82)	0.19
S										(0.78)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 67. Reliability coefficients and inter-scale correlations for European/Other

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.35	0.25	0.29	0.32	0.39	0.5	0.29	0.19	0.27
WSF		(0.76)	0.16	0.15	0.25	0.26	0.31	0.42	0.42	0.28
V			(0.75)	0.2	0.19	0.16	0.18	0.19	0.15	0.14
H				(0.83)	0.17	0.19	0.16	0.24	0.21	0.14
DB					(0.44)	0.26	0.18	0.25	0.17	0.27
BR						(0.68)	0.22	0.25	0.21	0.24
SC							(0.83)	0.27	0.21	0.14
US								(0.76)	0.56	0.27
CM									(0.72)	0.22
S										(0.71)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Deprivation Quintile

The reliability coefficients and inter-scale correlations for all scales by age group are presented in Table 68 to Table 72. In all deprivation quintiles on all scales the Cronbach's alpha reliability coefficient scales was greater than all correlations between each scale and all other scales. Thus there is evidence that each of the scales has measured unique reliable variance, representing distinct concepts. Across all age groups, the inter-scale correlations ranged from 0.06-0.61.

For all deprivation quintiles, the highest inter-scale correlation was between the US and CM scales (0.55, 0.61, 0.56, 0.55 and 0.56 respectively for those in quintiles 1 to 5). For all age

groups, there was moderately high inter-scale correlations was between the WPF and SC scales (0.50, 0.53, 0.46, 0.50 and 0.50 respectively for those in quintiles 1 to 5).

Table 68. Reliability coefficients and inter-scale correlations for Quintile 1

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.9)	0.3	0.21	0.24	0.28	0.35	0.5	0.22	0.15	0.23
WSF		(0.77)	0.22	0.12	0.16	0.22	0.27	0.45	0.44	0.3
V			(0.66)	0.21	0.14	0.17	0.13	0.25	0.19	0.14
H				(0.79)	0.14	0.12	0.12	0.21	0.2	0.16
DB					(0.35)	0.24	0.09	0.25	0.17	0.22
BR						(0.66)	0.19	0.21	0.18	0.21
SC							(0.79)	0.1	0.16	0.09
US								(0.74)	0.55	0.29
CM									(0.73)	0.24
S										(0.69)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 69. Reliability coefficients and inter-scale correlations for Quintile 2

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.36	0.28	0.27	0.35	0.44	0.53	0.34	0.19	0.28
WSF		(0.74)	0.17	0.13	0.27	0.31	0.32	0.45	0.45	0.28
V			(0.66)	0.21	0.21	0.23	0.19	0.22	0.15	0.17
H				(0.81)	0.14	0.22	0.15	0.25	0.26	0.12
DB					(0.46)	0.28	0.27	0.29	0.19	0.27
BR						(0.67)	0.3	0.29	0.22	0.26
SC							(0.85)	0.37	0.25	0.19
US								(0.8)	0.61	0.31
CM									(0.71)	0.22
S										(0.69)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 70. Reliability coefficients and inter-scale correlations for Quintile 3

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.91)	0.37	0.23	0.29	0.3	0.4	0.46	0.3	0.18	0.29
WSF		(0.73)	0.16	0.17	0.21	0.22	0.32	0.47	0.42	0.33
V			(0.7)	0.16	0.11	0.1	0.14	0.18	0.15	0.14
H				(0.84)	0.15	0.21	0.11	0.21	0.21	0.13
DB					(0.42)	0.16	0.18	0.19	0.13	0.27
BR						(0.64)	0.11	0.22	0.24	0.25
SC							(0.85)	0.27	0.19	0.16
US								(0.75)	0.56	0.27
CM									(0.73)	0.22
S										(0.71)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 71. Reliability coefficients and inter-scale correlations for Quintile 4

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.34	0.29	0.27	0.33	0.35	0.5	0.27	0.16	0.26
WSF		(0.78)	0.16	0.18	0.29	0.21	0.26	0.38	0.36	0.27
V			(0.82)	0.21	0.22	0.16	0.24	0.19	0.13	0.13
H				(0.81)	0.18	0.13	0.18	0.27	0.18	0.15
DB					(0.49)	0.29	0.23	0.29	0.16	0.29
BR						(0.74)	0.25	0.23	0.15	0.26
SC							(0.82)	0.22	0.14	0.14
US								(0.77)	0.55	0.26
CM									(0.77)	0.22
S										(0.72)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Table 72. Reliability coefficients and inter-scale correlations for Quintile 5

Scale	WPF	WSF	V	H	DB	BR	SC	US	CM	S
WPF	(0.92)	0.36	0.18	0.27	0.34	0.41	0.5	0.25	0.2	0.25
WSF		(0.79)	0.15	0.18	0.29	0.32	0.27	0.43	0.49	0.32
V			(0.73)	0.19	0.19	0.17	0.06	0.17	0.14	0.15
H				(0.87)	0.17	0.19	0.12	0.23	0.2	0.16
DB					(0.54)	0.32	0.14	0.25	0.16	0.32
BR						(0.74)	0.21	0.27	0.21	0.29
SC							(0.83)	0.23	0.19	0.12
US								(0.78)	0.56	0.27
CM									(0.74)	0.19
S										(0.73)

Note: Correlations in brackets are the Cronbach's alpha coefficient for the scale

Discriminative validity

There were 132 adults in the sample whose response to all chronic disease questions was either "Don't know", "Not specified' or "Refused". These respondents were excluded from the following analysis. For the remaining 12,397 respondents, the prevalence of chronic disease in adults in the 2002/03 NZHS was 57.7% (7,086).

The adults in the chronic disease group rated their health consistently lower than those without a chronic disease for all scales. Table 73 presents the means and standard deviations for those in the chronic disease and no chronic disease group, and the effect sizes comparing the two groups across the scales. The effect size was calculated to compare the means of those with and without chronic diseases across all scales, relative to the standard deviation in the chronic disease group.

Table 73. Discriminative ability of the WHO Long Form (NZ Version) scales between subgroups with and without chronic diseases reported

Scale	Chronic disease Mean (SD)	No chronic disease Mean (SD)	Effect size	Mann Whitney U test p-value[1]
Physical functioning (WPF)	84.2 (21.2)	96.0 (9.0)	0.56	<.0001
Social functioning (WSF)	92.9 (12.9)	96.3 (8.9)	0.26	<.0001
Vision (V)	95.5 (12.1)	98.0 (6.8)	0.21	<.0001
Hearing (H)	91.8 (15.8)	96.0 (10.7)	0.27	<.0001
Digestion & bodily excretions (DB)	94.4 (9.3)	97.1 (6.2)	0.29	<.0001
Breathing (BR)	93.8 (13.1)	98.0 (6.6)	0.32	<.0001
Self-care (SC)	98.2 (8.3)	99.8 (2.2)	0.19	<.0001
Understanding (US)	91.3 (13.7)	94.6 (10.0)	0.24	<.0001
Communication (CM)	95.2 (10.9)	96.9 (8.2)	0.16	<.0001
Sleep (S)	76.9 (23.6)	85.4 (18.9)	0.36	<.0001

[1] Two sided Mann-Whitney U test of differences between the subgroup with chronic disease(s) compared with the subgroup with no chronic disease.

Comparing the chronic disease to the no chronic disease group, the effect sizes were small for all scales except the WPF scale, which had a medium effect size (0.56). All differences between the median scale score of those with chronic diseases compared to those without chronic diseases were highly significant ($p<.0001$).

Summary

The best and worst psychometric results for population subgroups are summarised in Table 74.

Table 74. Summary of tests of scales associated with best and worst results

Criterion (desired result)	WPF	WSF	V	H	DB	BR	SC	US	CM	S
Item completeness: % complete data (100%)										
Best (highest %)	100	100	100	100	100	100	100	100	100	100
Worst (lowest %)	100	100	100	100	100	100	100	100	100	100
Total NZ population	100	100	100	100	100	100	100	100	100	100
Item-internal consistency: Item-scale correlation (>0.4)										
Highest	0.85	0.81	0.90	0.92	0.79	0.88	0.95	0.82	0.83	0.83
Lowest	0.35	0.33	0.53	0.70	0.34	0.62	0.14	0.67	0.69	0.69
Total NZ population	0.44-0.82	0.6-0.78	0.76-0.83	0.75-0.89	0.47-0.65	0.71-0.84	0.62-0.82	0.73-0.77	0.77-0.81	0.74-0.81
Scaling success rate: % (100%)										
Best (highest rate)	100	100	100	100	100	100	100	100	100	100
Worst (lowest rate)	100	100	100	100	100	100	100	100	100	100
Total NZ population	100	100	100	100	100	100	100	100	100	100
Reliability: Cronbach's alpha (>0.7 group, >0.9 individual)										
Highest	0.94	0.85	0.82	0.88	0.55	0.81	0.94	0.81	0.84	0.78
Lowest	0.89	0.68	0.51	0.79	0.35	0.64	0.71	0.74	0.69	0.67
Total NZ population	0.92	0.77	0.74	0.83	0.48	0.71	0.83	0.77	0.74	0.71
Floor effects: % scoring 0 (<25%)										
Best (lowest %)	0	0	0	0	0	0	0	0	0	0
Worst (highest %)	0.5	0.2	1.8	0.7	0.5	0.4	0.4	0.4	0.2	1.3
Total NZ population	0.2	0	0.3	0.1	0	0.1	0.1	0.1	0	0.9
Ceiling effects: % scoring 100 (<25%)										
Best (lowest %)	3.6	62.3	72.4	39.3	42.3	61.2	81.2	47.9	56.0	22.5
Worst (highest %)	66.7	75.5	87.2	87.9	69.3	85.9	97.1	69.4	83.0	78.6
Total NZ population	44.7	68.8	80.6	71.0	61.6	76.1	94.9	56.6	77.4	30.4

Item completeness and item internal consistency

There was no missing data for any scale across all subgroups and for the total population. Item-internal consistency, measured by item-scale correlations, was high across all subgroups and scales, with only a few falling below the desired level. The scales which did not meet this criterion were WPF for 15-24 year olds, Māori and Asian subgroups, WSF for the 75+ subgroup, DB for the NZDep Q1 subgroup and SC for the Asian subgroup. Scaling success rates met the desired level of 100% across all scales and subgroups, reflecting good construct validity.

Reliability

Internal consistency reliabilities met the criterion by between group comparisons of 0.70 for all scales and subgroups except the V scale for Asian people and 45-64 year olds, the DB scale for the total NZ population, females and those age 75+ years, the BR scale for those in NZDep Q1, the CM scale for those aged over 45 years and the S scale for those aged 75 years and older. The PF scale and the SC scale met the criterion for between individual comparisons of 0.9.

Floor and ceiling effects

Floor effects were not observed for any scale or subgroup. Substantial ceiling effects were observed for all scales and subgroups except the WPF and S scales. The strongest ceiling effects were observed for the SC followed by the HR scale for Asian people. For the sleep scale the females scored at the ceiling substantially more than any other subgroup.

National Norms

Norms across the WHO Long Form (NZ Version) scales are presented for the total population and the population subgroups by sex, age group, ethnic group and deprivation quintile. A table and figure is presented for the norms for each population subgroup. Error bars in the figures represent 95% confidence intervals for the norm.

Total population

Table 75 presents the WHO Long Form (NZ Version) norms for the total population.

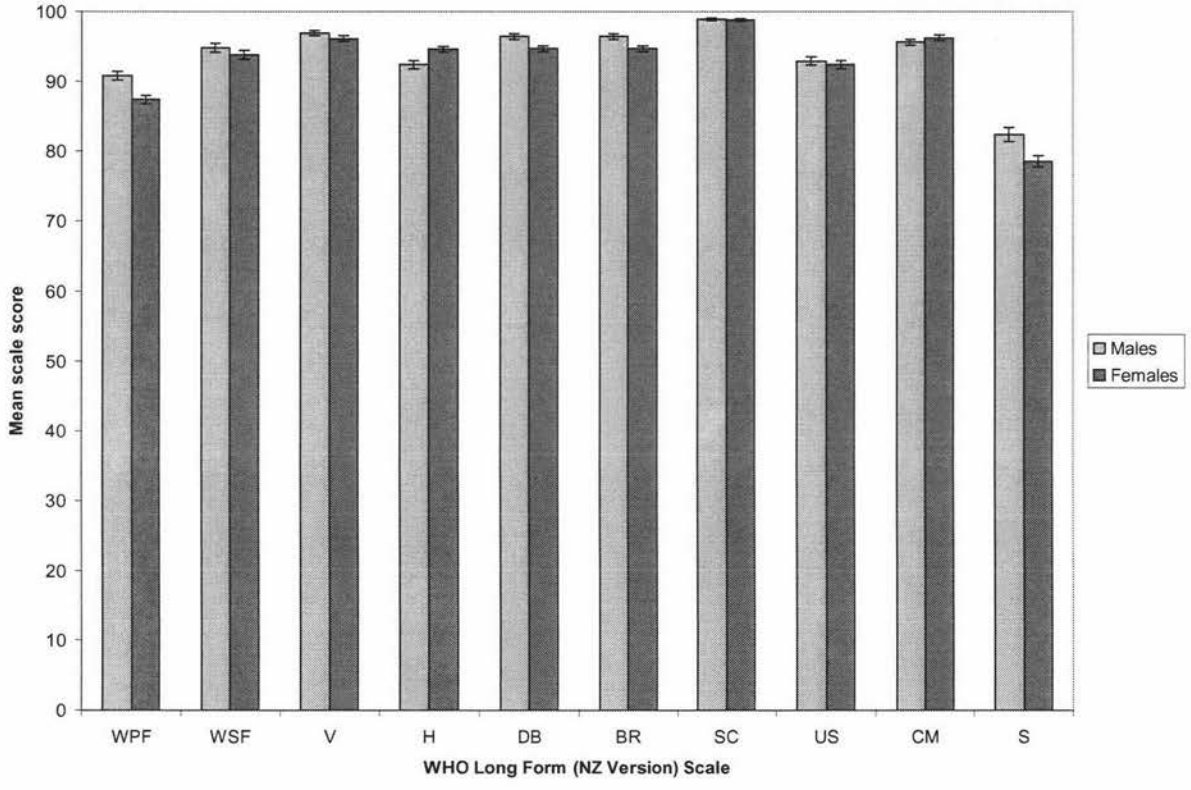
Table 75: WHO Long Form (NZ Version) norms for the total population

Scale	Mean (standard error)	Number of respondents
Physical functioning (WPF)	89 (0.2)	12529
Social functioning (WSF)	94.3 (0.2)	12527
Vision (V)	96.5 (0.1)	12529
Hearing (H)	93.5 (0.2)	12529
Digestion & bodily excretions (DB)	95.5 (0.1)	12529
Breathing (BR)	95.5 (0.1)	12529
Self-care (SC)	98.8 (0.1)	12529
Understanding (US)	92.7 (0.2)	12528
Communication (CM)	95.9 (0.2)	12527
Sleep (S)	80.5 (0.3)	12527

Sex and age group

Figure 3 presents the WHO Long Form (NZ Version) norms by sex and Figure 4 by age group. Males rated their health significantly better than females in four domains; physical functioning (WPF), digestion and bodily excretions (DB), breathing (BR) and sleep (S). Females scored significantly higher than males on one scale, hearing. The largest difference between males and females was observed for the sleep scale (S).

Figure 3: WHO Long Form (NZ Version) norms by sex



In general, the WHO Long Form (NZ Version) by age group showed decreasing self-reported health with increasing age, these differences being strongest for the physical health related scales, in particular the physical functioning scale (WPF) (Figure 4). On the communication scale (CM), scores increased with increasing age, except for the oldest age group, but not all groups were significantly different from the next highest age group. Scores on the social functioning scale (WSF) were stable across the age groups.

Figure 4. WHO Long Form (NZ Version) norms by life cycle age group

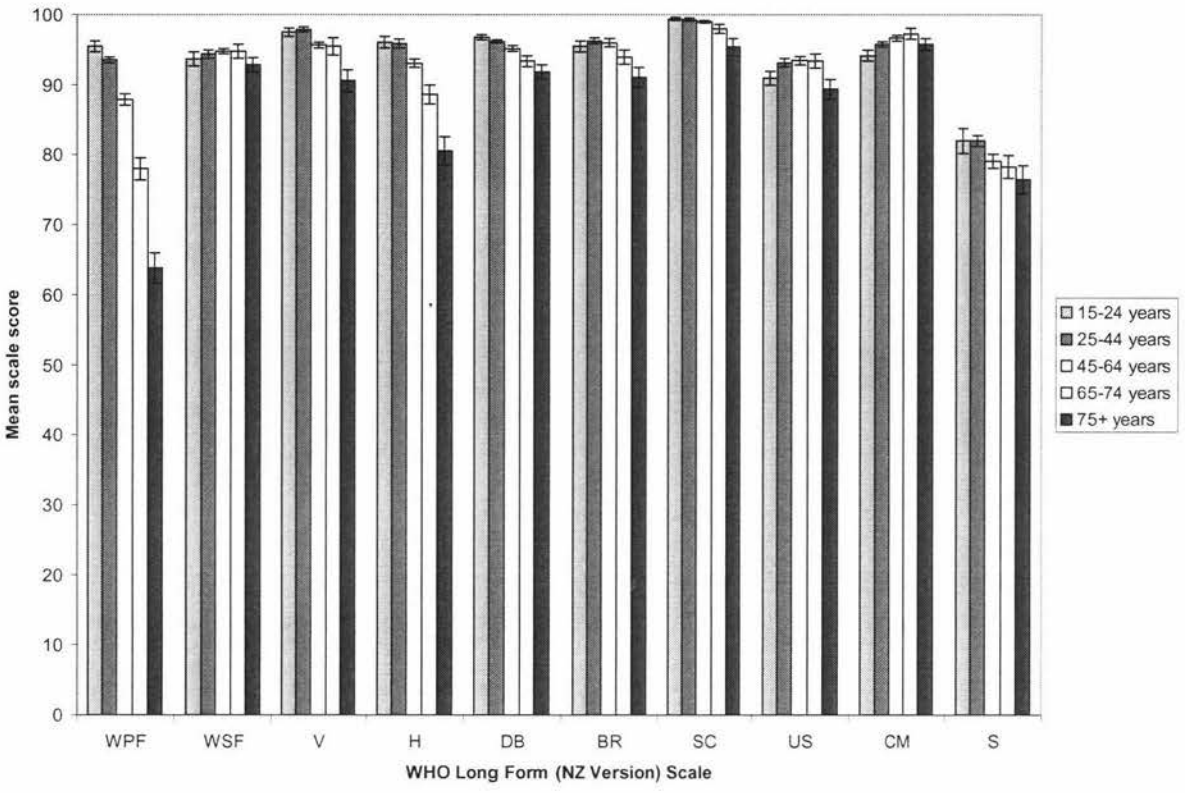


Table 76 and Table 77 present the WHO Long Form (NZ Version) norms by age group for males and females respectively.

Table 76. WHO Long Form (NZ Version) norms by age group for males (standard errors)

Age Group (Years)	WPF	WSF	V	H	DB	BR	SC	US	CM	S	Number of respondents
15-24	96.8 (0.6)	94.9 (0.7)	98.1 (0.3)	95.8 (0.5)	97.7 (0.3)	97 (0.5)	99.4 (0.2)	91.4 (0.7)	94.1 (0.6)	84.7 (1.4)	638
25-44	94.8 (0.3)	94.8 (0.4)	98.2 (0.4)	95.3 (0.5)	97.1 (0.2)	97.4 (0.2)	99.3 (0.2)	93.5 (0.4)	95.4 (0.3)	83.4 (0.6)	1776
45-64	89.4 (0.5)	95.3 (0.4)	95.8 (0.3)	91.5 (0.6)	96.1 (0.3)	96.4 (0.4)	98.9 (0.2)	93.9 (0.4)	96.7 (0.3)	81.1 (0.8)	1559
65-74	80.9 (1.2)	94.9 (0.8)	95.7 (1.0)	85.1 (1.2)	94.2 (0.6)	94.6 (0.8)	98.4 (0.4)	93.4 (0.7)	96.7 (0.7)	81.1 (1.1)	507
75+	67.1 (1.9)	92.3 (0.8)	91.6 (1.4)	77.8 (1.4)	91.9 (0.8)	91.2 (1.0)	96.1 (0.8)	88.9 (1.3)	94.2 (0.8)	78 (1.7)	389
											4869

Table 77. WHO Long Form (NZ Version) norms by age group for females (standard errors)

Age Group (Years)	WPF	WSF	V	H	DB	BR	SC	US	CM	S	Number of respondents
15-24	94.3 (0.5)	92.4 (0.6)	96.9 (0.4)	96.5 (0.5)	95.8 (0.4)	94 (0.6)	99.4 (0.2)	90.7 (0.7)	94.2 (0.6)	79.2 (1.2)	928
25-44	92.5 (0.4)	94 (0.4)	97.6 (0.2)	96.5 (0.3)	95.4 (0.2)	95.3 (0.4)	99.3 (0.1)	93 (0.4)	96.2 (0.3)	80.6 (0.6)	3259
45-64	86.4 (0.6)	94.3 (0.3)	95.6 (0.3)	94.7 (0.4)	94.4 (0.3)	95.5 (0.3)	99 (0.2)	93.1 (0.4)	96.8 (0.3)	77.1 (0.7)	2159
65-74	75.6 (1.2)	94.7 (0.6)	95.3 (0.6)	91.4 (0.8)	92.8 (0.5)	93.5 (0.6)	97.7 (0.5)	93.5 (0.7)	97.7 (0.3)	75.9 (1.1)	729
75+	61.2 (1.5)	93.4 (0.7)	89.8 (1.0)	82.8 (1.4)	91.9 (0.6)	91 (0.9)	94.9 (0.7)	89.9 (0.8)	97.1 (0.4)	75.4 (1.4)	579

Ethnic Group

Table 78 presents the WHO Long Form (NZ Version) norms by ethnic group.

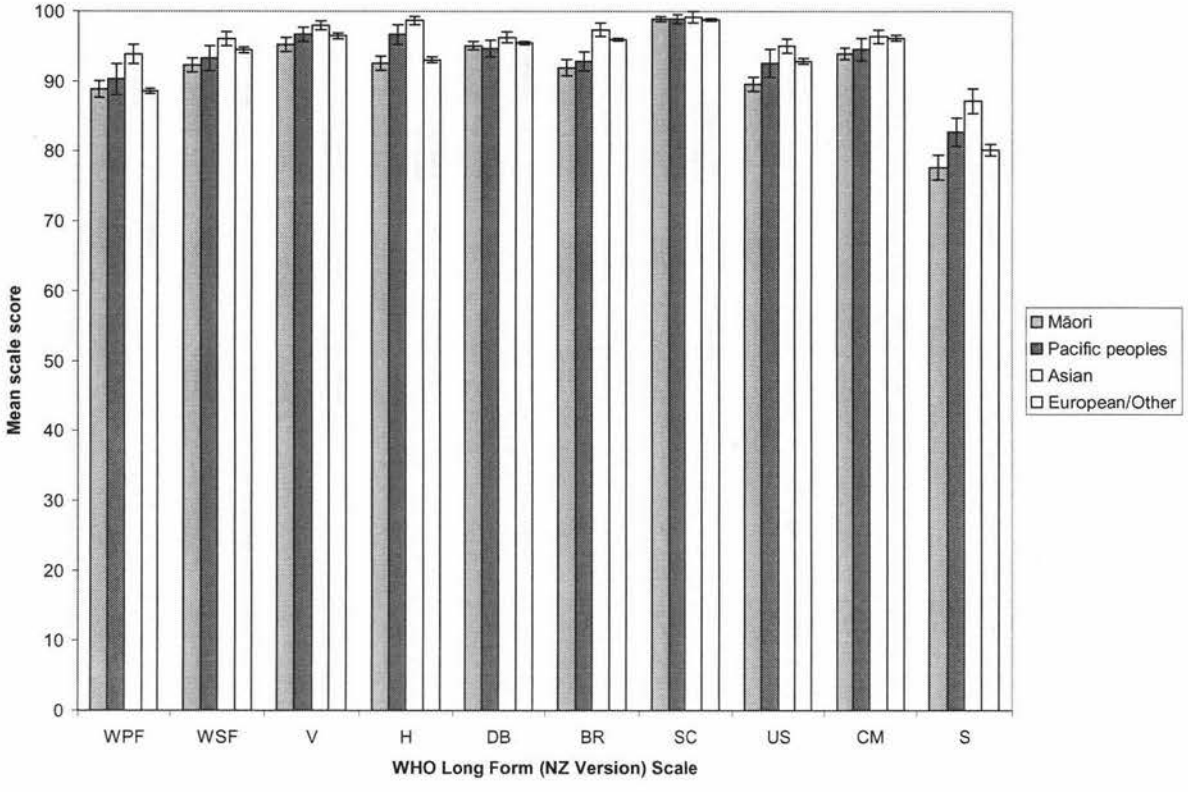
Table 78. WHO Long Form (NZ Version) norms for by ethnic group (standard errors)

	WPF	WSF	V	H	DB	BR	SC	US	CM	S	Number of respondents
Ethnic Group											
Māori	88.9	92.3	95.3	92.6	95.1	92.0	98.9	89.6	94.0	77.7	4120
	(0.6)	(0.5)	(0.5)	(0.5)	(0.3)	(0.6)	(0.2)	(0.5)	(0.4)	(0.9)	
Pacific	90.3	93.3	96.7	96.7	94.7	92.9	98.9	92.6	94.6	82.8	908
	(1.1)	(0.9)	(0.5)	(0.7)	(0.6)	(0.7)	(0.3)	(1.0)	(0.8)	(1.0)	
Asian	93.9	96.1	98.0	98.7	96.3	97.4	99.2	95.1	96.4	87.2	1172
	(0.7)	(0.5)	(0.3)	(0.3)	(0.4)	(0.5)	(0.4)	(0.5)	(0.5)	(0.9)	
European/Other	88.6	94.5	96.5	93.1	95.5	96.0	98.8	92.9	96.2	80.2	6329
	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.4)	

Asian adults scored significantly higher than European/Other adults on six of the ten scales; physical functioning (WPF), social functioning (WSF), hearing (H), breathing (BR), understanding (US), and sleep (S). Asian adults scored significantly higher than Māori and Pacific adults on the breathing (BR) and sleep (S) scales.

The European/Other ethnic group rated their health significantly higher than both Māori and Pacific peoples on one scale, breathing (BR). Compared to Māori alone, the European/Other ethnic group rated their health significantly better on the social functioning (WSF), breathing (BR), understanding (US) and communication (CM) scales. Figure 5 presents the norms by ethnic group.

Figure 5. WHO Long Form (NZ Version) norms by ethnic group



Deprivation Quintiles

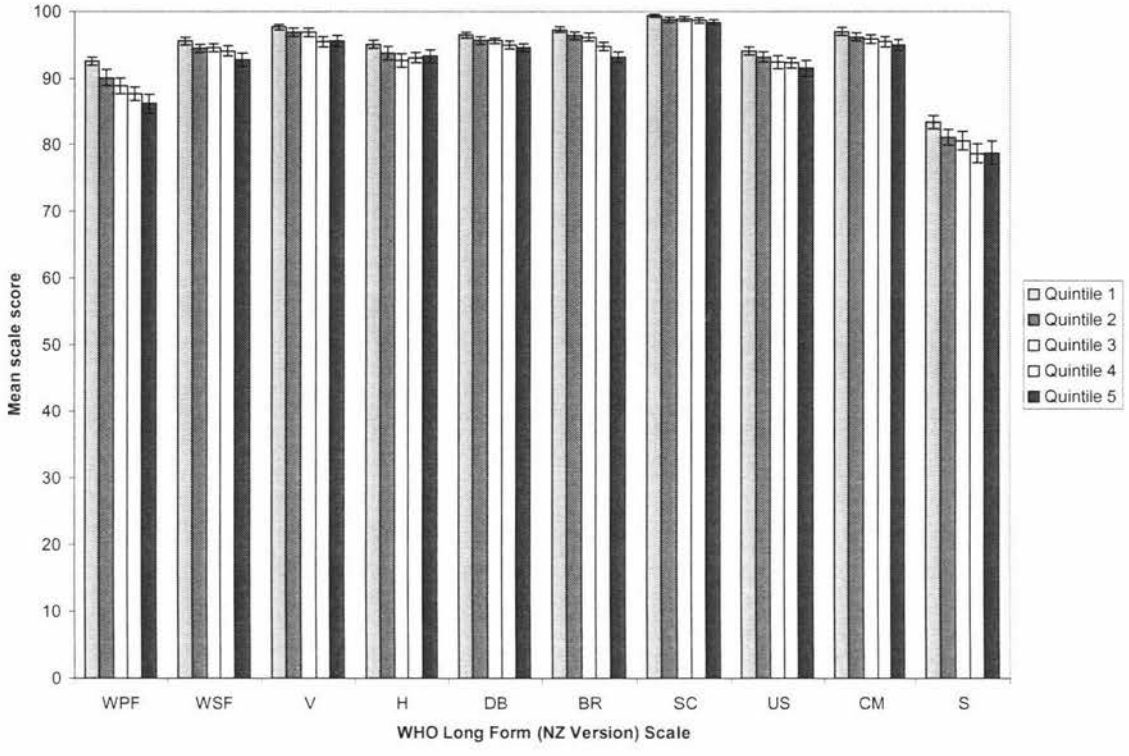
Table 79 and Figure 6 presents the WHO Long Form (NZ Version) norms by Deprivation quintile.

Table 79. WHO Long Form (NZ Version) norms for by deprivation (NZDep01) quintile (standard errors)

	WPF	WSF	V	H	DB	BR	SC	US	CM	S	Number of respondents
NZDep01 Quintile											
Quintile 1 (=least deprived)	92.6	95.6	97.7	95.1	96.5	97.3	99.4	94.1	97.0	83.4	1705
	(0.3)	(0.3)	(0.2)	(0.3)	(0.2)	(0.2)	(0.1)	(0.3)	(0.3)	(0.5)	
Quintile 2	90.1	94.5	96.9	93.8	95.7	96.4	98.8	93.2	96.2	81.1	1573
	(0.6)	(0.3)	(0.3)	(0.5)	(0.3)	(0.3)	(0.2)	(0.4)	(0.3)	(0.6)	
Quintile 3	88.9	94.6	96.9	92.7	95.7	96.2	98.9	92.4	95.9	80.6	1854
	(0.6)	(0.3)	(0.3)	(0.5)	(0.2)	(0.3)	(0.2)	(0.5)	(0.3)	(0.7)	
Quintile 4	87.7	94.1	95.5	93.1	95.0	94.8	98.7	92.3	95.5	78.7	2389
	(0.5)	(0.4)	(0.4)	(0.4)	(0.3)	(0.3)	(0.2)	(0.4)	(0.4)	(0.7)	
Quintile 5 (=most deprived)	86.2	92.8	95.6	93.3	94.6	93.2	98.4	91.5	95.0	78.8	4979
	(0.7)	(0.5)	(0.4)	(0.5)	(0.3)	(0.4)	(0.2)	(0.6)	(0.4)	(0.9)	

In general, the WHO Long Form (NZ Version) by deprivation quintile showed decreasing self-reported health with increasing deprivation. However, most of these differences were not statistically significant. Adults in Quintile 1(least deprived) rated their health significantly better than those in Quintile 5 (most deprived) on all scales. Scores were stable across the quintiles on the self-care (SC) scale.

Figure 6. WHO Long Form (NZ Version) Profile by deprivation quintile



Chapter 5 - Discussion

Introduction

Due to the rising burden of chronic disease and a decrease in mortality from infectious diseases, the demand health-related quality of life (HRQL) measures has increased. Self-reported health measures provide information about a wider range of health outcomes than objective measures of health status, such as mortality and hospitalisation rates. National health surveys play a role in monitoring population health. There are certain questions about the health of the population that can only be answered through health surveys (Scott et al., 1999).

An efficient way to measure health status or health-related quality of life is through scoring standardised responses to standardised questions (Lohr, 1989). Numerous questionnaires have been developed to measure HRQL. The Medical Outcomes Study Short Form 36 (SF-36), which was developed in the United States, is the most commonly used instrument (Bowden & Fox-Rushby, 2003) (Scott et al., 1999).

Questionnaires can have different meanings in different cultures and between countries, and also within countries between population subgroups. Thus it is important to assess the psychometric performance of instruments for the New Zealand population and subgroups within it, before national norms can be used in practice (Scott et al., 1999, Gandek, 1998 #18).

The New Zealand Health Monitor (NZHM) is the organised, co-ordinated and integrated survey programme of the Ministry of Health in New Zealand . The New Zealand Health Survey (NZHS) is one of the chief surveys of the NZHM. The aim of the New Zealand Health Monitor (NZHM) is to only collect data that cannot be collected more effectively and efficiently through other collection procedures, for example administrative data collection or epidemiologic study. Population based data is collected that is needed to guide decision making and assist with evidence-based policy making in the health sector (Ministry of Health, 2005).

One of the categories of information collected in the NZHM is health outcomes, and within this there is the subcategory of health status. The two parts of health status related to this thesis are subjective (self-rated) health and functional limitation. The ICF provides the

framework to describe the critical elements of non-fatal health outcomes captured by health status instruments (World Health Organization, 2001). A thorough review of existing instruments was conducted using the International Classification of Functioning, Disability and Health (ICF) Framework (Ustin et al., 2001). The ICF, approved by the WHO in May 2001, identifies 21 key domains of health (World Health Organization, 2001). The aim of the NZHM is to collect data on most if not all of these 21 ICF dimensions.

The WHO Multi-country Survey Study was set up to develop instruments that would allow for the measurement of health, responsiveness and other health-related dimensions in a way which allowed cross national comparisons. The WHO Long Form was developed as the health module in the WHO Multi-country Survey Study. The aim of this module was to develop valid, reliable and comparable instruments to portray a principal set of health domains. The WHO Long Form is made up of 20 health domains, some overlapping with the eight SF-36 domains (Ustin et al., 2001). The WHO Long Form did not have a set scoring system for scales, unlike the SF-36 instrument.

The SF-36 has been previously tested and validated in New Zealand in the 1996/97 NZHS (Scott et al., 1999). The WHO Long form questions had not previously been used in NZ and were used for the first time in NZ in the 2002/03 NZHS. There was a need to develop an approach to scoring the questions and summing items to form scales. The validation of these questions was the focus of this thesis.

The target population of the 2002/03 NZHS was the New Zealand adult population aged 15 years and over living in permanent private dwellings, approximately 2.6 million people, according to the 2001 Census of Populations and Dwelling (2001 Census) (Statistics New Zealand). The sample frame was area-based using meshblocks as the primary sampling units (PSUs) (Ministry of Health, 2004).

The 2002/03 NZHS used a complex sample design. A total of 12, 929 people responded to the survey, with 12,529 respondents being included in the CURF dataset available for research. A stratified design was used with strata defined according to the ethnicity variable defined by Question 11 on the Individual Form of the 2001 Census, with oversamples of Māori, Pacific and Asian peoples. The mode of collection was face-to-face interviewing by trained interviewers (Ministry of Health, 2004).

Most of the questions in the 2002/03 NZHS were sourced from international or local health surveys, or were developed by researchers with expertise in the topic area. The survey had

four health-related and one demographic module. The four health-related sections in the questionnaire were chronic diseases, health service utilisation, risk and protective factors and self-reported health status (Ministry of Health, 2004).

The weighting estimation was carried out by the Ministry of Health once the quality checks were conducted. Each person was assigned a weight, which represented how many population units they represented. The survey weights then allowed estimates to be calculated for the whole population (Ministry of Health, 2004).

The 2002/03 New Zealand Health Survey self-reported health status (HRQL) module was a combination of the SF-36 and the WHO Long Form (NZ Version). One of the aims of the 2002/03 NZHS relevant to this thesis was "to measure the health status of New Zealand adults, including their self-reported physical and mental health status, and the prevalence of selected health conditions" (Ministry of Health, 2004 p1). Another relevant aim was to "examine differences between population subgroups (as defined by sex, ethnicity, age and the New Zealand Deprivation Index 2001 (NZDep2001)" (Ministry of Health, 2004 p1).

There were five key aims specific to the current thesis. First, to group the WHO Long Form items in the 2002/03 NZHS into scales for each health domain and develop standard scoring protocols for each scale. Second, to test the reliability of the scales using standard psychometric tests for the total NZ population and for major population subgroups. Third, to test the validity of the scales using the standard psychometric tests for the total NZ population and for major population subgroups. Fourth, to construct norms for the WHO Long Form scales for the NZ population. And finally, to provide recommendations for the health status component of future NZ health surveys.

The health status section of the 2002/03 NZHS measures health-related quality of life (HRQL). The questions were derived from the SF-36 and the WHO Long Form questionnaire on health status. There were 16 health and health-related domains covered. The health domains covered in the 2002/03 NZHS were general health, vision, hearing, digestion, breathing, pain, sleep, energy and vitality, understanding, communication, physical functioning, self-care. The health-related domains covered in the 2002/03 NZHS were mental health, role-physical and role-emotional (usual activities), and social functioning.

The psychometric performance of the scales was tested for the major subgroups in the population defined in terms of sex, ethnicity, age and socio-economic status. Socio-economic

status was measured using the NZ Deprivation Index 2001 (NZDep01) (Salmond & Crampton, 2002).

First, scaling recommendations were made for all new scales from the WHO Long Form (NZ version). The methods used for the SF-36 scale construction were the preferred methods. Once scales were constructed, the psychometric properties of the items and scales were assessed for the total population and for major subgroups in the population.

The first level of analysis was the item. The following item level characteristics were examined: extent of missing data; frequency distribution; mean; and standard deviation.

The next level of analysis was the scale. The frequency distribution for each scale was examined, the percent of missing data assessed, and the mean, standard deviation and the percentage of scale scores at the floor (minimum) and ceiling (maximum) was calculated.

Item-scale correlations were computed after correcting for the item-scale overlap. The item-scale correlations were calculated, and assessed to see if they were all greater than 0.4 and for equality of these correlations. The correlations of the items with each of the nine other scales were calculated. The scaling success rate was calculated for each scale. This is equal to the number of significantly higher correlations with item's own scale divided by the total correlations with all other nine scales. The significance level used to test this assumption was equal to two standard errors, an approximation for the 95% confidence interval. Reliability was assessed using the internal consistency method. The reliability of the scales was estimated by calculating Cronbach's alpha coefficient for each scale.

To assess the construct validity of the scales, the correlations of the scales with each other were calculated. First, they were compared to the reliability coefficient for each scale which captures the correlation of a scale with itself. The patterns of correlations between scales were also examined as a further means of assessing construct validity to understand the inter-relationships between the scales and their underlying constructs.

Further validity tests were performed as follows. Two of the new scales, the physical functioning and social functioning scales, were composite measures of SF-36 and WHO items. The performance of these two new scales was compared to the SF-36 scales. The discriminative validity of the scales was assessed by comparing scale scores for those in the sample with and without chronic diseases, to see if the scale scores discriminated between these two groups. Discriminative validity was assessed with effects sizes and Mann Whitney

U tests, as done by Raat et al (2005). Finally norms were constructed for all scales for the total population and population subgroups.

Summary of findings

Scoring

Items were all given an equal weighting within all the scales, since unequal weighting of items rarely improves the performance of the scale enough to justify the added complexity associated with giving items different weightings. Also, the item-internal consistency correlations for all scales were roughly equal, suggesting an equivalent relationship of all items and scales. After summing, the scale totals were then transformed to give scores on all scales from 0-100, where 100 represented the best possible health. This is the same process that has previously been used for the SF-36 scales (Ware et al., 1993).

Reliability analysis

There was no missing data for any scale across all subgroups and for the total population. Item-internal consistency was high across all subgroups and scales, with only a few falling below the desired level. All items had significantly stronger associations with their own scale than any scale measured by the scaling success rate. Scaling success rates met the desired level of 100% across all scales and subgroups, reflecting good construct validity.

Internal consistency reliabilities met the standard for between group comparisons of 0.7 for most scales and subgroups. The scale with the worst internal consistency reliability was the DB scale, which did not meet this criterion. The PF scale and the SC scale met the criterion for between individual comparisons of 0.9. Floor effects were not observed for any scale or subgroup. Substantial ceiling effects were observed for all scales and subgroups except the WPF and S scales.

In terms of the comparative analyses of the SF-36 PF and SF scales and the Composite WPF and WSF scales, the following results were produced. The descriptive statistics and internal Cronbach's alphas for each pair of scales measuring the same constructs were near identical. The correlations between each set of scales were 0.99 and 0.91 respectively for physical functioning and social functioning.

Validity analysis

In terms of construct validity, for all scales and population subgroups, the Cronbach's alpha reliability coefficient was greater than all correlations between each scale and other scales. Thus each of the scales measured a unique reliable variance and measured distinct concepts for the total populations and population subgroups.

With regards to discriminative validity, the chronic disease and without chronic disease groups were compared by means of effect sizes and Mann-Whitney U tests. The adults in the chronic disease group consistently rated their health lower than those without a chronic disease for all scales. Comparing the chronic disease and no chronic disease group, the effect sizes were small for all scales except the WPF scale, which had a medium effect size. Using the Mann-Whitney U test, all differences between the median scale scores of those with chronic diseases compared to those without were highly significant.

National norms

The norms that were developed for the scales demonstrated the following patterns. Males rated their health significantly better than females in four domains: physical functioning (WPF), digestion and bodily excretions (DB), breathing (BR) and sleep (S). Females scored significantly higher than males on one scale, hearing. The largest difference between males and females was observed for the sleep scale (S). Asians rated their health higher than the other ethnic groups, but these differences were not statistically significant. The European/Other ethnic group rated their health significantly higher than both Māori and Pacific peoples for one scale. Compared to just the Māori ethnic group, the European/Other ethnic group rated their health significantly better on four scales. The norms are useful for comparing "well" and "sick" populations eg, as above comparing those with and without chronic disease.

Strengths and weaknesses of study

The measurement of reliability, i.e. whether scales measure the concept of interest consistently, is relatively straightforward, and there are several measures of the reliability of items and scales. For example, in the current study the measures for reliability used were: completeness of data and descriptive statistics at item and scale level; item internal consistency correlations; and the summary measure Cronbach's alpha coefficient for each scale. The techniques used to assess the reliability of health status measures thus lies within

the bounds of the study, and it is relatively simple to employ these statistical techniques to assess whether the questionnaires, or individual scales have satisfactory reliability properties. One of the strengths of the current study was the thorough assessment of reliability for the new WHO Long Form scales used in the 2002/03 NZHS.

However the assessment of validity i.e. whether a scale or questionnaire measures what it purports to measure, is not nearly as straightforward. There are far fewer readily accessible techniques for assessing validity, most focus on reliability. A weakness of the current study was the assessment of the validity of the WHO Long Form scales. The types of validity assessed in the current study were 'construct validity' (scaling success and interrelationships between scales) and 'discriminative validity'. Thus the assessment of validity was not nearly as thorough as the assessment of reliability. However, this weakness applies to most studies of this type, rather than this particular study alone.

The type of validity that would ideally be measured is 'criterion validity'. Unfortunately there is no straightforward method for assessing criterion validity. The reason for this is that there is no "gold standards" available in the area of health-related quality of life. This is also outside the bounds of the current research. Whether the findings from using a particular questionnaire are consistent with other similar questionnaires or other research findings on the health of the population of interest, goes outside the bounds of such surveys as the 2003/03 NZHS. Whether criterion validity can ever be conclusively assessed is a matter for debate in itself, and the validation of a series of scales in a questionnaire is an evolutionary process that continues as long as new information is obtained about the interpretation of scale scores.(Sullivan & Karlsson, 1998)

For some of the scales presented, such as vision, it would be possible to check results from the questionnaire against objective tests of vision. But this would not be possible for all scales, as there are not objective measures for many of them, such as social functioning. Also, it would be important to clarify what is desired to measure, as the subjective nature of the measures are part of what is intended to be measured, i.e. an individual's perception of their level of functioning and how this impacts upon their health.

Overall, the current study uses standard techniques for validating health status measures. The same methods that were used to validate the SF-36 in New Zealand, and internationally in the IQOLA project, were used to validate the WHO Long Form scales. As the health status component is constructed from the SF-36 and the WHO Long Form, it is useful to have

results of psychometric testing for both questionnaires in similar format. Thus the comparability of psychometric testing for these two sets of scales is a strength of the current study, in relation to the use of results from the NZHS, both in the past and future.(Bowling, 2005)

Comparisons with other international studies

In general findings from the current study are consistent with those of other studies internationally. The major difference from previous validation studies of HRQL instruments both in NZ and overseas (Scott et al., 1999; Sullivan et al., 1995), was in the proportion of missing data which was present for the WHO Long Form (NZ version) questions compared to the IQOLA Project, the WHO Survey and the previous 1996/97 NZHS. (Scott et al., 1999; Sullivan et al., 1995; Ustin et al., 2001). Item completeness is a major measure of assessing the acceptability of a questionnaire. In the 2002/03 NZHS Health Status module, including the WHO Long Form questions, missing data was virtually non-existent, compared to about 2-4% in the SF-36 in the 1996/97 NZHS (Scott et al., 1999). The proportion of missing data was associated with socio-economic status, as had been previously reported in Sweden. (Sullivan et al., 1995) One possible explanation for this major difference is possible impact of mode of questionnaire administration between the two health surveys. (Streiner & Norman, 2003). The Health Status module of the 1996/97 NZHS was completed in writing by the respondent at the end of the interview, whereas in the 2002/03 survey, the module was completed by face-to-face interview in the 2002/03 NZHS (Ministry of Health, 1999a) and (Ministry of Health, 2004).

International comparisons of the psychometric properties and norms for the WHO Long form scales are not yet possible. Unlike the widespread use of the SF-36 across the world in population health surveys and clinical studies, the WHO Long form was used for the first time in a population health survey in NZ. If these questions and scales are used in other countries in future, it will be possible to compare performance and norms across countries.

Summary and implications

In summary, this thesis developed a method for producing scale scores for domains of health not previously measured in New Zealand Health Surveys, providing greater coverage of domains from the ICF. There were virtually no missing data for all items and subgroups within the questions used to develop the scales. The scaling approach was consistent with

that for the SF-36, allowing the new scales to be presented alongside the SF-36 scales. All scales for the total population and major population subgroups met the required criterion for satisfactory psychometric properties, with the exception of DB scale. For the DB scale, the Cronbach's alpha was lower than that required for between group comparisons. The composite PF and SF scales performed no better than the existing SF-36 scales and were highly correlated with these scales.

After their first use in the 2002/03 NZHS, the new WHO-Long Form (NZ Version) questions now have scoring methods available to produce scales for the new domains of health covered which can be used in future analyses. The advantage of the techniques employed to produce scale scores is that they use the same method of scoring to the SF-36 so can be presented alongside the SF-36 scores in the Health Status module of the NZHS. Thus within the Health Status module of the NZHS, the majority of domains from the ICF are now covered.

The scale scores allow for the possibility of future analyses relating these new domains of HRQL eg, understanding and sleep, to other aspects of health in the NZHS eg, multivariate analyses relating HRQL domains to health risk and health service utilisation. This thesis has thoroughly tested the psychometric properties of the items and scales for the total population and major subgroups of interest within the population. All scales except the DB scale had satisfactory properties not only for the total population but also the population subgroups.

The new scales have now been validated for the total population and subgroups in the same way in which the SF-36 was validated in the 1996/97 NZHS. Overall, WHO Long Form questions performed well, with results comparable to those obtained for the SF-36 instrument (Scott et al., 1999). Again data was heavily skewed towards the health end of the scales, but this was more severe than for the SF-36. New techniques were used to test the validity of the instrument, compared to those used for the SF-36 validation (Scott et al., 1999). These new techniques were those employed to test discriminative validity (Ratt et al., 2005). The issue with criterion validity testing is the lack of a "gold standard" to compare the scale scores against. Conclusions and recommendations for future surveys

Notwithstanding the limitations of this study, key findings of interest are that the new WHO Long Form questions can be used to form scales that cover physical functioning (WPF), social functioning (WSF), vision (V), hearing (H), digestion and bodily excretions (DB), breathing (BR), self-care (SC), understanding (US), communication (CM) and sleep (S). The majority of the questions and scales work for the NZ population and subgroups. All but one

of the scales, DB, have satisfactory psychometric properties for the total population and major subpopulation groups of interest. The BR scale had moderately good psychometric properties. The respondent burden is an important consideration for the NZHS, thus it cannot be argued that enough is gained from adding questions to the PF and SF domains, thus it would be recommended that the SF-36 scales are used to measure there two domains of health. Thus it would be recommended that the DB and BR scales are dropped, and no additional questions are added to the PF and SF domains, instead the existing SF-36 domains used. The new WHO Long Form scales can now be presented alongside the SF-36 scales and used in future analyses looking at interrelationships between factors such as health risk and health status.

Most scales appear to perform satisfactorily for the population as a whole and major subgroups, but it would be useful to conduct factor analyses of the scales to assess whether the interrelationships between domains differ substantially by subgroup, as was the case for ethnicity for the SF-36 (Scott et al., 2000). A further aspect relating to ethnicity in the NZHS, is to consider whether the domains covered in the health status module cover all domains within Māori models of health e.g. wairua and whanau (Durie, 1994). The broadening of the model of health underpinning the survey to encompass Māori models of health and thus measurement within this module could be considered in future health surveys.

There are a couple of issues that could be explored in future surveys. The first is psychometric analysis relating to respondent burden. Whilst it is desired that as many of the ICF domains of health (21) are measured as possible in the health status module of the NZHS, the time the questionnaire takes to administer and thus respondent load or burden must be borne in mind. From the analysis carried out in the current study, it is not possible to assess whether subscales could be used (formed from the most powerful items of their parent scales) satisfactorily for some domains, thus shortening the number of questions but maintaining suitable psychometric properties. As the demand to collect more and more information in the NZHS increases, this type of analysis may be desired so as to minimise respondent burden. The benefits of shorter measures for research and policy include reduced burden on respondents and cost of surveys and also ease of interpretation. (Bowling, 2005)

The second issue relates to the mode of administration of the health status portion of the NZHS. To check for the comparability of results obtained using interviewer administration and self-administration, both techniques could be employed for a subset of participants. The results from each mode of administration could then be compared.

References

Aaronson, N. K., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D., Perrin, E., et al. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*, 193-205.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *British Medical Journal, 314*, 572.

Bowden, A., & Fox-Rushby, J. (2003). A systematic and critical review of the process of translation and adaption of generic health-related quality of life measures in Africa, Asia, Eastern Europe, The Middle East, South America. *Social Science and Medicine, 57*, 1289-1306.

Bowling, A. (1997). *Measuring Health: A review of quality of life measurement scales* (2nd ed.). Buckingham: Open University Press.

Bowling, A. (2005). Just one question: If one question works, why ask several? *Journal of Epidemiology and Community Health, 59*(5), 342-345.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioural Sciences*.New York: Academic Press.

Durie, M. (1994). *Whaiora Maori: Health development*.Auckland: Oxford University Press.

Evans, R. G., & Stoddart, G. L. (1994). Producing Health, Consuming Health Care. In R. G. Evans, M. L. Barer & T. R. Marmor (Eds.), *Why are some people healthy and others not? The Determinants of Health of Populations* (pp. 27-64). New York: Aldine De Gruyter.

Gandek, B., Ware, J., Aaronson, N. K., Alonso, J., Apolone, G., Bjoner, J., et al. (1998). Tests of Data Quality, Scaling Assumptions, and Reliability of the SF-36 in Eleven Countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*(11), 1149-1158.

Kott, P. S. (1998). *Using the Delete-A-Group Jackknife Variance Estimator in NASS Surveys* (No. No. RD-98-01): U.S. Department of Agriculture, National Agricultural Statistics Service.

Lohr, K. N. (1989). Advances in Health Status Assessment: Overview of the Conference. *Medical Care, 27, 3, Supplement*.

Ministry of Health. (1999a). *Taking the Pulse - the 1996/97 New Zealand Health Survey*.Wellington: Ministry of Health.

Ministry of Health. (1999b). *Our Health, Our Future: The Health of New Zealanders 1999*.Wellington: Ministry of Health.

Ministry of Health. (2002). *The New Zealand Health Monitor: A 10-year cycle of health-related surveys*.Wellington, New Zealand: Ministry of Health.

Ministry of Health. (2004). *A Portrait of Health: Key Results of the 2002/03 New Zealand Health Survey*.Wellington, New Zealand: Ministry of Health.

Ministry of Health. (2004b). *2002/03 New Zealand Health Survey: National Confidentialised Unit Record File (CURF) User Document - Version 1a*.Wellington: Ministry of Health.

Ministry of Health. (2005). *The New Zealand Health Monitor: Updated strategic plan*.Wellington: Ministry of Health.

Ministry of Health. (Unpublished report). *2006/07 New Zealand Health Survey: Draft Content Guide*.Wellington: Ministry of Health.

Patrick, D., & Erickson, P. (1993). *Health Status and Health Policy: Allocating Resources to Health Care*.New York: Oxford University Press.

Ratt, H., Botterweck, A. M., Landgraf, J. M., Hoogeveen, W. C., & Essink-Bot, M. L. (2005). Reliability and validity of the short form of the child health questionnaire for parents (CHQ-PF28) in large random school based and general population samples. *Journal of Epidemiology and Community Health, 59*, 75-82.

Salmond, C., & Crampton, P. (2002). *NZDep2001 Index of Deprivation*.Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences.

Sanson-Fisher, R. W., & Perkins, J. (1998). Adaptation and Validation of the SF-36 Health Survey for Use in Australia. *Journal of Clinical Epidemiology, 51*(11), 961-967.

Scott, K. M., Sarfati, D., Tobias, M. I., & Haslett, S. J. (2000). A challenge to the cross-cultural validity of the SF-36 health survey: factor structure in Maori, Pacific and New Zealand European ethnic groups. *Social Science and Medicine, 51*, 1655-1664.

Scott, K. M., Tobias, M. I., & Sarfati, D. (1999). SF-36 health survey reliability, validity and norms for New Zealand. *Australian and New Zealand Journal of Public Health, 23*(4), 401-406.

Statistics New Zealand. (2005). Meshblock 2005. Retrieved 12 April 2005, 2005, from http://www.stats.govt.nz/statistical-methods/classifications/ameshblock_2005.htm

Streiner, D. L., & Norman, G. R. (2003). *Health Measurement Scales: A practical guide to their development and use* (Third ed.). Oxford: Oxford University Press.

Sullivan, M., & Karlsson, J. (1998). The Swedish SF-36 Health Survey III. Evaluation of Criterion-Based Validity: Results from Normative Population. *Journal of Clinical Epidemiology, 51*(11), 1105-1113.

Sullivan, M., Karlsson, J., & Ware, J. E. (1995). The Swedish SF-36 Health Survey - I. Evaluation of Data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Social Science and Medicine, 41*(10), 1349-1358.

Ustin, T. B., Chatterji, S., Villanueva, M., Bendib, L., Celik, C., Sadana, R., et al. (2001). *WHO Multi-country Survey Study on Health and Responsiveness: GPE Discussion Paper 37*.Geneva: WHO.

Ware, J., & Gandek, B. (1998). Methods for Testing Data Quality, Scaling Assumptions, and Reliability: The IQOLA Project Approach. *Journal of Clinical Epidemiology, 51*(11), 945-952.

Ware, J., Kosinski, M., Gandek, B., Aaronson, N., Apolone, G., Bech, P., et al. (1998). The Factor Structure of the SF-36 Health Survey in 10 Countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*(11), 1159-1165.

Ware, J., & Sherbourne, C. D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual Framework and Item Selection. *Medical Care, 30 (6)*, 473-481.

Ware, J., Snow, K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey: Manual & Interpretation Guide*.Boston, Massachusetts: The Health Institute, New England Medical Center.

World Health Organisation. (1958). *The First Ten Years of the World Health Organisation*.Geneva: World Health Organisation.

World Health Organization. (2001). *International Classification of Functioning, Disability and Health*.Geneva: World Health Organization.

Appendix I – 2002/03 NZHS Relevant sections of questionnaire

HEALTH STATUS

GENERAL HEALTH

Q.187 In general, how would you say that your health is? *(Card 187)*

- ☐ 1 Excellent
- ☐ 2 Very good
- ☐ 3 Good
- ☐ 4 Fair
- ☐ 5 Poor

Q.188 Compared to one year ago, how would you rate your health in general now? *(Card 188)*

- ☐ 1 Much better now than one year ago
- ☐ 2 Somewhat better now than one year ago
- ☐ 3 About the same as one year ago
- ☐ 4 Somewhat worse now than one year ago
- ☐ 5 Much worse now than one year ago

How TRUE or FALSE is **each** of the following statements for you?

Q.189 I seem to get sick a little easier than other people. *(Card 189)*

- ☐ 1 Definitely true
- ☐ 2 Mostly true
- ☐ 3 Don't know
- ☐ 4 Mostly false
- ☐ 5 Definitely false

Q.190 I am as healthy as anybody I know.

- ☐ 1 Definitely true
- ☐ 2 Mostly true
- ☐ 3 Don't know
- ☐ 4 Mostly false
- ☐ 5 Definitely false

Q.191 I expect my health to get worse.

- ☐ 1 Definitely true
- ☐ 2 Mostly true
- ☐ 3 Don't know
- ☐ 4 Mostly false
- ☐ 5 Definitely false

Q.192 My health is excellent.

- ☐ 1 Definitely true
- ☐ 2 Mostly true
- ☐ 3 Don't know
- ☐ 4 Mostly false
- ☐ 5 Definitely false

VISION

Q.193 During the past 4 weeks, how much difficulty (wearing glasses or contact lenses if you usually do) did you have in seeing and recognising a person you know across the road, i.e. from a distance of about 20 metres. *(Card 193)*

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot see

Q.194 And how much difficulty in seeing or recognising a person you know across the room, i.e. from a distance of about 5 metres?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot see

Q.195 Reading a book or newspaper?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot see

HEARING

The next few questions are about your hearing.

Q.196 I would like to ask you about your use of special or technical equipment or services for people who are deaf or hard of hearing. Do you use any of these items? *(Card 196)*

Q.197 Is there any equipment or services for people who are deaf or hard of hearing which you need for yourself, but do not have? *(Card 196) (Record below)*

	Q.196 Use now	Q.197 Need
A hearing aid with T-switch (Telephone switch)	☐ 01	☐ 01
Another type of hearing aid	☐ 02	☐ 02
A telecommunications device such as a teleprinter or TTY, specifically because of your hearing difficulty	☐ 03	☐ 03
Teletext, specifically because of your hearing difficulty	☐ 04	☐ 04
Hearing loop, FN or Infrared system	☐ 05	☐ 05
A sign language interpreter	☐ 06	☐ 06
Flashing alarms or visual alarms	☐ 07	☐ 07
A volume control telephone	☐ 08	☐ 08
A computer to communicate, specifically because of your hearing difficulty	☐ 09	☐ 09
A fax machine to assist, specifically because of your hearing difficulty	☐ 10	☐ 10
Some other equipment or service that I have not mentioned *(Specify)* Q196 _____ ☐ _____ ☐ _____ ☐ Q197 _____ ☐ _____ ☐ _____ ☐	☐ 98	☐ 98
No special equipment or services to assist with a hearing difficulty	☐ 99	☐ 99

Q.198 During the past 4 weeks, how much difficulty (using a hearing aid if you usually do) did you have in hearing what is said in a conversation with one other person in a quiet room? *(Card 198)*

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot hear

Q.199 Hearing someone talking on the other side of the room in a normal voice?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot hear

Q.200 Hearing what is said in a group conversation with at least 3 other people?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty / cannot hear

DIGESTION & BODILY EXCRETIONS

Q.201 During the past 4 weeks, how much of the time did you have indigestion, e.g. burning in the stomach, "heartburn"? *(Card 201)*

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.202 Have constipation (difficulty with passing bowel motions)?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.203 Have difficulty passing urine (in other words, peeing, passing water or urinating)?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.204 Have difficulty controlling urine (in other words, incontinence)?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

BREATHING

Q.205 During the past 4 weeks, how much of the time did you get short of breath with mild exercise, such as walking on the flat? *(Card 201)*

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.206 Get short of breath at rest?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.207 Experience coughing or wheezing for ten minutes or more at a time?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

PAIN AND DISCOMFORT

Q.208 How much bodily pain have you had during the **past 4 weeks**? *(Card 208)*

- ☐ 1 No bodily pain → *Go to instructions before Q.212*
- ☐ 2 Very mild
- ☐ 3 Mild
- ☐ 4 Moderate → *Go to Q.209*
- ☐ 5 Severe
- ☐ 6 Very severe

Q.209 Where did you have pain? *(Card 209) (Record below. Multiple response.)*

Q.210 *If more than one site, ask:* "Which affected you the most?" *(Single response)*

Location	Q.209 Where	Q.210 Most
Head (headache, migraines)	☐ 01	☐ 01
Neck	☐ 02	☐ 02
Back	☐ 03	☐ 03
Stomach or abdomen	☐ 04	☐ 04
Joints like arms, hands, legs, or feet	☐ 05	☐ 05
Face or jaw or the joint just below the ear	☐ 06	☐ 06
Chest	☐ 07	☐ 07
Anywhere else *(Specify)* Q.209 _____ ☐ _____ ☐ _____ ☐ Q.210 _____ ☐	☐ 98	☐ 98

Q.211 During the **past 4 weeks**, how much did **pain** interfere with your normal work (including both work outside the home and housework)? *(Card 211)*

☐ 1 Not at all
☐ 2 A little bit
☐ 3 Moderately
☐ 4 Quite a bit
☐ 5 Extremely

MENTAL HEALTH

*These questions are about how you feel and how things have been with you **during the past 4 weeks.** For each question, please give the one answer that comes closest to the way you have been feeling.*

Q.212 How much of the time during the **past 4 weeks h**ave you been a happy person? *(Card 212)*

☐ 1 All the time
☐ 2 Most of the time
☐ 3 A good bit of the time
☐ 4 Some of the time
☐ 5 A little of the time
☐ 6 None of the time

Q.213 Have you felt calm and peaceful?

☐ 1 All the time
☐ 2 Most of the time
☐ 3 A good bit of the time
☐ 4 Some of the time
☐ 5 A little of the time
☐ 6 None of the time

Q.214 Have you felt so down in the dumps that nothing could cheer you up?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.215 Have you felt down?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.216 Have you been a very nervous person?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

SLEEP

Q.217 During the past 4 weeks, how much of the time did you have a problem with falling asleep?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.218 Waking up frequently during the night?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.219 Waking up too early in the morning?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

ENERGY AND VITALITY

Q.220 How much of the time during the **past 4 weeks, d**id you feel full of life?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.221 Did you have a lot of energy?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.222 Did you feel worn out?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.223 Did you feel tired?

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

UNDERSTANDING AND REMEMBERING

Q.224 During the past 4 weeks, how much difficulty did you have in concentrating on doing something for at least 10 minutes? *(Card 224)*

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.225 Remembering to do important things?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.226 Analysing and solving problems in day-to-day life?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.227 Learning a new task? (For example, learning how to get to a new place.)

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

COMMUNICATING

Q.228 During the past 4 weeks, because of your physical or emotional health, how much difficulty did you have in generally understanding what people say? **(This is not related to hearing difficulty.)** *(Card 224)*

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.229 Starting and maintaining a conversation? (not related to speaking difficulty)

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.230 During the past 4 weeks, how much difficulty did you have in speaking clearly?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

PHYSICAL FUNCTIONING (MOBILITY AND DEXTERITY)

*The following questions are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?*

Q.231 **Vigorous** activities, such as running, lifting heavy objects, participating in strenuous sports *(Card 231)*

- ☐ 1 Yes, limited a lot
- ☐ 2 Yes, limited a little
- ☐ 3 No, not limited at all

Q.232 **Moderate** activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf

- ☐ 1 Yes, limited a lot
- ☐ 2 Yes, limited a little
- ☐ 3 No, not limited at all

Q.233 Lifting or carrying groceries

- ☐ 1 Yes, limited a lot
- ☐ 2 Yes, limited a little
- ☐ 3 No, not limited at all

Q.234 Climbing **several** flights of stairs

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.235 Climbing **one** flight of stairs

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.236 Walking **more than one kilometre**

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.237 Walking **half a kilometre**

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.238 Walking **100 metres**

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.239 Bending, kneeling or stooping

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.240 Standing up from sitting down

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.241 Placing your hands behind your head

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.242 Using your hands and fingers (picking up small objects or opening or closing containers)

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

HEALTH RELATED DOMAINS

Self care

The following questions ask about caring for yourself.

Does **your health now limit you** in these activities? If so, how much?

Q.243 Bathing yourself *(Card 231)*

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.244 Dressing yourself

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.245 Grooming yourself (for example, combing your hair)

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.246 Eating (for example, cutting up food, using a knife and fork)

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.247 Using the toilet

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

Q.248 Staying by yourself for a few days

☐ 1 Yes, limited a lot

☐ 2 Yes, limited a little ☐ 3 No, not limited at all

USUAL ACTIVITIES

The following questions ask about your work or other regular daily activities such as housekeeping or looking after a child or other person.

During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities, **as a result of your physical health**?

Q.249 Cut down on the **amount of time** you spent on work or other activities

☐1 Yes ☐2 No

Q.250 **Accomplished less** than you would like

☐1 Yes ☐2 No

Q.251 Were limited in the **kind** of work or other activities

☐1 Yes ☐2 No

Q.252 Had **difficulty** performing the work or other activities (for example, it took extra effort)

☐1 Yes ☐2 No +

During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)?

Q.253 Cut down the **amount of time** you spent on work or other activities

☐1 Yes ☐2 No

Q.254 **Accomplished less** than you would like

☐1 Yes ☐2 No

Q.255 Didn't do work or other activities as **carefully** as usual

☐1 Yes ☐2 No

SOCIAL FUNCTIONING

The following questions ask about your relationships with other people.

Q.256 During the **past 4 weeks**, to what extent has your **physical health or emotional problems** interfered with your normal social activities with family, friends, neighbours, or groups? *(Card 256)*

- ☐ 1 Not at all
- ☐ 2 A little bit
- ☐ 3 Moderately
- ☐ 4 Quite a bit
- ☐ 5 Extremely

Q.257 During the **past 4 weeks**, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting friends, relatives, etc.)? *(Card 257)*

- ☐ 1 All the time
- ☐ 2 Most of the time
- ☐ 3 A good bit of the time
- ☐ 4 Some of the time
- ☐ 5 A little of the time
- ☐ 6 None of the time

Q.258 During the past 4 weeks, because of your **physical or emotional health**, how much difficulty did you have in dealing with people you do not know? *(Card 258)*

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.259 Maintaining a friendship?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty

Q.260 Getting along with people who are close to you?

- ☐ 1 No difficulty
- ☐ 2 A little bit of difficulty
- ☐ 3 Moderate difficulty
- ☐ 4 Quite a bit of difficulty
- ☐ 5 Extreme difficulty