# Computer-assisted Approaches to the Collection of Quality of Life Data in Oncology

Adam Barnett Smith

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

March 2004

## Acknowledgements

A large part of this thesis makes use of two large datasets which have been collated from a series of studies undertaken between 1996 and 2003 by the Psycho-Social Oncology Research Groups and St. James's Hospital (Leeds) and the Western General Hospital (Edinburgh). Although I contributed to the early part of the data collection, I am in particular grateful to Dr. Penny Wright, Laura Booth, Ann Gillibrand, Maggie Kiely, Pamela Lynch (Leeds), Jo Chappell and Vanessa Strong (Edinburgh) who were responsible for the bulk of the data collection. For the remainder, I was solely responsible for the data collection for the comparison of the computer-assisted and standard electronic questionnaire study. I am very grateful to Valerie Walker, clinical nurse specialist, for allowing me to disrupt her Tuesday morning clinic and enabling me to approach her patients.

I am indebted to my supervisors Profs. Peter Selby and Doug Altman for their support, advice and comments on all the drafts of the thesis I sent their way. I would also like to express my sincere gratitude to Dr. Galina Velikova, who not only read, corrected and commented on each of the drafts of the chapters, but also proofread the thesis in its entirety. I also am very grateful to Penny Wright for her practical advice and encouragement, as well as to all of the Psycho-Social Oncology team for their support throughout.

Prof. Alan Tennant provided invaluable advice regarding Rasch analyses and models, as well as regarding the software used for the analysis in the thesis.

I would not have attempted to write the thesis had it not been for my family: My parents, for encouraging me, and especially, Andrea and Abigail, Emily, James and Tom, who endured my long absences and absent-mindedness with love and patience, and unfailingly supported me.

This thesis is dedicated to the loving memory of my grandparents, and to Andrea and the kids.

Computer-assisted Approaches to the Collection of Quality of Life Data in Oncology

by Adam Barnett Smith

Doctor of Philosophy

Abstract

March 2004

The assessment of cancer patients' quality of life (QOL) has been increasing in both importance and relevance in recent years, and is becoming more integrated into clinical practice. This has been greatly facilitated by the development of standard QOL instruments. However, the standard questionnaires may overlook certain aspects of QOL or focus on areas which do not present a problem to patients.

The aims of this thesis were to increase the relevance of QOL instruments to patients by developing systems that allow patients to select relevant domains from questionnaires and secondly, to minimise patient burden by reducing the number of questions presented to patients.

Initially, a computer-assisted version of the EORTC QLQ-C30 was compared with a standard electronic version of the questionnaire. Patients completed both forms on the same day. The results demonstrated that although patients completed the computer-assisted questionnaire more quickly, there was poor exact agreement, between the two forms. However, general agreement was good (i.e. > 70%) for all symptom scales, but not for the majority of the functioning scales. In addition, patients tended to report higher levels of symptoms and poorer functioning on the standard questionnaire.

Studies were then developed and conducted using Factor and Rasch analyses on a series of standard questionnaires, namely the HADS, the EORTC QLQ-C30, and the FACT-G, in order to assess their structure and the performance of each item. The results from HADS scale demonstrated a two-factor structure

corresponding to anxiety and depression, and an overall psychological distress measure. In addition to confirming this structure, the Rasch analysis identified one misfitting item for each of the full HADS-scale and two subscales.

For the EORTC QLQ-C30 the results demonstrated a four-factor structure corresponding to a physical functioning factor, a factor covering social and role functioning, and including pain and fatigue symptoms, a third factor covering the emotional and cognitive functioning domains, and finally a factor covering the remaining symptoms. The Rasch analysis demonstrated good fit for all items of the Emotional Functioning, and Fatigue scales, and only one misfitting item from the Physical Functioning scale.

The results for the FACT-G demonstrated four factors corresponding to the four FACT-G subscales, although all subscales contained at least two misfitting items.

The misfitting items from the HADS were systematically removed from the HADS and its subscales, and the screening efficacy of the scales re-evaluated against psychiatric interview data (PSE/SCAN). The results demonstrated no loss in screening efficacy when these items were removed.

In the final study scores from the corresponding scales of the EORTC QLQ-C30 and FACT-G were converted to log-odds (logit) scores and agreement between the scales was calculated. The results demonstrated high levels of agreement between three of the scales, namely Physical and Emotional Functioning and overall quality of life, and good levels of agreement for the other two scales (Role and Social Functioning).

In conclusion, the utility of Rasch models in identifying items for removal from instruments in order to reduce patient burden was demonstrated in this thesis. This work provides a foundation for the subsequent development of computer-adaptive questionnaires.

## Table of Contents

# List of Tables

**Chapter 8**



**Chapter 9**

## List of Figures

**Chapter 6**



**Chapter 7**

# List of Equations

## Abbreviations

| | |
|---|---|
| BDI | Beck's Depression Inventory |
| EORTC QLQ-C30 | European Organisation for Research and Treatment of Cancer |
| FACIT | Functional Assessment of Chronic Illness Therapy |
| FACT-G | Functional Assessment of Cancer Therapy – General |
| FLIC | Functional Living Index – Cancer |
| HADS | Hospital Anxiety and Depression Scale |
| LASA | Linear Analogue Self-Assessment Scale |
| PSE | Present State Examination |
| QOL | Quality of Life |
| ROC curve | Receiver Operating Characteristic curve |
| SCAN | Schedule for the Clinical Assessment in Neuropsychiatry |
| SEIQoL | Schedule for the Evaluation of Individual Quality of Life |
| SF-36 | Short Form 36 |
| SSTI | Spielberger State-Trait Inventory |
| WHO | World Health Organisation |

## Publications

Smith, A.B., Selby, P.J., Velikova, G., Stark, D., Wright, E.P., Gould, A., & Cull. (2002). The Factor structure of the HADS Questionnaire from a large cancer population. *Psychology and Psychotherapy* 75, 165 – 176.

Smith, A.B., Velikova, G., & Selby, P. (in press, 2004). Computer-assisted Questionnaires may facilitate Quality-of-Life Assessment: At a cost. *Computer in Human Behavior*

# 1. Literature Review and Hypothesis

## 1.1. Quality of Life

### *1.1.1. Introduction & Overview*

Patients with cancer face the physical manifestations of these life-threatening diseases and a relatively high risk of premature death. In addition the disease and its treatment can result in impairment of their lives across the whole spectrum of dimensions including physical, emotional and social aspects. Cancer can therefore very substantially reduce the quality of patients' lives. In the 1970s and 1980s, increasing recognition that quality of life was an important factor for cancer patients led to the development of a portfolio of approaches to measuring quality of life in oncology, in order to allow comparisons between different patient groups, measure changes over time, and to enhance our ability to describe for cancer patients the consequences of the diseases and their treatment (Aaronson et al., 1993; Cella et al., 1993). Since then the use of quality of life measurements has become a relatively standard practice in cancer clinical trials (e.g. Fayers et al., 1997; Osoba, 1999; Staquet, Berzon, Osoba, & Machin, 1996).

More recently, it has been recognised that our ability to measure at least some aspects of quality of life in a reproducible and psychometrically sound way could enhance the care of cancer patients by ensuring that healthcare professionals were better informed and systematically appraised of the impact of the disease and its treatment on aspects of quality of life. However, there were very substantial barriers that existed which limited our ability to introduce quality of life measurement into clinical practice in cancer care. The availability of appropriate measurement instruments, their evaluation in a clinical practice setting, the logistic problems posed by pen-and-paper approaches which generated huge amounts of data to be entered into computers, checked and quality-assured, were among the limitations. However, the hypothesis that appropriate measurement of quality of life would improve the well being of cancer patients when these data were used to enhance their interaction with healthcare professionals was attractive to a number of research groups.

*1.1.2. Definition of Quality-of-Life*

Despite the fact that the measurement of quality of life (QOL) has become increasingly important in routine practice in oncology clinics (Cull, Stewart & Altman, 1995; Detmar et al., 2002; Ganz, 1994), and despite the growing awareness of QOL by clinicians and the increased use of QOL measures in clinical practice, in particular over the last thirty years, QOL remains an elusive topic to define.

According to the World Health Organization (1946) health can be defined as "a state of complete physical, mental, and social well-being, and not merely the absence of disease". Calman (1984) defined QOL as the difference between hopes and expectations, and the patient's current state. Other definitions include "the subjective evaluation of life as a whole" (de Haes, 1988), and "patients' appraisal of and satisfaction with their current level of functioning compared with what they perceive to be possible or ideal" (Cella and Cherin, 1988). Finally, as pointed by Velikova et al. (Velikova, Stark, and Selby, 1999) "QOL encompasses all aspects of patients' well-being".

Quality of life can therefore be considered as a multidimensional concept (e.g. Aaronson et al., 1993; Cella, 1994; Velikova, Stark, and Selby, 1999), encompassing the patients physical functioning (e.g. the ability to carry out daily activities, and mobility), psychological (e.g. anxiety and depression, and social functioning (social interactions, hobbies and leisure activities), as well as including disease-related and or treatment-related symptoms (such as pain, nausea and vomiting, sleep disruption, hair loss).

In summary then QOL measurement has emphasised the subjective nature of the concept, and requires patients to reflect on and evaluate their QOL. These are facets, which in turn are reflected in the questionnaires that have been developed.

*1.1.3. Quality-of-Life and Cancer*

The use of questionnaires to capture quality of life information from oncology patients can probably be traced back to the early pioneering work of Priestman and Baum (1976) using Linear Analogue Self-Assessment Scales (LASA) to assess quality of life of women with

advanced breast cancer. Since this time the importance of QOL questionnaires in oncology has been increasingly recognised by clinicians, and indeed is becoming more integrated into routine clinical practice (Cull, Stewart & Altman, 1995; Ford, Fallowfield, & Lewis, 1994; Velikova et al., 2004), as well as this quality of life measurement is also an important adjunct as an outcome measure in clinical trials (Fayers et al., 1997; Ganz, 1994).

Although cancer specific QOL questionnaires have been developed, such as the Functional Living Index – Cancer (FLIC, Schipper et al., 1984), the LASA (Coates et al., 1983; Selby et al., 1984), the Rotterdam Symptom Checklist (RSCL, de Haes, van Knippenberg, and Neijt, 1990) and the Spitzer QOL Index (Spitzer et al., 1981), the development of QOL measures has tended to adopt a modular approach with core questionnaires being designed consisting of subscales covering the major QOL domains (e.g. physical functioning, role and social functioning, emotional, as well as symptoms scales). These core questionnaires can then be augmented with supplementary questionnaires which contain disease-specific modules and/or treatment specific modules.

Two widely used QOL questionnaires which have been developed by adopting this approach are the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30, Aaronson et al., 1993), and the Functional Assessment of Cancer Therapy questionnaire, which has recently been adapted for use with chronic diseases and renamed the Functional Assessment of Chronic Illness Therapy questionnaire (FACT-G/FACIT, Cella et al., 1993).

The EORTC QLQ-C30 has modules covering head and neck, lung, breast, oesophageal, pancreatic and colorectal cancers, as well as myeloma (Sprangers et al., 1993; 1998). Similarly, the FACT also has supplementary modules covering the major cancers, as well as modules for patients with anaemia, fatigue or those undergoing biological treatment or bone marrow transplants.

In addition to the questionnaires described above which focus largely on the physical and social aspects of Quality of Life, a number of other questionnaires are also employed in oncology for identifying psychological distress in cancer patients. Recent studies have

estimated that prevalence of anxiety in oncology patients ranges from between 7% to 23% (Stark and House, 2000), and prevalence of depression ranges between 7% and 47% (Sellick and Crooks, 1999). A number of questionnaires have been used for screening for psychological distress in cancer patients, including the Hospital Anxiety and Depression Scale (HADS, Sigmond and Snaith, 1983) which is a 14-item questionnaire with 7 questions covering anxiety and 7 depression; Beck's Depression Inventory (BDI, Beck, Ward, Mendelson et al., 1961), with 21 questions; Spielberger State-Trait Inventory (Spielberger, Gorsuch, Luchene et al., 1983), which consists of 40 questions with 20 questions for "state" anxiety and 20 for "trait" anxiety; and the Zung Self-rating Depression Scale (Zung, 1965) which includes 20 questions on depression.

However, there are some drawbacks to these questionnaires. For instance, many of these standard QL instruments were originally developed for use in clinical trials and group comparisons and therefore contain a large number of questions which may prove to be too time-consuming and impractical to be used by patients in clinic. Shorter forms of questionnaires could be developed although the major drawback of shorter questionnaires is that detail and precision is lost particularly at the level of the individual patient (Ware, Bjorner, and Kosinski, 1999). Furthermore, there is a danger with a reduced number of questions that the questionnaire may either overestimate ("floor effect") or underestimate ("ceiling effect") patients' abilities (e.g. Ware et al., 2003). This problem also occurs with standard versions of questionnaires where often very few (e.g. <5) items or questions are utilised for each QOL domain. In addition, another problem with these standard questionnaires is that because the number of questions is fixed the patients may be asked questions which are not relevant to them or conversely the questionnaire will not explore problem areas in greater detail (Lai et al., 2003; Revicki and Cella, 1997; Ware et al., 2003). Finally, there is an issue as to whether QOL questionnaires developed for group comparisons (e.g. clinical trials) could be used for individual patient monitoring (Joyce, Hickey, McGee and O'Boyle, 2003; McHorney & Tarlov, 1995), and indeed whether quality

of life instruments, which rely on data summed from various questions, can truly reflect individual concerns (Leplege and Hunt, 1997).

Questionnaires such as the Schedule for the Evaluation of Individual QOL (SEIQoL, Browne et al., 1997; Hickey et al., 1996; McGee et al., 1991; O'Boyle et al., 1992; Waldron et al., 1999) and the Patient Generated Index or PGI (Macduff and Russell, 1998) address the issue of relevance by requiring the patients to nominate areas of their life which currently impact on their quality of life. The problem with questionnaires such as these is that the lack of standardization makes comparisons between individuals extremely difficult. Additionally, these questionnaires do not offer a solution for how to monitor individuals over time if, for instance, priorities change over time, and how this change is to be assessed and compared with previous the measure. Furthermore, these questionnaires have not been subjected to the rigorous psychometric testing that the standard questionnaires have undergone (e.g. Aaronson et al., 1993; Cella et al., 1993).

One method which may potentially reconcile the competing demands of more relevant, yet shorter questionnaires which are still psychometrically sound is computer adaptive testing (e.g. Wainer, 1990). Computer adaptive testing relies on statistical methodologies such as Rasch models (Rasch, 1960/1980), and the next section will describe these methodologies in the context of fundamental measurement.

## 1.2. Fundamental Measurement: Rasch Analysis

The idea of attempting to derive measures in the social sciences equivalent to measurement in physical sciences is not a new one. However, whereas the history of measurement in the physical sciences can be traced back as far as Aristotle's times (Michell, 1990), the idea of measurement in social sciences probably finds it origins in the early work by Fechner and others on "psychophysics" in the late 19$^{th}$ century (e.g. Nunnally and Bernstein, 1994). That this concept of measurement in the social sciences was critical can be seen by a quote from Cattell from that time:

> "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement", (Cattell, 1890, p. 373).

The modern, prevalent of notion what constitutes measurement in the social sciences (e.g. Michell, 1990, 1999) is encapsulated in Stevens's definition that:

> "...[W]e may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules", (Stevens, 1946, p. 677).

In its strictest sense, however, assigning numbers or numerals according to rules is not equivalent of measurement as conceived in the physical sciences, where the concept of measurement is taken to mean that the quantities[1] being measured are related by ratios expressed as real numbers, and that measurement is the attempt to uncover these numerical relationships (e.g. Michell, 1999).

Campbell (1920) drew the distinction between "fundamental scales", as characterised in the physical sciences by scales measuring concepts, such as length, weight and electrical resistance (Stevens, 1946), which Campbell refers to as "A-magnitudes" and are sometimes known as "fundamental magnitudes" (Reese, 1943), and "derived scales", which refer to measurement of concepts, such as density, which cannot be measured directly, but are derived from mathematical relationships between fundamental magnitudes (Stevens, 1946). These derived measures are referred to as "derived magnitudes" or "B-magnitudes" by Campbell (1920).

Campbell (1920) determined that in order for measurement in the social sciences to emulate that of the physical sciences the units of measurement must be able to be concatenated or added together, such as joining together rods to concatenate length or

1. There is additional debate surrounding the issue about whether psychological concepts are quantifiable, and indeed whether this "quantifiability" is a prerequisite for measurement (e.g. Michell, 1990, 1996; Nunnally and Bernstein, 1994).

bricks to concatenate weight (Campbell, 1920). This additivity principle was, for Campbell (1920) the *sine non qua* for "fundamental measurement", which would place measurement in social sciences on par with measurement in the physical sciences.

Other authors developed the work of Campbell and added further features to the requirements for satisfying fundamental measurement. For instance, to the requirement of additivity Fisher (1920) added the sufficiency condition or the notion of "sufficient statistics" for parameters estimated. This condition stipulated that for fundamental measurement to apply, or for a scale to act as a fundamental measure parameters needed to be estimated independently from other parameters in the model. A further requirement of fundamental measurement, the concept of "divisibility", where parameters are infinitely divisible and which forms the mathematical foundation behind both additivity and sufficiency was derived by Levy (Levy, 1937 as cited by Wright, 1997).

However, perhaps most of the development of the requirements for fundamental measurement was carried out by Thurstone and his colleagues in the 1920s and 30s (e.g. Thurstone, 1925, 1926, 1928, 1931; Thurstone and Chave, 1929). The conditions for fundamental measurement articulated by Thurstone included:

a). Unidimensionality: Measurement of any "object" or "entity" should only include one "attribute" of the object/entity being measured. Thurstone refers to this as the "universal characteristic of all measurement", (Thurstone, 1931, p. 257);

b). Linearity: Fundamental scales should be measuring an abstract, i.e. not directly measurable concept which is part of a "linear continuum" (Thurstone and Chave, 1929);

c). Invariance: The measurement process should be repeatable at different parts along the scale without modification to the scale (Thurstone, 1931);

d). Independence and Test-free measurement: The measurement instrument, e.g. items from a questionnaire, must be independent of the object of measurement, e.g. individuals responding to those questions (Thurstone, 1928, p. 547). Similarly, questions should be able to be omitted at different levels of the scale without affecting the individual score or measure. Furthermore, it should not be required to submit every subject to the whole range of the scale. The starting and end point should not directly affect the individual score or measure derived (Thurstone, 1926).

Guttman (1950) extended the requirements for fundamental measurement by adding the "conjoint transitivity" condition, where if an individual endorses a more extreme statement, they should also endorse all less extreme statements for the statements to be considered as part of a scale. Guttman (1950) referred to the items of the scale as having a "common content" if an individual with a higher score (or an individual ranked higher) than an other, scores as high or higher on every item of the scale.

There have been several attempts at developing measurement models that fulfil the requirements for fundamental measurement, including Thurstone's comparative judgment (Thurstone, 1927), multidimensional scaling, and unfolding theory (Coombs, 1964). However, in the 1960s the Danish mathematician, Georg Rasch, developed a series of probabilistic models which fulfilled the requirements for fundamental measurement, namely additivity and sufficiency (which Rasch referred to as "specific objectivity", e.g. Rasch, 1960/1980).

Rasch (1960/1980) using a dichotomous model (and the Poisson distribution) to estimate children's reading ability, realised that a person's success or failure on a test item can simply be defined as a function of their ability (B) and the difficulty of the item (D), which is expressed mathematically as:

$\mu(\theta) = \xi / \delta$, or $f(P) = b/d$, where P is the probability of a correct answer, f is a function of P, and b and d are the person ability and item difficulty estimates respectively (Rasch,

1960/1980, p. 118). This equation fulfils the condition for divisibility of the parameters (e.g. Levy, 1937 cited in Wright, 1997). Rasch then transforms this model into an additive model, thus fulfilling the additivity condition, by taking logarithms, so that

$$\log(f(P)) = \log(P/(1-P)) = \log b - \log d = B - D.$$

These two parameters can be estimated independently from the distribution of the responses by individuals to the items (person ability estimates) and the distribution of the responses across persons (item difficulties), in which way the sufficient statistics condition is met.

In addition to this Rasch (1960/1980) demonstrated that if P is the probability of success on a given item, and

$P = e^{(b-d)} / 1 + e^{(b-d)}$, then $P_{10}$ can be defined as the probability of person 1 succeeding and person 2 failing, and $P_{01}$ as the opposite, so that

$$P_{10} / P_{01} = e^{(b1 - b2)} \text{ or taking logs, } \log(P_{10} / P_{01}) = b1 - b2.$$

If d in this example is the difficulty of a given item it can be seen that the difficulty of the item drops out of the calculations, and that the distance between person 1 and 2 in terms of their ability remains the same regardless of the item difficulty, that is it is constant across the measurement continuum or invariant.

Subsequent developments of the model have included the development rating scale models (Andrich, 1978), as well as fit statistics, which test that the assumption of unidimensionality is being met by the items (Wright and Masters, 1982; Wright and Panchakesan, 1969).

The Rasch models (Rasch, 1960/1980) were developed in order to fulfil the conditions for fundamental measurement. In this respect they differ from classical test theory

(e.g. Guilford, 1954) where item difficulties and person ability estimates are derived from samples, and there is therefore no independence of parameter estimation, i.e. the parameters are both sample and test dependent.

Another set of models, which are often considered to conform to the requirements of fundamental measurement are item-response theory models (Birnbaum, e.g. in Lord and Novick, 1968). Indeed, Rasch models are frequently and erroneously defined as a subset of item-response theory models (e.g. Embretson and Reise, 2000; Lai et al., 2003). However, as Wright has pointed out (Wright, 1992, 1997) there are a number of shortcomings in item-response theory models, which mean they do not conform to the requirements of fundamental measurement.

Item-response theory models were developed from logistic models (e.g. Birnbaum, 1968). In addition to estimates for item difficulties and person abilities, similar to the Rasch model, item-response theory models contain parameters for item discrimination (referred to as the two parameter model) and item discrimination and a guessing parameter (the three parameter model). The introduction of the item discrimination parameter into the equation ensures that item-response models are not additive, do not have sufficient statistics and finally ensures that item and person parameters cannot be estimated independently (e.g. Wright, 1997). Consequently these models cannot be used as fundamental measures.

In this section the conditions for fundamental measurement were set out. In particular, two conditions namely independence of item and person parameter estimation and invariance across the measurement continuum were highlighted as critical features of certain probabilistic models know as Rasch models (Rasch, 1960/1980). It is these two features of the Rasch model which makes it attractive for computer-adaptive testing (Wainer, 1990). In the next section the development of computer-adaptive testing is described followed by the development of computer-assisted questionnaires in oncology.

### 1.3. Computer-adaptive testing

In classical test theory, the reliability of a test (Nunnally and Bernstein, 1994), that is the extent to which results from a test are repeatable or reproducible is contingent on test length with longer tests assumed to produce a more accurate reflection of an individual's "true score" (e.g. Nunnally and Bernstein, 1994). Reliability is often derived from test-retest studies or studies involving parallel forms of the test. The way in which reliability differs with test length can be derived from the Spearman-Brown (Prophecy) Formula (e.g. Suen, 1990).

The Spearman-Brown formula for dichotomous tests is given below:

New reliability = $\frac{n * r_t}{1+(n-1)*r_t}$, where n = ratio of new items on test to old items, and $r_t$ is the test reliability.

For instance, if we take reliability of a 10 item test to be 0.50 then doubling the number of items to 20 will increase the reliability coefficient to 0.67,

i.e. New reliability = $\frac{2*0.50}{1+0.50} = \frac{1}{1.50}$ = 0.67. Similarly, halving the number of items to 5, will decrease the reliability coefficient to 0.33, i.e.

New reliability = $\frac{0.5*0.5}{1+(0.5-1)*0.50} = \frac{0.25}{1+(-0.25)}$ = 0.33.

It this way it can be seen that in classical test theory test reliability and test length are co-dependent. However, since Rasch models can be employed to derive item-free estimates of person measurement it follows that reliability and test length are not co-dependent in the Rasch model, and that therefore computerised (adaptive) systems can be developed where item pools of items with known item difficulty parameters can be used to present fewer items to test-takers. This could lead to shorter tests (Wright, 1967), and potentially a more "accurate" reflection of the underlying latent trait derived from different sets of items.

Computer adaptive tests work on the principle of adapting the questions or items presented to patients to the responses made by the patients. Patients are usually presented with a item with the mean level of difficulty, or with the assumption of a mean or zero person ability and based on the response from the patient to this item the computer then selects the next item to present from a pool of items with known parameters (otherwise know as the item pool or item bank). This next item and successive items may either by "more difficult" or "easier", i.e. have greater or smaller item difficulty parameters depending on how the patient responds to previous items (Embretson and Reise, 2000; Hambleton, Swaminathan, and Rogers, 1991; Suen, 1990; Wainer, 1990).

Two methods are commonly employed for scoring individuals responses, namely maximum likelihood estimation (this procedure is described in more detail for the Rasch model in Chapter 3), and expected a posteriori estimation. Since maximum likelihood estimation can only be carried out once the patient has made a correct and an incorrect response, i.e. has endorsed and not endorsed items, the computer programme often increments or decreases the patient's ability estimate by a fixed level or step until these two conditions have been met (Embretson and Reise, 2000). Subsequent items are selected from the pool of items which maximise the likelihood function. Generally speaking the testing comes to an end once the standard error measurement falls below a certain predetermined level (Suen, 1990).

Item banks consist of large numbers of items, typically in excess of 100 items and possibly as many as 1000 items, which cover the full range of item parameters (Hambleton et al., 1991). However for test situations involving clinical decisions, e.g. whether a patient may be clinically anxious, items pools can be constructed using items with difficulty parameters clustered around the clinical cutoff point, providing greater test information around cutoff point. This form of computer adaptive test is referred to as "clinical decision adaptive testing " (Waller and Reise, 1989).

Computer adaptive tests can reduce the number of items presented to individuals by 50% (Embretson and Reise, 2000; Suen, 1990; Wainer, 1990). These tests therefore can

potentially reduce the burden placed on patients significantly, whilst maintaining or indeed improving the person ability estimates. However, constructing large item banks requires large data collection exercises, possibly including international collaboration between research groups (Revicki and Cella, 1997). It is perhaps given these facts that there has been little research published to date on the use of these tests in quality of life research. Although more recently Bjorner, Kosinski and Ware (2003) have reported the development of a computer adaptive test to assess the impact of headaches, and Lai and colleagues have reported the start of the development of a computer adaptive fatigue questionnaire (Lai et al., 2003).

### 1.4. Touchscreen technology and data collection

The process of collecting and disseminating quality of life information to clinicians has been greatly facilitated by advancements in technology. The majority of studies undertaken which have compared quality of life data collected by computers (such as touchscreen monitors, hand-held devices, etc.) have shown that the computerised or electronic questionnaires produce reliable and valid results. Furthermore, these studies have almost all demonstrated patient preference for the computerised questionnaires over the traditional pen-and-paper versions of questionnaires.

For instance, in a between subjects design by Schmitz et al. (2000), psychosomatic patients attending out-patient clinics were either assigned to an experimental group, who completed a computer-administered version of the Symptom Checklist (SCL-90-R) or a control group who completed the pen-and-paper version of the questionnaire. The results only demonstrated significant differences between subscales of the SCL-90-R (Anger-Hostility and Obsessive-Compulsive) with the experimental group scoring higher than controls on both of these subscales. The remaining differences between the test scores from both groups were small.

In a randomised cross-over trial, Pouwer et al. (1998) compared the data from two questionnaires, the Well-being Questionnaire and the Diabetes Treatment Satisfaction

Questionnaire completed by diabetic outpatients. Patients completed the questionnaires on computer and by pen-and-paper one week apart. The order for the mode of presentation was randomised. The results from the study demonstrated no significant differences between the majority of scores from the different methods of presentation. Two items from the Well-being Questionnaire were significant, however the authors concluded that this could be explained as a "chance-finding" (Pouwer et al., p. 37), given that the test-retest reliability indices for all items were high (range 0.61 – 0.85). Thirty-nine percent of patients reported a preference for the computerised questionnaires, compared to 46% who reported no preference.

Patient preference for computerised questionnaires was also explored by Drummond et al. (Drummond, Ghosh, Ferguson, Brackenridge & Tiplady, 1995) in a study with patients recruited from a gastro-intestinal clinic. The data were collected using hand-held computers in a randomised cross-over study where patients completed two questionnaires (Gastro-intestinal Symptom Rating Scale, and the Psychological General Well-being Index). The data collected from both forms of questionnaire were very similar, and the majority of patients (57%) expressed a preference for the electronic questionnaire.

More recently, Ryan et al. (Ryan, Corry, Attewell, Smithson, 2002) demonstrated preference ratings of 71% in favour of a computerised version of the SF-36 in a randomised cross-over controlled trial with chronic pain sufferers and healthy controls. Patients and healthy controls completed both versions of the SF-36 on the same day with a five minute interval between presentations. There was no significant difference in time taken to complete the two versions of the questionnaires, although there was an order effect for administration mode, with the first questionnaire in either format being completed slower than the second. There were no statistical differences between scores from the two versions of the questionnaire when Type I error rates were controlled for multiple testing (Bonferroni correction). In addition, exact agreement between the questionnaires ranged from 64% to 93%.

Similarly, in a randomised crossover study of patients with gastro-oesophageal reflux disease Kleinman et al. (Kleinman, Leidy, Crawley, Bonomi and Schoenfeld, 2001) found high levels of reliability and validity for the versions of two quality of life questionnaires (QOLRAD and SF-36) completed either on a touchscreen computer or by pen-and-paper, i.e. high Cronbach's alpha and inter-class correlations for both versions of the questionnaire, as well as moderate to good correlations between the two questionnaires completed in both modes of administration (0.65 – 07.32).

In a separate study which made use of internet technology, Bliven et al. (Bliven, Kaufman, and Spertus, 2001) patients attending a cardiology clinic completed both pen-and-paper and a computerised version of the RAND-36, a general health status measure, on a touchscreen computer. A subset of these patients, with coronary heart disease, also completed the Seattle Angina Questionnaire. The database used for collecting and storing the data was linked to the hospital intranet to allow instant access to the data for clinicians.

No difference was found in completion time between the two versions of the questionnaires when the time taken to complete the secure log-in procedure for the computerised questionnaires was controlled for. There were high correlations between the two versions of the scales from the Seattle Angina Questionnaire (0.84-0.93), and moderate to good correlations between the two versions of the domains of the RAND-36. In addition to these results 89% of patients participating in the study reported a preference for the computerised questionnaire. This preference was, amongst other things, not significantly related to age, sex or patients' previous computer experience.

The results described above from studies involving general medical and psychiatric outpatients and computerised questionnaires have to some extent also been replicated in oncology. One of a set of two series of studies employing computerised questionnaires reported in oncology was by Taenzer et al. (Taenzer, Speca, Atkinson, Bultz, Page, Harasym, and Davis, 1997). In the first study, which investigated the feasibility of using computers in data collection, breast cancer patients were asked to complete the EORTC-QLQ questionnaire on computer. Patients reported finding the questionnaire easy to use

(89%), and easy to understand (92%). In addition, 80% of the women indicated that they liked this method of data collection.

Taenzer et al. (1997) carried out a second study, which was a randomised control trial with a second sample of breast cancer patients who completed both the computerised version of this questionnaire and the paper version. Results from this study demonstrated high correlations between the two versions of the questionnaire for all subscales (> 0.65) with the exception of one subscale, namely Dyspnoea. Additionally, the mean percentage exact agreement for all items was 89%.

Buxton et al. (Buxton, White and Osoba, 1998) explored the feasibility of collecting quality of life data from a heterogeneous sample of cancer patients using touchscreen computers. As for the Taenzer et al. (1997) study patients in this study were asked to complete the EORTC QLQ-c30 questionnaire. The median time taken to complete the questionnaire was 5 minutes, and the majority of patients found the system easy to use. Ninety-six percent of patients responded that they were willing to complete similar questionnaires at future visits. These results were replicated in a study with a small sample of cancer patients by Carlson et al. (2001), and for a larger sample of cancer patients by Allenby et al. (2002). Similarly, Newell et al. (1997) also found high levels of acceptability of computerised questionnaires (the Hospital Anxiety and Depression Scale and the Cancer Needs Questionnaire) in a large sample (>250 patients) of cancer patients with heterogeneous diagnoses.

The issue of reliability, validity and acceptability to patients of computerised questionnaires has been extensively explored in a series of studies by our group (Cancer Research UK Psychosocial Oncology Group) at St. James's University Hospital, Leeds.

The main focus of research of our group has been and is on the development, evaluation and introduction into clinical oncology practice of patient-centred measurement of symptoms, functioning, emotional distress, social problems and quality of life. The purpose of this is to inform and enhance the interaction between cancer patients and health care professionals. The programme has evolved over many years, starting from the development

and assessment of computer touch-screen technology, through to evaluating patient compliance and moving towards introducing the measurement in clinical practice and measuring the benefits to the process of care and patient well-being.

The first study carried out to investigate the feasibility and reliability of the computerised questionnaires was described in a paper by Velikova et al. (1999). In this study patients' responses to an electronic version of the HADS and the EORTC QLQ-c30, presented on a touchscreen computer, were compared to responses to the paper versions of these instruments. Patients completed the two forms of the questionnaires on the same day with approximately 3 hours between each questionnaire. The order of presentation of the questionnaires was randomised. Feasibility was assessed by analyses of covariance for order effects and mode-order interactions. In addition the time taken to complete the questionnaires was also recorded. Reliability was measured by mean differences between the modes of presentation. In addition patients' preference for either mode of questionnaire was also elicited.

A total of 149 patients completed the study. Just over half of the patients (52%) expressed a preference for the electronic questionnaires, compared to 24% who preferred the paper questionnaire. Another 24% expressed no preference for either method. On average patients completed the electronic questionnaires quicker (approx. 8 minutes compared to 10 minutes), although there was an order effect with patients completing the electronic questionnaire quicker if the paper questionnaires had been presented first. For the majority of the subscales of the questionnaires there were no statistically significant differences. However, there were significant differences between scores from the Emotional Functioning, Nausea and Vomiting, Fatigue and Appetite scales with patients reporting better emotional health, and fewer symptoms on the electronic version of the questionnaire compared to the paper questionnaire. These differences were small (<5%) and were within the Type I error rates.

An additional study was conducted (e.g. Velikova et al., 1999) to investigate the test-retest reliability of the electronic questionnaires. A total of 80 patients completed the

electronic versions of the HADS and EORTC QLQ-c30 on the same day with an approximate interval of 3 hours between each presentation. Reliability was measured as the exact agreement, as well as the global agreement (i.e. responses within one response category) between scores from the two presentations.

The results demonstrated very good agreement (global agreement > 0.80) between 13 of the 15 scales of the EORTC QLQ-c30, and the other two scales of the EORTC QLQ-c30 and the HADS subscales all demonstrated good global agreement (0.60 – 0.80).

A separate study (Cull, et al. 2001) has investigated the validity of using electronic questionnaires to detect levels of psychological distress amongst cancer patients.

The feasibility of using electronic questionnaires to capture quality of life data in oncology clinics was explored in a recent publication by Wright et al. (2003). In study 1 a consecutive cohort of patients attending oncology clinics in two centres (Leeds and Edinburgh) were recruited into the study and followed-up for a period of six months. The patients were instructed to complete the questionnaires on their own initiative when attending clinic.

The results from this study showed that an initial 272 patients (84%, 272/324 of those approached) consented to participate at baseline. However, this number quickly dropped and the median compliance over the six month period was 40%.

A second study (Wright et al., 2003) investigated levels of compliance of completion of electronic quality of life questionnaires when introduced as routine practice over a 12 week period (in Leeds and Edinburgh, as well as at district hospital, i.e. Airedale). In this study a total of 1271 assessments were completed out of the 1826 patients visits. The mean overall compliance was 72%, although this depended on clinic location with the nurse-led adjuvant chemotherapy clinic demonstrating the highest compliance (mean of 93%), compared to mean compliance of around 65% for the other clinics and hospitals. These studies clearly showed that high compliance with regular quality of life assessment can be achieved only if the procedure is fully integrated into routine patient care and the

measurement is performed as part of the usual clinic assessment (such as patient registration, blood tests, doctor review, etc.).

Velikova et al. (2002) carried out a pilot intervention study to investigate the impact of immediate feedback of quality of life information on the issues discussed in the medical consultation. A total of 28 patients completed the electronic versions of the HADS and the EORTC QLQ-c30 on a touchscreen computer on two occasions. Results of the questionnaires were only provided to the clinicians (in both graphical and numerical formats) at the second visit (intervention visit). After each visit patients' satisfaction with the visit, attitude to the quality of life data and the content of the consultations was recorded. Clinicians were interviewed after each consultation, as well as at the end of the study.

Overall there was an increase in the number of issues discussed in the intervention visit. Additionally, patients reported that they believed clinicians enquired more about their daily activities, emotional well-being and limitations in doing work or their leisure activities when the quality of life data was made available to clinicians. The majority of patients believed the questionnaires were useful to inform their clinicians of how they felt physically and emotionally, and most reported that they were willing to complete the questionnaires at each clinic visit.

The clinicians reported that the quality of life data enhanced communications with their patients and also contributed to some clinical management decisions (e.g. stopping chemotherapy, adjustment of symptomatic drugs, blood transfusions, and life-style counselling). However, the discussion of the quality of life results with the patients may have lengthened the consultation (by between 1 and 5 minutes), although this was considered acceptable by the clinicians.

The next logical step in this programme was to examine the effects of regular use of standard QL questionnaires in oncology practice on a larger scale (Velikova et al., 2004). In a randomised study including 286 patients and 28 oncologists, the hypothesis was tested that regular collection and transfer of QOL data to practicing oncologists may have positive impact on the process of medical care and may result in benefits for the patients. The study

employed a prospective randomised design with 3 groups - an intervention group (regular completion of EORTC QLQ-C30 and HADS questionnaire on touch-screen computer over 6 months and feeding back results to clinicians); an attention-control group (regular completion of QL questionnaires without feedback of information to clinicians); and a control group (no completion of QL questionnaires in clinics). Primary outcomes were patient well being, measured by Functional Assessment of Cancer Therapy-General questionnaire (FACT-G), and doctor-patient communication and clinical management, measured by content analysis of tape-recorded consultations. Secondary outcomes were other process measures (tests, drugs, medical records), continuity of care and patient satisfaction. The results of the study suggested an impact of the intervention on the content of patient-doctor communication with more frequent enquiry about non-specific symptoms, including fatigue (60% of intervention consultations vs. 42% of attention-control vs. 48% of control consultations), insomnia (31% vs. 8% vs. 18% respectively), lack of appetite (48% vs. 35% vs. 25%) and a trend for more frequent discussion of emotional issues (54% vs. 46% vs. 41%). No significant effect on patient management was found. Using mixed-effects modelling to analyse the longitudinal QOL outcomes data, a significant improvement in patient well-being over time was observed for patients in both intervention and attention-control group (who completed the QOL questionnaires on a regular basis) in comparison with the control group. No significant difference was found between the intervention and the attention-control group. The QOL differences between intervention and control group were clinically significant. Forty percent of the patients in the intervention group showed clinically meaningful improvement in QOL (FACT-G change > 7 points), in comparison with 32% in the attention-control and 24% in the control group. The number needed to "treat" for one patient to benefit was 4.2. Similar results, with main differences between the intervention and control, but not between intervention and attention-control, were observed separately for Physical well-being and Functional well-being. However, for patient Emotional well-being an effect was observed only for in the intervention group, but not in attention-control group. An improvement was found in patient perceptions of the continuity of their care. All patients were highly satisfied

with the quality of medical care and no between-group differences were observed (Velikova et al., 2004). It was concluded, that regular assessment of cancer patients' QOL had a positive impact on doctor-patient communication and resulted in benefits for some patients, who had better HRQL and emotional functioning. The feedback from both physicians and patients was generally very positive. 92% of patients indicated they would be happy to use the QOL measurement in their usual care. It should be noted, that 37% of patients felt that some questions were irrelevant to their present situation.

In summary, all these studies have demonstrated that patients have had no problems using the touchscreen computer technology to answer QOL instruments. Indeed, the majority of patients preferred using the computerised questionnaires to the more traditional paper-and-pen versions. The touchscreen systems were at least as reliable as the more conventional methods of data collection and additionally could also be used for detection of psychological distress. Patient compliance was studied and it was demonstrated that it is feasible to use the computer touch-screen systems to generate data on a high proportion of large numbers of patients attending oncology clinics. However, better overall compliance was generated by an approach which incorporated data collection into routine clinical practice. In an intervention study regular measurement of QOL and feedback of results to oncologists had a positive impact on doctor-patient communication and patient emotional and overall well-being. The research work also identified problems related to the use of standard rather than individualised questionnaires, concerns about use of measures developed for group comparisons for monitoring of individuals, patient burden, possible floor and ceiling effects of the questionnaires.

The above studies span over a period of 8 years, during which I had a key role in the team in programming the touch-screen questionnaires and maintaining the equipment, participating in the data collection, analysis and writing of the manuscripts. My research interest evolved from this role and focused on investigating different approaches of using computer programming and modern statistical methods to make the QOL questionnaires less burdensome and more relevant to individual patients.

## 1.5. Hypothesis and Aims

In this thesis I have applied computer-based technologies and recently developed statistical methods (together with traditional methods) to evaluate the performance of three key Quality of Life questionnaires in cancer patients in order to identify ways of improving the questionnaires and optimising their use.

The aim of the thesis is twofold. The first aim is to increase the relevance of Quality of Life instruments to patients by developing systems that allow patients to select items or domains from questionnaires. The second aim is to reduce patient burden by reducing the number of questions presented to patients. This theme is approached in two ways and by two different methodologies in the thesis:

- Firstly, by using an experimental approach with the use of computer-assisted programmes, which allow patients to select areas of concern, and
- Secondly a more theoretical and statistical approach investigating the use of Rasch models to improve the questionnaires presented to patients by identifying and removing uninformative items and to provide the foundations for the development of computer-adaptive systems.

This work has provided a foundation to the wider programme of our group, which seeks to apply Quality of Life questionnaires in clinical practice in oncology and to the hypothesis that their application will improve patient care and well-being.

## 1.6. Structure of Thesis

The structure of the rest of the thesis is as follows:

1). Chapter 2 describes the three quality of life instruments used in the work and the development of a computer-assisted questionnaire for presenting these questionnaires to patients, which allows patients to select areas or quality of life domains which are of concern to them.

2). Chapter 3 describes the statistical methods employed in the thesis. In particular, the Rasch model, and the principal components analysis are described in detail. Additionally, factor analytic methods are contrasted with the Rasch analysis of residuals.

3). Chapter 4 describes an experimental study carried out comparing the responses of patients to a standard quality of life questionnaire (EORTC QLQ-C30) with their responses to the computer-assisted version of the same questionnaire allowing them to select areas of concern (as developed in Chapter 2). This study addresses the first aim of the thesis, using an experimental approach (as described above).

The following chapters address the second aim of the research using a more theoretical and statistical approach.

4). Chapter 5 describes the analysis of the Hospital Anxiety and Depression Scale using traditional statistical approaches and using Rasch models. The aim of this chapter is to compare the results derived from the two methodologies and to identify items from the questionnaires (through Rasch models) which could be potentially removed from the questionnaire.

5). Chapters 6 and 7 describe the analysis of the European Organisation for the Treatment and Research of Cancer Core questionnaire (EORTC QLQ-C30) and the Functional Analysis of Cancer Treatment questionnaire (FACT-G) respectively, using traditional statistical approaches and using Rasch models, and are similar in approach to Chapter 5. The aim of the chapters is to compare the results derived from the two methodologies and to identify items from the questionnaires (through Rasch models) which could be potentially removed.

7). Chapter 8 describes a study where the items from the HADS which were identified as misfitting (Chapter 5) were removed from the questionnaire. The sensitivity and specificity of the reduced HADS were then re-evaluated by comparing the area-under-the-curve (AUC) of the receiver-operating characteristic curves (ROC) against mental health diagnoses derived from an additional psychiatric interview (SCAN/PSE). The purpose of this study was to assess how much the screening efficacy of HADS was affected by shortening the instrument.

8). Chapter 9 describes a study comparing the agreement between two quality of life instruments (EORTC QLQ-C30 and FACT-G) using measures derived through Rasch analysis.

9). Chapter 10 is the concluding chapter and summarises the contribution of my work to the overall programme. Furthermore, it discusses the implications of the results from both theoretical and practical perspectives. It acknowledges some limitations of this research and looks at future directions of work.

## 2. Instruments and Software Development

This chapter describes the quality-of-life instruments employed in the subsequent studies and the development of the software for presenting the questionnaires on a touch screen computer.

Three questionnaires were included in this work, namely the Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983), the EORTC Quality of Life Questionnaire (Aaronson et al., 1993), and the Functional Assessment of Cancer Treatment (Cella et al., 1993). All three questionnaires were selected on the basis of their common and widespread usage in oncology for assessing patient quality of life (EORTC QLQ-C30 and FACT-G) and screening for psychological distress (HADS).

### 2.1. Instruments

#### *2.1.1. Hospital Anxiety and Depression Scale (HADS)*

Originally designed to assess psychological distress of patients in medical and surgical settings, the Hospital Anxiety and Depression Scale (HADS, Zigmond & Snaith, 1983) has now been evaluated and validated for different medical and psychiatric patient populations (Spinhoven et al., 1997; White et al. 1999), and non-medical populations (Dagnan, Chadwick & Trower, 2000; Lisspers, Nygren, and Söderman, 1997).

The HADS is a 14-item scale that requires respondents to endorse a verbal response which is scored as an index of the severity of anxiety or depression (see Appendix 1). The scores are then summed to produce two subscales corresponding to Anxiety (HADS-A), and Depression (HADS-D). In addition to the subscale totals, an overall total can be derived to indicate the level of psychological distress. Zigmond and Snaith (1983) advocated cutoffs between 8 and 10 for 'possible cases', and scores of 11 or more for 'definite cases'. The rates of prevalence of psychological distress reported using the HADS differ markedly. For instance, Hall et al. (Hall,

A'Hern, Fallowfield, 1999) report a rate of 13.5% for anxiety and 7.5% for depression in breast cancer patients using a cutoff of 11. These rates increased to 39.4% and 16.5% for anxiety and depression respectively using a threshold of 7. Whereas, Hopwood, Howell & Maguire (1991) reported that 27% of their sample of women with breast cancer had a probable case of affective disorder using a HADS threshold of 11. Similarly, Hopwood and Stephens (2000) reported levels of depression at 33% and anxiety at 34% in a sample of patients with lung cancer.

A number of studies have reported the efficacy of the HADS as a screening instrument for mental health problems (Abiodun, 1994; Hall, A'Hern, & Fallowfield, 1999; Hopwood, Howell, & Maguire, 1991; Ibbotson et al., 1994; Lewis & Wessely, 1990; Razavi et al., 1990; Silverstone, 1993; Spinhoven et al., 1997). These studies demonstrate that the HADS is a more consistent measure for detecting generalized anxiety disorders (sensitivity ranging from 59%-93%, and specificity ranging from 73%-90%), compared to depressive disorders (sensitivity ranging from 14%-90%, and specificity from 73%- 100%). The combined HADS scores perform similarly in detecting either depressive or anxiety disorders with sensitivity ranging from 20% to 92%, and specificity from 74% to 95%.

The original two-factor structure of the HADS, corresponding to the Anxiety and Depression subscales, has been confirmed by a number of subsequent studies (e.g. Dagnan, Chadwick, & Trower, 2000; Lisspers, Nygren, & Söderman, 1997; Moorey et al., 1991; Spinhoven et al., 1997, and White et al., 1999). Although two studies have demonstrated different factor structures (Andersson, 1993) and (Lewis, 1991), both these studies involved small sample sizes which may have contributed to a distorted factor structure.

More recent interest in the HADS has centred on its relationship to the tripartite theory of anxiety and depression (Clark & Watson, 1991), and whether the factor structure corresponds to this model. In a recent study Dunbar et al. (2000)

have proposed a three-level factor structure for the HADS corresponding to Clark and Watson's (1991) tripartite theory of anxiety and depression. Clark and Watson's model is comprised of three factors, psychological distress or negative affectivity (NA), autonomic anxiety, and depression as anhedonia. Dunbar et al. (Dunbar, Ford, Hunt & Der, 2000) suggest that four items from the Anxiety subscale (items 1, 5, 7, and 11) represent negative affectivity, and the other three items correspond to autonomic anxiety. Dunbar et al. consider the Depression subscale to correspond to anhedonia. The study collected HADS data from a large community sample. Although, confirmatory factor analyses suggested that there was evidence supporting the two-factor structure, the tripartite model provided a better fit of their data. In fact Dunbar et al.'s (2000) data suggested that the HADS conformed to a hierarchical model, as proposed by Clark, Watson and Mineka (1994), where the two secondary factors, anhedonia and autonomic anxiety, are subordinate to the higher factor, psychological distress or NA.

*2.1.2. European Organisation for Research and Treatment of Cancer Core Questionnaire version 3 (EORTC QLQ-C30)*

The EORTC QLQ-C30 was developed in the late 1980's and early 1990's as a modular approach to evaluating quality of life in clinical trials (Aaronson et al., 1993). The idea behind this modular approach to questionnaire design was to produce a core questionnaire covering the major domains in quality of life, e.g. physical, social and emotional which are relevant to cancer patients, and to supplement it with additional modules which are either diagnosis- and / or treatment-specific (Aaronson et al., 1993).

The first core questionnaire, the EORTC QLQ-c36 (Aaronson et al., 1991), was developed in 1987. After psychometric testing a number of non-informative items

were discarded, and the final form of the questionnaire, namely the EORTC QLQ-C30 was developed (Appendix 2).

The EORTC QLQ-C30 is a 30-item instrument that measures health-related QOL in five functional domains and one general quality of life domain (Physical, Emotional, Role, Social, and Cognitive Functioning, and Global Quality-of-Life), and seven symptom scales (Fatigue, Pain, Dyspnoea, Nausea & Vomiting, Constipation, Diarrhoea, Insomnia) and a scale relating to Finance. Patients' responses are scored on a four-point Likert scale (i.e. Not at all, A little, Quite a lot, Very much), for all scales with the exception of Global Quality-of-Life, which is scored on a seven point scale with two anchor points: 1 – "very poor" and 7 – "excellent". The time frame for the EORTC QLQ-C30 is the preceding week (except for the Physical Functioning scale). The raw scores are converted to summated scales which are scored from 0 to 100, where 100 indicates the best functioning for Functional Scales, but worst symptomatology for Symptom Scales.

The EORTC QLQ-C30 version 3, is the most current version of the core questionnaire, and differs little from earlier versions (such as version 1.0, 2.0 and +3) with the exception that the Physical Functioning scale (in versions 1.0, 2.0 and +3) and Role Functioning scale (version 2.0) are not scored dichotomously, and the wording of the Role Functioning scale, i.e. "Are you limited in any way in doing either your work or doing household jobs?" and "Are you completely unable to work at a job or to do household jobs?" (versions 1.0 and +3), was changed to "Were you limited in doing either your work or other daily activities?" and "Were you limited in pursuing your hobbies or other leisure time activities?". In addition, version +3 of the instrument also included a third question as part of the Global Quality of Life scale namely, "How would you rate your overall physical condition during the past week?" (Osoba et al., 1997).

The initial validation and psychometric assessment of the EORTC QLQ-C30 (version 1.0, Aaronson et al., 1993) was carried out on an international sample of 305 patients who completed the questionnaire before and during treatment. Reliability coefficients (Cronbach's alpha) were high (>0.70) for all scales in the treatment phase with the exception of the Role Functioning scale. Validity of the instrument was measured against clinical status, and demonstrated that patients with poorer performance status reported significantly worse physical, role and cognitive functioning, lower quality of life scores, and scored higher levels of symptoms for all symptom scales. Inter-scale correlations ranged from modest to good which led to the conclusion that the scales were assessing distinct components of the quality of life domains. Subsequent studies have replicated the validity and internal consistency and reliability of the instrument in a variety of cancer patient diagnoses, including heterogeneous samples (Osoba et al., 1997; Velikova et al., 1999), metastatic prostate cancer (Sharp et al., 1999), malignant melanoma (Sigurdardottir et al., 1993), small cell lung cancer (Bergman et al., 1994) and patients receiving palliative radiotherapy (Kaasa et al., 1995).

The factorial structure of the EORTC QLQ-C30 was investigated in a study by Ringdal and Ringdal (1993). The scalability of each of the scales, i.e. whether a particular structure is represented by the data, was evaluated using Mokken's scaling models (Mokken, 1982). In addition, the internal consistency of the instrument scales was also investigated using Mokken's scaling models (Mokken & Lewis, 1982), as well as Cronbach's alpha. The results of this study demonstrated that all scales with the exception of the Cognitive Functioning scale had good levels of scalability. Furthermore, the internal consistency was good for all scales except for the Role and Cognitive Functioning scales. Ringdal et al. (1999) have replicated these results in a large sample of cancer patients with varied diagnoses and treatments. All scales

demonstrated good levels of scalability. However, both the Role and Cognitive Functioning scales showed lower levels of reliability (0.63 and 0.64 respectively).

Subsequent work has assessed the test-retest reliability of the instrument. Hjermstad, Fossa, Bjordal and Kaasa (1995) measured the test-retest reliability of the paper version of the EORTC QLQ-C30 (version 1.0) in patients who were off-treatment and who were attending outpatient clinics. A total of 190 out 262 patients agreed to complete the questionnaire twice with an interval of four days between completion. Percentage agreement between the scores was high for all symptom scales (75%) with the exception of scores from the Fatigue and Pain scales (54% and 65% respectively). For the functioning scales percentage agreement was high for the Physical and Role Functioning scales. Agreement for other scales ranged from 51% (Global Quality of Life) to 69% (Cognitive Functioning). Correlation coefficients were uniformly high (>0.70) for all scales.

In addition to this reliability study, Velikova et al. (1999) have also demonstrated good levels of test-retest reliability for an electronic version of the EORTC QLQ-C30 (version 3.0). Patients completed the electronic questionnaire on a touchscreen computer with an optimum delay of around 3 hours between presentations. Percentage global agreement (responses within 1 response category, Velikova et al., 1999) ranged from 75% (Emotional Functioning) to 100% (for Physical Functioning, as well as the symptom scales, excluding Fatigue, Pain and Nausea and Vomiting). Additionally, all correlation coefficients were high (>0.75).

The initial development of modules included work on head and neck cancers (Bjordal & Kaasa, 1992; Bjordal et al., 1994) and lung cancer (Bergman et al., 1994), however a comprehensive list of modules has now been developed by the EORTC Quality of Life Study Group covering all major cancer sites (http://www.eortc.be/ql).

The EORTC QLQ-C30 is widely used in cancer clinical trials and is familiar to most oncologists through its application in trials and publication of results in the medical literature.

*2.1.3. Functional Assessment of Cancer Therapy – General Questionnaire (FACT-G)*

The FACT-G was originally developed by David Cella and colleagues using semi-structured interviews of patients and oncology professionals to generate instrument items (Cella et al., 1993). The items generated from this process were then subsequently evaluated and resulted in a final, 28-item version of the instrument (FACT-G). A factor analysis of the logit transformed scores revealed a six factor structure which was condensed by the research group into five factors corresponding to: Physical Well-being, Social Well-being, Emotional Well-being, Functional Well-being and relationship with doctor (Cella et al., 1993). These were summed to provide an overall or total score (Appendix 3). In addition to this each of the scales also contained a final question asking the patients how the individual scale affected their quality of life, e.g. "How much does your **PHYSICAL WELL-BEING** affect your quality of life?"

Psychometric analyses of the instrument demonstrated that Cronbach's alpha was high for the total scale (0.89) indicating high levels of reliability. Similarly, test-retest reliability coefficients ranged between 0.82 (Emotional Well-being and Relationship with doctor) to 0.88 (Physical Well-being). Test-retest reliability for the total score was 0.92.

There has been very little additional validation work carried out on the FACT-G. However, Winstead-Fry and Schultz (1997) conducted a validation study of the FACT-G (version 2) on a sample of 344 cancer patients living in rural areas (i.e. non-metropolitan) in the US. The factor analysis of the scores (transformed to logits, as

per Cella et al., 1993) revealed the same five subscales. Furthermore, Cronbach's alpha levels were within the same range as reported by Cella et al (1993).

Kemmler et al. (2002) investigated the structure of the FACT-G (version 2) using multidimensional scaling. This analysis revealed that most subscales, but particularly Physical and Social Well-being, as well as the Relationship with doctors scales, demonstrated high levels of consistency with items from each subscale clustering together. Items from the Functional Well-being scale showed higher degrees of scatter, and there was an amount of overlap between Emotional and Functional Well-being.

The current version of the FACT-G is version 4 and consists of four scales, Physical Well-being (PWB), Social & Family Well-being (SFWB), Emotional Well-being (EWB), and Functional Well-being (FWB) domains, which are rated on a five-point Likert scale (i.e. Not at all, A little bit, Somewhat, Quite a bit, Very much). The scales are derived by summing the raw scores, and range from 0 to 28 (or 0 to 24 for Emotional Well-Being). Higher scale scores indicate better health or functioning. The timescale for the FACT-G is the past 7 days. The "Relationship with doctor scale" has been removed from this version of the instrument. In addition, patients are not required to provide an evaluation of how each subscale has impacted on their overall quality of life.

A number of site- and disease-specific modules have been developed for the FACT-G and these include modules for anaemia and fatigue (Cella, 1997), colorectal cancer (Ward et al., 1999), breast cancer (Brady et al., 1997) and lung cancer (Cella et al., 1995).

Similar to the EORTC QLQ-C30, the FACT-G is increasingly used in cancer clinical trials, particularly in North America and most oncologists would be familiar with the instrument through participation in trials and publication of results in the medical literature.

## 2.2. Software Development

### *2.2.1. Overview*

The purpose of this part of my work was to design a computer programme to present patients with the thirty questions from the EORTC-QLQ C30 (version 3.0) on a touchscreen computer and also to present a screen to patients which enabled them to select scales from the standard question that had been problematical in the last week.

The programme was designed to be capable of presenting both versions of the questionnaire to the patients and be able to store patients' responses and simple demographic details, such as surname and unique hospital identifier in an MS-Access database.

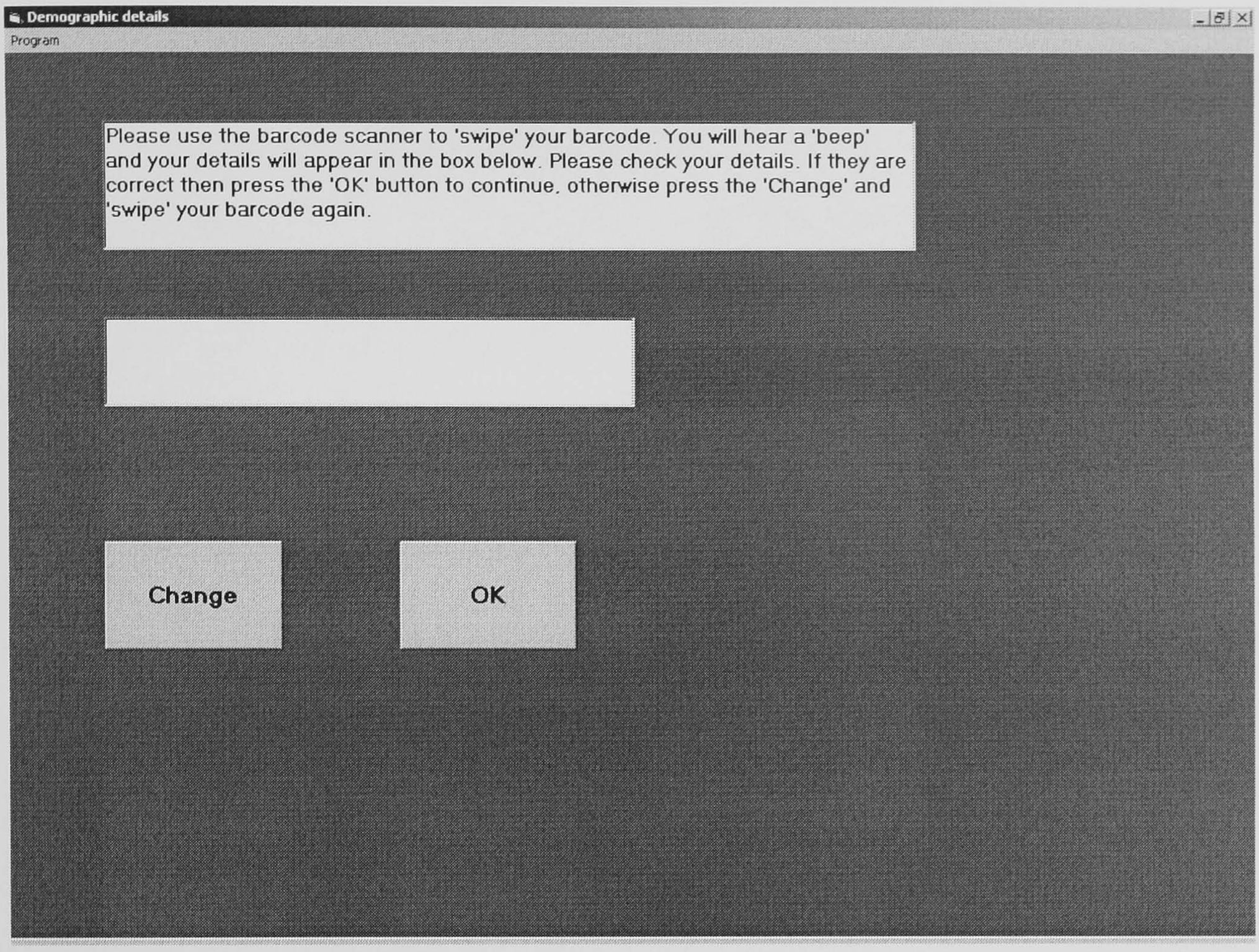
### *2.2.2. Tools*

In order to facilitate programming Microsoft Visual Basic 6 was used to write the software for the computer programme. This was linked to a Microsoft Access 97 database, which was used to store the data generated by the patients.

### *2.2.3. Design*

A total of five screens were designed for the programme: 1). A start-up screen, containing brief instructions on how to enter the patient's details using the barcode scanner, and a textbox for the patient's details, along with a 'Change' button to clear the textbox and allow patients to re-enter their details and an 'OK' button to continue with the programme (Fig. 2.2.1);

**Figure 2.2.1. Instruction screen for entry of patients' details**




2). An introduction screen, which for the standard questionnaire contained the instructions and description provided by the EORTC, and for the selection-questionnaire contained the standard description and an additional set of instructions, informing them how to select the relevant scales and also that they were not obliged to select anything, if they had no problems (Fig. 2.2.2 and 2.2.3 respectively);

**Figure 2.2.2. Introduction screen for standard version of the EORTC QLQ-C30**

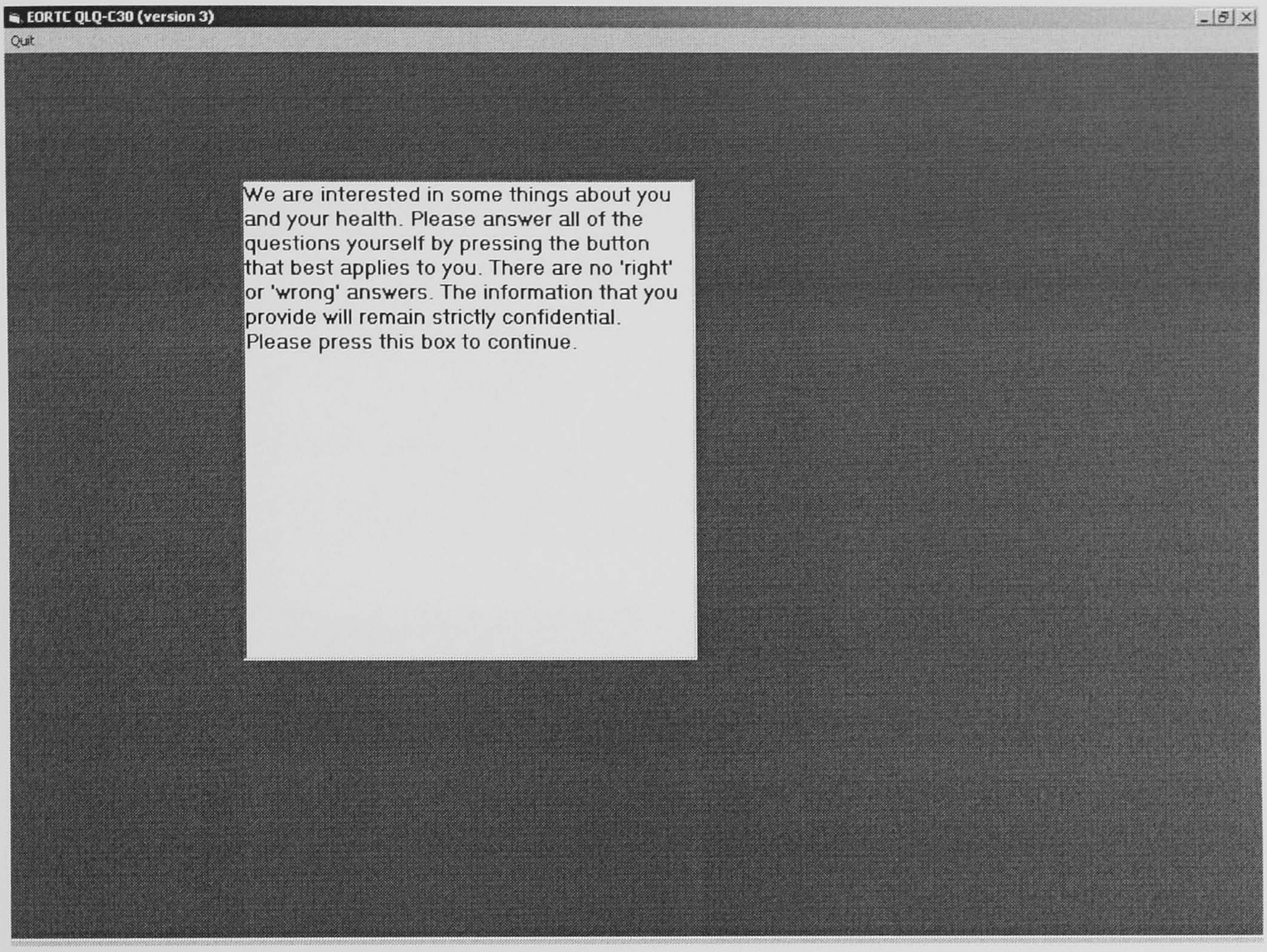
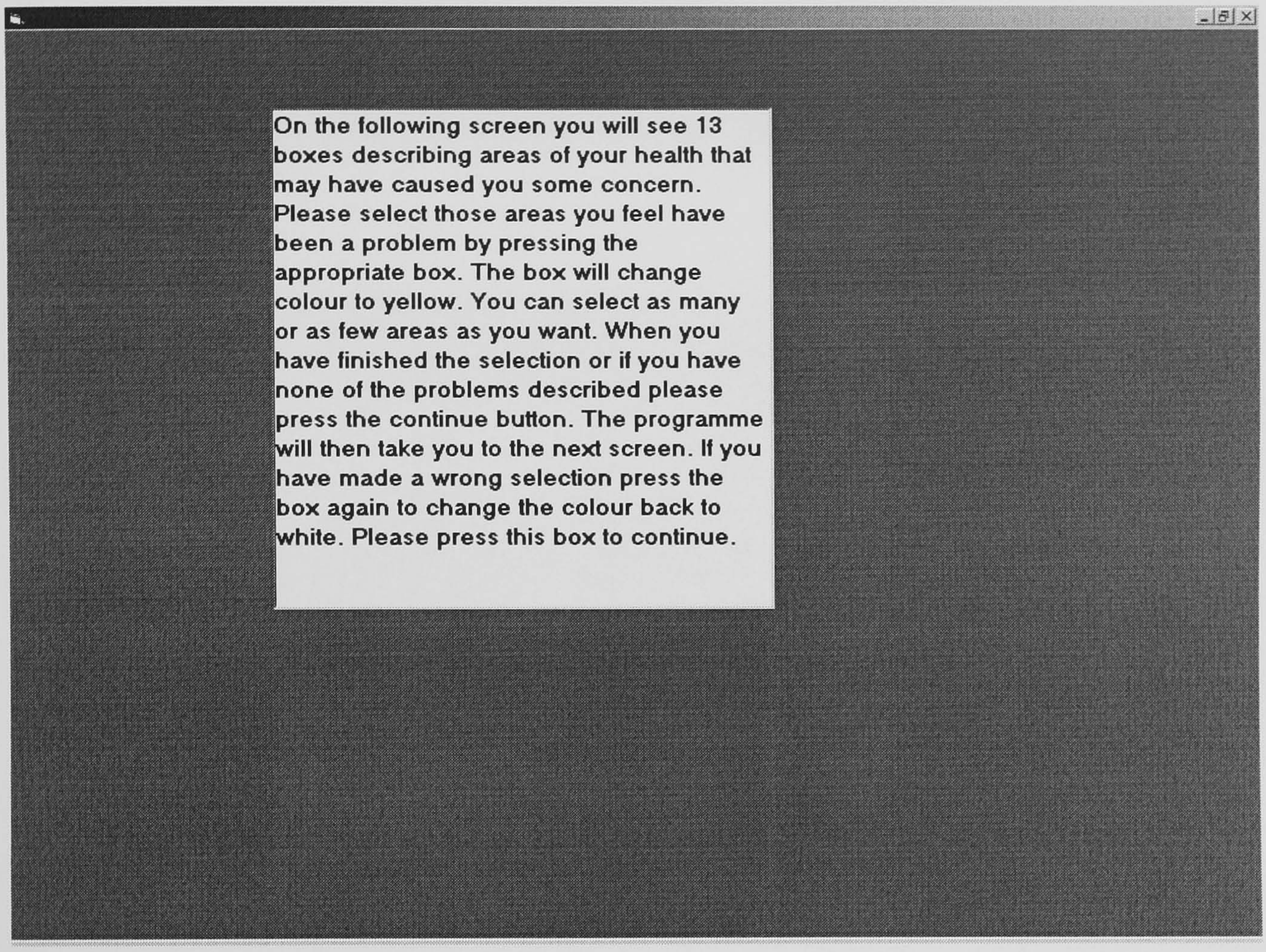
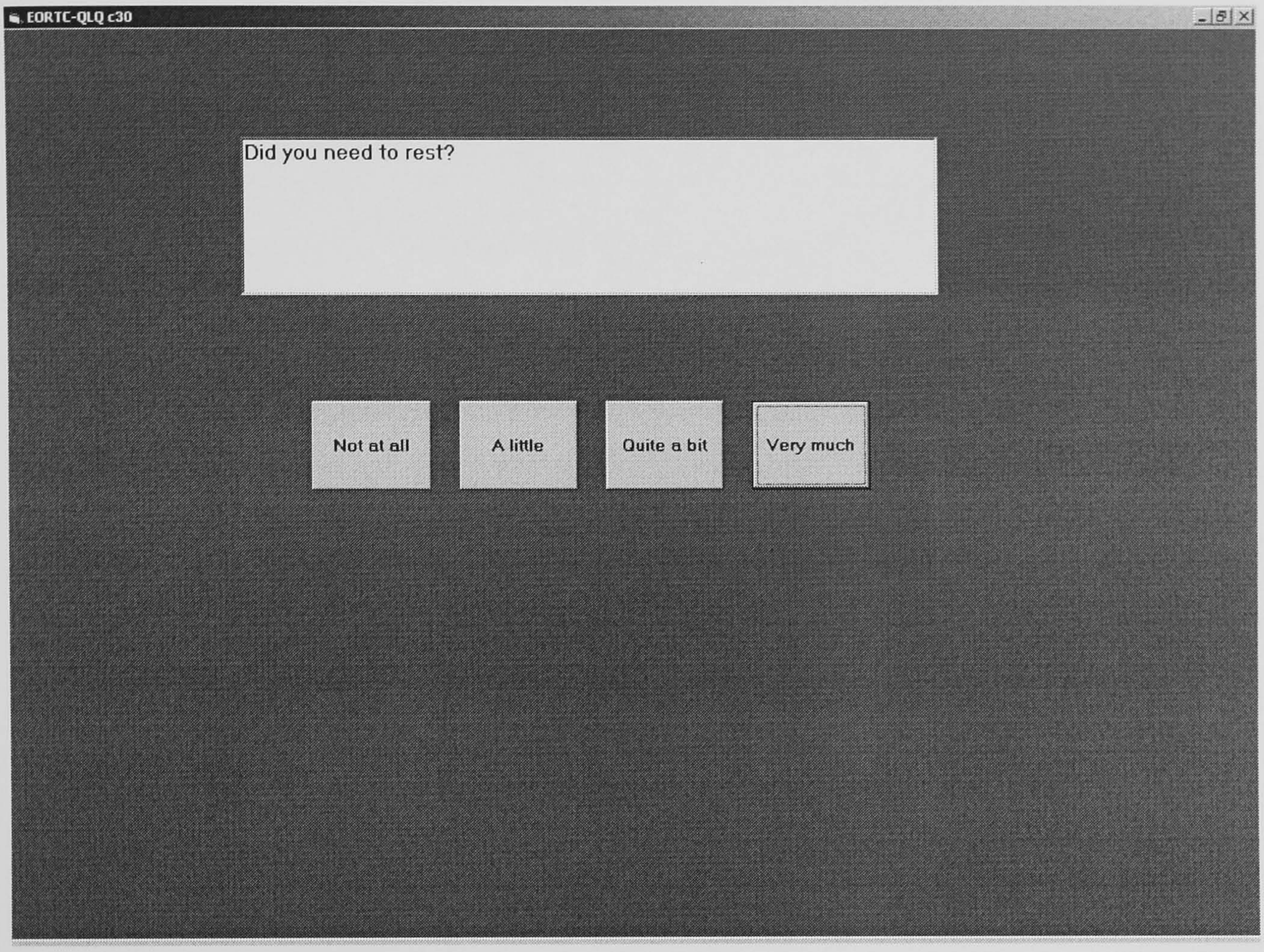


**Figure 2.2.3. Introduction screen for computer-assisted version of the EORTC QLQ-C30**



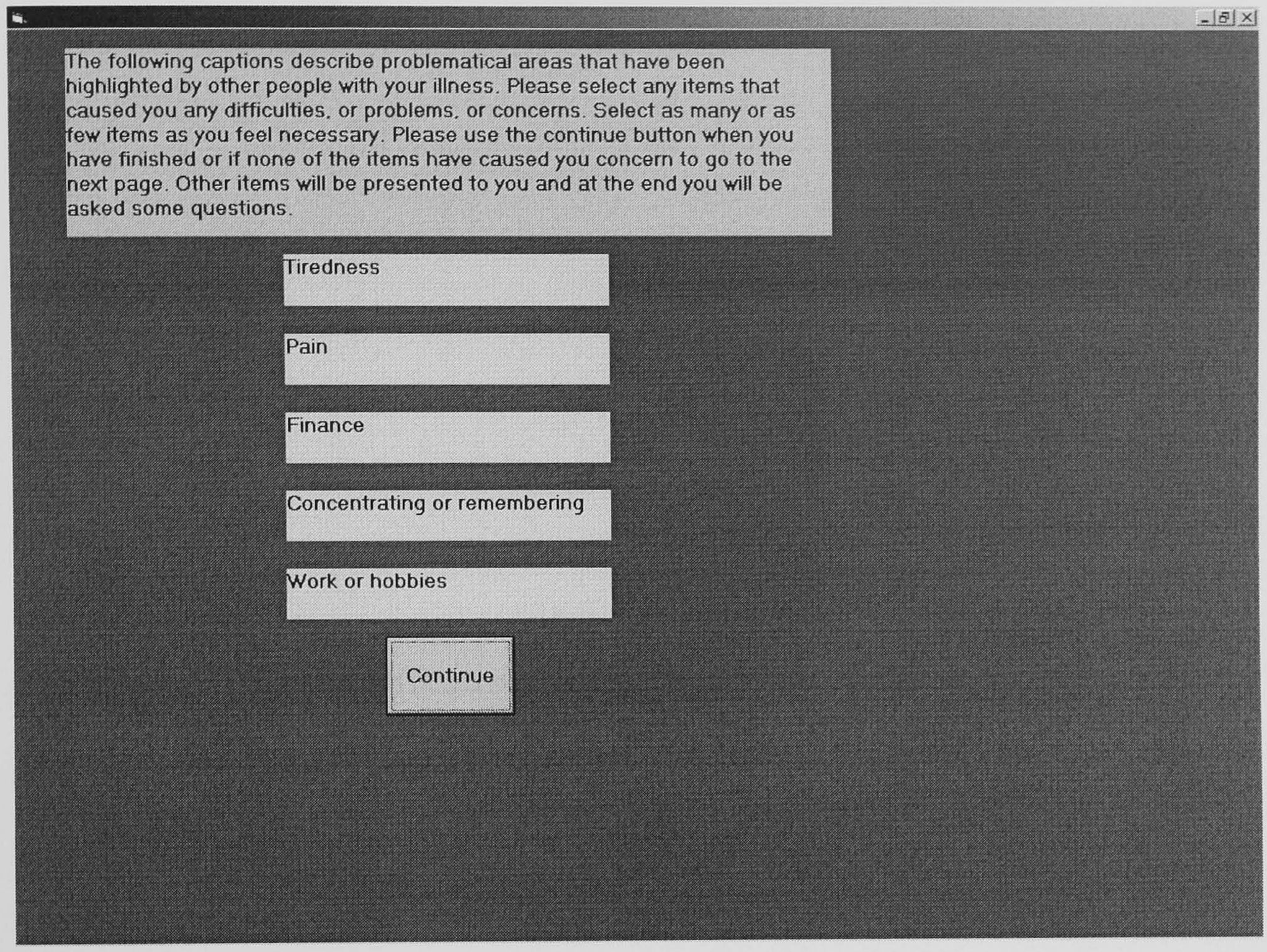

3). A screen was designed to present the questions clearly to the patients and consisted of a single textbox for the questions and row of four/or seven buttons (Fig. 2.2.4), which were labelled (Not at all, A little, Quite a bit, Very much; and 1 to 7 for Global Health questions);

**Figure 2.2.4. Screen following selection of "Tiredness"**



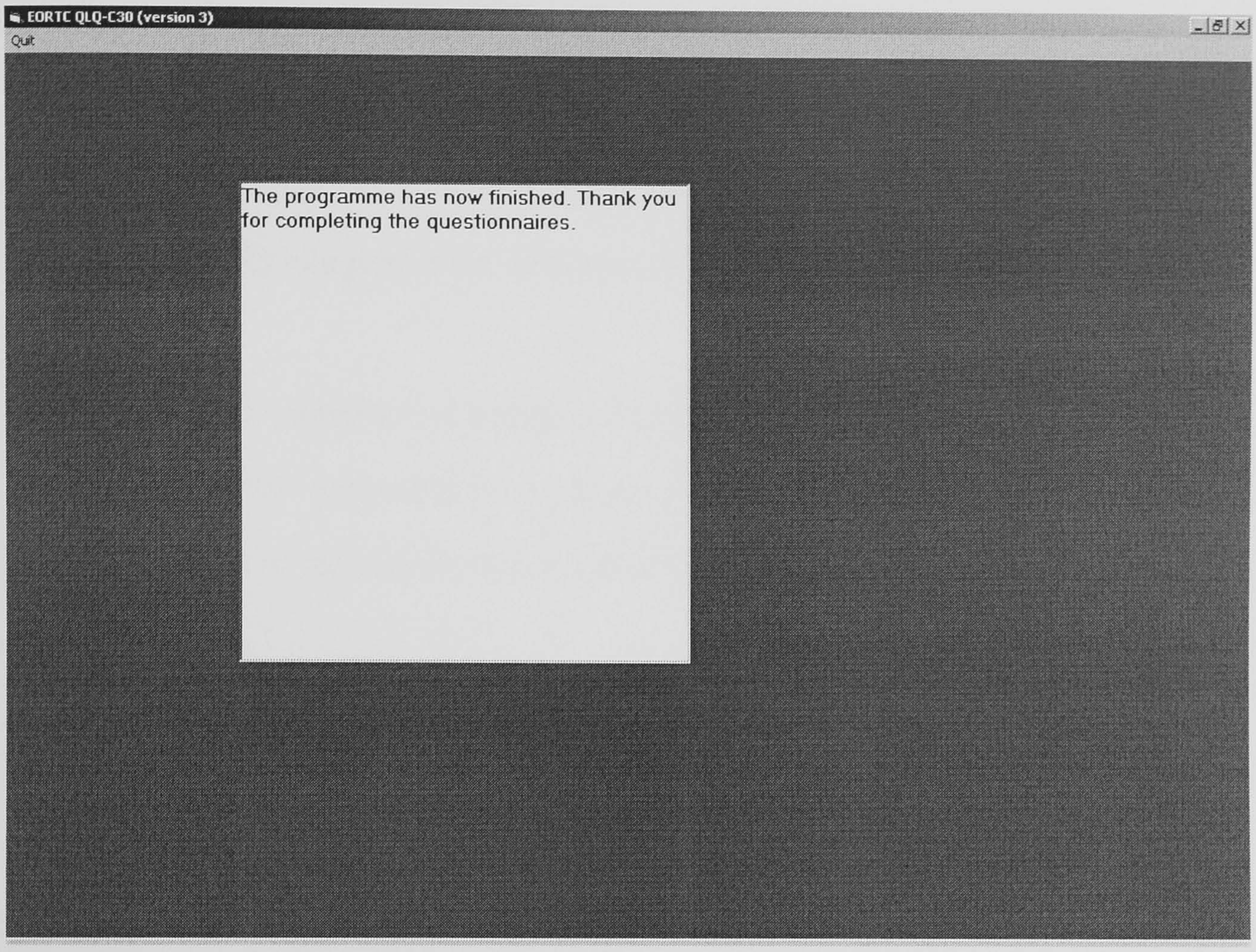

4). A screen was also designed (Fig. 2.2.5) to allow patients to select from the thirteen EORTC-QLQ scales (Anxiety and Depression were combined from the Emotional Functioning, Nausea and Vomiting symptoms were also combined, and Constipation and Diarrhoea were combined as well). The labels relating to these scales contained a brief description, e.g. 'Difficulties with physical activities' (Physical Functioning), 'Pain', or 'Financial difficulties', and changed from white to yellow when selected by the patients (or back to white when double-clicked).

**Figure 2.2.5. Selection screen from the CA-questionnaire**




A final screen was designed to thank patients for completing the questionnaire (Fig. 2.2.6).

**Figure 2.2.6. "Thank you" screen**




The type of questionnaire could be chosen from a pull-menu in the left-hand corner of the start-up screen, and the programme could also be closed down using the Quit option from this menu.

The font, MS-Sans Serif, was chosen to be easily readable by the patients, and similarly the font size 14 and bold were also deemed to facilitate reading. The questions were presented in textboxes with white background, against a screen background of dark blue/green for contrast.

Patients were presented on both versions of the questionnaire with the two questions from the Global Health Scale of the EORTC-QLQ (i.e. How would you rate your overall **health/quality of life** during the past week?) as a measure of reliability of responses.

An MS-Access 97 database was created to store the data generated from patients' responses to the questionnaires and to calculate the subscales' scores.

### *2.2.5. Coding*

The coding for the programme is described in detail in Appendix 4.

The next chapter describes a series of studies comparing the standard form of the EORTC QLQ-C30 presented on a touch screen computer against a computer-assisted version of the questionnaire, which allowed patients to select specific areas of concern.

# 3. Statistical Methods

This chapter describes the statistical methods employed in the analysis in Chapters 5 to 9. The Rasch model and the principal components analysis are discussed in detail. Factor analytic methods are contrasted with the Rasch analysis of residuals. More details on specific analyses (such as, Area-under-the-curve, specificity and sensitivity calculations, and correlations) are provided in the subsequent chapters.

## 3.1 Rasch Models

In traditional test theory item difficulty, e.g. the probability of subjects responding yes or no to items, or selecting a category from a number of response options, is calculated from the number of responses or proportion of responses in the sample (Suen, 1990). In other words a p-value is calculated, where p reflects the ratio of responses to a given option over the total number of responses. In this instance a high p-value would indicate an "easy" item, since most subjects were able to answer or endorse it, whereas a low p-value would indicate a "difficult" item. However, the major drawback of this approach is that estimation of item difficulty is sample dependent: the p-value for any given item will be larger if drawn from a more able population (e.g. a healthier population), than if drawn from a less able population. A similar approach can also be applied to estimating person ability (e.g. quality of life, physical health). Any given estimate of an individual's ability on a latent (i.e. not directly observable) trait will be dependent on the range of difficulties of the items presented.

Rasch models (Rasch, 1980) overcome this problem of sample dependency by estimating person ability ($\beta$ or B) and item difficulty ($\delta$ or D) independently (Wright and Masters, 1982). The raw data are the sufficient statistics for estimating these parameters, that is the models only use the raw scores from individuals for estimating item difficulties, and the response sets across items for person ability estimates

(Wright and Masters, 1982). In order to achieve the separation of item and person parameter estimations, the Rasch models rely on two assumptions, namely: unidimensionality and local dependence.

Rasch models assume that a uniform latent trait or construct underlies the data being investigated (McNamara, 1996), e.g. mathematical knowledge, physical health. This assumption is then tested using fit statistics and / or principal components analysis of residuals (see below). Local independence is related to unidimensionality, and refers to the assumption that the single latent trait (i.e. the unidimensionality) accounts for all the variance in the data, that is the association between the variables in a dataset should disappear once the Rasch model has been controlled for (Bond and Fox, 2000). It is possible to have unidimensionality, but not local dependence, however if local independence is proven then there must also be unidimensionality in the data set.

If the assumptions have been met, then the (log) probability of a person responding to an item can then be expressed as the difference between the individual's ability (B) and the item difficulty (D). For instance, for an item with a dichotomous response, the probability of answering yes (or 1), rather than no (or 0) can be expressed as:

Log(Pni1 / Pni0) = Bn – Di, where Pni1 is the probability of responding "yes", Pni0 is the probability of responding 0, and Bn is the person's ability estimate and Di is the item difficulty estimate.

In total, there are five Rasch models (Wright and Masters, 1982). The following sections describe these Rasch models in more detail, and the equations underpinning the item and person separation, as well as the calculations utilised for deriving the estimates[1].

---

[1] All figures and formulae are taken from Wright & Master's "Rating Scale Analysis" (1982).

### *3.1.1. Dichotomous Model*

The dichotomous model is simplest Rasch model. This model was originally developed by Georg Rasch (1980) for analysing dichotomous data, i.e. data in which only two responses are available, such as pass/fail, yes/no, 0/1, etc. The dichotomous model is shown in Equation 3.1.1.

**Equation 3.1.1 – Dichotomous Rasch Model**

$$\Phi n1 = \frac{\pi_{ni1}}{\pi_{ni0}+\pi_{ni1}},$$

where $\pi_{ni0}$ is the probability of scoring 0, and $\pi_{ni1}$ is the probability of scoring 1.

The general form is given by Equation 3.1.2

**Equation 3.1.2 – General Form of the Dichotomous Rasch Model**

$$\Phi ni1 = \frac{\exp(\beta_n-\delta_{i1})}{1+\exp(\beta_n-\delta_{i1})},$$

where $\Phi ni1$ is the person n's probability of scoring 1 rather than 0 on item i, $\beta n$ is the ability of person n, and $\delta i1$ is the difficulty of the one step in item i.

Figure 3.1.1 shows the item operating curve for the dichotomous model. It can be seen from this figure that the probability of responding increases as ability (β) increase along the x-axis.

**Figure 3.1.1. Item Operating Curve for a Dichotomous Item**

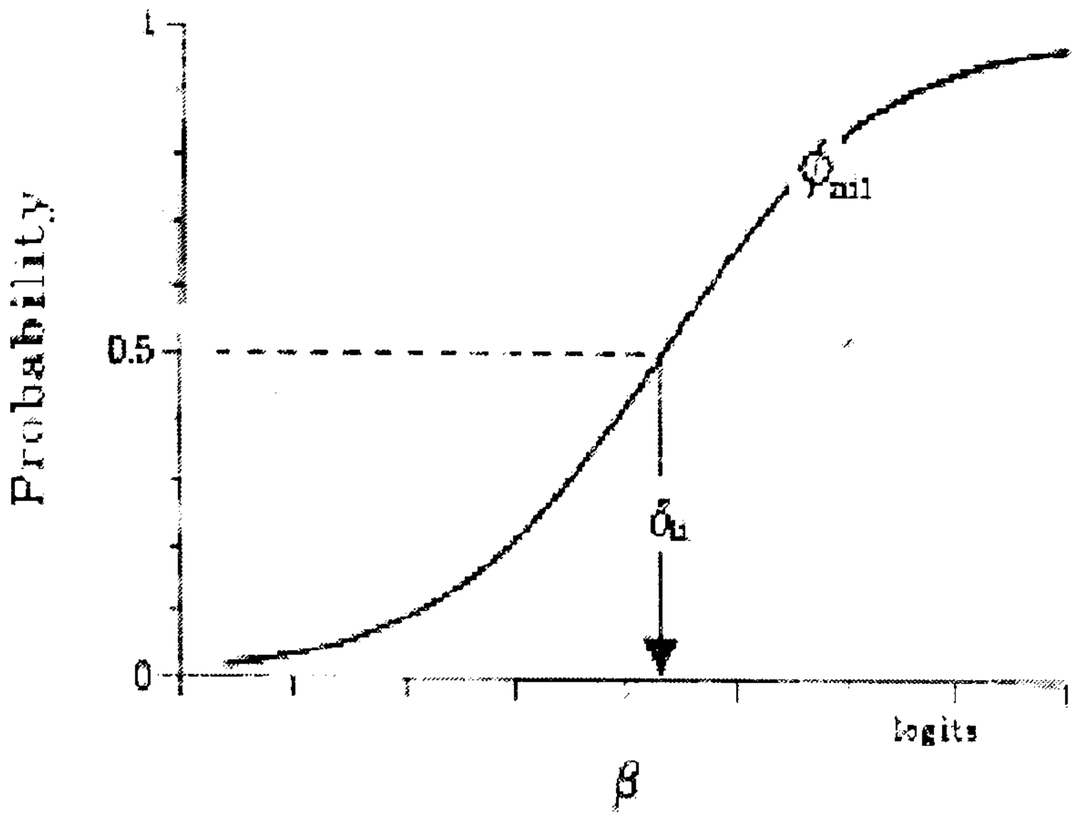
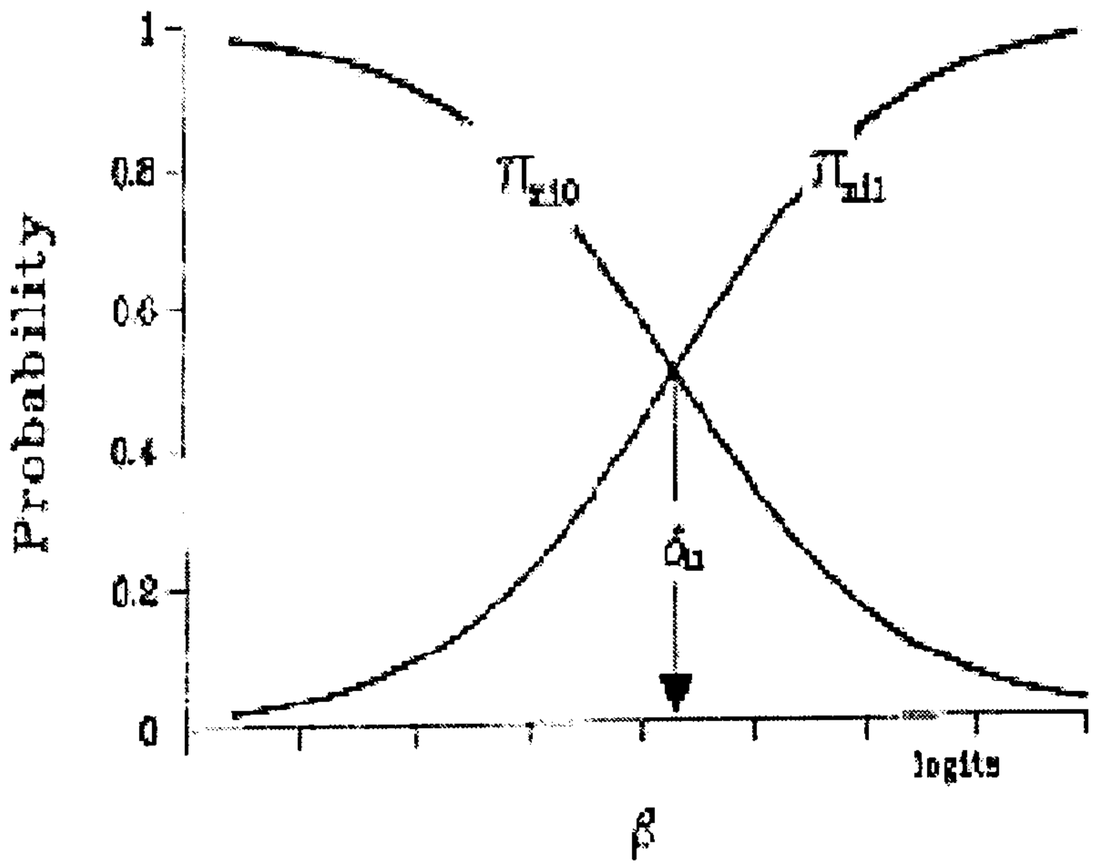



Figure 3.1.2 shows the category probability curve for the dichotomously scored item. It can be seen from the graph that the probability of responding 0 to the item ($\pi$ni0) decreases with ability ($\beta$), and conversely that the probability of responding 1 ($\pi$ni1) increases. The intersection of the two curves determines the item difficulty, $\delta$.

**Figure 3.1.2 Category Probability Curves for a Dichotomous Item**



*3.1.2. Partial credit model*

The Partial Credit model is used for instances where an individual may receive credit for completing part of a response, where they may not have completed all of the responses (Wright and Masters, 1982).

The partial credit model can be broken down into steps or dichotomies, per subset of the overall response, and can therefore be considered as a specific form of the dichotomous model. For instance for a two-step response, the probability of responding to the first and second step are given as below respectively:

**1).** $\Phi ni1 = \exp(\beta n - \delta i1)/1 + \exp(\beta n - \delta i1)$

**2).** $\Phi ni2 = \exp(\beta n - \delta i2)/1 + \exp(\beta n - \delta i2)$

The difference between this model and the dichotomous model is that $\pi_{ni0} + \pi_{ni1} < 1$. In other words the sum of the probabilities of the steps is less than or equal to one, and that there is more than one step:

Therefore the General Partial Credit Model is given by Equation 3.1.3:

**Equation 3.1.3 – General Partial Credit Model**

$$\Phi nik = \exp(\beta n - \delta ik)/1 + \exp(\beta n - \delta ik)$$

for k = 1, 2,….,mi number of steps of difficulties $\delta_{i1}$, $\delta_{i2}$, $\delta_{i3}$,…$\delta_{im}$.

The requirement of the model is that at least one score must be made by the person of the $m_i + 1$ possible scores on item i

i.e. $\sum_{k=0}^{m_i} \pi_{nik} = 1$

The probability of person n scoring x on item i is then shown below (Equation 3.1.4),

**Equation 3.1.4 – Probability of Scoring of on Partial Credit Model**

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta n - \delta ij)}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta n - \delta ij)}$$

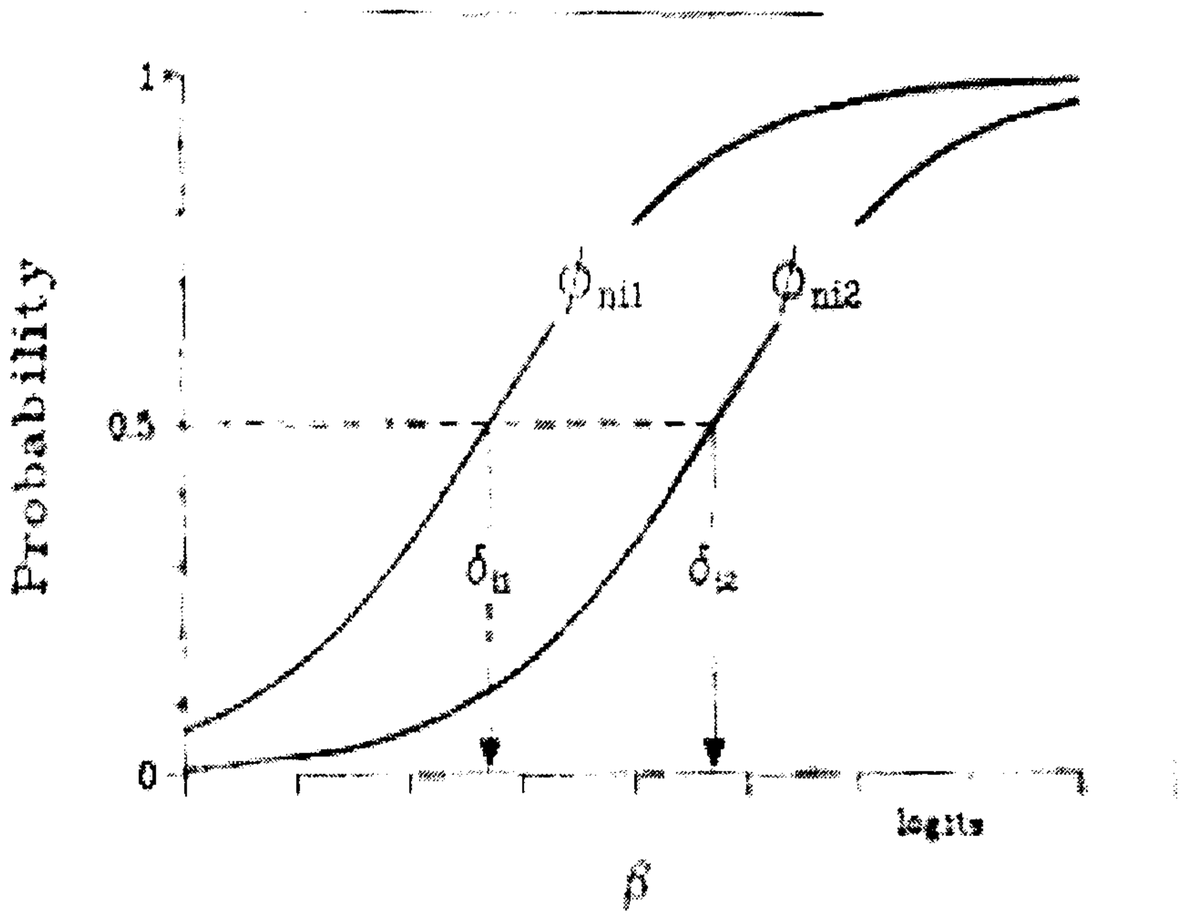
for x = 0, 1, …,$m_i$ ,where x is the count of the completed steps, where

$\delta i0 \equiv 0$, therefore $\sum_{j=0}^{k} (\beta n - \delta ij) = 0$, and $\exp \sum_{j=0}^{k} (\beta n - \delta ij) = 1$.

The numerator describes the difficulties ($\delta_{i1}$, $\delta_{i2}$, $\delta_{i3}$,…$\delta_{ix}$) of the number of x completed steps, whereas the denominator describes the sum of all possible steps ($m_i + 1$).

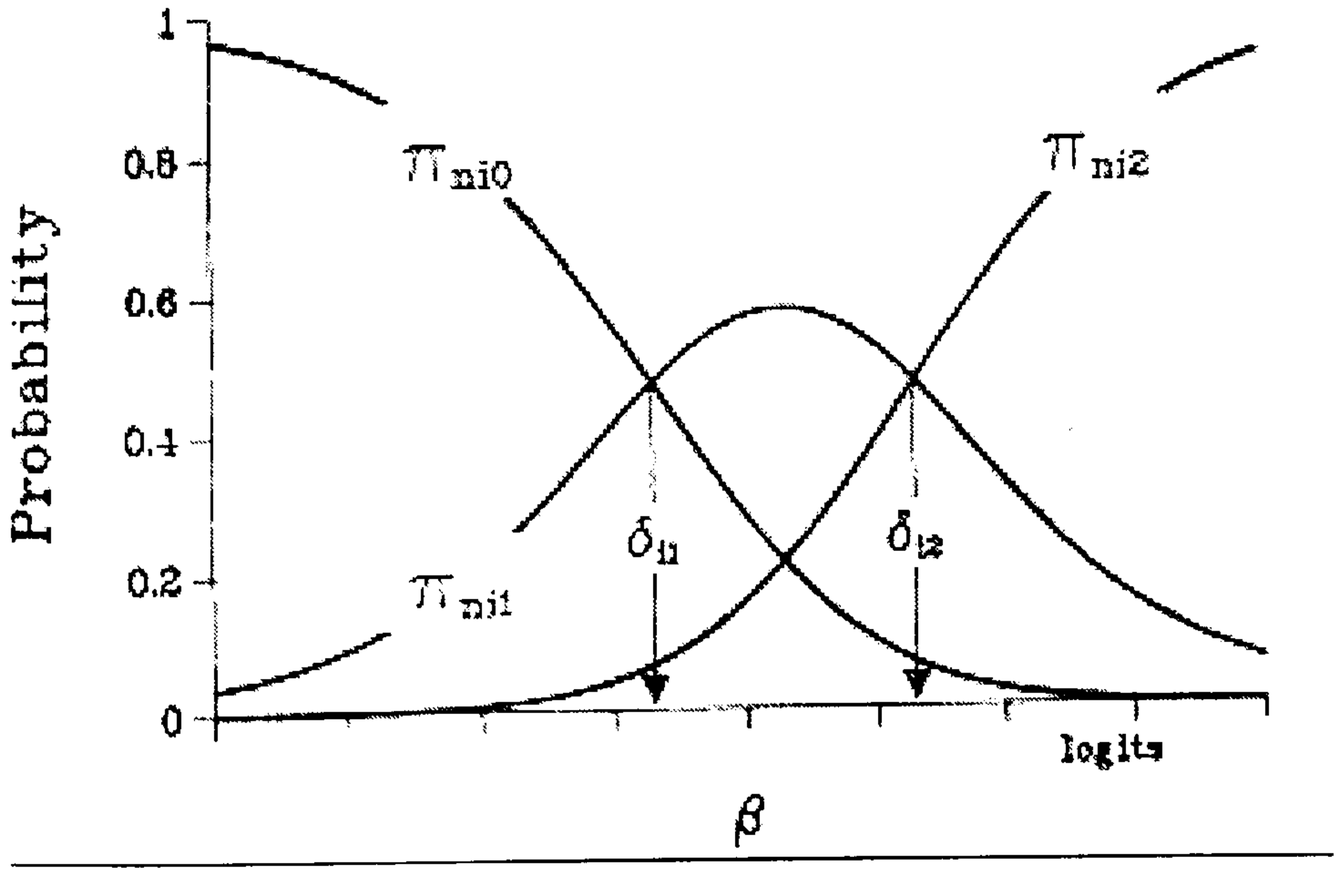
Figure 3.1.3 shows the item operating curves for a partial credit model for two steps. It can be seen that this is an extension of the dichotomous model, and that the probability of responding to either category increases as the person ability increases.

**Figure 3.1.3. Item Operating Curve for a Two- Step Partial Credit Model**




Similarly, from Figure 3.1.4, the category probability curve, it can be seen that the probability of completing either step 1 or 2 increases with ability. The intersection of πni0 and πni1 and πni1 and πni2, i.e. the probability of completing step 1 rather than 0, and step 2 rather than 1, form the difficulties for step 1 and 2 respectively.

**Figure 3.1.4 category Probability Curve for Two Step Partial Credit Model**



*3.1.3. Rating Scale, Binomial and Poisson Models*

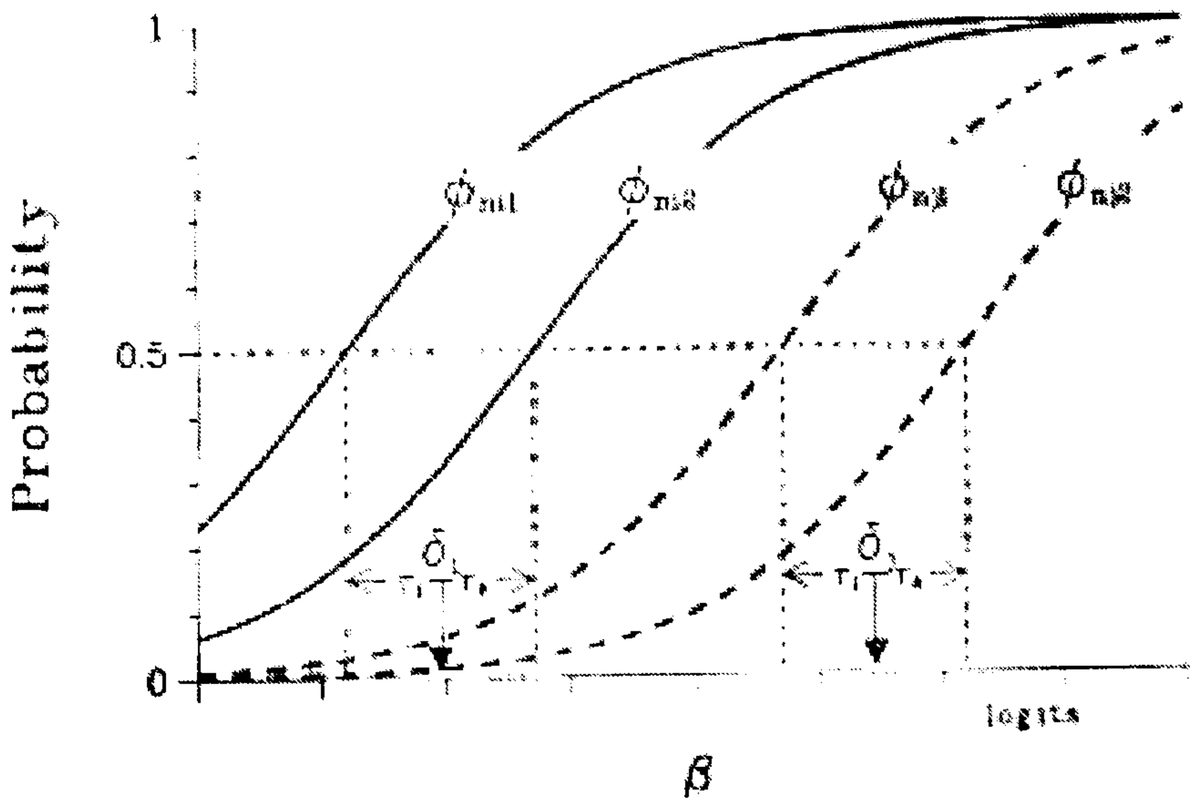
The Rating scale model (e.g. Andrich, 1978a, b) is employed where individuals have to select a response from a series of response options, such as a Likert scale. For instance, to take an example from the EORTC QLQ-c30, "Not at all", "A little", "Quite a bit", and "Very much" (Aaronson et al., 1993).

These response options or categories can be considered as ordered "steps", where the patient chooses or completes the $k^{th}$ step rather than (k-1) step. This can be expressed as,

$$\delta_{ik} = \delta_i + \tau_k$$

where $\delta_i$ is the location (scale value) of item i on the variable and $\tau_k$ is the location of the $k^{th}$ step for each item in respect to the item location. This is shown more clearly in Figure 3.1.5:

**Figure 3.1.5 Item Operating Characteristic Curve for the Rating Scale Model**



In practice, the threshold parameters ($\tau_1,\tau_2, \tau_3, \ldots, \tau_m$) are estimated once for all of the items. The threshold parameters are estimated for each pair of ogives (logistic curves) by setting $\tau_1$ to equal 0, which has the effect of centring $\delta$ for each pair of ogives.

The rating scale model (Andrich 1978a, 1978b) can then be written as:

$$\Phi nik = \frac{\pi nik}{\pi nik-1 + \pi nik} = \frac{\exp[\beta n - (\delta i + \tau k)]}{1 + \exp[\beta n - (\delta i + \tau k)]} \quad \text{for k = 1,2,...m.}$$

or,

$$\pi nix = \frac{\exp \sum_{j=0}^{x} [\beta n - (\delta i + \tau j)]}{\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} [\beta n - (\delta i + \tau j)]} \quad \text{where x = 0,1,...m.}$$

Which is the probability of person n selecting category x to item I, and where $\tau 0 \equiv 0$,

therefore $\exp \sum_{j=0}^{k} [\beta n - (\delta i + \tau j)]$ = 1.

This model can be used in order to estimate $\beta$n for each person n, and $\delta$i for each item i. Additionally, m response thresholds ($\tau_1,\tau_2, \tau_3, \ldots, \tau_m$) can be estimated for m + 1 response categories.

Other Rasch models include the Binomial Trials model and Poisson model. For the previous models (specifically the Partial Credit and the Rating Scale Models) the items were completed in a specific order and the persons' scores were taken as the number of steps completed.

This is not the case for the Binomial Trials and Poisson Models where the order of successes or failures on an item is assumed not to be important, and in which each outcome is considered to be independent.

The Binomial Trials model is used when the number of successes (or failures) x is counted for m trials, where m is the number of independent attempts at each item. The model is used for tests of psychomotor skills (Wright and Masters, 1982), e.g. the number of times a target is hit successfully.

The Binomial Trials model can be developed from the Dichotomous model for one attempt at item I, e.g.

$$P = \frac{\exp(\beta_n - \delta i)}{1+\exp(\beta_n - \delta i)}$$

Since it is assumed that the attempts are independent, the probability of success on x attempts and failures on m-x is given by: $P^x (1-P)^{m-x}$

There are $\binom{m}{x}$ number of permutations of success x in m attempts, therefore the probability of success in x attempts is given by:

$$\pi_{nix} = \binom{m}{x} P^x (1-P)^{m-x}$$

Substituting for P,

$$\pi_{nix} = \binom{m}{x} \frac{\exp[x(\beta n - \delta i)]}{[1+\exp((\beta n - \delta i)]^m}$$

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} [\beta n - (\delta i + cj)]}{\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} [\beta n - (\delta i + cj)]}$$

x = 0, 1, …,m, and where $c_j$ = log[j/m-j+1)].

The Poisson Model is a variation of the Binomial Trials model where there is no upper limit on the number of events (either successes or failures), and where the probability of failure or success is small.

The Poisson Model was developed by Rasch (1980) and used for analysing errors and speed of reading. The probability of success or failure is x out of m attempts is given by the binomial expression,

$$\pi_{nix} = \binom{m}{x} P^x (1-P)^{m-x}$$

where the expected number of successes / failures on item i is $\lambda_{ni}$ = mP.

If $\lambda_{ni}$ remains constant as both m increases and P decreases, then the probability can be replaced by the Poisson expression:

$$\pi_{nix} = \frac{\lambda_{ni}^{x}}{x! \exp(\lambda_{ni})}$$

This gives the probability of an individual making x successes (or failures) on an item i when there is no upper limit. This is a function of $\lambda_{ni}$, which in turn are functions of $\beta_n$ and the item difficulty $\delta_i$.

If we substitute $\lambda ni = \exp(\beta n - \delta i 1)$ then,

$$\pi_{nix} = \frac{\exp[x(\beta n - \delta i 1)]}{x!\exp[\exp(\beta n - \delta i 1)]}$$

### *3.1.4 Characteristics of Rasch Models*

All Rasch models can be derived from the general form:

$$\Phi nix = \frac{\pi nix}{\pi nix - 1 + \pi nix} = \frac{\exp(\beta n - \delta ix)}{1 + \exp(\beta n - \delta ix)}$$ , for x = 1,2,..., $m_i$

This defines the probability of a person n scoring x rather than x-1, as a function of the person parameter or ability $\beta$, and the item parameter or item difficulty $\delta$. Furthermore, the parameter estimates and probabilities are derived from the raw data or counts (Wright and Masters, 1982). The raw person scores and item scores are "minimally sufficient statistics for person and item parameters" (Wright and Masters, 1982, p. 59). In other words, the raw score or steps completed by a person is sufficient to estimate the parameter $\beta$. Similarly, the scores on each item are sufficient statistics for estimating $\delta$.

The general equation does not contain a parameter for the slope or item discrimination, unlike item-response theory models (e.g. Samejima, 1969), e.g. (from Hambleton, Swaminathan, and Rogers, 1991, p. 15):

$$Pi(\theta) = \frac{\exp Dai(\theta - bi)}{1 + \exp Dai(\theta - bi)}$$ , for I = 1, 2,...,n.

which gives the equation for a two-parameter model, where D is a scaling factor (=1.7), and the additional parameter a is the item discrimination parameter.

Consequently all items modelled using Rasch models have the same slope, which means that all person parameters and item parameters are point locations on the same latent trait, and can be expressed in the same scale units (Wright and Masters, 1982, p. 55).

Another feature of Rasch models is the independence of parameters, a derivation of which is given below.

If the probability of a person n making a particular set of responses or response vector ($x_{ni}$) to an L-item test is given by Equation 3.1.5,

**Equation 3.1.5.**

$$P\{(x_{ni});\ \beta_n,\ ((\delta_{ij}))\} = \exp \sum_{i=1}^{L} \sum_{j=0}^{xni} (\beta n - \delta ij) \Big/ \prod_{i=1}^{L} [\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta n - \delta ij)]$$

And the probability of a person n making a score r, where r is defined as the total count of the number of steps completed, is given by,

**Equation 3.1.6.**

$$P\{r;\ \beta_n,\ ((\delta_{ij}))\} = \sum_{(xni)} \exp \sum_{i}^{L} \sum_{j=0}^{xni} (\beta n - \delta ij) \Big/ \prod_{i=1}^{L} [\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta n - \delta ij)]$$

Then the probability of obtaining a given set of responses ($x_{ni}$) given a score of r can be derived by dividing Equation 3.1.5 by 3.1.6, which results in:

**Equation 3.1.7.**

$$P\{(x_{ni});\ \beta_n,\ ((\delta_{ij}))\ |\ r\} = \exp(-\sum_{i}^{L} \sum_{j=0}^{xni} \delta ij) \Big/ \sum_{(xni)}^{r} \exp(-\sum_{i}^{L} \sum_{j=0}^{xni} \delta ij)$$

It is immediately apparent from this equation that the person parameter or ability estimate $\beta$ is absent. This means that in Rasch models, the probability of a particular response or set of responses is derived independently of the ability of the person producing the responses, and only depends on the difficulties, $\delta$, of the items concerned.

The equation for the conditional probability of the entire response set (of an individual) is given below:

$$P\{(x_{ni});\ \beta_n,\ ((\delta_{ij}))\ |\ r_n\} = \prod_{n=1}^{N} \exp(-\sum_{i}^{L} \sum_{j=0}^{xni} \delta ij) \Big/ \sum_{(xni)}^{r} \exp(-\sum_{i}^{L} \sum_{j=0}^{xni} \delta ij)$$

Once again the person ability parameters do not appear in the equation, which means that the person parameters are estimated independently of the person scores, but are derived from the item parameters or difficulties.

For the item parameters a similar type of analysis is applied using the following equation, which gives the probability of an N-person set of responses to an item i:

**Equation 3.1.8.**

$$P\{x_{ni}); (\beta_n),(\delta_{ij}) = \left[\exp(\sum xni\beta n)\right]\left[\exp(-\sum_{n=1}^{N}\sum_{j=0}^{xni}\delta ij)\right] / \prod_{n=1}^{N}[\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\beta n - \delta ij)]$$

The probability of obtaining a given set of responses S to item i is given by

**Equation 3.1.9.**

$$P\{S); (\beta_n),(\delta_{ij}) = \left[\sum_{(xni)}^{(S)}\exp(\sum_{n=1}^{N} xni\beta n)\right]\left[\exp(-\sum_{n=1}^{N}\sum_{j=0}^{xni}\delta ij)\right] / \prod_{n=1}^{N}[\sum_{k=0}^{mi}\exp\sum_{j=0}^{k}(\beta n - \delta ij)]$$

The probability of a set of responses given S is then found by dividing Equation 3.1.8 by Equation 3.1.9:

**Equation 3.1.10.**

$$P\{x_{ni}); (\beta_n) \mid (S)\} = \exp(\sum xni\beta n) / \sum_{(xni)}^{(S)}\exp(\sum_{n=1}^{N} xni\beta n)$$

It can then be seen from Equation 3.1.10 that the item difficulty parameters do not appear. In other words the information regarding the item difficulties can be derived solely from the vector of responses of the items completed.

The derivation of the equations 3.1.5 to 3.1.10 highlights an important feature of the Rasch models which distinguish it from classical test theory models, namely that the probability of the data given a person score r can be derived solely from the item difficulty parameters, d, conversely that the data given scores of S can be derived from just the person parameters or ability: This means that item parameters can be estimated independently of any given data, and likewise that person abilities can be estimated independently of item difficulties. In other words, Rasch models allow for "sample free" estimates of item difficulties and "test free" estimates of ability (Wright and Masters, 1982).

## 3.2. Estimation Procedures

Several methods exist for estimating item and person parameters, however only two methods, PROX and UCON (Wright and Masters, 1982) will be discussed in this section, since these are the two procedures employed in the software (Winsteps, Linacre, 2003) employed in the analyses in Chapters 5 – 7.

### *3.2.1 PROX*

The PROX procedure allows item and person parameters to be calculated by hand (e.g. Wright and Masters, 1982). In the Winsteps programme PROX is used to derive the initial item and person parameter estimates, which are then refined by maximum likelihood estimation (see 3.1.3.2).

The PROX procedure is described below for estimating item difficulties and person parameters for dichotomous items:

1. Firstly, perfect or extreme scores, i.e. maximum and minimum scores are removed, since these are not useful for item parameter estimation because no information is provided about differences between items since an individual has the same response for each item.

2. Then proportions are calculated, e.g. for GHQ12 scored dichotomously, 100 patients, the total possible maximum score for each item is 100, the proportion of the maximum value and the inverse of this are then calculated, e.g. if 35 patients score 1 for this item (and 65 score 0), then

P = 35 / 100 = 0.35, and 1- P = (100-35)/100 = 65/100 = 0.65

3. The logit is then calculated for each item by taking natural log of P/(1-P), for item above,

Logit = Ln(P/(1-P) = Ln(0.35/0.65) = -0.62

The logits for each item are then multiplied by –1 to set or anchor the mean of the item parameter estimates to zero.

4. The mean of the item estimates is then calculated, which is then subtracted from the initial logit to adjust the estimates for sample effects (make them independent of sample effects, since if a sample of patients with worse health had been questioned scores would have been higher, and vice versa for a sample of healthier patients).

5. Steps 2 and 3 are repeated for patients, although the logits derived are not multiplied by –1, e.g. a patient agreeing with 7 out of a possible of 12 items from the GHQ12 (when scored dichotomously) would have an initial ability estimate of Ln(P/(1-P) = 0.32. Similarly, the initial ability estimate for a patient scoring just 2 would be –1.61 (Ln(0.17/0.83)).

6. The item estimates are then multiplied by an expansion factor Y (to control for sample dispersion), derived from the item variance (U) and the variance in person estimates (V), e.g.

$$Y = [(1+V/2.89)/(1-UV/8.35)]^{1/2}$$

To arrive at the final item parameter estimates. Similarly the person estimates are also multiplied by an expansion factor, X, to control for dispersion in item estimates:

$$X = [(1+U/2.89)/(1-UV/8.35)]^{1/2}$$

This procedure highlights two factors referred to earlier of Rasch models, namely that the total scores for the items are sufficient statistics for person ability estimates, and conversely that person scores are sufficient statistics for item difficulty estimates. Secondly, that the estimation procedures for item and person parameters are independent.

*3.2.2. UCON*

The UCON procedure is an unconditional joint maximum likelihood estimation procedure (Wright and Panchapakesan, 1969). Maximum likelihood estimates refer to finding the value for ability scores (for individuals) that maximises the likelihood of the responses or response sets observed, or conversely, the likelihood of the item parameters for the data. It is limited in that it is not able to produce person estimates for extreme scores (i.e. perfect scores or zero responses) and the item estimates for these scores must be eliminated before the estimation procedure (e.g. Hamilton, Swaminathan, and Rogers, 1991). However, it is a common procedure employed in determining parameter estimates for the Rasch model.

Maximum likelihood estimations for the ability and item parameters are generated from two sets of equations. The next section will describe the estimation procedure first for the ability parameters, then the item parameters, and finally the joint estimation procedure will be explained.

The probability of an individual with a given ability $\beta$ producing a ( dichotomous) response set to a series of questions n can be expressed as:

**Equation 3.2.1**

$$P(U_1, U_2, \ldots U_n | \beta) = \prod_{j=1}^{n} P(U_j | \beta)^{2},$$

where Uj refers to the response 1 or 0.

Since Uj can only be 1 or 0, Equation 3.2.1 can also be expressed as,

**Equation 3.2.2**

$$P(U_1, U_2, \ldots U_n | \beta) = \prod_{j=1}^{n} [P(U_j | \beta)]^{Uj} [1 - P(Uj | \beta)]^{1-Uj}$$

When the response is observed the equation can be expressed as a likelihood, rather than probability, and Equation 3.2.2 then becomes:

**Equation 3.2.3.**

$$L(u_1, u_2, \ldots u_n | \beta) = \prod_{j=1}^{n} [P(u_j | \beta)]^{uj} [1 - P(uj | \beta)]^{1-uj}$$

In order to simplify calculations, Equation 3.2.3. is transformed using natural logarithms:

---

[2] Formulae in this section are taken from Hamilton, Swaminathan and Rogers (1991)

**Equation 3.2.4.**

$\ln L(\mathbf{u}|\beta) = \sum_{j=1}^{n} uj \ln Pj + (1 - uj) \ln(1 - Pj)$ , where **u** is the vector of responses.

The procedure normally employed to calculate maximum likelihood estimations is known as the iterative Newton-Raphson procedure (e.g. Embretson and Reise, 2000). This procedure makes use of the fact that the slope or tangent of the curve of the ability levels plotted against the log-likelihood estimates for a given response set will be zero at the maximum log-likelihood value. Given this fact equation 3.2.4 can be partially differentiated with respect to $\beta$ and solved for zero to derive the ability estimates (the second derivative can be used to calculate the standard error of the ability estimate).

A similar procedure is employed to derive item parameter estimates. Equation 3.2.5 describes the responses of N individuals to an item:

**Equation 3.2.5.**

$L(u_1, u_2, \ldots u_n | \beta, \delta) = \prod_{i=1}^{N} P_i^{ui} Q_i^{1-ui}$ , where Q = (1-P($U_j$|β)).

In this procedure the first derivative of the likelihood function with respect to the item parameter is found by setting equation 3.2.5 to equal zero and then solving.

Both procedures described above are employed when either the item parameters or ability parameters respectively are known. However, it is often the case that neither parameters are known and therefore have to be estimated jointly.

The likelihood function for joint estimation is given below in Equation 3.2.6 for N individuals responding to n items:

**Equation 3.2.6.**

$L(u1,u2,\ldots un \mid \beta, \delta) = \prod_{i=1}^{N}\prod_{j=1}^{n} P_{ij}^{uij} Q_{ij}^{1-uij}$ , where ui is the response set of person i to n items, and $\beta$ is the set of N ability estimates, and $\delta$ is the set of item parameters for n-items.

The procedure for joint maximum likelihood estimation occurs in two stages. Firstly, initial values for the ability estimates are chosen, for instance by taking the log of the ratio of correct to incorrect responses (for the dichotomous model), and then the initial item parameters are estimated by assuming the person or ability parameters are known. The second stage involves the initial item parameters being treated as known and fed back into the estimation procedure to derive the next estimations for the ability parameters. This procedure is repeated until values for the estimates do not change or a predetermined number of iterations has been performed.

## 3.3. Evaluation of data

The item difficulty ($\delta$) and person estimates ($\beta$) can be evaluated using fit statistics. In addition, unidimensionality of the questionnaires or set of items can also be assessed by a factor analysis (principal components analysis) of the residuals of the Rasch model. Finally, item invariance can be assessed using differential item analysis. These methods are described in more detail below.

### *3.3.1 Fit statistics*

Once the item and ability parameters have been estimated as described above in section 3.2., the parameters can then be assessed to check the extent to which they fit the Rasch Model.

This section describes two fit statistics, the infit mean square and outfit mean square, employed in the Winsteps (Linacre & Wright, 2001) programme which was used in later chapters for the Rasch analysis.

Both statistics are based on the residuals derived from the difference between the expected value and the observed score. For instance, for the Rating Scale model, the expected value of a response x, can be calculated from:

$Eni = \sum_{k=0}^{m} k\pi_{nik}$ , where $\pi_{nik}$ is the probability of responding to category k for item i, the score residual can then be calculated by subtracting the expected score E from x the observed score. This residual is then standardised by dividing the score residual by the square root of the variance (W) of x, e.g.

$$W_{ni} = \sum (k - Eni)^2 \pi_{nik}$$

The unweighted mean square or outfit statistic can then be calculated by summing the squares of the standardised residuals and dividing by the total (N) number of persons, i.e.

Unweighted mean square or outfit statistic = $u_i = \sum_{n=1}^{N} z_{ni}^2 / N$ .

However the unweighted mean square statistic is sensitive to unexpected responses from outliers (Bond and Fox, 2001; Wright and Masters, 1982), i.e. individuals for whom a particular item is either too easy or too difficult. To overcome this problem the squared residuals can be weighted by the variance for each item. The resultant statistic is known as the weighted mean square or infit statistic:

Weighted mean square or infit statistic = $v_i = \sum_{n=1}^{N} z_{ni}^2 W_{ni} / \sum_{n=1}^{N} W_{ni}$ .

Both statistics have an expected value of 1 and range from zero to positive infinity, i.e. they only take positive values.

Fit statistics from different items can be compared by standardising either the infit or outfit statistics by transforming them into an approximately normalised t distribution using the Wilson-Hilferty (Wilson and Hilferty, 1931) transformation (see below for the transformation for the infit statistic):

$t_i = (v^{1/3} - 1)(3/q) + (q/3)$, where q is the variance of the weighted mean square.

The expected value for t is 0 and the standard deviation one. These t statistics are referred to as outfit and infit t in Winsteps. A similar analysis can be employed to derive fit statistics for the person ability estimates by summing over items, rather than persons.

There has been and there continues to be a considerable debate around the issue of which is the most appropriate fit statistics to use, what range of fit statistics to be employed when evaluating fit, and how fit statistics should be interpreted.

Given that the expected value for both infit and outfit statistics is 1, then fit statistics greater than 1 can be interpreted as demonstrating more variation between the model and the observed scores, e.g. a fit statistics of 1.25 for an item would indicate 25% more variation (or "noise") than predicted by the Rasch model. Conversely, an item with a fit statistic of 0.70 would indicate 30% less variation (or "overlap") than predicted. Items demonstrating more variation than predicted by the model can be considered as not conforming to the unidimensionality requirement of the Rasch model.

Smith and his colleagues (Smith, 1988; Smith, 1991; Smith, Schumacker and Bush, 1998; Smith and Suh, 2003) have explored the association between fit

statistics and sample sizes and number of items per questionnaire. Their results from a number of studies involving simulated data sets – data sets where the data fit the Rasch model (e.g. in the Smith et al. (1998) paper, a data set was created where person abilities were normally distributed with a mean of 0 and a standard deviation of 1, and where item difficulties were uniformly distributed from –2.0 to +2.0 logits) – have demonstrated that as the sample size increases the range of both the infit and outfit statistics decreases. For instance, although no differences were found for the number of items used, the researcher did discover that as the sample size increased from 150 persons to 1000 persons, the range decreased from 0.72 (0.73 – 1.45) to 0.25 (0.89 – 1.14) for the unweighted mean square (outfit statistic, Smith et al., 1998). Similarly, the range for the infit (weighted mean square), which was narrower than the range for the outfit statistic, decreased from 0.29 (0.86 – 1.15) to 0.10 (0.95 – 1.05) for the same sample sizes. In addition to this Smith et al. (1998) found that the Type I error rate for the weighted mean square varied considerably depending on sample size. Using cutoffs of 0.7, 0.8, 0.9 and 1.1, 1.2 and 1.3 Smith et al. (1998) found that the Type I error rate fell from around 8% for a cutoff of (less than) 0.90 and (greater than) 1.1, to below 1% for values beyond this, as sample sizes increased. The Type I error rate for the unweighted mean square was maintained at around 5% using the ranges 0.80 – 1.3 for 150 persons, 0.90 – 1.2 for 500, and 0.90 – 1.1 for 1000 persons. The mean squares for the outfit statistic were also not distributed evenly around the mean (1) with more extremes occurring above 1 than below. Changes associated with increasing sample sizes were also found for the standardised t-statistic. However, although the Type I error rate decreased from around 3% for both the weighted and unweighted t-statistics at sample sizes of 150 to around 1% for larger sample sizes the differences observed as the samples increased were not as pronounced as those for the weighted and unweighted mean squares.

On the basis of this Smith et al. (1998) suggested that the standardised t-statistic rather than the weighted and unweighted mean squares should be used to identify misfit, given that this statistic appears to be less sensitive to changes in sample size. Additionally, they suggest that critical values or cutoffs of 1.16 for sample sizes of 150, 1.09 for 500 and 1.06 should be employed for the weighted mean square, and cutoffs of 1.48, 1.27 and 1.19 should be used for the same sample sizes for the unweighted mean square.

However, the main criticism of the work by Smith and colleagues is that the results arise from data sets where both items and person ability parameters fit the Rasch model perfectly. As Linacre (RMT, 1999) has pointed out "no data ever fit... [the Rasch model] perfectly" (p. 706). It is not clear, and there has been no research published to date, on the interaction between the weighted and unweighted mean squares and the standardised t-statistic and sample sizes for data that do not fit the Rasch model perfectly. In addition, more recently, Lai et al. (2003) have pointed out, on the basis of work carried out on real (as opposed to simulated) data, that some of these statistics, notably the t-statistic, is sample size dependent and may produce spurious results for large samples. Similarly, Linacre (2002) has pointed out that mean square values larger than 2 may be produced by only one or two outliers, and that the misfit for standardised t-statistics greater than 3 may actually be small for large sample sizes.

Given the uncertainty surrounding the relationship between the fit statistics and sample sizes, specifically for data which do not fit the Rasch model, and the absence of any definitive solution to the problem, the range of 0.70 to 1.30 for both the weighted (infit) and unweighted (outfit) mean square statistics, suggested by Wright, Linacre, Gustafson and Martin-Löf (1994) will be used throughout the analyses (Chapters 5 to 7) to identify misfit. This criterion range has been employed in a number of studies (e.g. Doward et al., 2003; Ryser et al., 1999) to identify misfitting items. Items with fit statistics less than 0.70 or greater than 1.30 will be

identified as having poor fit, i.e. exhibiting redundancy or excessive noise, respectively. Furthermore, following remarks by Lai et al (2003) and Linacre (2002) the standardised t-statistics for the weighted and unweighted mean squares will not be used to identify misfit.

*3.3.2. Reliability*

Cronbach's alpha is the reliability index used most frequently in traditional test theory for assessing reliability, where reliability in this context refers to the reproducibility of the data (e.g. Linacre, 1997):

$\alpha = \frac{K}{K-1}(1-\frac{\sum \sigma_i^2}{\sigma_x^2})$, where $\sigma_i^2$ is the variance of the ith item and $\sigma_x^2$ is the total variance of the scores, and K is the number of items.

The Rasch models employ two types of reliability indices based on the same concept as Cronbach's alpha, one for person estimates, referred to as the person separation reliability ($R_p$), and the other for item difficulty estimates referred to as the item separation reliability (Bond and Fox, 2001; Wright and Masters, 1982).

The person separation reliability, that is the percentage of observed variance in the responses that is reproducible is calculated as follows:

$R_p = \frac{SA_p^2}{SD_p^2}$, where $SD_p^2$ refers to the total person variance, i.e. how much individuals differ on the measure of interest, and $SA_p^2$ is the adjusted person variability and refers to the amount of variance that can be reproduced by the Rasch model. The adjusted person variability is arrived at by subtracting the error variance (i.e. the variance not explained by the Rasch model or $SE_p^2$) from the total variance. The

person separation reliability ranges between 0 and 1, and is independent of the test items (Wright and Masters, 1982).

Another person reliability index, namely the person separation index ($G_p$) can be calculated by dividing the adjusted person variability by the average measurement error, i.e. taking the square root of $SA_p^2$ and $SE_p^2$:

$$G_p = \frac{SA_p}{SE_p}$$

Since this is a standardised statistic and is not restricted to a range between 0 and 1, it can be used for comparisons of reliabilities across different analyses. A useful feature of the person separation index is that it can be used to calculate the number of distinct ability groupings or strata in the data. For instance, if a separation of three standard errors is used to distinguish between ability strata, then the number of strata in the sample can be calculated from ([4G + 1]/3) i.e. for person separation index of 2 there will be 3 distinct ability strata in the sample. The minimum person separation index is 2 (e.g. Prieto, Alonso and Lamarca, 2003), i.e. the instrument should be able to distinguish between 2 distinct person ability strata.

Separation reliability and separation indices can also be derived for items (i.e. $R_i$ and $G_i$) as above by replacing person variance with item variance.

### *3.3.3. Differential Item Functioning*

Since one of the most important features of the Rasch models is the separation of item and ability parameter estimation, it follows that estimation of item difficulties should remain constant, or invariant, not only across different person abilities, but also across different groups individuals, e.g. males and females, cancer patients and general medical inpatients, etc. (Bond and Fox, 2001).

Differential item functioning (DIF) refers to instances where the item (difficulty) estimates differ depending on the sample used for the estimation. Definitions of DIF include, "[the] simple observation that an item displays different statistical properties in different group settings", (Angoff, 1993, p. 4), and "...is operationally defined as statistically different item difficulty estimates for the same item in subpopulations of interest" (Smith, 1992, p. 86).

Two features of DIF highlighted by Smith (1992) are that 1) performance on items is subject to sources of variation other than that intended, and 2). The unintended source of variation systematically affects identifiable subgroups.

Therefore the key feature of differential item functioning is independence of item parameter estimation from sample characteristics, i.e. item parameter estimation should remain invariant irrespective of the composition of the sample (or group) from which it is estimated. This follows from the local independence and unidimensionality requirements of the Rasch models, that the underlying latent trait must be a single construct, which in turn implies that when the variance explained by the model is partialled out the items in the scale are independent – i.e. performance on one item does not inform how an individual might perform on another item in the scale once the latent trait is controlled for.

Item invariance is an important property and can be used along with the fit statistics and the principal components analysis of the residuals to identify items which do not conform to the unidimensionality criterion of the Rasch models. For instance, if item difficulty estimation is not independent of the sample, then this could

potentially confound the interpretation of results, e.g. if item parameter estimates are dependent on the sample from which they are drawn then results could be confounded since performance on an item or items cannot be explained in terms of a person's location on a single latent trait alone, but other confounding factors could be employed in the explanation.

For instance, to take an example from the Physical Functioning scale of the EORTC QLQ-C30, such as "Do you have trouble taking a long walk?" It could be pre-supposed that there might be differences between lung cancer and breast cancer patients' responses to this question, perhaps with breast cancer patients more likely to respond positively to this question than lung cancer patients.

However in order to ascertain for certain that there are true differences between lung cancer and breast cancer patients in responding to this question, the item difficulty estimate – i.e. the location of the item on the continuum, must be independent of whether it was calculated from a sample of lung or breast cancer patients. If it is truly sample independent – i.e. if there is no differential item functioning or bias – then any differences found between the two groups in response to this question can be ascribed to true differences. On the other hand if there is DIF or bias, the results are potentially confounded – other factors may be influencing results, e.g. male / female differences, stage at diagnosis, age, etc.

The general procedure (Wright and Masters, 1982) employed to identify a lack of item invariance, or differential item functioning, is to estimate item parameters separately for groups of individuals separately, e.g. d1 and d2, along with their variances, s1 and s2, and then to derive the standardised difference,

$z = (d1-d2)/(s_1^2 + s_2^2)^{1/2}$ which can be evaluated can be evaluated against the criterion 1.96 (for $\alpha = 0.05$) for large (N>100) sample sizes.

A significant difference, i.e. a z value greater than 1.96 would indicate that the item was not invariant across ability groups, and was demonstrating differential item functioning. This fact could be used as a criterion for removing items from a scale. However, Linacre (1994) has suggested that for sample sizes greater than 150, a difference between item estimates of less than 0.50 logits is stable at a confidence interval of 99%.

Removing items which demonstrate DIF, could potentially remove confounding factors from the results. For instance, from the example above for the Physical Functioning and two groups of cancer patients, any differences found between the groups with a "new" scale, i.e. a scale where items demonstrating DIF have been removed, allows these differences to be interpreted as "true" differences between the groups, given that the remaining items on the questionnaire form a unidimensional construct, and therefore other reasons for these differences should not have to be advanced. The fact that items with DIF have been removed, however, does not influence whether a scale is able to discriminate between groups in a given sample, which depends on the scale construction, and in particular on the items in the scale. The item invariance requirement simply refers to the estimation procedure for item difficulty parameters. Removing items which show an invariance or bias depending on which sub-samples are used to estimate the parameters, i.e. removing items with DIF removes the confounding factors, and provides a clearer picture of how the items perform for the different groups, and how the items may discriminate between these groups, free of any bias in estimation.

Since differential item functioning is often investigated for sets of items (e.g. items from a questionnaire), where item invariance is explored simultaneously for each item, the multiple testing could affect (i.e. inflate) the family-wise (Type I) error rate (Howell, 2002). In order to control for this a Bonferroni adjustment (Howell, 2002) could be employed by dividing $\alpha$ by the number of comparison being made

(Groenvold, Bjorner, Klee, & Kreiner, 1995), for instance for five items $\alpha' = \alpha/5$ = 0.05/5 = 0.01, the critical value for this then becomes 2.33 (rather than 1.96).

*3.3.4. Item and Test Information Curves*

The final features of Rasch models to be described are the item and test information curves. Information in this instance refers to "psychometric information", which was defined by Fisher (cited in Baker, 2001) as the reciprocal of the precision with which a parameter can be estimated, in other words it refers to the variability of the estimates around the parameter, i.e.

$$\text{Information, } I = \frac{1}{\sigma^2}$$

For Rasch models the information function can be derived for the ability estimates by plotting the reciprocal of the variance of the ability estimates for each item for each level of ability. For the dichotomous model the item information can be derived from:

$$I(\beta) = P(\beta)\, Q(\beta)$$

Therefore, the smaller the variance of the estimates, the greater the precision, and consequently the greater the information provided. Conversely the greater the variance the less information can be provided for any given level of the ability latent trait.

The item information functions can be summed across items to be produce a test information function describing the item across the set of questions:

$I(\beta) = \sum_{i=1}^{N} I_i(\beta)$, where $I(\beta)$ is the amount of test information at any given level of ability ($\beta$).

### 3.4. Principal Components (Factor) Analysis

In essence factor analysis refers to a set of statistical procedures for simplifying datasets, commonly by extracting factors from correlation matrices (e.g. Kline, 1992, 1997). A factor refers to a dimension or construct which can be inferred from the set or sets of variables.

The analysis in Chapters 5 to 7 makes use of a factor analysis technique known as principal components analysis. This method was chosen above other available techniques since it makes no prior assumptions regarding the factor structure in the sample (Kline, 1997) and explains all of the variance in the correlation matrix unlike other factor analysis methods (Kline, 1997; Nunnally and Bernstein, 1994).

Principal components analysis is an iterative procedure, which is used to simplify a correlation matrix of variables by explaining it in terms of underlying factors. An initial vector is calculated by summing the correlation coefficients for each column of the correlation matrix to from a vector $\mathbf{U}_1$. The elements in $\mathbf{U}_1$ are then squared and summed. The elements of the vector $\mathbf{U}_1$ are then divided by the square root of this total (this process is known as "normalising" $\mathbf{U}_1$) to produce the first trial vector $\mathbf{V}_1$. The correlation matrix, **R**, is then multiplied by the first trial vector, $\mathbf{V}_1$ to produce a new vector $\mathbf{U}_2$. $\mathbf{U}_2$ is then normalised as described above to produce $\mathbf{V}_2$. This process is repeated until the **V** vectors converge, i.e. the difference between estimates become negligible, at which stage $\mathbf{V}_{n-1}$ becomes the first characteristic vector or eigen vector. Eigen vectors refer to vector column of weights each element of which is applied to one of the variables in the correlation matrix. The normalised vector $\mathbf{U}_n$ is referred to as the first eigen value, $\lambda_a$. Eigen values refer to the proportion of variance explained by each factor. Factor loadings, i.e. correlations of variables with the factor, are then calculated by multiplying the elements of $\mathbf{V}_{n-1}$ by the square root of the eigen value, $\lambda_a$ to form the first component or factor. This

process is repeated for the next component until all variance has been accounted for. However, subsequent extractions are calculated using the residual correlation matrix, i.e. the correlation matrix which remains once the first factor has been partialled out. The residual matrix is derived by multiplying each pair of factor loadings for each pair of variables, and by setting the diagonals to equal the square root of the factor loadings. These values are then subtracted from the original correlation matrix, and all residuals, with the exception of the diagonals, are multiplied by –1 to avoid the problem of the column sums equalling zero. In this matrix the diagonals reflect the proportion of variance remaining once the first factor has been partialled out, and the other cells reflect the partial covariance between variables.

The principal components analysis described above refers to an unrotated solution of the correlation matrix in which the correlations between individual pairs of variables was explored. This analysis can be extended to incorporate sets of variables by rotating factors to change the factor loadings and therefore the meaning of the factors, although the variance explained by the rotated and unrotated factor structures remains the same. Two rotation methods can be identified, namely, orthogonal rotation, in which factors are assumed a priori to be uncorrelated and oblique in which this restriction does not apply. Considering the correlations between factors from quality of life data (e.g. Aaronson et al., 1993; Smith et al., 2002) oblique rotations will be employed in the analysis in Chapters 5 to 7.

Two common criteria employed to identify the number of factors or components derived by factor analysis are the Kaiser-Guttman criterion and the scree-plot (Nunnally and Bernstein, 1994). The Kaiser-Guttman criterion simply defines that factors with eigen values of 1 or greater should be retained. However, the problem with this criterion is that it is dependent on the number of items: the greater the number of items, the less variance needs to be accounted for to reach criterion (Nunnally and Bernstein, 1994). This criterion therefore tends to overestimate the number of factors.

In the scree-plot the eigen values as simply plotted against their ordinal number, and the point at which the values drop off, i.e. the transition point is used to identify the number of "real" factors. This criterion has the advantage over the Kaiser-Guttman criterion of suggesting fewer factors, particularly when correlations are low and/or the number of variables is large (Nunnally and Bernstein, 1994).

Factor analysis, specifically (unrotated) principal components analysis can also be employed within Rasch analyses to identify multidimensional within data sets. However, Wright (1996) has identified a number of shortcomings of "traditional", i.e. non-Rasch principal components analysis (PCA) in identifying violations of unidimensionality: 1). In traditional PCA the raw scores are used to calculate the factors. Since raw scores are non-linear, i.e. not converted to logits, this analysis only approximates a Rasch dimension; 2). Similarly, the residuals from the first factor extraction are also non-linear.

In order to overcome the problem of non-linearity Wright (1996) has suggested performing a principal components analysis of the residuals following the Rasch analysis. In this situation the Rasch dimension is equivalent to the first principal component. The residuals are derived from the difference between the observed value and expected value computed according to the Rasch model. These are then standardised by dividing them by the variance of the expected values. Pearson correlations between the standardised residuals are then computed for each pair of items and then a principal components factor analysis is carried out on the item correlation matrix. Since one of the assumptions behind Rasch models is that items are locally independent, any associations left in the correlation matrix should be random or "noise" if the original Rasch factor is unidimensional. Therefore any factors identified from this matrix reflect violations in the unidimensionality and local independence assumptions.

This form of principal components analysis can be applied to both item and person ability estimates. As with the non-Rasch PCA, the eigenvalues derived from

the analysis refer to the amount of variance explained. Smith and Miao (1994) have suggested, on the basis of simulated data sets, that factors greater than 1.4 may suggest multidimensionality in the sample. More recently Linacre (2002) has suggested that eigenvalues greater than 2 may imply that substructures or dimensions exist in the samples.

## 4. Computer-Assisted Questionnaires

The measurement of quality of life is becoming increasingly important in routine practice in oncology clinics (Cull et al., 1995; Detmar et al., 2002, Velikova et al., in press). Quality of life questionnaires typically address a range of issues such as physical, psychological and social concerns experienced by the patient, as well as symptoms. These measures can be used in the clinical consultation process, and may even act to highlight issues and concerns of patients to their clinicians (Cull et al., 1995; Ford et al., 1994).

Over several decades a range of questionnaires has been developed that may either be generic, disease or topic specific. Among these several are very comprehensive producing a near complete picture of the assessed aspect of a patient's life, but may include in the order of 100 questions (e.g. the Sickness Impact Profile: Bergner et al., 1981). The time and effort taken by patients to complete these has led to a move towards shorter questionnaires within the order of 30 questions and in cancer research two of these have achieved prominence (EORTC QLQ-C30 and FACT-G). These questionnaires can be used in research and in clinical practice with meaningful and useful results (e.g. Cella et al., 2002, Velikova et al, 2002). However, they are still moderately time consuming and the conventional structure of the questionnaire means that patients are often having to answer questions which are irrelevant to them as individuals. Thus for example a patient who has no problem in a particular area, such as physical health, will have to complete five questions on the EORTC QLQ-C30, pertaining to physical functioning.

Another limitation of questionnaires such as these is that they cannot be adapted to the patients' degree of impairment. Furthermore, there is a danger with a reduced number of questions that the questionnaire may either overestimate ("floor effect") or underestimate ("ceiling effect") patients' abilities. In standard

questionnaires, the number of questions is fixed and the patients may be asked questions which are not relevant to them or conversely the questionnaire will not explore problem areas in greater detail.

The advent of computerised systems for collecting patients' quality of life data presents an opportunity not only to adapt the questionnaires to accommodate floor and ceiling effects but also, potentially, to eliminate unnecessary questions and therefore shorten the time taken and the burden on the patients.

## 4.1. Aim

In this study, we have evaluated this approach and used the EORTC QLQ-C30 as a model for our purpose. We have sought to reduce this questionnaire to the minimum number of questions compatible with sampling each key area using a computer touchscreen presentation system. We adopted an experimental approach which allowed patients to select problematic areas and then only answer items in those areas. We have compared these results to those obtained by a conventional presentation of the whole questionnaire on a computer touchscreen. This study presents the first empirical and experimental approach of my work (see Chapter 1).

## 4.2. Method

### *4.2.1. Participants*

Patients were recruited from oncology wards and clinics at St. James's University Hospital, Leeds. Patients were provided with written information concerning the study and asked to provide written consent before participation. Exclusion criteria were inability to understand written and spoken English, visual impairments, and pre-existing psychological morbidity (either detailed in the medical notes or expressed by clinical or nursing). The study received ethical approval from St. James's Local Research Ethics Committee.

*4.2.2. Instruments*

The standard EORTC QLQ-C30 and an electronic version of the questionnaire were used in this study. Both instruments are described in detail in Chapter 2.

*4.2.3. Procedure*

The patients completed a standard EORTC QLQ-C30 questionnaire and a computer-assisted questionnaire (CA-questionnaire) in a randomised order on the same day. Three slight modifications of the CA-questionnaire were employed in the study. The first CA-questionnaire design presented patients with a single screen consisting of thirteen descriptions (see Table 4.2.1.) corresponding to the summated scales of the standard questionnaire (e.g. the Physical Functioning scale was described as "Difficulties with physical activities"). Patients were instructed to select those areas which had been a problem or a concern. The second design was identical to the first except that a textbox was presented on the screen reminding patients to select as many concerns or problems as they thought necessary. The final design reduced the number of descriptions presented to the patients to either four or five. This decision was based on the experimental findings of the limitations of human information processing. These results suggest that the number units of information or "chunks" which humans can process is limited to around 7 give or take 2 (Cowan, 2001; Miller, 1956). Patients were not made aware of the content of the CA-questionnaire before selecting options from the screen.

For all versions of the programme patients were able to select or modify as many choices as they wished. The CA-questionnaire subsequently presented the questions from the standard questionnaire which had been highlighted.

The patients were also presented with the Global Quality-of-Life scale on both versions of the questionnaire as a measure of reliability of responses.

**Table 4.2.1. List of descriptions used for the selection screen of the CA-questionnaire**

| EORTC QLQ-C30 scale | Description (Difficulties with.../ Did any of these cause you difficulties?) |
|---|---|
| **Physical Functioning** | Physical activities |
| **Role Functioning** | Work or hobbies |
| **Emotional Functioning** | Feeling anxious or depressed |
| **Social Functioning** | Family or social activities |
| **Cognitive Functioning** | Concentrating or remembering things |
| **Fatigue** | Tiredness |
| **Nausea & vomiting** | Feeling sick or vomiting |
| **Pain** | Pain |
| **Constipation** | Bowels |
| **Diarrhoea** | Bowels |
| **Appetite** | Lack of appetite |
| **Sleeplessness** | Trouble sleeping |
| **Dyspnoea** | Shortness of breath |

*4.2.4 Statistical analysis*

The sample size required was derived using the means, standard deviations and domain-domain correlations from previously published work (Velikova et al., 1999). Using these figures for one of the domains, e.g. Fatigue (mean difference between paper version and computer version of 3.3, and a standard deviation of the difference of 14.4) the standardised difference (d) can be calculated by taking the ratio of the difference to the standard deviation. The sample size to required test the difference between two matched samples can then be calculated using, $N = (\delta/d)^2$, where $\delta$ can be found from statistical tables for a given power and level of $\alpha$ (e.g. Howell, 2002), and d is the standardised difference. For a level of $\alpha$ of 0.05 and a power of 0.60 a total of 90 patients is required.

The age and diagnosis of patients was recorded. Age differences between participators and non-participators was assessed by t-tests. The time taken to complete both versions of the questionnaire was recorded, and differences in completion times were compared using multivariate analyses of variance (MANOVA with one between subjects factor, study design, with three levels, and one within subjects factor, questionnaire type, with two levels). The Global Quality-of-Life scores from both types of the questionnaires were compared using t-test. Differences in the scores between the two instruments were assessed by paired t-tests.

The standard questionnaire was scored using the scoring algorithms published by the EORTC Quality of Life study group (see Chapter 2), where higher scores on the functioning scales indicates a better level of functioning, and high scores on the symptom scores indicates a high level of symptoms. Since it was not possible to provide a score for those patients who did not select a given option from the CA-questionnaire directly, for instances such as these a score of 100 or 0 was scored for each (equivalent) scale not selected by the patient for the functioning and symptom scales respectively. For instances, if a patient were to select all the options corresponding to the symptom scales from the standard questionnaire, but none from the functioning scales, then scores of 100 would be entered for the latter scales, along with the responses made by the patient to the symptom scales.

Exact and global agreement between the two forms of questionnaire was calculated in accordance with our previous work. Exact agreement was calculated as the percentage of identical responses to the same questions, and 'global' agreement as the proportion of agreement within one response category Velikova et al. (1999). The difference or agreement between scores on the two types of questionnaire was calculated by subtracting the scores on the computer-assisted questionnaire from the scores from the standard questionnaire.

In addition, the effects of order of presentation were assessed by comparing differences between scores between the standard and computer-assisted

questionnaire by order in which these were completed using a one-way analysis of variance with one between subjects factor, order with two levels. This allowed an assessment to be made of whether completion of the standard questionnaire first acts as a cue for selection from the CA-questionnaire.

## 4.3. Results

### *4.3.1. Participants*

In total 110 patients were approached and asked to participate in the study. Eighty-eight patients (69 females and 19 males) agreed to participate in the study. Information regarding the date of birth of four patients (1 male and 3 female patients) was not available. The average age of the remaining patients was 57.9 years (range: 28.1 – 82.2 years). The average age of females was 57.7 years (range: 28.1 – 82.2 years). The average age of males was 58.8 years (range: 29.0 – 78.1). A breakdown of diagnoses by patient and gender is given in Table 4.3.1.

**Table 4.3.1. Diagnosis by patient gender**

| | Male | Female |
|---|---|---|
| Breast | 0 | 41 |
| Ovary | 0 | 19 |
| Uterus | 0 | 2 |
| Liver | 1 | 0 |
| Colorectal | 3 | 2 |
| Sarcoma | 1 | 0 |
| Bladder | 3 | 0 |
| Renal | 2 | 0 |
| Testis | 3 | 0 |
| Lymphoma | 1 | 1 |
| Melanoma | 3 | 2 |
| Unknown | 2 | 2 |
| | | |
| Total | 19 | 69 |

The average age of 22 patients (13 females and 9 males) who did not agree to participate in the study was 61.7 years (range: 40.5 years – 93.3 years). The

average age of females from this group was 63.1 years (range: 40.5 years – 84.1 years), and males 59.8 years (range: 43.4 years – 93.3 years). The age differences between the participators and non-participators was not statistically significant ($t < 1$).

The number of patients entered into each design were as follows: design 1, 32 patients (11 male and 21 female), design 2, 21 patients (4 male and 17 female), and design 3, 35 patients (4 male and 31 females).

*4.3.2. Time taken to complete questionnaires*

The average time taken to complete the questionnaires by study design is given in Table 4.3.2. There were no main effects for study design ($F < 1$), and the interaction between study design and questionnaire was also not significant. However, there was a main effect for time taken to complete each questionnaire ($F(1, 85) = 96.16$, $p < 0.01$), with the time taken on the CA-questionnaire less than half of that for the standard questionnaire (standard questionnaire 3 min 13 sec vs. CA-questionnaire 1 min 19 sec).

**Table 4.3.2. Time taken (in minutes) to complete questionnaires**

| | Standard questionnaire | | CA-questionnaire | |
|---|---|---|---|---|
| Design | Mean | range | Mean | range |
| 1 | 3.4 | 1.6 – 11.8 | 1.6 | 0.3 – 6.7 |
| 2 | 3.0 | 0.7 – 7.5 | 1.4 | 0.5 – 5.3 |
| 3 | 3.5 | 1.5 – 10.7 | 1.0 | 0.1 – 4.2 |
| | | | | |
| Overall | 3.3 | 0.7- 11.8 | 1.3 | 0.1 – 6.7 |

*4.3.3. Scores and Agreement between Quality-of-Life Measures*

Since all patients completed the Global Quality-of-Life (QOL) questions twice as a measure of reliability of their responses, the differences between the scores (for each questionnaire) were calculated. There were no significant differences between the

QOL scores from the three study designs (F(2, 85) = 1.36, n.s.) and therefore the EORTC QLQ-C30 scores were collapsed across the different study designs (Table 4.3).

In general the agreement scores from the Functional scales of the EORTC QLQ-C30 demonstrated negative means, indicating that patients were reporting lower levels of functioning on the standard questionnaire, compared to their responses on the CA-questionnaire. Similarly, patients reported greater levels of symptoms on the standard questionnaire, than on the CA-questionnaire as evidenced by the positive scores for the mean differences between the questionnaires.

The differences between the means for the two questionnaires were all statistically significant, with the exception of the means for Role Functioning and Social Functioning, Global Quality-of-Life (Table 4.3.3). Despite this the mean differences for all the scales fell within one standard deviation. Although, four particular scales showed higher levels of mean difference between the scores from the two questionnaires, i.e. Social Functioning (not statistically different), Nausea and Vomiting, Dyspnoea and Diarrhoea, the mean difference scores for these scales also fell within one standard deviation. Furthermore, since Global Quality-of-Life was used as a reliability index, high agreement between scores indicates a high level of reliability from responses to both forms of questionnaires.

The percentage exact agreement was low for all scales with the lowest exact agreement shown on the functioning scales (e.g. Emotional Functioning, 17%), and slightly better agreement (e.g. Appetite, 67%) on the symptom scales (Table 4.3.4).

In terms of global agreement, the Symptom scales (except Nausea and Vomiting, and Pain) and Global Quality-of-Life Scale demonstrated greater levels of agreement with 75% or more of responses falling within one category between questionnaires. This level agreement was only demonstrated by the remaining

**Table 4.3.3. Mean (standard deviations) of the EORTC QLQ-C30 scores**

| Scales | Standard questionnaire | | CA-questionnaire | | Difference between scales* | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. | t-test (d.f. = 87) | p |
| PF | 70.30 | 23.20 | 82.65 | 24.60 | -12.35 | 19.43 | 5.96 | p < 0.01 |
| RF | 81.63 | 24.64 | 79.55 | 34.81 | 2.08 | 40.89 | < 1 | p > 0.05 |
| EF | 70.64 | 22.21 | 79.36 | 28.20 | -8.71 | 27.99 | 2.92 | p < 0.01 |
| CF | 67.05 | 25.89 | 87.50 | 23.06 | -3.22 | 30.83 | 7.06 | p < 0.01 |
| SF | 85.23 | 23.49 | 88.45 | 26.67 | -20.45 | 27.18 | < 1 | p > 0.05 |
| QL | 52.08 | 25.78 | 53.60 | 25.45 | -1.52 | 9.67 | 1.47 | p > 0.05 |
| FA | 38.01 | 24.97 | 43.31 | 28.69 | -5.30 | 25.44 | 1.96 | p < 0.05 |
| NV | 38.83 | 27.42 | 11.17 | 22.56 | 27.65 | 32.35 | 8.02 | p < 0.01 |
| PA | 38.07 | 30.84 | 22.35 | 30.94 | 15.72 | 30.78 | 4.79 | p < 0.01 |
| DY | 39.77 | 34.60 | 19.70 | 30.17 | 20.08 | 36.63 | 5.14 | p < 0.01 |
| SL | 40.53 | 32.93 | 26.14 | 32.54 | 14.39 | 40.37 | 3.35 | p < 0.01 |
| AP | 25.00 | 31.66 | 15.91 | 29.46 | 9.09 | 24.09 | 3.54 | p < 0.01 |
| CO | 25.76 | 27.56 | 12.12 | 25.86 | 13.64 | 35.96 | 3.56 | p < 0.01 |
| DI | 30.68 | 28.24 | 8.71 | 21.14 | 21.97 | 35.34 | 5.83 | p < 0.01 |
| FI | 9.09 | 18.03 | 3.03 | 12.00 | 6.06 | 15.61 | 3.64 | p < 0.01 |

*(Score on Standard questionnaire) – (Score on CA-questionnaire).

**PF – Physical Functioning; RF – Role Functioning; EF – Emotional Functioning; CF – Cognitive Functioning; SF – Social Functioning; QL – Global Quality of Life; FA- Fatigue; NV – Nausea and Vomiting; PA – Pain; DY – Dyspnoea; SL – Sleeplessness; AP – Appetite; CO – Constipation; DI – Diarrhoea; FI – Finance.

Symptom scales and Functioning scales when two or three adjacent response categories were taken into account.

**Table 4.3.4. Cumulative percentage agreement between questionnaires per Category of Scale Scores**

| | PF | | EF | QL | | FA |
|---|---|---|---|---|---|---|
| Category | | Category | | | Category | |
| 0.00 | 20.77 | 0.00 | 17.34 | 58.97 | 0.00 | 21.38 |
| 6.67 | 48.09 | 8.33 | 31.35 | 81.86 | 11.11 | 48.48 |
| 13.33 | 64.18 | 16.67 | 58.09 | 94.04 | 22.22 | 66.99 |
| 20.00 | 81.86 | 25.00 | 73.03 | 95.70 | 33.33 | 89.40 |
| 26.67 | 85.19 | 33.33 | 85.78 | 98.96 | 44.44 | 94.20 |
| 33.33 | 86.78 | 41.67 | 88.41 | | 55.56 | 97.37 |
| 40.00 | 91.07 | 50.00 | 93.75 | 100.00 | 66.67 | 98.96 |
| 46.67 | 95.29 | 58.33 | 95.83 | | 77.78 | 100.00 |
| 60.00 | 98.96 | 66.67 | 97.92 | | 100.00 | 100.00 |
| 73.33 | 100.00 | 75.00 | 100.00 | | | |
| | | | | | | |
| | RF | CF | SF | NV | PA | |
| Category | | | | | | |
| 0.00 | 32.73 | 22.10 | 44.59 | 9.63 | 15.83 | |
| 16.67 | 52.75 | 56.73 | 70.73 | 46.96 | 59.21 | |
| 33.33 | 72.98 | 83.57 | 85.29 | 71.86 | 79.43 | |
| 50.00 | 81.36 | 91.95 | 91.04 | 84.74 | 90.99 | |
| 66.67 | 85.58 | 96.25 | 96.88 | 91.67 | 97.92 | |
| 83.33 | 94.58 | 97.29 | 98.96 | 97.92 | 100.00 | |
| 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | |
| | | | | | | |
| | DY | SL | AP | CO | DI | FI |
| Category | | | | | | |
| 0.00 | 34.99 | 33.35 | 67.06 | 46.63 | 39.62 | 81.41 |
| 33.33 | 77.82 | 76.02 | 93.70 | 84.04 | 83.49 | 97.92 |
| 66.67 | 95.83 | 94.25 | 98.96 | 97.29 | 94.58 | 100.00 |
| 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | |

*Abbreviations are as for Table 4.3

Table 4.3.5. demonstrates that with the exception of Fatigue, Pain and Sleeplessness the majority, or 60% or more of patients did not select the corresponding option relating to the quality-of-life domain from the computer-assisted questionnaire. Interestingly almost 50% of patients did select Emotional Functioning on the CA-questionnaire. However, for the remaining Functional scales the majority of patients did not select the corresponding option on the CA-questionnaire.

**Table 4.3.5. Table showing number and percentage of patients not selecting option from Computer-assisted questionnaire**

| | PF | RF | EF | CF | SF |
|---|---|---|---|---|---|
| **N** | 53 | 61 | 49 | 63 | 72 |
| % | 60.92 | 70.11 | 56.32 | 72.41 | 82.76 |
| | | | | | |
| | | | | | |
| | FA | NV | PA | DY | SI |
| **N** | 15 | 65 | 49 | 54 | 45 |
| % | 17.24 | 74.71 | 56.32 | 62.07 | 51.72 |
| | | | | | |
| | | | | | |
| | | | | | |
| | AP | CO | DI | FI | |
| **N** | 64 | 68 | 71 | 81 | |
| % | 73.56 | 78.16 | 81.61 | 93.10 | |

Figures 4.3.1. and 4.3.2. show the cumulative percentage of agreement in terms of categories selected on the standard questionnaire for those patients who did not select the scales, i.e. those patients who, in effect, reported perfect functioning or total absence of symptoms on the computer-assisted questionnaire. It can be seen for instance, that only for the Role or Social Functioning scales did 50% of the patients not selecting problems from the computer-assisted questionnaire, also score 100 on their responses from the standard questionnaire. A similar pattern was also only demonstrated for the Symptom scales for Nausea and Vomiting, and Finance.

**Figure 4.3.1. Cumulative percentage of responses by agreement in categories for Functioning Scales (for non-selection on CA questionnaire)**

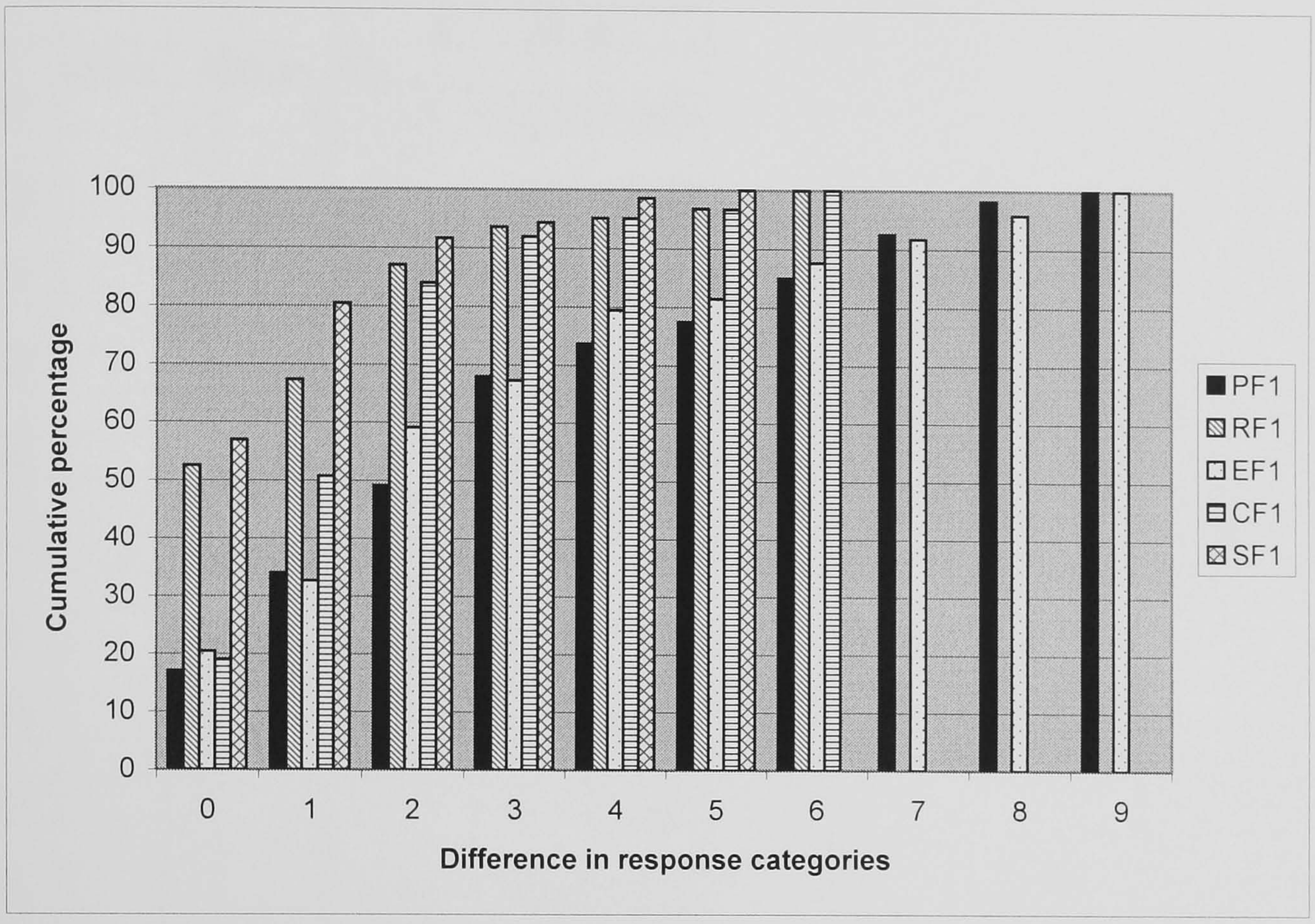



**Figure 4.3.2. Cumulative percentage of responses by agreement in categories for Symptom Scales (for non-selection on CA questionnaire)**

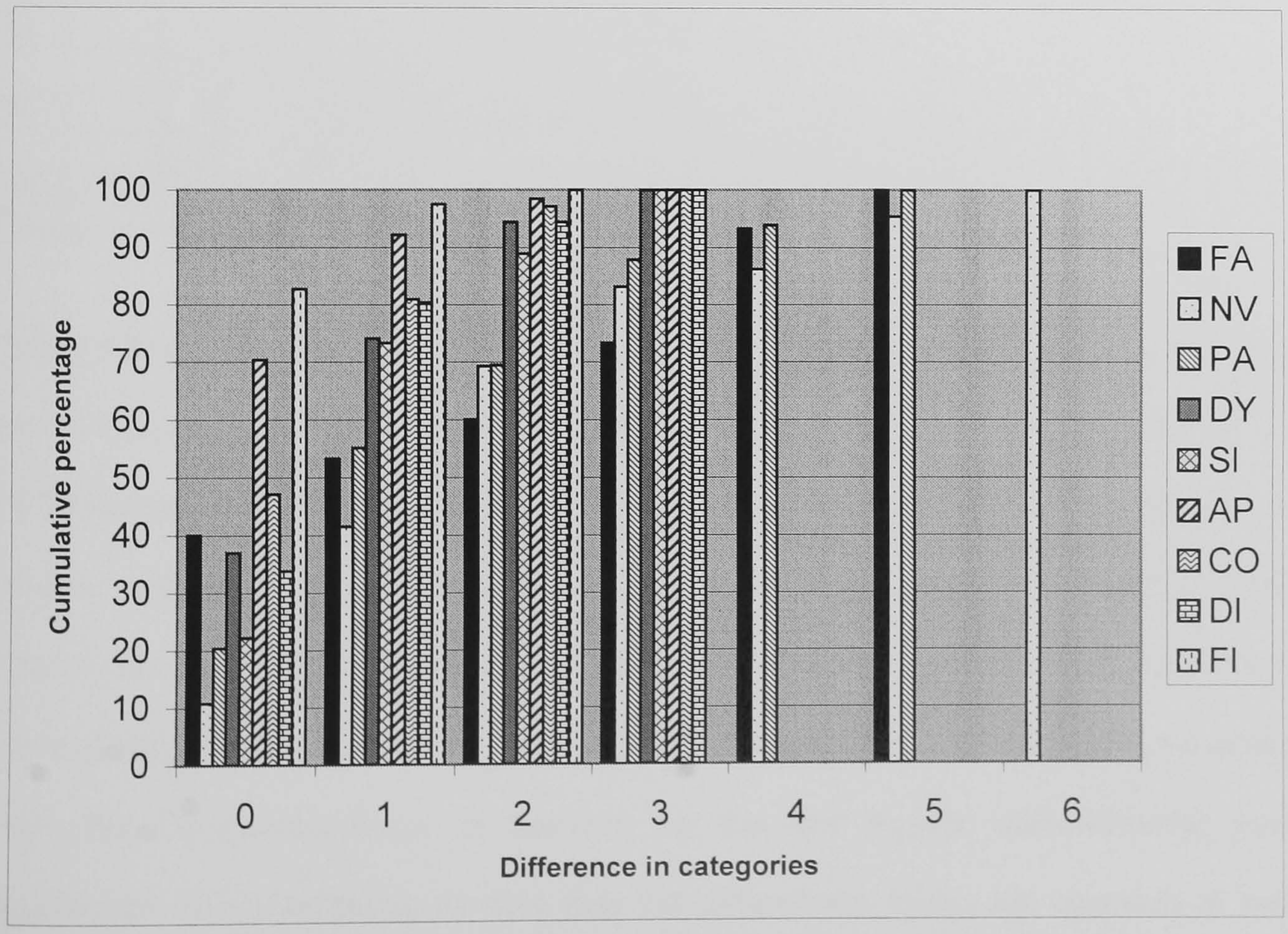


**Table 4.3.6. Mean level of agreement between Standard and CA-questionnaires grouped by patients for scales from the EORTC QLQ-C30**

| | | | | | t | df | p |
|---|---|---|---|---|---|---|---|
| | Group | N | Mean | S.D. | | | |
| PF | 1 | 34 | 1.96 | 11.90 | 6.39 | 85 | 0.0001 |
| | 2 | 53 | -21.01 | 18.65 | | | |
| RF | 1 | 26 | 49.36 | 31.79 | 10.60 | 85 | 0.0001 |
| | 2 | 61 | -18.03 | 24.96 | | | |
| CF | 1 | 24 | 1.39 | 32.20 | 4.64 | 85 | 0.0001 |
| | 2 | 63 | -26.98 | 22.49 | | | |
| EF | 1 | 38 | 8.99 | 24.61 | 6.60 | 85 | 0.0001 |
| | 2 | 49 | -23.64 | 21.41 | | | |
| SF | 1 | 15 | 44.44 | 32.53 | 9.14 | 85 | 0.0001 |
| | 2 | 72 | -12.96 | 19.42 | | | |
| FA | 1 | 72 | -12.19 | 21.44 | -5.43 | 85 | 0.0001 |
| | 2 | 15 | 21.48 | 23.93 | | | |
| NV | 1 | 22 | 1.52 | 37.41 | -4.64 | 85 | 0.0001 |
| | 2 | 65 | 35.64 | 26.82 | | | |
| PA | 1 | 38 | -2.19 | 30.55 | -5.30 | 85 | 0.0001 |
| | 2 | 49 | 28.91 | 24.24 | | | |
| DY | 1 | 33 | 0.00 | 39.97 | -4.18 | 85 | 0.0001 |
| | 2 | 54 | 31.48 | 29.97 | | | |
| SL | 1 | 42 | -12.70 | 33.70 | -7.48 | 85 | 0.0001 |
| | 2 | 45 | 38.52 | 30.11 | | | |
| AP | 1 | 23 | 0.00 | 24.62 | -2.31 | 85 | 0.05 |
| | 2 | 64 | 13.02 | 22.71 | | | |
| CO | 1 | 19 | -29.82 | 31.22 | -7.39 | 85 | 0.0001 |
| | 2 | 68 | 25.00 | 27.84 | | | |
| DI | 1 | 16 | -22.92 | 37.94 | -6.43 | 85 | 0.0001 |
| | 2 | 71 | 30.52 | 28.03 | | | |
| FI | 1 | 6 | 0.00 | 21.08 | -0.99 | 85 | 0.32 |
| | 2 | 81 | 6.58 | 15.29 | | | |

Group 1 = Patients selecting corresponding option on CA-questionnaire
Group 2 = Patients not selecting corresponding option on CA-questionnaire

Table 4.3.6. shows the mean level of agreement (between scores from the standard and CA-questionnaires) between patients grouped by whether they selected quality of life domains on the CA-questionnaire. It can be seen that with the exception of four scales, namely Role Functioning and Social Functioning, Constipation and Diarrhoea, the level of agreement between the two versions of the questionnaire were closer to exact agreement for those patients selecting the corresponding option from the CA-questionnaire. In addition, for the four scales demonstrating poor agreement, it is interesting to note that the differences found are opposite to that

observed for the overall sample, i.e. greater problems in functioning and symptoms were recorded on the CA-questionnaire compared to the standard questionnaire for patients selecting the domain from the CA-questionnaire.

Finally, the effects of the order of presentation on agreement scores were evaluated, in order to assess whether completion of either form of the questionnaire may influence completion of the other questionnaire by "priming" the patient or providing them with cues regarding aspects of their quality of life, which may subsequently influence their response to the second questionnaire (Table 4.3.7).

The results of the analysis of order effects comparing differences between the standard and computer-assisted version of the questionnaire grouped by order of presentation showed that although there were minor discrepancies between differences for the two presentation orders for all scales, all discrepancies were in the same direction. Moreover, a univariate ANOVA comparing differences by order effects demonstrated no significant differences for order of presentation.

**Table 4.3.7. Means and results of ANOVA between Agreement Scores by Order of Presentation**

| | | Mean | Std. Deviation | ANOVA | |
|---|---|---|---|---|---|
| Scale | Order | | | F | Sig. |
| PF | 1 | -10.76 | 19.51 | 0.37 | 0.55 |
| | 2 | -13.33 | 20.21 | | |
| RF | 1 | 3.41 | 39.96 | 0.09 | 0.77 |
| | 2 | 0.78 | 42.72 | | |
| EF | 1 | -11.93 | 28.27 | 0.74 | 0.39 |
| | 2 | -6.78 | 27.71 | | |
| CF | 1 | -21.59 | 28.66 | 0.65 | 0.42 |
| | 2 | -16.67 | 28.17 | | |
| SF | 1 | -4.17 | 33.74 | 0.11 | 0.74 |
| | 2 | -1.94 | 28.22 | | |
| QL | 1 | -0.19 | 11.85 | 1.48 | 0.23 |
| | 2 | -2.71 | 6.74 | | |
| FA | 1 | -4.04 | 23.97 | 0.77 | 0.38 |
| | 2 | -8.79 | 26.51 | | |
| NV | 1 | 29.92 | 25.81 | 0.68 | 0.41 |
| | 2 | 24.03 | 39.39 | | |
| PA | 1 | 18.56 | 34.52 | 0.96 | 0.33 |
| | 2 | 12.02 | 27.30 | | |
| DY | 1 | 16.67 | 40.98 | 0.53 | 0.47 |
| | 2 | 22.48 | 33.11 | | |
| SL | 1 | 13.64 | 43.92 | 0.00 | 0.97 |
| | 2 | 13.95 | 37.96 | | |
| AP | 1 | 6.06 | 19.39 | 1.97 | 0.16 |
| | 2 | 13.18 | 27.35 | | |
| CO | 1 | 15.91 | 34.09 | 0.55 | 0.46 |
| | 2 | 10.08 | 38.86 | | |
| DI | 1 | 23.48 | 37.06 | 0.52 | 0.47 |
| | 2 | 17.83 | 35.89 | | |
| FI | 1 | 6.06 | 16.51 | 0.00 | 0.97 |
| | 2 | 6.20 | 15.01 | | |

*Order 1 = Standard questionnaire followed by Computer-assisted questionnaire; Order 2 = Computer-assisted questionnaire followed by Standard questionnaire.
**Scale abbreviations are as for Table 4.3.3.
***Difference calculated by (Standard scores) –(Computer-assisted scores)

## 4.4. Discussion

Quality of life assessment of patients is important in oncology clinics. However questionnaires tend to be lengthy and since questionnaires are often designed for clinical trials they tend not be relevant to individual patients. There is therefore a

need to design questionnaires that can both be specific and relevant to individual patients, whilst allowing comparisons to be made across patients.

This study attempted to evaluate a computer-assisted (CA) questionnaire that allowed patients to select areas of concern from a standard questionnaire. Comparisons were made to patients' responses on the standard questionnaire.

The results demonstrated that patients were able to complete the CA questionnaire in roughly half the time it took them to complete the standard questionnaire. However this was at a cost of accuracy (in terms of agreement between responses to the types of questionnaires) when comparing responses to a fixed length standard questionnaire. Significant differences were found between all but 3 scales of the questionnaires, indicating poor exact agreement. The data illustrate the substantial influence of the mode of presenting questions and emphasize the critical need to evaluate changes in presentation.

Velikova et al. (1999) have pointed out that higher agreement would be expected for shorter scales (e.g. Appetite) consisting of a single question, rather than longer scales (e.g. Physical Functioning), and this was indeed found in the results. However, although the shorter symptom scales demonstrated better global agreement when adjacent categories were taken into account, global agreement remained poor for the longer symptom scales (e.g. Pain and Fatigue) and Functioning scales. These scales required 2 or even 3 additional response categories to be taken into account before global agreement levels reached 75%.

Overall the results showed that the majority of patients were not selecting domains from the CA-questionnaire selection screen. This was particularly the case for the Symptom scales. However, a more detailed analysis of the agreement scores which split the respondents into two groups depending on whether patients had also selected the domain from the CA-questionnaire, revealed an interesting pattern of responses. These results showed predominantly that the "inconsistencies" in responses between the standard and CA-questionnaire were limited to those

individual patients who had not selected the corresponding domains from the CA-questionnaire. Those patients who had selected domains from the CA-questionnaire showed better agreement in scores from the two versions of the EORTC QLQ-C30. In addition, where the agreement between scores was poor patients reported poorer functioning and a greater level of problems with symptoms on the CA-questionnaire.

In contrast to the overall pattern of results and those from patients not selecting domains from the CA-questionnaire, the overall direction or bias of the differences between responses to the questionnaires indicated that these patients (that is patients selecting options from the CA-questionnaire) scored higher levels of functioning and lower levels of symptoms on the CA-questionnaire compared to the standard. Clearly these patients are under-reporting concerns on the CA-questionnaire, if we take the full questionnaire as standard, although it is not known why the patients are doing this. The Global Quality-of-Life scale, which was used as reliability index, demonstrated good levels of exact agreement and global agreement, therefore patients are responding reliably to this scale, and we could speculate that patients should also be responding reliably to the other scales. However, this does not explain the differences found in responses to the questionnaires. It would appear that perhaps the format of the questionnaires affected patient responses' for those patients not selecting domains from the CA-questionnaire. The fact that better functioning and fewer symptoms were recorded on the CA-questionnaire for these patients demonstrates that they selected fewer items from the selection screen. It may well be that patients did not identify these areas on the selection screen as areas of concern or indeed as problems. In this sense the wording of the items on the selection screen are critical, and patients may not have selected items because of interpreting the questionnaire format differently. Therefore patients may have been selecting domains which were of importance to them, and which may impact on their quality of life.

However, the differences found between the means of the scales all fell within one standard deviation of the individual scales, demonstrating that despite the poor level of exact and global agreement, the differences observed between the two types of questionnaires were relatively small.

The results of the overall pattern of agreement scores are similar to those of others, e.g. Watson et al. (1992) report the results of a meta-analysis of nine studies comparing responses from the computerised version of the Minnesota Multiphasic Personality Inventory (MMPI) with those from the paper version completed by the same patients. The results of the meta-analysis demonstrated that the computerised MMPI significantly underestimated scores compared with those from the paper-and-pencil version of the MMPI on most domains, although these differences were small. More recent research by Boyes et al. (2002) has demonstrated that the format in which computerised questionnaires are presented may affect responses. Patients in this study completed a paper version of the Supportive Care Needs Survey (SCNS), and a computer version of these questionnaires which was either a close as possible to the paper version, or which was designed to allow patients to select for themselves whether help was needed in specific domains. The results of the study by Boyes et al. (2002), were similar to those of this study, namely that higher levels of agreement and exact agreement were demonstrated for the computerised questionnaire which was consistent with the paper version. These authors conclude that the closer the format of the computerised questionnaire is to that of the paper questionnaire the closer the scores produced.

Additionally, there are a number of problems with this study. Firstly, the sample size was small. This means that the power of the study was low. A recalculation of the power requirements indicated that between 120 and 150 patients would be required to raise the power of the study to above 0.70. Furthermore, small sample sizes are potentially subject to larger variances, which could also have

affected the results. Secondly the sample used was predominantly females with breast cancer, which limits the generalisation of the results.

In conclusion, the CA-questionnaire allowed a quick individual assessment to be made, which provided a global impression or 'snapshot' of the health status of the patient. For the majority of patients the scores generated this way would correspond roughly to the scores generated from the full standard version of the questionnaire within one standard deviation, and it may provide a means whereby subsequent questions could be presented to investigate problem areas more thoroughly. Therefore, overall the computer-assisted questionnaires may present a more realistic picture of patients' problems, compared to the standard questionnaires which may reflect patients' symptoms. This is certainly the case for the minority of patients who selected domains from the CA-questionnaires, and whose scores demonstrated high levels of agreement between the two forms of the questionnaire. For these patients scores on the two versions of the questionnaire were interchangeable for most of the quality of life domains.

It remains for future research to attempt to establish why patients respond differently to CA-questionnaires. A promising avenue of research, which may help to shed light on this phenomenon is referred to as the Cognitive aspects of survey methodology (CASM, e.g. McColl, Meadows and Barofsky, 2003). CASM is concerned with exploring the cognitive processes involved when patients respond to questionnaires. Although a number of competing models exist the most commonly employed model is Tourangeau and colleagues' "Four Stage Model" (e.g. Jobe, 2003; Tourangeau, Rips and Rasinski, 2000). However, all models include as a minimum four processing stages (from Jobe, 2003, p. 219): 1). Comprehension; 2). Retrieval of information; 3). Judgement, including use of heuristics; and 4). Response.

Future work could explore the interaction between features of the design of the CA-questionnaire and stages of processing from the Four-Stage Model

(Tourangeau et al., 2000), and assess whether this interaction provides an explanation for patients not selecting domains from the CA-questionnaire.

Owing to the difficulties with the computer-assisted approach as highlighted by the limited correspondence between scores derived from this method and those from standard questionnaires demonstrated in this study, and confirmed in the literature, the computer-assisted methodology was not explored further. However, electronic questionnaire whether they be standard questionnaires (e.g. Velikova et al., 1999) or CA-questionnaires as explored here, provide a useful tool not only for assessing patients' quality of life, but also as a process for facilitating doctor-patient communication. For instance, a number of recent studies (Detmar et al., 2002; Velikova et al., 2002; Velikova et al. 2004) has shown that the availability of quality of life results prior to clinical consultation facilitated doctor-patient interaction, by improving doctors' awareness of patient's quality of life, and helping them identify issues for discussion. Furthermore, patients felt questionnaires were useful for informing doctors of their problems. In addition a more recent study has highlighted that recording quality of life information prior to consultation can in turn have a positive influence on patients' quality of life (Velikova et al., 2004).

Subsequent chapters will detail the results of the Rasch analyses of questionnaires. These analyses were carried out in order to identify possible items for removal from the quality of life instruments to reduce the number of items presented to patients through the selection of misfitting items. In addition, traditional psychometrics which were also carried out on questionnaires will be compared with results from Rasch analysis.

## 5. Factor and Rasch Analysis of the Hospital Anxiety & Depression Scale

The Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983) has been widely used as a self-report instrument for screening for psychiatric distress in psychiatric and medical patient populations.

The two-factor structure of the questionnaire (i.e. anxiety and depression) has been confirmed in a number of studies (e.g. Dagnan, Chadwick, & Trower, 2000; Lisspers, Nygren, & Söderman, 1997; Moorey et al., 1991; Spinhoven et al., 1997, White et al., 1999), although there is some evidence for different factor structures (Andersson, 1993; Lewis, 1991). Previous research has also indicated that some items from Anxiety subscale may load onto the Depression subscale (e.g. Moorey et al., 1991).

Additionally reports of the screening efficacy of the HADS have shown the instrument's ability to detect cases or anxiety and/or depression may be limited (Abiodun, 1994; Hall et al., 1999; Hopwood et al., 1991; Ibbotson et al., 1994; Lewis & Wessely, 1990; Razavi et al., 1990; Silverstone, 1993; Spinhoven et al., 1997).

This chapter describes a traditional psychometric analysis of the Hospital Anxiety & Depression Scale (Zigmond & Snaith, 1983), including a first and second-order factor analysis, as well as a Rasch analysis of the total scale and the individual subscales HADS-Anxiety and HADS-Depression. The aims of the study are to explore the factor structure from a heterogeneous cancer patient population using factor analysis and to carry out Rasch analysis on the HADS to assess the unidimensionality of the instrument and subscales, as well as to identify misfitting items, which potentially can be removed to make the questionnaire more suitable for routine clinical applications.

## 5.1. Factor Analysis of the HADS

### *5.1.1. Aim*

This study investigated the factor structure of the HADS in a large heterogeneous sample of 1474 cancer patients. Factor analyses were carried out to investigate the factor structure of the instrument across gender and different age groups, and with a subgroup of patients with metastatic cancer. A second-order factor analysis was also performed on the total dataset to explore whether HADS conforms in general to the tripartite model, and specifically to a hierarchical model (e.g. Clark and Watson, 1991, see Chapter 2).

### *5.1.2 Method*

Patient data was collated from a total of five studies which have been carried out by the ICRF (now Cancer Research UK) Psychosocial Oncology Groups, at St. James's University Hospital, Leeds and Western General Hospital, Edinburgh over the past four years. The data from two studies have been previously reported (Cull et al., 2001; Velikova et al., 1999). All patients completed the HADS on a computer with a touchscreen monitor.

Ethical approval for the studies had been given by the local hospital ethics committees in Leeds and Edinburgh.

### *5.1.3 Participants*

A total of 1474 patients participated in the studies. The majority of patients were recruited from outpatient oncology clinics (609 patients from St. James's Hospital, Leeds, and 785 patients from the Western General Hospital, Edinburgh). 80 patients were recruited from a general oncology ward (Cookridge and St. James's Hospitals).

The average age of the sample was 55.9 years (range = 15.3 - 99.6). The total number of males taking part was 632, with an average age of 54.5 years (range = 15.3 - 91.7), and the number of females was 842, average age = 57.0 years (range

= 17.8 - 99.6). Some data from five patients was missing; the remaining data from these patients has been included in the analyses. Table 5.1.1 gives the diagnoses.

**Table 5.1.1 Diagnoses of patients**

| **Diagnosis** | Number of patients (percentage) |
|---|---|
| Breast | 379 (25.7) |
| Gastro-intestinal | 152 (10.3) |
| Lung | 46 (3.1) |
| Male genito-urinary | 262 (17.8) |
| Female genito-urinary | 307 (20.8) |
| Lymphoma | 45 (3.1) |
| Other | 283 (19.2) |

*5.1.4 Methodology*

A principal components analysis was carried out on the data. Factors were identified using a scree plot and Kaiser's criterion of eigenvalues greater than 1. Subsequently a factor analysis was carried out on the rotated data. An orthogonal rotation, e.g. oblimin was employed since the previous studies have suggested a correlation between anxiety and depression.

*5.1.5 Results*

*5.1.6 HADS scores by Age and Gender*

The sample was split into three groups of approximately the same size based on age (group 1, <50 years, n = 492; group 2, ≥ 50 and <65, n = 519; and group 3, ≥ 65, n = 454) to allow comparisons of scores across different ages.

The mean score on the Anxiety subscale for all patients was 6.05 (s.d. 4.03) with a range between 0 and 21. The mean score for the Depression subscale was lower at 4.38 (s.d. 3.73) with a range between 0 and 20.

A breakdown of the scores by gender demonstrated higher scores for both Anxiety (mean 6.65, s.d. 3.99) and Depression (mean 4.61, s.d. 3.93) for women compared to men (mean 5.22, s.d. 3.73, and mean 4.06, s.d. 3.73, respectively). This

pattern was also observed for each age group. Table 5.1.2 shows the mean HADS scores by age and gender.

**Table 5.1.2 Mean HADS scores by age and gender (standard deviation)**

| **Age** | **<50** | **=> 50 & <65** | **=> 65** | **Max. score** | **Mean** |
|---|---|---|---|---|---|
| **HADS-A:** | | | | | |
| females | 6.74 (4.01) | 6.97 (4.01) | 6.07 (3.90) | 18 | 6.65 (3.99) |
| males | 5.56 (4.14) | 5.31 (3.79) | 4.75 (3.77) | 21 | 5.22 (3.94) |
| **HADS-D:** | | | | | |
| females | 4.08 (3.75) | 4.69 (3.81) | 4.99 (3.52) | 20 | 4.61 (3.73) |
| males | 3.48 (3.75) | 4.33 (3.66) | 4.54 (3.68) | 20 | 4.06 (3.73) |

Although the summated anxiety and depression scales were not normally distributed, and were also highly skewed (skewness .63 (s.e. 0.064)) for Anxiety, and .62 (s.e. 0.064) for Depression), Levene's test of equality of error variances revealed no significant differences for either subscale ($F < 1$). Therefore, a multivariate analysis of variance was performed on the subscale scores with two 'between group' factors (gender with two levels, and age with the three levels), which showed significant effects for Anxiety and Depression by gender ($F(1, 1464) = 42.66$, $p < .0001$) and ($F(1, 1464) = 5.54$, $p < .05$) respectively, and by age group ($F(2, 1459) = 5.19$, $p < .05$) and ($F(2, 1459) = 9.05$, $p < .001$) respectively. The interaction between gender and age group for the two subscales was not statistically significant ($F < 1$).

*Post hoc* Bonferroni contrasts demonstrated significant differences ($p < .05$) between the Anxiety subscales for group 1 and 3, and group 2 and 3. Indicating that the oldest patients were experiencing significantly lower levels of anxiety than the youngest group of patients. The contrasts for the Depression subscale indicated significant differences between groups 1 and 2, and groups 1 and 3. The contrast between group 2 and 3 was not statistically different. These results demonstrated that levels of depression increased with age.

**Table 5.1.3 Mean HADS scores by cancer site**

| | HADS-A | | HADS-D | | HADS-total | |
|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Breast | 7.14 | 4.49 | 3.76 | 3.47 | 10.90 | 7.37 |
| Gastro-intestinal | 6.28 | 4.38 | 5.13 | 3.92 | 11.40 | 7.50 |
| Lung | 4.82 | 3.22 | 5.09 | 3.91 | 9.91 | 6.35 |
| Male Genitourinary | 4.71 | 3.82 | 2.92 | 3.44 | 7.63 | 6.48 |
| Female Genitourinary | 6.39 | 3.91 | 4.39 | 3.76 | 10.79 | 6.77 |
| Other | 5.67 | 3.94 | 4.26 | 3.59 | 9.92 | 6.70 |

A breakdown of HADS-A, HADS-D and HADS scores by cancer site is shown in Table 5.1.3. It can be seen that higher levels of anxiety were reported by breast cancer patients, as well as patients with gastro-intestinal and female genitourinary cancers. In contrast patients with lung cancers and male genitourinary cancers reported lower levels of anxiety. A univariate ANOVA with one between subjects factor (diagnosis with six levels) demonstrated significant differences between levels of anxiety by diagnosis ($F(5, 1469) = 5.85$, $p < 0.001$). A post hoc bonferroni test revealed only significant differences between the male genitourinary cancer group and both the breast cancer and female genitourinary cancer groups.

On the other hand breast cancer patients, as well as male genitourinary cancer patients reported low levels of depression, whereas both the gastro-intestinal and lung cancer patients demonstrated higher levels. A univariate ANOVA with one between subjects factor (diagnosis with six levels) demonstrated significant differences between levels of depression by diagnosis ($F(5, 1469) = 5.56$, $p < 0.001$). A post hoc bonferroni test revealed significant differences between the male genitourinary cancer group and the female genitourinary cancer, gastro-intestinal cancer and other cancer groups.

Finally, high levels of psychological distress were recorded for the breast, gastro-intestinal and female genitourinary cancer groups. Once again a univariate ANOVA with one between subjects factor (diagnosis with six levels) demonstrated significant differences between levels of psychological distress by diagnosis (F(5,

1469) = 6.27, $p < 0.001$). A post hoc bonferroni test revealed significant differences between the male genitourinary cancer group and the female genitourinary cancer, gastro-intestinal cancer and breast cancer groups.

*5.1.7 Depression and Anxiety*

Employing a threshold score of 8, as recommended by Zigmond and Snaith (1983) resulted in 33.3 % (491/1469) of patients being identified as at least 'possible cases' of anxiety, and 19.8 % (291/1469) as 'possible cases' for depression. For each subscale the proportion of patients scoring greater than 8 was higher among women than men (women: anxiety 334/842; depression 80/842, vs. men: anxiety 156/627, depression 110/627). These gender differences were statistically significant (anxiety: $\chi^2 = 64.66$, d.f. = 1 $p < .005$, and depression, $\chi^2 = 16.9$, d.f. = 1, $p < .005$).

*5.1.8 Factor Analysis*

The correlation matrix for the HADS items is shown in Table 5.1.3. Table 5.1.4 shows the inter-item reliability for the HADS. The various subgroups show the same Cronbach's alpha, which is around 0.83 for the Anxiety subscale and 0.79 for Depression. Both these are within the acceptable limits (e.g. Nunnally and Bernstein, 1994). Table 5.1.5 demonstrates the item-total correlations and the changes to the reliability coefficient when items are removed from the subscales (the revised Cronbach's alpha).

**BLANK IN ORIGINAL**

## Table 5.1.4 Inter-item correlation matrix for HADS

| | | HADS-A | | | | | | | HADS-D | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Item 1 | Item 3 | Item 5 | Item 7 | Item 9 | Item 11 | Item 13 | Item 2 | Item 4 | Item 6 | Item 8 | Item 10 | Item 12 |
| HADS-A | Item 1 | . | . | . | . | . | . | . | . | . | . | . | . | . |
| | Item 2 | .50 | . | . | . | . | . | . | . | . | . | . | . | . |
| | Item 3 | .55 | .61 | . | . | . | . | . | . | . | . | . | . | . |
| | Item 4 | .36 | .32 | .36 | . | . | . | . | . | . | . | . | . | . |
| | Item 5 | .41 | .49 | .44 | .33 | . | . | . | . | . | . | . | . | . |
| | Item 6 | .33 | .32 | .36 | .29 | .25 | . | . | . | . | . | . | . | . |
| | Item 7 | .52 | .58 | .57 | .30 | .50 | .37 | . | . | . | . | . | . | . |
| HADS-D | Item 1 | .20 | .22 | .27 | .38 | .21 | .14 | .20 | . | . | . | . | . | . |
| | Item 2 | .23 | .21 | .24 | .36 | .23 | .09 | .20 | .36 | . | . | . | . | . |
| | Item 3 | .43 | .39 | .47 | .39 | .27 | .26 | .37 | .38 | .39 | . | . | . | . |
| | Item 4 | .30 | .26 | .35 | .29 | .16 | .24 | .26 | .47 | .22 | .39 | . | . | . |
| | Item 5 | .26 | .21 | .25 | .23 | .11 | .21 | .25 | .26 | .21 | .35 | .29 | . | . |
| | Item 6 | .37 | .32 | .37 | .42 | .27 | .18 | .28 | .51 | .46 | .49 | .41 | .40 | . |
| | Item 7 | .27 | .25 | .27 | .35 | .24 | .21 | .24 | .26 | .31 | .35 | .25 | .29 | .40 |

* all correlations are significant at $p < 0.001$

## Table 5.1.5 Inter-item reliability coefficients (Cronbach's alpha) of the HADS

| | HADS- A | HADS-D |
|---|---|---|
| **Total** | 0.83 | 0.79 |
| **Age group 1** | 0.85 | 0.82 |
| **Age group 2** | 0.83 | 0.79 |
| **Age group 3** | 0.81 | 0.74 |
| **Females** | 0.82 | 0.78 |
| **Males** | 0.83 | 0.79 |
| **Split-half 1** | 0.84 | 0.80 |
| **Split-half 2** | 0.82 | 0.78 |

**Table 5.1.6 Item-total correlations and the revised Cronbach's α for HADS-A & HADS-D**

| HADS-Anxiety: | | | Items from HADS-Depression | | |
|---|---|---|---|---|---|
| | | | | | |
| **Item-total Correlation** | **Revised Cronbach's α** | | **Item-total Correlation** | **Revised Cronbach's α** | |
| | | | | | |
| | | | | | |
| **Item 1** | .63 | .80 | **Item 1** | .56 | .75 |
| **Item 2** | .67 | .79 | **Item 2** | .46 | .77 |
| **Item 3** | .68 | .79 | **Item 3** | .58 | .75 |
| **Item 4** | .44 | .83 | **Item 4** | .50 | .77 |
| **Item 5** | .56 | .81 | **Item 5** | .43 | .78 |
| **Item 6** | .43 | .83 | **Item 6** | .67 | .73 |
| **Item 7** | .67 | .79 | **Item 7** | .44 | .77 |

The literature suggests a correlation between the anxiety and depression factors (Clark & Watson, 1991) therefore a principal components analysis with an oblique rotation was used for the factor analysis. The factor analysis of the entire dataset revealed a two-factor structure (Table 5.1.6). The first factor explained 37.69% of the variance (eigenvalue of 5.27) and the second factor accounting for 11.49% of the variance (eigenvalue of 1.61). The remaining factors had eigenvalues less than 1, and were therefore not selected for subsequent analysis (Figure 1). The rotated factor structure revealed two factor structures corresponding to the Anxiety and Depression subscale. Only one item, (item 4, "I can sit at ease and feel relaxed") from the Anxiety subscale loaded more strongly on the Depression subscale. The two factors were significantly correlated ($r = 0.52$).

**Table 5.1.7 Rotated Factor Structure for the entire dataset (n=1474)**

| | Factor 1 | Factor 2 |
|---|---|---|
| **Items from HADS-Anxiety subscale:** | | |
| Item 1 – "Tense" | 0.70 | 0.09 |
| Item 2 – "Frightened" | 0.82 | -0.04 |
| Item 3 – "Worrying" | 0.76 | 0.09 |
| Item 4 – "Relaxed" | 0.20 | 0.52 |
| Item 5 – "Butterflies" | 0.72 | -0.06 |
| Item 6 – "Restless" | 0.57 | -0.02 |
| Item 7 - "Panic" | 0.85 | -0.08 |
| | | |
| **Items from HADS-Depression subscale:** | | |
| Item 1 – "Enjoy things" | -0.17 | 0.81 |
| Item 2 – "Laugh" | -0.12 | 0.70 |
| Item 3 – "Cheerful" | 0.23 | 0.57 |
| Item 4 – "Slowed down" | 0.05 | 0.60 |
| Item 5 – "Appearance" | 0.02 | 0.54 |
| Item 6 – "Enjoyment" | -0.02 | 0.81 |
| Item 7 – "Enjoy book" | 0.06 | 0.55 |

Factor analyses of the data by gender, age-group and a split-half reliability sample, where the dataset was divided into two subsets and the factor structure analysed for both sets, revealed similar factor structures. The factor loadings for the first two age groups and females were reversed compared to the entire dataset and the other factor analyses. However, the factor structures remain the same. In addition, extent of disease had also been recorded for a subset of the patients. The data from 197 patients with metastatic disease were also analysed and the same factor structure was again demonstrated.

Since the factor analyses revealed that the two subscales were strongly correlated a second-order factor analysis and a Schmid-Leiman transformation were carried out on the data. The first-order factors had factor loadings of 0.59 on a second-order factor, and around 70% of the variance of these factors was explained by the second-order factor. As can be seen from Table 7, psychological distress accounts for about a third of the common variance.

**Table 5.1.8 - Factor structure following second-order factor analysis and transformation**

| | Psychological Distress | Anxiety | Depression |
|---|---|---|---|
| **Items:** | | | |
| | | | |
| **Anxiety:** | | | |
| **Item 1** | 0.46 | 0.57 | 0.06 |
| **Item 2** | 0.47 | 0.66 | -0.02 |
| **Item 3** | 0.50 | 0.62 | 0.06 |
| **Item 4** | 0.42 | 0.16 | 0.42 |
| **Item 5** | 0.39 | 0.58 | -0.05 |
| **Item 6** | 0.33 | 0.46 | -0.01 |
| **Item 7** | 0.45 | 0.69 | -0.06 |
| | | | |
| **Depression:** | | | |
| **Item 1** | 0.38 | -0.14 | 0.66 |
| **Item 2** | 0.31 | -0.14 | 0.57 |
| **Item 3** | 0.47 | 0.19 | 0.46 |
| **Item 4** | 0.38 | 0.03 | 0.49 |
| **Item 5** | 0.33 | 0.02 | 0.44 |
| **Item 6** | 0.47 | -0.02 | 0.66 |
| **Item 7** | 0.36 | 0.05 | 0.45 |
| **Eigenvalue** | 2.38 | 2.27 | 2.21 |
| **% common variance** | 34.70 | 33.07 | 32.23 |

*5.1.9 Discussion*

Previous studies have reported a single factor, as well as two, three and four factor structures for the HADS. The results from this study demonstrated a two-factor structure for HADS in a very large heterogeneous sample of cancer patients. The factor structure remained when the sample was divided into three age groups, and comparisons between males and females, as well patients with metastatic cancer, revealed the same two factors approximately corresponding to anxiety and depression.

Given the heterogeneity of the sample there is a potential for the factor structure of the HADS to differ markedly by disease site. However, the factor structure remained constant across age groups, between males and females, as well as when extent of disease was take into consideration, even though age effects,

gender differences and differences between cancer site were observed for the scores from the subscales of HADS. Older patients generally reported higher levels of depression, but lower levels of anxiety than younger patients, and females in the second age group (between 50 and 65 years of age) reported the highest levels of anxiety. Similarly, higher levels of psychological distress were observed for breast cancer patients, whereas male genitourinary cancer patients reported the lowest levels of anxiety, depression and psychological distress.

Differences such as these in levels of psychological distress by disease site and age have been reported in the literature (e.g. Zabora et al., 2001) and therefore care must be taken when interpreting levels of distress recorded from a large heterogeneous sample such as this, particularly if comparisons between subgroups of patients differentiated by diagnosis are not being made. In addition, a corollary of this is that given the differences found between age groups, as well as cancer site, there may be problems generalising findings to other cancer groups.

There may be a number of explanations for the observed differences in factor structures found in other studies (e.g. Andersson, 1993; Lewis, 1991), and those reported here. For instance, the sample size used by Andersson (1993) was relatively small for factor analysis (n=163). Also differences in reported factor structures have been found with groups from non-medical community samples. The HADS was designed originally for use for in hospital clinics and wards, and the results from studies using patient groups find similar factor structures (Spinhoven et al., 1997; White et al., 1999). It may therefore be that non-patient groups respond differently to the HADS than patient groups, and that different measures may be needed to identify non-hospital based cases of psychological distress (Groenvold et al., 1999).

The two factors were strongly correlated (.52), which taken together with a second-order factor analysis confirmed a single higher-order factor, corresponding to psychological distress or Negative Affectivity (NA) in the tripartite model, and two

subordinate factors, Anhedonia and Autonomic Anxiety. Although Clark and Watson's (1991) original model placed the three factors on equal footing, more recent formulations of the model (e.g., Clark et al., 1994) have suggested a hierarchical structure. Therefore, these results provide evidence that the HADS corresponds to the tripartite structure, but that this structure differs from the model demonstrated by Dunbar et al. (2000).

The difference between this study and the study by Dunbar et al. (2000) lies in the statistical models used. Dunbar et al. (2000) employed confirmatory factor analysis, whereas the method chosen for this study was the same as the method used by Clark et al. (1994), i.e. hierarchical or second-order factor analysis.

Both confirmatory factor analysis (CFA) and hierarchical factor analysis (HFA) are special forms of structural equation modelling. These models depend on correlations between variables to define or surmise relationships between latent (i.e. unobserved) and observed variables. In many ways the two methods are similar, the main difference between the models being that in CFA the latent model is fitted to a correlation matrix, and the 'goodness of fit' of the model is then evaluated, whereas HFA relies on the rotation of the correlation matrix to produce first and subsequent order factors (Loehlin, 1998). Since both statistical techniques depend on correlation between the observed and latent variables, neither model is truly causal: both models, may posit causality, but it remains for subsequent empirical work to either confirm or dispute this.

The difference between Dunbar et al.'s (2000) study and this study in terms of the results, is subtle. The results from both studies demonstrate that a hierarchical tripartite structure probably underlies the HADS. However, whereas Dunbar et al. (2000) suggest that the HADS-A scale is split into Negative Affectivity and Autonomic Anxiety (AA) items, the results from this study suggest that the NA is a common factor whose variance is shared by both AA and Anhedonic Depression, and which is not specific to either subscale.

## 5.2. Rasch Analysis of the HADS

### *5.2.1 Aim*

The aim of this study was to examine the Hospital Anxiety and Depression scale using Rasch models, in order to identify misfitting items which could be removed, and to assess the possibility of using the items in computer-adaptive testing.

### *5.2.2. Methodology*

The data employed in this section were the same as described in section 5.1.3. The scores from the total HADS, and the two subscales HADS-Anxiety and HADS-Depression were converted to interval-level logit (log-odds) scores using the *Winsteps* software (Linacre & Wright, 2000), and the Rating Scale Model for polytomous data (Andrich 1978a, b) as described in Chapter 1.2.

The fit statistics were calculated for the scales, and the differences between adjacent scores were plotted for each scale. Additionally, the item difficulty and person ability estimates were derived, as well as fit statistics (infit and outfit) for the items for both the full scale, as well as the subscales. Furthermore, a principal components analysis was performed for the full scale and subscales, and test characteristic curves were also calculated.

### *5.2.3. Results for HADS –total*

#### *5.2.3.1. Analysis of Unidimensionality*

The fit statistics for the HADS-scale can be seen in Table 5.2.1. The majority of items from this scale demonstrate a good fit (i.e. infit greater than 0.70 and less than 1.30, Wright et al., 1994) suggesting that the items define a common construct. Three items from this scale, namely Depression subscale item 2 (“I can laugh and see the funny side of things “), 5 (“I have lost interest in my appearance“) and item 7 (“I get sudden feelings of panic“) had infit statistics greater than 1.3 indicating that they did

not fit the model well. Furthermore, item 1 from the Anxiety subscale ("I feel tense or 'wound up'") showed an amount of redundancy with an infit statistic smaller than 0.70.

**Table 5.2.1. HADS Scale - Unidimensionality measures**

| Entry | Measure | Count | Raw Score | In. MSQ | In. ZSTD | Out. MSQ | Out ZSTD | Name |
|---|---|---|---|---|---|---|---|---|
| 1 | -0.45 | 1423 | 1335 | 0.62 | -9.9 | 0.74 | -7.04 | ANX1 |
| 2 | -0.39 | 1423 | 1294 | 0.97 | -0.8 | 0.96 | -0.97 | ANX2 |
| 3 | -0.72 | 1423 | 1533 | 0.78 | -6.69 | 0.79 | -5.8 | ANX3 |
| 4 | -0.49 | 1423 | 1364 | 0.8 | -5.93 | 0.88 | -3.15 | ANX4 |
| 5 | 0.07 | 1423 | 1007 | 0.93 | -1.76 | 0.96 | -0.98 | ANX5 |
| 6 | -0.62 | 1423 | 1456 | 1.25 | 6.31 | 1.31 | 7.26 | ANX6 |
| 7 | 0.27 | 1423 | 892 | 0.9 | -2.42 | 0.85 | -3.3 | ANX7 |
| 8 | -0.25 | 1423 | 1206 | 1.26 | 6.24 | 1.24 | 5.33 | DEP1 |
| 9 | 0.91 | 1423 | 589 | 1.44 | 8.36 | 1.48 | 6.74 | DEP2 |
| 10 | 0.97 | 1423 | 566 | 0.76 | -5.59 | 0.67 | -5.87 | DEP3 |
| 11 | -1.23 | 1423 | 1931 | 1.11 | 3.12 | 1.13 | 3.3 | DEP4 |
| 12 | 0.75 | 1423 | 659 | 1.56 | 9.9 | 1.45 | 6.7 | DEP5 |
| 13 | 0.29 | 1423 | 883 | 1.01 | 0.22 | 0.91 | -1.85 | DEP6 |
| 14 | 0.9 | 1423 | 595 | 1.41 | 7.78 | 1.31 | 4.5 | DEP7 |

*In MSQ – Infit Mean Square Statistic; In ZSTD – Standardised Infit Mean Square Statistic; Out MSQ – Outfit Mean Square Statistic; Out ZSTD – Standardised Outfit Mean Square Statistic (see Chapter 3).

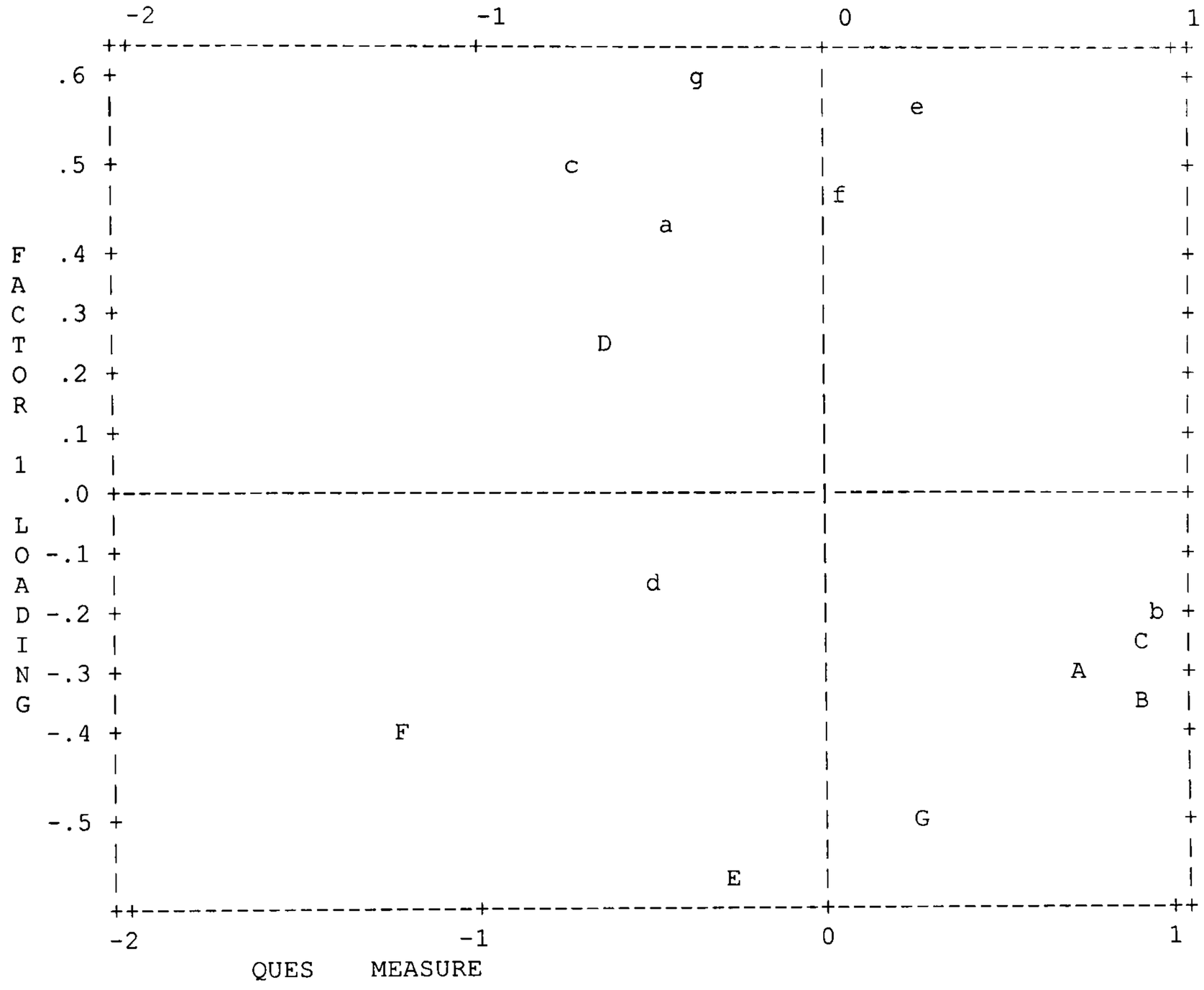
In addition, the principal components analysis (PCA) of the HADS-scale resulted in 2.42 eigenvalues from the analysis of the residuals. The results of the PCA can be seen in Table 5.2.2. Following Smith & Miao's (1994) demonstration that values greater than 1.4 indicate structure in the residuals, it can be seen that an inspection of the factor loadings suggests two factors in addition to the Rasch factor (the unidimensional factor corresponding to psychological distress), roughly corresponding to the HADS-Anxiety and HADS-Depression subscales. It can be seen from Table 5.2.2. that item 4 from the HADS-Anxiety subscale loads onto the HADS-Depression subscale. These results are identical to the outcomes from Study 5.1. The factor plot can be seen in Figure 5.2.1.

**Table 5.2.2. Factor 1 from Principal Component Analysis of standardised residuals for the HADS-scale (sorted by loading). Factor 1 explains 2.42 of 14**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | QUES | |
|---|---|---|---|---|---|---|---|
| 1 | .59 | -.39 | .97 | .96 | g | 2 | ANX2 |
| 1 | .58 | .27 | .90 | .85 | e | 7 | ANX7 |
| 1 | .49 | -.72 | .78 | .79 | c | 3 | ANX3 |
| 1 | .48 | .07 | .93 | .96 | f | 5 | ANX5 |
| 1 | .41 | -.45 | .62 | .74 | a | 1 | ANX1 |
| 1 | .23 | -.62 | 1.25 | 1.31 | D | 6 | ANX6 |
| 1 | -.56 | -.25 | 1.26 | 1.24 | E | 8 | DEP1 |
| 1 | -.52 | .29 | 1.01 | .91 | G | 13 | DEP6 |
| 1 | -.37 | -1.23 | 1.11 | 1.13 | F | 11 | DEP4 |
| 1 | -.35 | .91 | 1.44 | 1.48 | B | 9 | DEP2 |
| 1 | -.28 | .75 | 1.56 | 1.45 | A | 12 | DEP5 |
| 1 | -.23 | .90 | 1.41 | 1.31 | C | 14 | DEP7 |
| 1 | -.20 | .97 | .76 | .67 | b | 10 | DEP3 |
| 1 | -.16 | -.49 | .80 | .88 | d | 4 | ANX4 |

*Ques – Items from each subscale

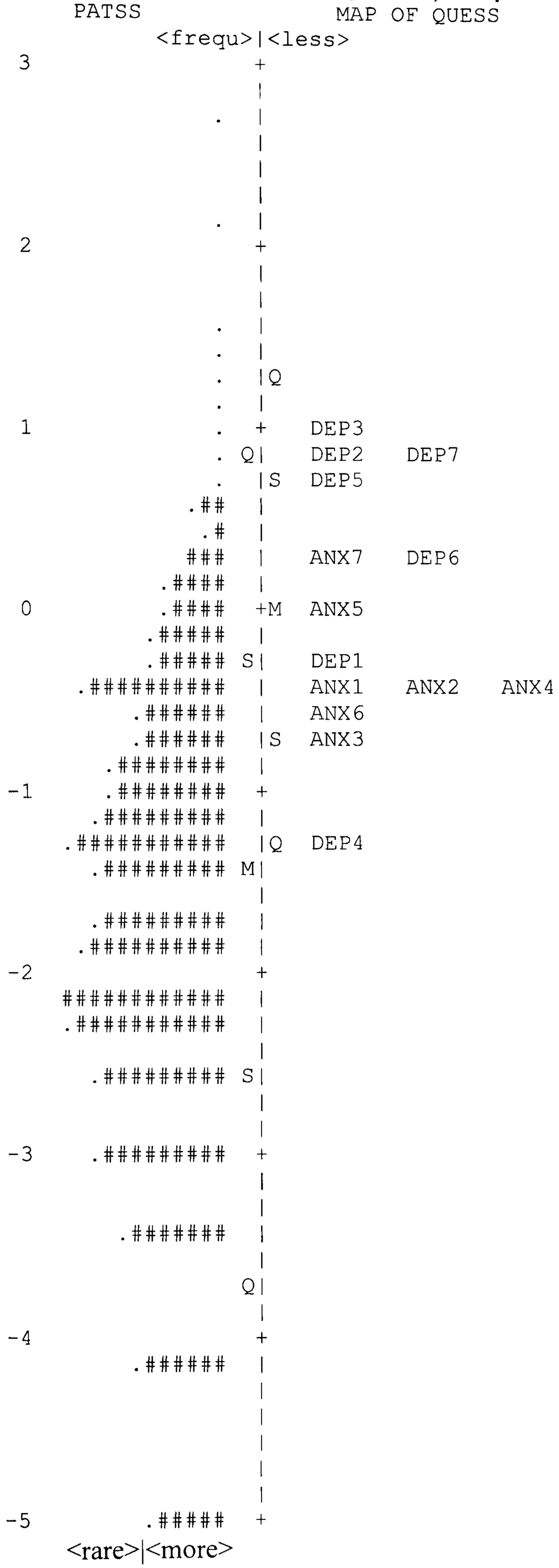
**Figure 5.2.1. Principal Components (Standardized Residual) Factor Plot of the HADS-Scale**



*5.2.3.2 Analysis of Item Locations*

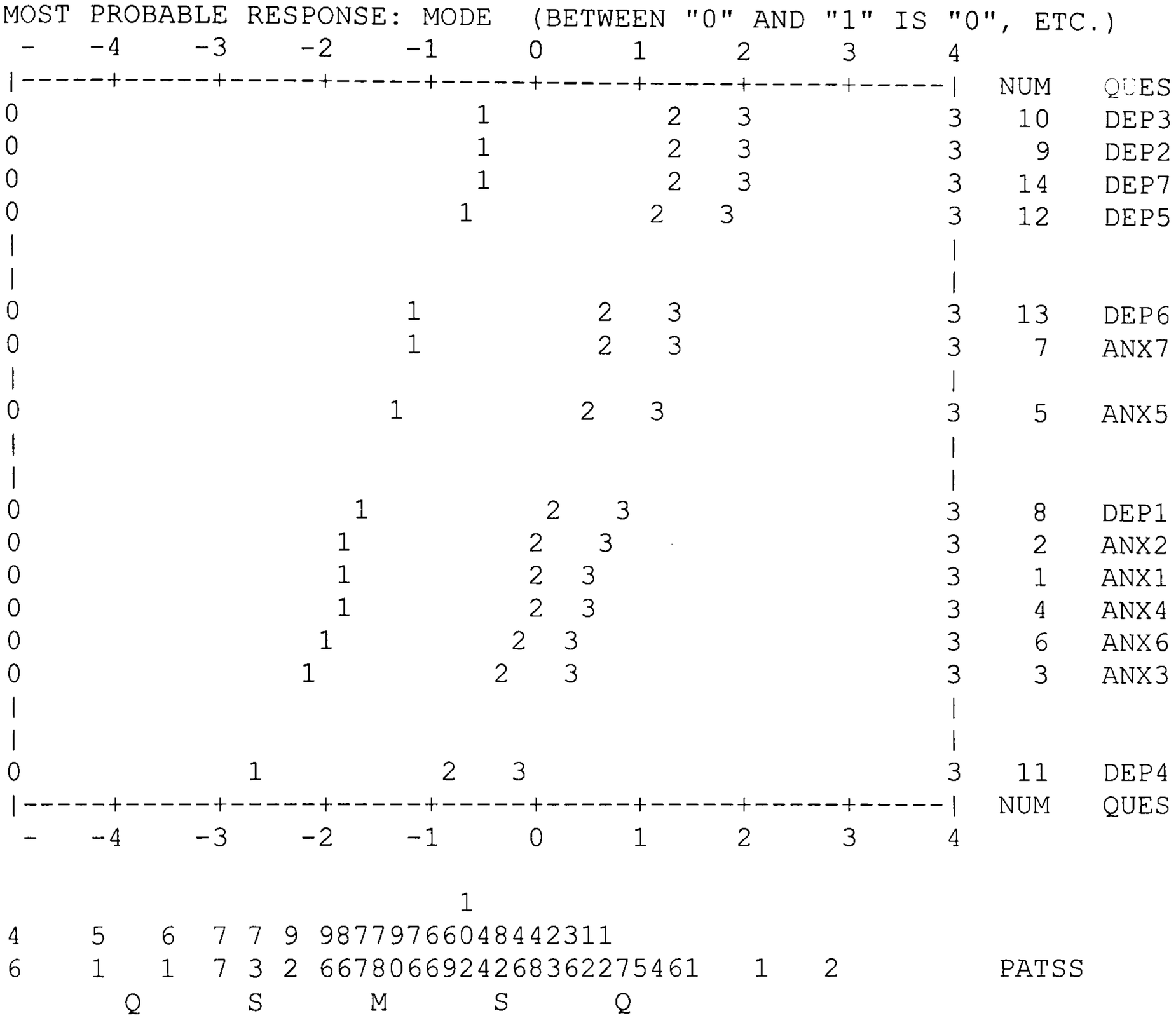
Figure 5.2.2. shows the location of all of the items along the logit scale. It can be seen from this figure that most items cluster around the middle of the scale, and that a number of items are situated at the same location, i.e. from the depression subscale item 2 and 7 ("I can enjoy a good book or radio or TV programme"), and from the anxiety subscale items 1 ("I feel tense or 'wound up'"), 2 ("I get a sort of frightened feeling as if something awful is about to happen"), and 4 ("I can sit at ease and feel relaxed"). It is interesting to note that item 7 ("I get sudden feelings of panic") from the anxiety subscale and item 6 ("I look forward with enjoyment to things") from the depression subscale also share the same location.

**Figure 5.2.2. Logit map of all items (QUESS) and patients (PATSS) for the HADS - Scale**



The most probable response table is shown in Figure 5.2.3.

**Figure 5.2.3. Most Probable Response for HADS Scale**




A summary for the items is given in Table 5.2.3. The separation index for the items is approximately 14, indicating 14 statistically distinct difficulty (D) strata.

Table 5.2.4 shows the summary of measured steps across all items. The number of observed counts for each category is considerably greater than the minimum of 10 recommended for each category (Linacre, 1999b), similarly the measures for each category increase monotonically. Step calibrations are also shown in Table 5.2.4. Although the threshold measures increase monotonically, and therefore demonstrating no disorder, whereas the difference between step 1 and step

**Table 5.2.3 Item Summary for the HADS-Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1093.6 | 1423.0 | .00 | .04 | 1.06 | .6 | 1.05 | .3 |
| S.D. | 402.9 | .0 | .67 | .01 | .28 | 6.2 | .26 | 5.0 |
| MAX. | 1931.0 | 1423.0 | .97 | .05 | 1.56 | 9.9 | 1.48 | 7.3 |
| MIN. | 566.0 | 1423.0 | -1.23 | .04 | .62 | -9.9 | .67 | -7.0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .05 | ADJ.SD | .67 | SEPARATION | 14.70 | QUES | RELIABILITY | 1.00 |
| MODEL RMSE | .04 | ADJ.SD | .67 | SEPARATION | 15.91 | QUES | RELIABILITY | 1.00 |
| S.E. OF QUES | MEAN | .19 | | | | | | |

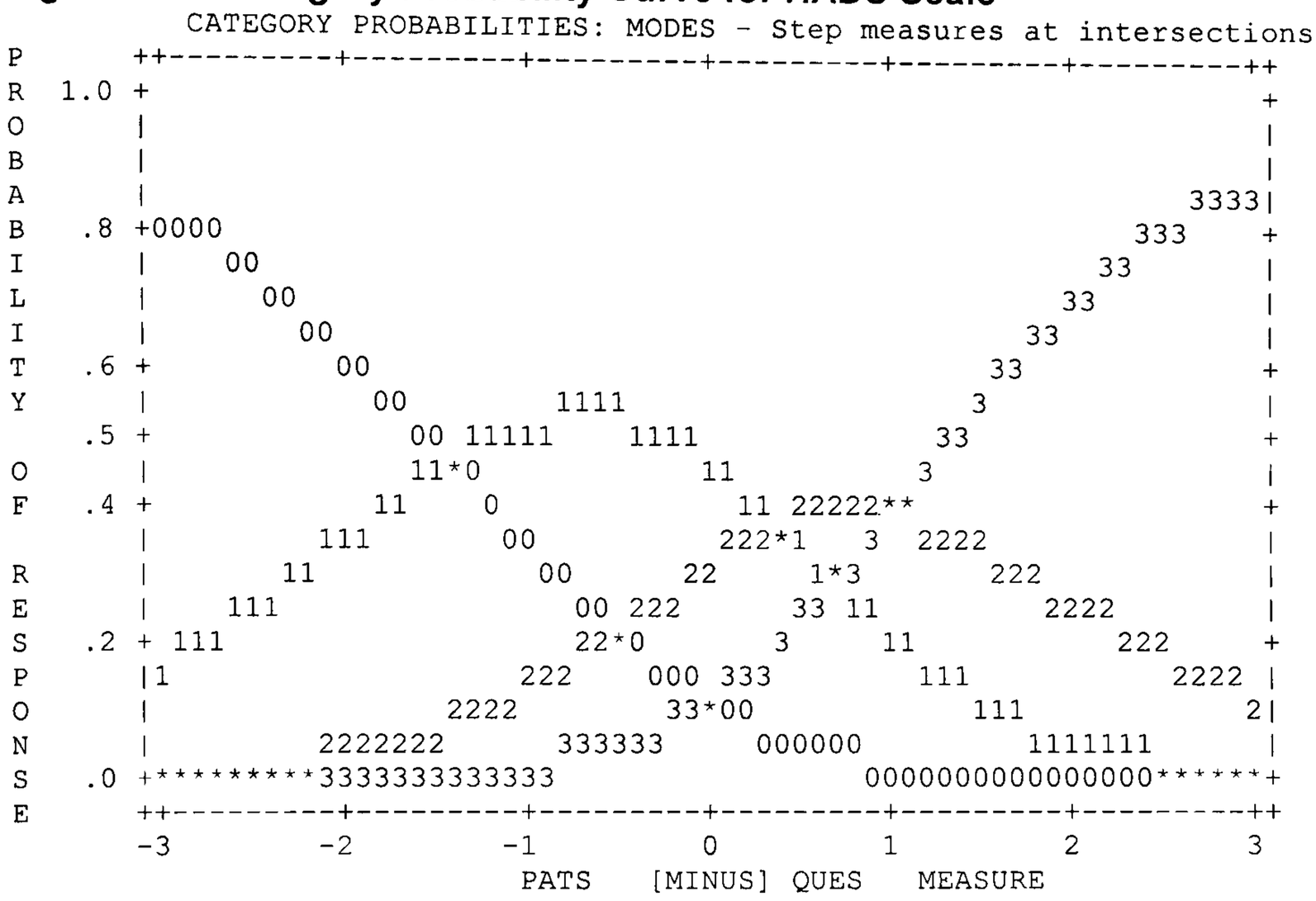
**Table 5.2.4 Summary of Measured Steps for the HADS-Scale**

| CATEGORY LABEL | OBSERVED COUNT | AVERAGE MEASURE | EXP. | COHERENCE EXP% | COHERENCE OBS% | INFIT MNSQ | OUTFIT MNSQ | STEP CALIBRATN |
|---|---|---|---|---|---|---|---|---|
| 0 | 9174 | -2.32 | -2.25 | 80% | 69% | .94 | .96 | NONE |
| 1 | 7221 | -1.01 | -1.12 | 53% | 69% | .88 | .75 | -1.42 |
| 2 | 2492 | -.16 | -.24 | 40% | 38% | .91 | .88 | .40 |
| 3 | 1035 | .13 | .53 | 66% | 8% | 1.44 | 2.08 | 1.02 |

AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.
EXP% = (expected & observed)/(all expected) [MEASURE->RATING?]
OBS% = (expected & observed)/(all observed) [RATING->MEASURE?]

2 is greater than 1.4 (Linacre, 1999a), that between steps 2 and 3 is considerably less. Similarly, the outfit mean square for category 3 is greater than 2, indicating that the category is introducing more noise into the measurement process, i.e. more misinformation than information is being produced (Linacre, 1999a).

**Figure 5.2.4 Category Probability Curve for HADS Scale**




The category probability curve is given in Figure 5.2.4. This demonstrates for instance that a patient measured as –0.5 logits along the psychological distress continuum is most likely to score 1 in response to a question, whereas an individual further along the continuum, say at +2.0 logits is most likely to respond with a 3 to a given question.

*5.2.3.3 Analysis of Person Locations*

The person measures from the HADS-scale can be seen in Table 5.2.5.

**Table 5.2.5. Person measures for the HADS – Scale**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 0 | -4.93 | 1.00 | 0.00 | 1.00 | 0.00 |
| 1 | -4.20 | 1.11 | 0.11 | 2.47 | 0.89 |
| 2 | -3.44 | 0.76 | -0.40 | 0.54 | -0.68 |
| 3 | -2.97 | 1.35 | 0.57 | 0.84 | -0.26 |
| 4 | -2.62 | 2.63 | 2.33 | 2.33 | 1.78 |
| 5 | -2.33 | 0.98 | -0.06 | 0.79 | -0.47 |
| 6 | -2.08 | 1.26 | 0.55 | 1.09 | 0.20 |
| 7 | -1.86 | 0.41 | -1.82 | 0.43 | -1.72 |
| 8 | -1.66 | 1.29 | 0.64 | 1.11 | 0.26 |
| 9 | -1.49 | 1.03 | 0.08 | 0.89 | -0.29 |
| 10 | -1.32 | 1.19 | 0.46 | 1.17 | 0.41 |
| 11 | -1.16 | 2.54 | 2.82 | 3.06 | 3.51 |
| 12 | -1.02 | 1.21 | 0.51 | 1.32 | 0.76 |
| 13 | -0.88 | 1.21 | 0.54 | 1.32 | 0.78 |
| 14 | -0.74 | 0.47 | -1.82 | 0.46 | -1.85 |
| 15 | -0.62 | 1.13 | 0.36 | 1.09 | 0.23 |
| 16 | -0.49 | 0.42 | -2.13 | 0.42 | -2.10 |
| 17 | -0.37 | 0.45 | -2.04 | 0.53 | -1.65 |
| 18 | -0.26 | 0.86 | -0.45 | 0.85 | -0.45 |
| 19 | -0.14 | 0.93 | -0.20 | 0.95 | -0.15 |
| 20 | -0.03 | 0.22 | -3.59 | 0.21 | -3.61 |
| 21 | 0.08 | 0.72 | -0.95 | 0.74 | -0.85 |
| 22 | 0.19 | 2.19 | 2.76 | 2.24 | 2.80 |
| 23 | 0.30 | 1.48 | 1.31 | 1.51 | 1.35 |
| 24 | 0.41 | 1.66 | 1.72 | 1.65 | 1.66 |
| 25 | 0.52 | 0.98 | -0.06 | 1.00 | 0.01 |
| 26 | 0.63 | 0.86 | -0.47 | 1.02 | 0.05 |
| 27 | 0.74 | 0.57 | -1.55 | 0.57 | -1.51 |
| 28 | 0.86 | 0.97 | -0.08 | 1.11 | 0.31 |
| 29 | 0.98 | 1.00 | -0.01 | 0.93 | -0.20 |
| 30 | 1.10 | 0.77 | -0.72 | 0.75 | -0.76 |
| 31 | 1.23 | 0.78 | -0.68 | 1.01 | 0.02 |
| 32 | 1.36 | 0.85 | -0.43 | 1.15 | 0.36 |
| 33 | 1.51 | 1.21 | 0.53 | 1.02 | 0.04 |
| 37 | 2.21 | 2.30 | 2.00 | 2.27 | 1.60 |
| 39 | 2.76 | 0.84 | -0.27 | 1.04 | 0.05 |

*MEASURE refers to the person measure estimates in logits

The differences between adjacent scores in logits are shown in Figure 5.2.5. The scores beyond 33 are not included in this figure. As the graph, shows differences

between adjacent scores are virtually identical from a score of 11 onwards, demonstrating that this portion of the HADS-scale is interval-based. There are increasing differences between adjacent scores less than 11.

The summary of person measures can be seen in Table 5.2.6, which shows good overall fit (1.04 infit and 1.05 outfit), and good reliability (0.79) for the person measures. The separation index of 1.96 indicates that there are approximately 2 distinct strata of ability (B).

**Table 5.2.6. Summary of Person Measures**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 10.8 | 14.0 | -1.45 | .46 | 1.04 | -.2 | 1.05 | -.2 |
| S.D. | 6.7 | .0 | 1.17 | .15 | .64 | 1.4 | .81 | 1.4 |
| MAX. | 39.0 | 14.0 | 2.76 | 1.03 | 4.61 | 5.8 | 8.24 | 6.1 |
| MIN. | 1.0 | 14.0 | -4.20 | .33 | .20 | -3.7 | .20 | -3.7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .53 | ADJ.SD | 1.04 | SEPARATION | 1.96 | PATS | RELIABILITY | .79 |
| MODEL RMSE | .48 | ADJ.SD | 1.06 | SEPARATION | 2.20 | PATS | RELIABILITY | .83 |
| S.E. OF PATS | MEAN | .03 | | | | | | |
| WITH | 46 EXTREME PATSS | = | 1469 PATSS | MEAN | -1.55 | S.D. | 1.30 | |
| REAL RMSE | .58 | ADJ.SD | 1.16 | SEPARATION | 2.00 | PATS | RELIABILITY | .80 |
| MODEL RMSE | .54 | ADJ.SD | 1.18 | SEPARATION | 2.19 | PATS | RELIABILITY | .83 |

The test information curve for HADS-Total is given in Figure 5.2.6, which shows that the greatest amount of information is provided in the –0.74 to + 1.10 range. This corresponds to patients' scores between 14 and 30, and 28.6% of patients scores fell in this range. The peak of the graph corresponding to the maximum amount information occurs at 0.41 logits or a raw score of 24.

**Figure 5.2.5. Differences in logits between adjacent scores of the HADS-Scale**

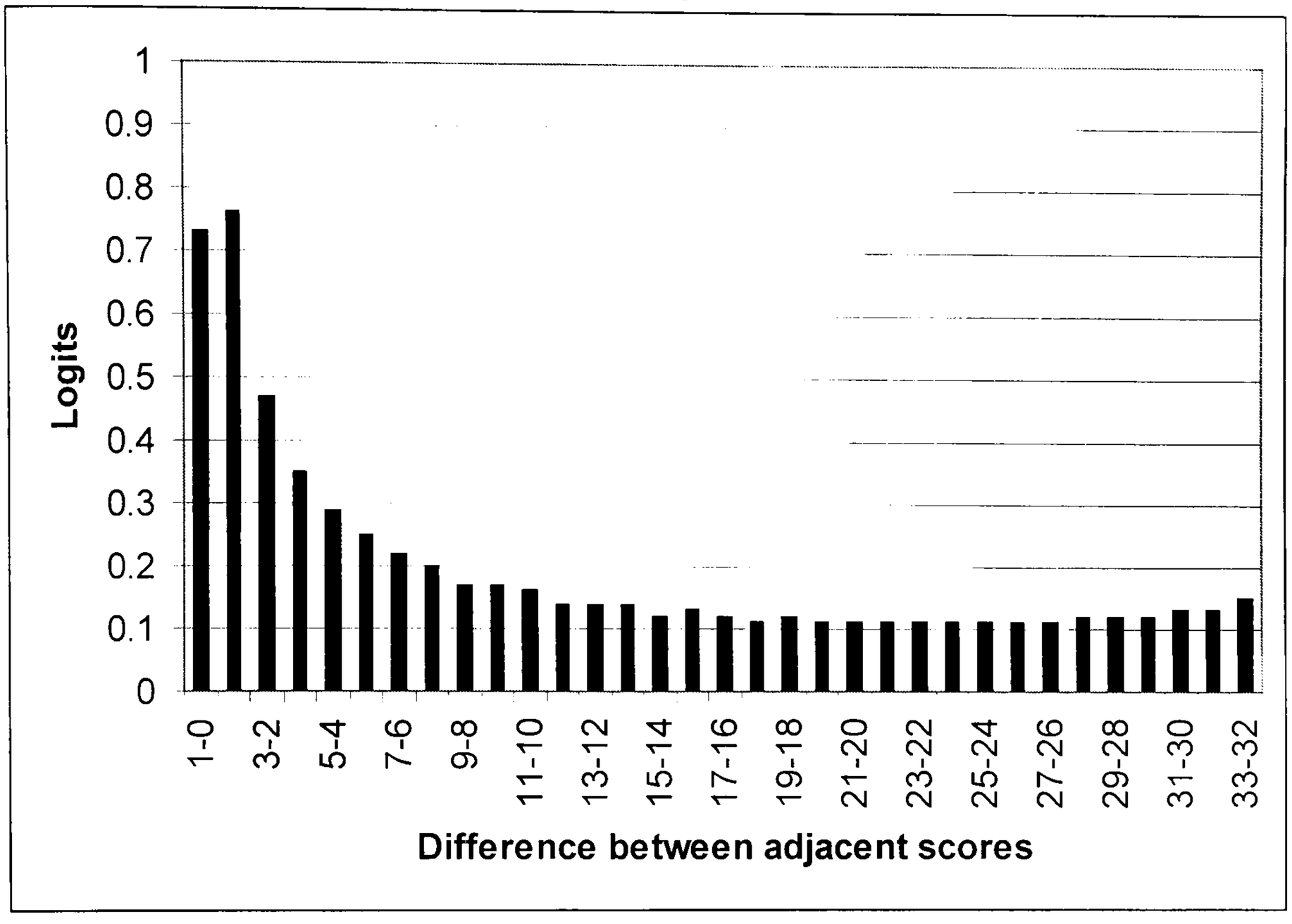
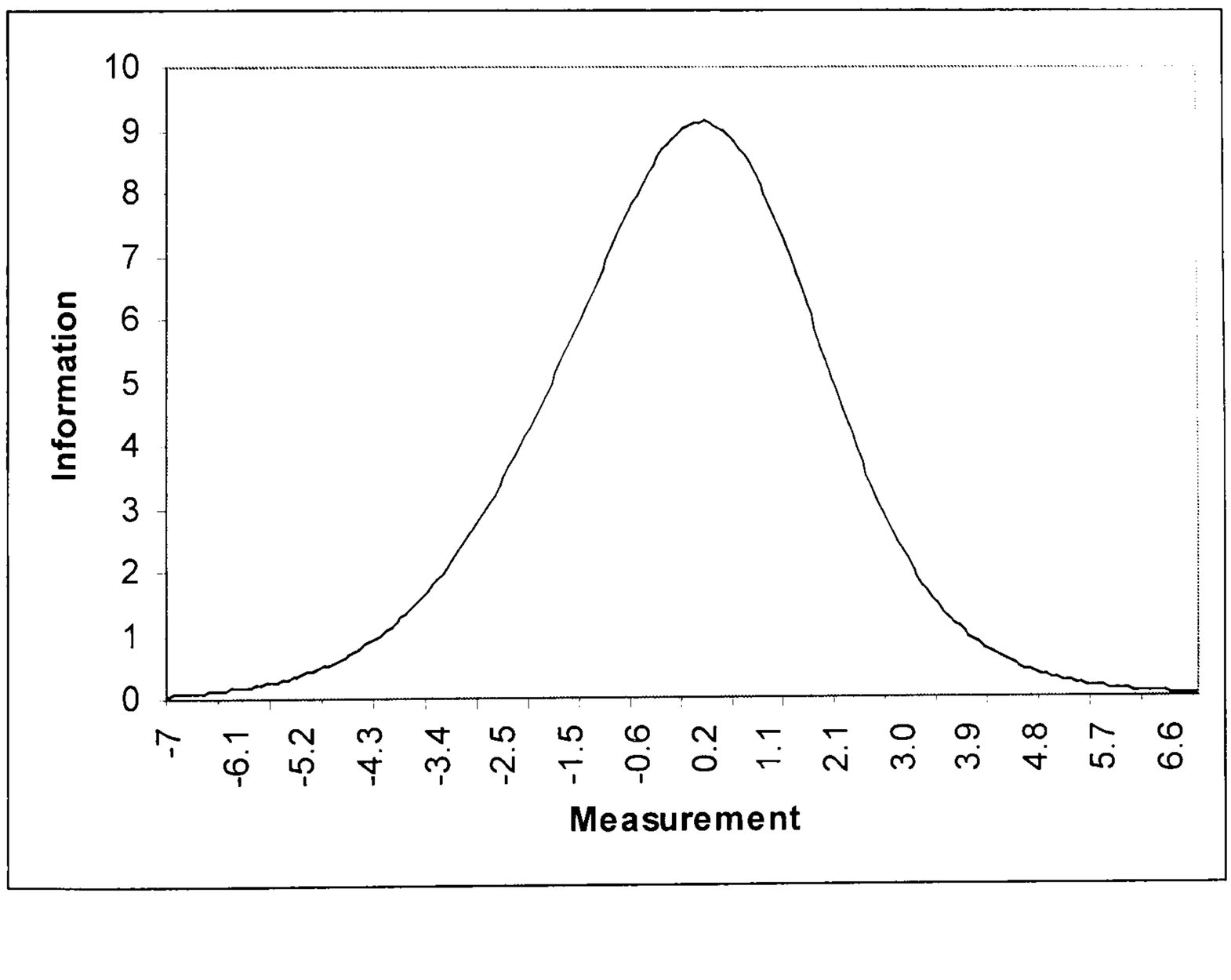



**Figure 5.2.6. Test Information Curve for the HADS-Scale**



*5.2.4 Differential Item Functioning of HADS-Total*

The HAD scale was investigated for differential item functioning (DIF), the results of which can be seen in Table 5.2.7.

**Table 5.2.7 Differential Item Functioning of the HADS**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -.48 | .05 | 2 | -.38 | .06 | -.11 | .08 | -1.36 | INF | 1 | A1 |
| 1 | -.47 | .05 | 2 | -.25 | .06 | -.22 | .08 | -2.74 | INF | 2 | A2 |
| 1 | -.81 | .05 | 2 | -.58 | .06 | -.23 | .08 | -3.02 | INF | 3 | A3 |
| 1 | -.43 | .05 | 2 | -.57 | .06 | .14 | .08 | 1.83 | INF | 4 | A4 |
| 1 | .01 | .05 | 2 | .15 | .07 | -.13 | .09 | -1.56 | INF | 5 | A5 |
| 1 | -.54 | .05 | 2 | -.73 | .06 | .19 | .08 | 2.57 | INF | 6 | A6 |
| 1 | .12 | .05 | 2 | .53 | .07 | -.41 | .09 | -4.52 | INF | 7 | A7 |
| 1 | -.19 | .05 | 2 | -.36 | .06 | .17 | .08 | 2.13 | INF | 8 | D1 |
| 1 | .99 | .06 | 2 | .79 | .08 | .20 | .10 | 1.97 | INF | 9 | D2 |
| 1 | 1.07 | .07 | 2 | .83 | .08 | .24 | .10 | 2.33 | INF | 10 | D3 |
| 1 | -1.20 | .05 | 2 | -1.28 | .06 | .08 | .07 | 1.17 | INF | 11 | D4 |
| 1 | .74 | .06 | 2 | .75 | .08 | .00 | .10 | -.04 | INF | 12 | D5 |
| 1 | .30 | .06 | 2 | .26 | .07 | .04 | .09 | .46 | INF | 13 | D6 |
| 1 | .95 | .06 | 2 | .80 | .08 | .15 | .10 | 1.49 | INF | 14 | D7 |

The sample was split into male and female groups (group 1 and group 2 respectively) for this analysis. Table 5.2.7 demonstrates that none of the items from the HADS exhibited differential item functioning. Although the t-statistic is significant for items 2, 3, 4, 6 and 7 from the Anxiety items, and items 1 and 3 from the Depression subscale, the difference between the item estimates for the samples is smaller than 0.50 logits (Wright and Panchapakesan, 1969).

## 5.3. Results for HADS – Anxiety subscale

*5.3.1. Analysis of Unidimensionality*

The fit statistics for the HADS-Anxiety (HADS-A) subscale show good fit for all but one item (item 6, "I feel restless as if I have to be on the move"), which has an outfit statistic greater than 1.3 (Wright et al., 1994).

The results from the principal components analysis of the HADS-A showed an eigenvalue of 1.59 remained once the Rasch factor had been extracted from the

data. This value is close to the criterion of 1.4 proposed by Smith & Miao (1994), and suggests that no further factor structures are present in the residuals.

**Table 5.3.1. HADS – Anxiety – Unidimensionality measures**

| Entry | Measure | Count | Raw Score | In. MSQ | In. ZSTD | Out MSQ | Out ZSTD | NAME |
|---|---|---|---|---|---|---|---|---|
| 1 | -0.15 | 1376 | 1332 | 0.7 | -8.94 | 0.77 | -6.62 | ANX1 |
| 2 | -0.07 | 1376 | 1291 | 0.98 | -0.55 | 0.96 | -1.14 | ANX2 |
| 3 | -0.53 | 1376 | 1530 | 0.84 | -4.4 | 0.86 | -3.99 | ANX3 |
| 4 | -0.21 | 1376 | 1361 | 1.2 | 4.89 | 1.25 | 6.04 | ANX4 |
| 5 | 0.54 | 1376 | 1004 | 0.98 | -0.57 | 0.95 | -1.23 | ANX5 |
| 6 | -0.39 | 1376 | 1453 | 1.44 | 9.9 | 1.45 | 9.9 | ANX6 |
| 7 | 0.81 | 1376 | 889 | 0.87 | -3.33 | 0.82 | -4.21 | ANX7 |

**Table 5.3.2. Factor 1 from Principal Component Analysis of Standardised Residuals for HADS – Anxiety (sorted by loading). Factor 1 explains 1.59 of 7**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | QUES |
|---|---|---|---|---|---|---|
| 1 | .09 | -.15 | .70 | .77 | a 1 | ANX1 |
| 1 | -.62 | -.07 | .98 | .96 | C 2 | ANX2 |
| 1 | -.44 | -.53 | .84 | .86 | b 3 | ANX3 |
| 1 | .64 | -.21 | 1.20 | 1.25 | B 4 | ANX4 |
| 1 | -.16 | .54 | .98 | .95 | D 5 | ANX5 |
| 1 | .60 | -.39 | 1.44 | 1.45 | A 6 | ANX6 |
| 1 | -.46 | .81 | .87 | .82 | c 7 | ANX7 |

The factor plot of the HADS-A can be seen in Figure 5.3.1.

*5.3.2. Analysis of Item Locations*

The logit map of items and patients can be seen in figure 5.3.2. As for the HADS-Scale the majority of items cluster around the middle of the scale, and there is an overlap between items 1 ("I feel tense or 'wound up'") and 4 ("I can sit at ease and feel relaxed").

**Figure 5.3.1. Logit map of all items and patients for the HADS-A**

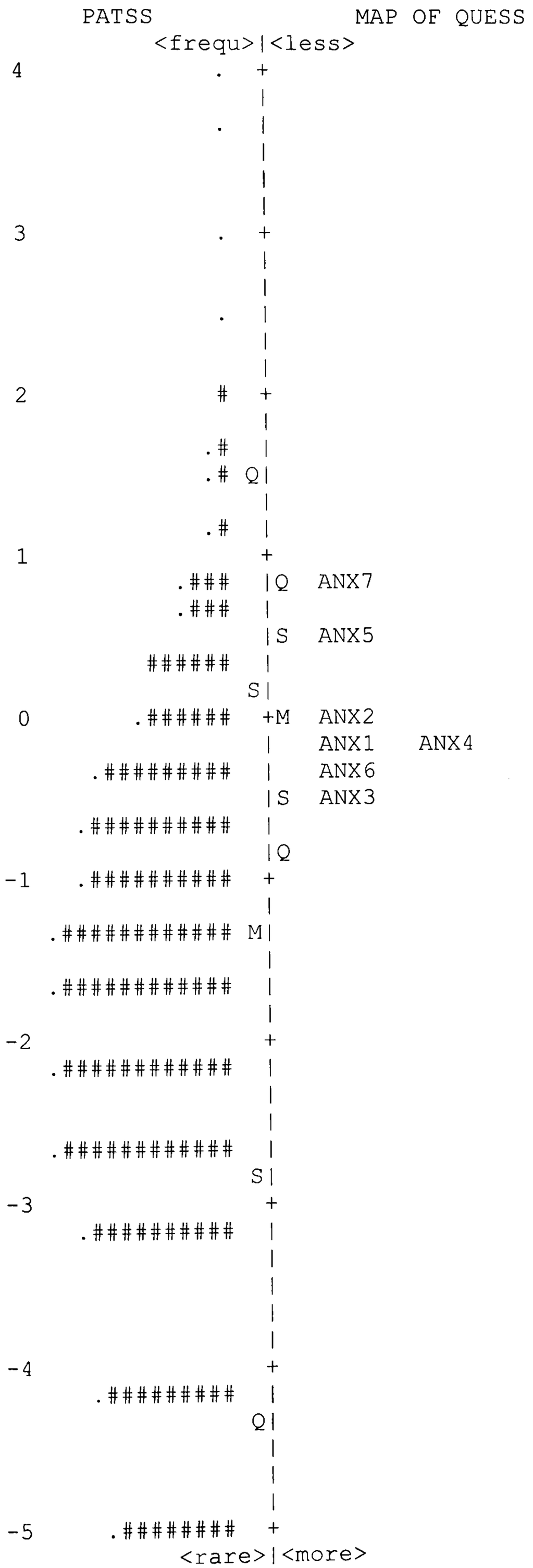
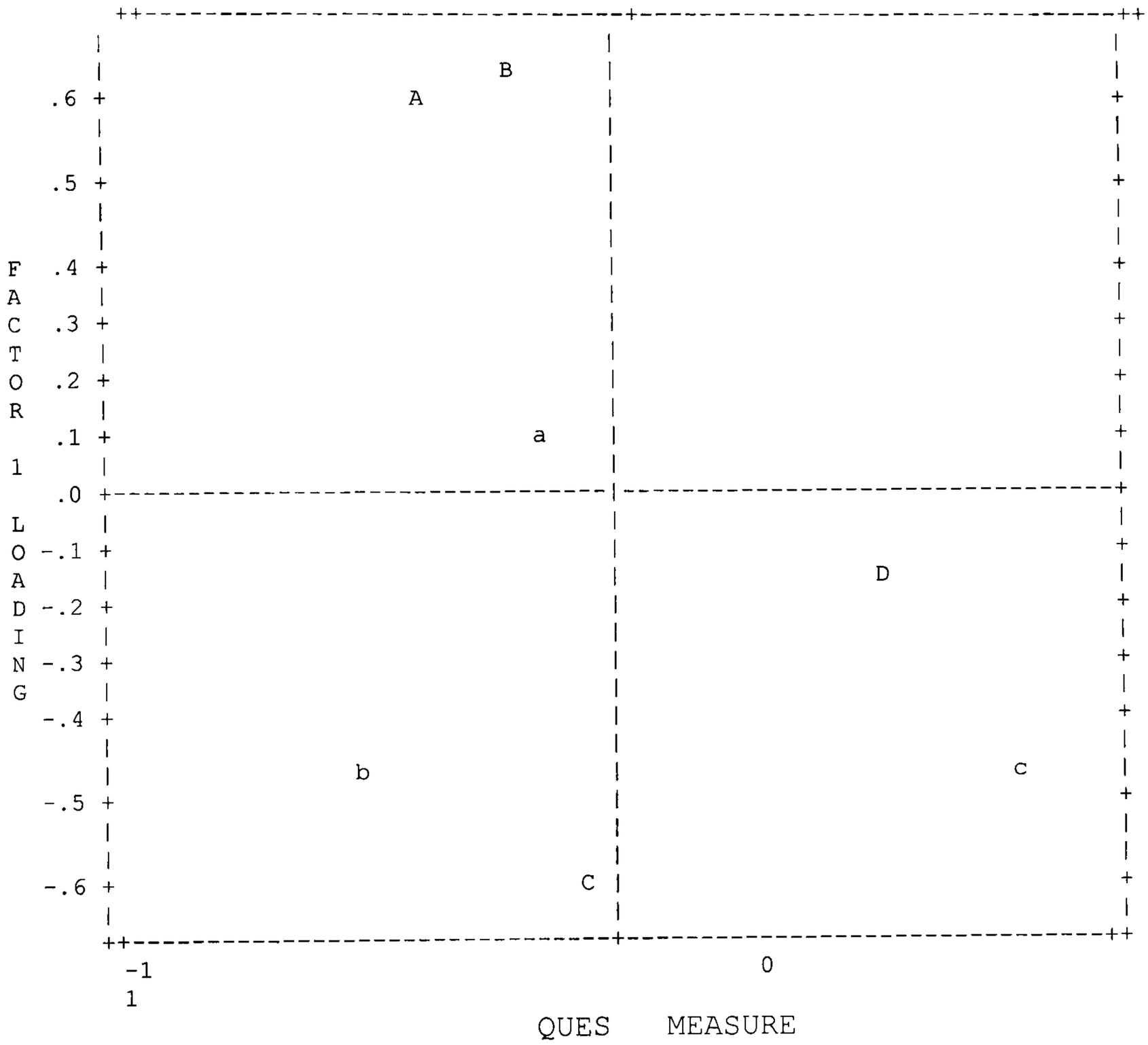


**Figure 5.3.2. Rasch analysis of Anxiety scores - Principal Components (Standardized Residual) Factor Plot**



A number of observations can be made from the results in both Table 5.3.1. and Figure 5.3.1. Both demonstrate that the "easiest" item, or item endorsed more easily, is item 3 ("Worrying thoughts go through my mind", measure of –0.59 logits). That is to say patients need not score very high on anxiety to answer positively to this question. Conversely, the "hardest" item, or item least likely to be endorsed is item 7 ("I get sudden feelings of panic", measure of 0.81), which implies that patients need to be experiencing high levels of anxiety before answering this question positively.

The most probable response table for the HADS-A is shown in Figure 5.3.3.

**Figure 5.3.3. Most Probable Response for HADS - A**

```
MOST PROBABLE RESPONSE: MODE  (BETWEEN "0" AND "1" IS "0", ETC.)
 -   -4    -3    -2    -1     0     1     2     3     4
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   QUES
0                      1               2         3     3    7   ANX7
|                                                    |
0                    1                 2       3     3    5   ANX5
|                                                    |
|                                                    |
0                 1                  2       3      3    2   ANX2
0                 1                  2       3      3    1   ANX1
0                   1                2       3      3    4   ANX4
0                  1                 2       3      3    6   ANX6
0                 1                  2       3      3    3   ANX3
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   QUES
 -   -4    -3    -2    -1     0     1     2     3     4

     1     1  1  1  1 1 1 1 1
9    0     1  3  3  4 3 1 1 0 6 6 33 1 21 1
2    3     3  8  6  2 6 8 2 1 8 6 86 6 03 1  5  3   1 1  PATSS
    Q       S       M       S       Q
```

As can be seen from Figure 5.3.3. as we read the graph from left to right, i.e. as the level of anxiety increases, the probability of responding with a higher score to each question increases. For instance, it can be seen that the most probable response from an individual patient scoring low on anxiety (e.g. –3.0) to item 3 ("Worrying thoughts go through my mind") is 1 or "From time to time but not too often". In contrast a patient scoring high on anxiety, e.g. +3.0, is more likely to respond with a 3 ("Very often indeed") to item 7. This figure corroborates the results of the item measures (Table 5.3.1 and Figure 5.3.1).

**Table 5.3.3 Item Summary for the HADS-A**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1265.7 | 1376.0 | .00 | .05 | 1.00 | -.4 | 1.01 | -.2 |
| S.D. | 216.9 | .0 | .45 | .00 | .23 | 5.8 | .23 | 5.5 |
| MAX. | 1530.0 | 1376.0 | .81 | .05 | 1.44 | 9.9 | 1.45 | 9.9 |
| MIN. | 889.0 | 1376.0 | -.53 | .04 | .70 | -8.9 | .77 | -6.6 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .05 | ADJ.SD | .45 | SEPARATION | 9.58 | QUES | RELIABILITY | .99 |
| MODEL RMSE | .05 | ADJ.SD | .45 | SEPARATION | 9.98 | QUES | RELIABILITY | .99 |
| S.E. OF QUES | MEAN | .19 | | | | | | |

**Table 5.3.4. Summary of Measured Steps for the HADS-A**

| CATEGORY LABEL | OBSERVED COUNT | AVERAGE MEASURE | EXP. | COHERENCE EXP% | COHERENCE OBS% | INFIT MNSQ | OUTFIT MNSQ | STEP CALIBRATN |
|---|---|---|---|---|---|---|---|---|
| 0 | 3248 | -2.63 | -2.59 | 73% | 59% | .98 | .98 | NONE |
| 1 | 4361 | -1.21 | -1.22 | 59% | 75% | .90 | .89 | -2.20 |
| 2 | 1570 | .12 | -.02 | 51% | 46% | .88 | .86 | .42 |
| 3 | 453 | .74 | 1.09 | 61% | 14% | 1.38 | 1.58 | 1.78 |

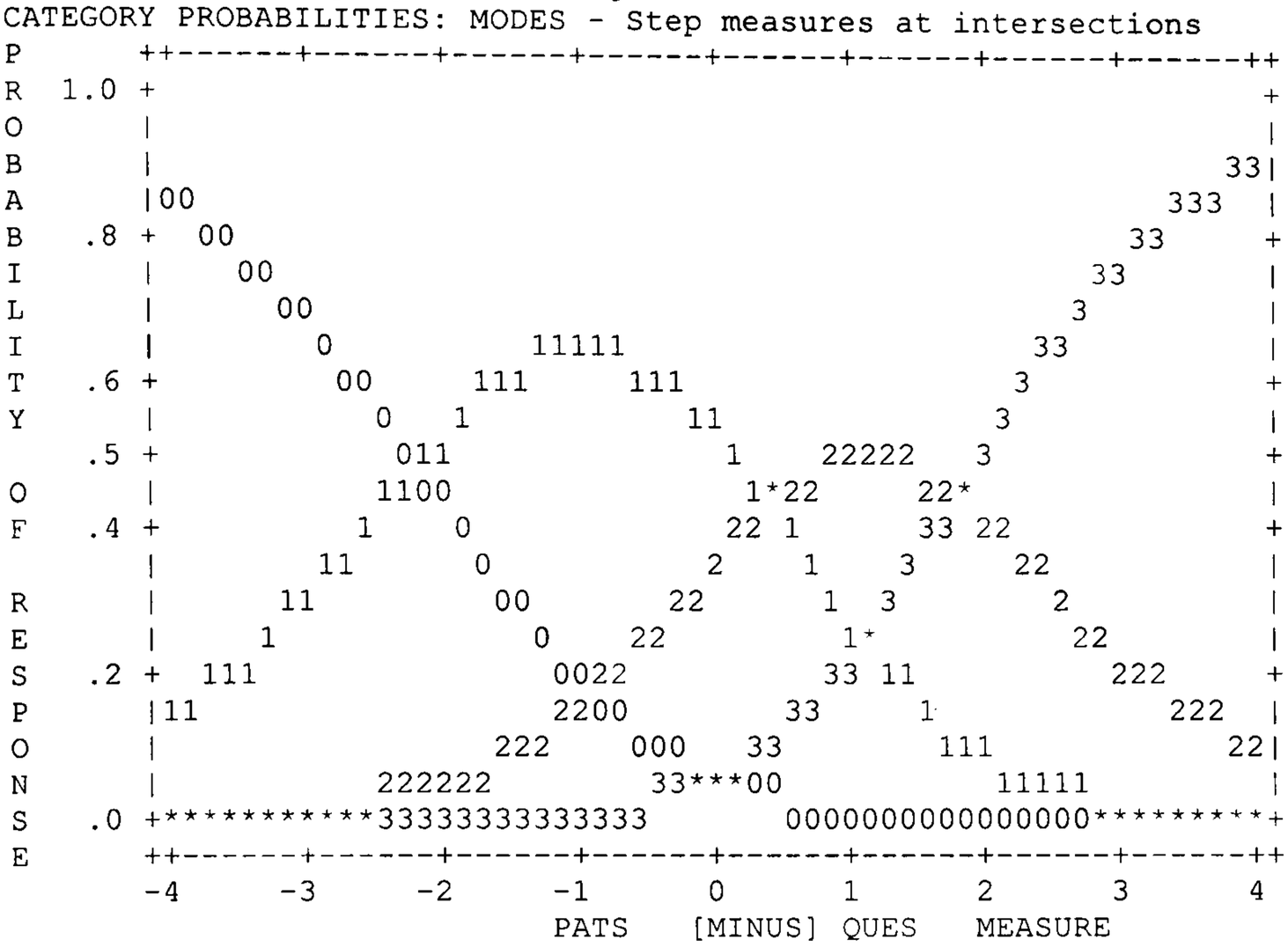
AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.
EXP% = (expected & observed)/(all expected) [MEASURE->RATING?]
OBS% = (expected & observed)/(all observed) [RATING->MEASURE?]

The item summary is shown in Table 5.3.4. The separation index for the items is 9.58, indicating that there are approximately 9 distinct difficulty (D) strata. The item reliability is high (0.99).

A summary of the step measures is shown in Table 5.3.5. The observed count is significantly higher than the minimum criterion for each category. Similarly, the step measures increase monotonically, as do the step threshold calibrations with a difference of at least 1.4 between all steps. The outfit statistics demonstrate good fit.

The category probability curve is shown below in Figure 5.3.4 and demonstrates that, for instance, a patient with low levels of anxiety, e.g. –3.0 logits is more likely to respond with a zero to any given question, whereas more anxious patients, e.g. measured at +3.0 logits are more likely to respond with a 3.

**Figure 5.3.4 Category Probability Curve for HADS Scale**




*5.3.3 Analysis of Person Locations*

The person measures from the HADS-A can be seen in Table 5.3.5. The differences between adjacent scores from the subscale expressed in logits can be seen in Figure 5.3.5.

**Table 5.3.5. Person measures for the HADS-A**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 0 | -4.86 | 1 | 0 | 1 | 0 |
| 1 | -4.09 | 0.97 | -0.04 | 0.86 | -0.16 |
| 2 | -3.23 | 0.83 | -0.33 | 0.76 | -0.45 |
| 3 | -2.65 | 0.62 | -0.94 | 0.6 | -0.99 |
| 4 | -2.17 | 0.48 | -1.38 | 0.48 | -1.37 |
| 5 | -1.74 | 1.63 | 1.02 | 1.69 | 1.13 |
| 6 | -1.34 | 0.42 | -1.36 | 0.43 | -1.35 |
| 7 | -0.97 | 0.53 | -1.02 | 0.54 | -0.99 |
| 8 | -0.62 | 0.74 | -0.52 | 0.75 | -0.49 |
| 9 | -0.29 | 0.42 | -1.4 | 0.4 | -1.46 |
| 10 | 0.02 | 0.43 | -1.45 | 0.4 | -1.51 |
| 11 | 0.31 | 0.25 | -2.29 | 0.25 | -2.28 |
| 12 | 0.59 | 0.38 | -1.78 | 0.38 | -1.76 |
| 13 | 0.86 | 0.94 | -0.14 | 0.93 | -0.16 |
| 14 | 1.14 | 1.78 | 1.44 | 1.76 | 1.42 |
| 15 | 1.42 | 0.34 | -2.07 | 0.34 | -2.05 |
| 16 | 1.73 | 0.8 | -0.48 | 0.78 | -0.52 |
| 17 | 2.06 | 0.42 | -1.54 | 0.46 | -1.36 |
| 18 | 2.45 | 1.08 | 0.13 | 1.08 | 0.14 |
| 19 | 2.96 | 0.48 | -0.97 | 0.4 | -1.08 |
| 20 | 3.75 | 1.05 | 0.05 | 1.26 | 0.22 |
| 21 | 4.49 | 1 | 0 | 1 | 0 |

It can be seen from Figure 5.3.5 that differences between the range of 10 to 15 are approximately equal (around 0.28 logits), beyond this range the differences increase sharply on both sides. This demonstrates that differences between scores in the third quarter of scores are roughly interval, whereas above and below this range larger, unequal differences occur.

Table 5.2.6. shows the summary for the person measures. It can be seen that the separation index for persons is 1.69, indicating that one distinct stratum of difficulty can be established. Reliability measures are good indicating reliabilities around 0.75.

**Figure 5.2.5. Differences in logits between adjacent scores of the HADS-A**

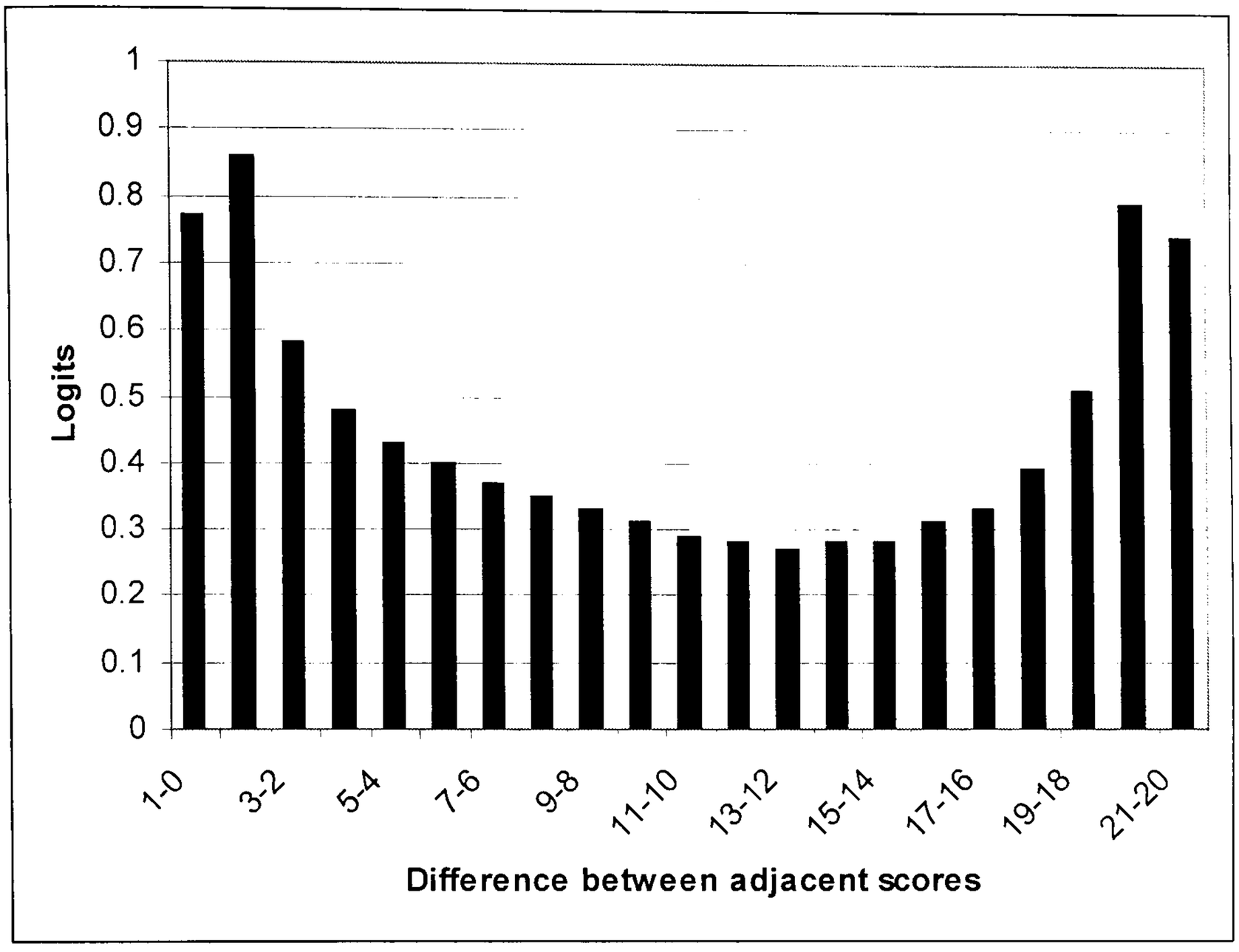


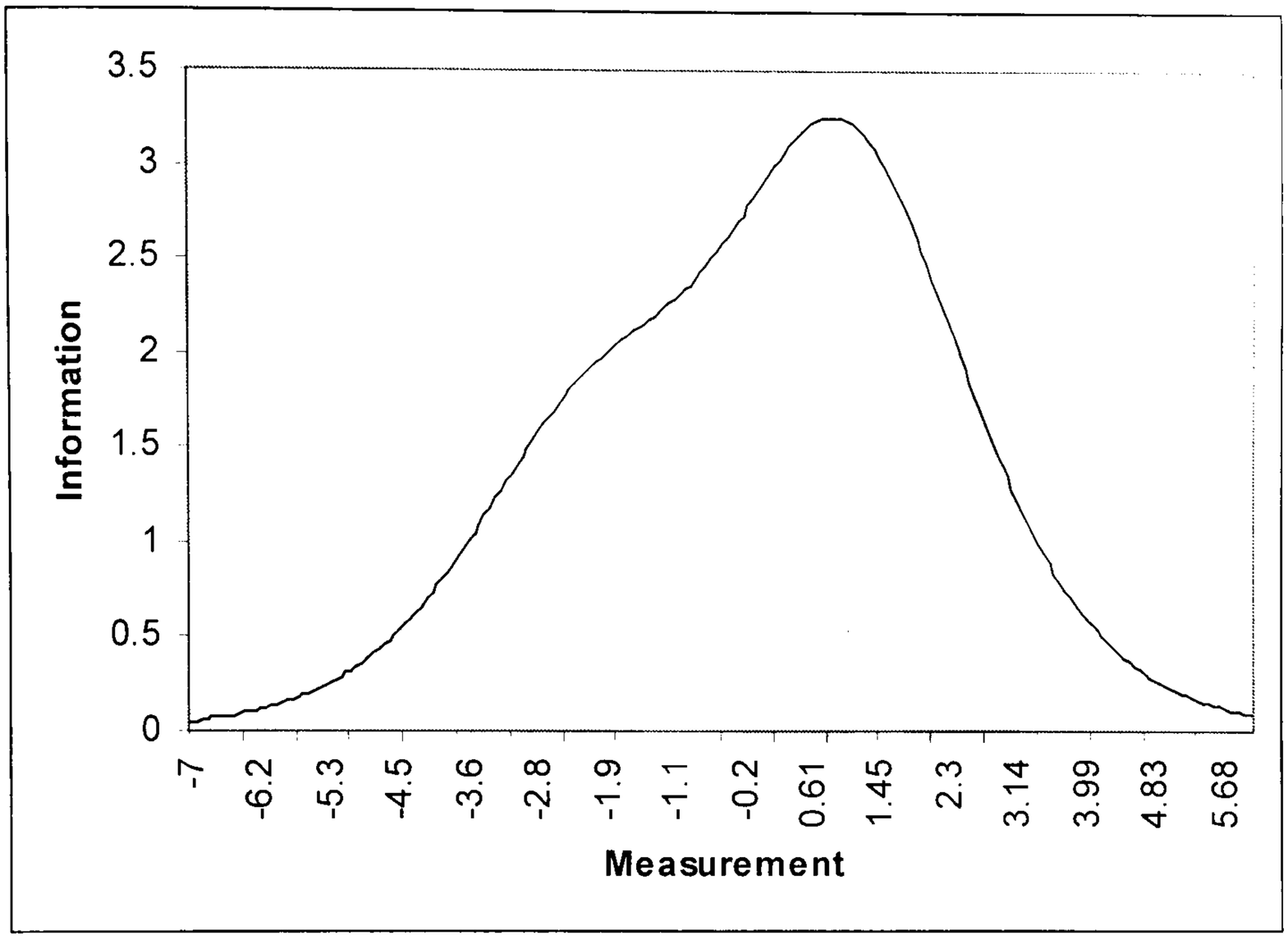

**Table 5.2.6. Summary of Person Measures for HADS-A**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 6.4 | 7.0 | -1.38 | .66 | 1.02 | -.2 | 1.01 | -.3 |
| S.D. | 3.8 | .0 | 1.48 | .14 | .75 | 1.3 | .74 | 1.3 |
| MAX. | 20.0 | 7.0 | 3.75 | 1.07 | 5.33 | 4.2 | 5.46 | 4.8 |
| MIN. | 1.0 | 7.0 | -4.09 | .52 | .08 | -3.5 | .08 | -3.5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .75 | ADJ.SD | 1.27 | SEPARATION | 1.69 | PATS | RELIABILITY | .74 |
| MODEL RMSE | .68 | ADJ.SD | 1.31 | SEPARATION | 1.93 | PATS | RELIABILITY | .79 |
| S.E. OF PATS MEAN | .04 | | | | | | | |
| WITH 93 EXTREME PATSS = 1469 PATSS | | | | MEAN | -1.59 | S.D. | 1.67 | |
| REAL RMSE | .82 | ADJ.SD | 1.45 | SEPARATION | 1.78 | PATS | RELIABILITY | .76 |
| MODEL RMSE | .75 | ADJ.SD | 1.49 | SEPARATION | 1.97 | PATS | RELIABILITY | .80 |

MAXIMUM EXTREME SCORE: 1 PATSS
MINIMUM EXTREME SCORE: 92 PATSS

**Figure 5.2.6 Test Information Curve for the HADS-A**




The test information curve for HADS-A is shown in Figure 5.2.6. As can be seen from the graph the most information is provided in the range from –0.97 to +1.73, corresponding to scores between 7 and 16. This range represents around 40% of patients' scores. The maximum point occurs at a measurement level of 1.10 or a raw score of around 14.

*5.3.4 Differential Item Functioning of HADS-Anxiety*

The HADS-A scale was investigated for differential item functioning, the results of which can be seen in Table 5.3.7.

**Table 5.3.7. Differential Item Functioning of HADS-A**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -.16 | .06 | 2 | -.14 | .07 | -.02 | .09 | -.21 | INF | 1 | A1 |
| 1 | -.13 | .06 | 2 | .04 | .07 | -.17 | .09 | -1.83 | INF | 2 | A2 |
| 1 | -.60 | .05 | 2 | -.41 | .07 | -.19 | .09 | -2.16 | INF | 3 | A3 |
| 1 | -.08 | .06 | 2 | -.40 | .07 | .32 | .09 | 3.52 | INF | 4 | A4 |
| 1 | .52 | .06 | 2 | .56 | .08 | -.04 | .10 | -.42 | INF | 5 | A5 |
| 1 | -.23 | .06 | 2 | -.62 | .07 | .39 | .09 | 4.39 | INF | 6 | A6 |
| 1 | .67 | .06 | 2 | 1.06 | .08 | -.40 | .10 | -3.84 | INF | 7 | A7 |

The sample was split into male and female groups (group 1 and group 2 respectively) for this analysis. Table 5.2.7 demonstrates that none of the items from the HADS-A exhibited differential item functioning. Although the t-statistic is significant for items 3, 4, 6 and 7, the difference between the item estimates for the samples is smaller than 0.50 logits (Wright and Panchapakesan, 1969).

## 5.4. Results for HADS – Depression subscale

### *5.4.1. Analysis of Unidimensionality*

The fit and location ("measure") statistics for the HADS-Depression subscale can be seen Table 5.4.1. Similar to HAD-Scale and HADS-A subscale the fit statistics for the HADS-D subscale indicate good fit (fit statistics < 1.3) for all items. Item 5 ("I have lost interest in my appearance") and 7 ("I can enjoy a good book or radio or TV programme") demonstrated fit statistics just greater than 1.3.

**Table 5.4.1. HADS-Depression - Unidimensionality measures**

| Entry | Measure | Count | Raw Score | In MSQ | In ZSTD | Out MSQ | Out ZSTD | Name |
|---|---|---|---|---|---|---|---|---|
| 1 | -0.61 | 1283 | 1206 | 0.98 | -0.45 | 0.93 | -1.69 | DEP1 |
| 2 | 0.62 | 1283 | 589 | 1.27 | 5.07 | 1.12 | 1.84 | DEP2 |
| 3 | 0.68 | 1283 | 566 | 0.76 | -5.17 | 0.78 | -3.7 | DEP3 |
| 4 | -1.68 | 1283 | 1931 | 1.01 | 0.29 | 1.07 | 1.7 | DEP4 |
| 5 | 0.44 | 1283 | 659 | 1.42 | 7.75 | 1.27 | 4.09 | DEP5 |
| 6 | -0.04 | 1283 | 883 | 0.77 | -5.84 | 0.68 | -7.14 | DEP6 |
| 7 | 0.6 | 1283 | 595 | 1.33 | 6 | 1.28 | 4.03 | DEP7 |

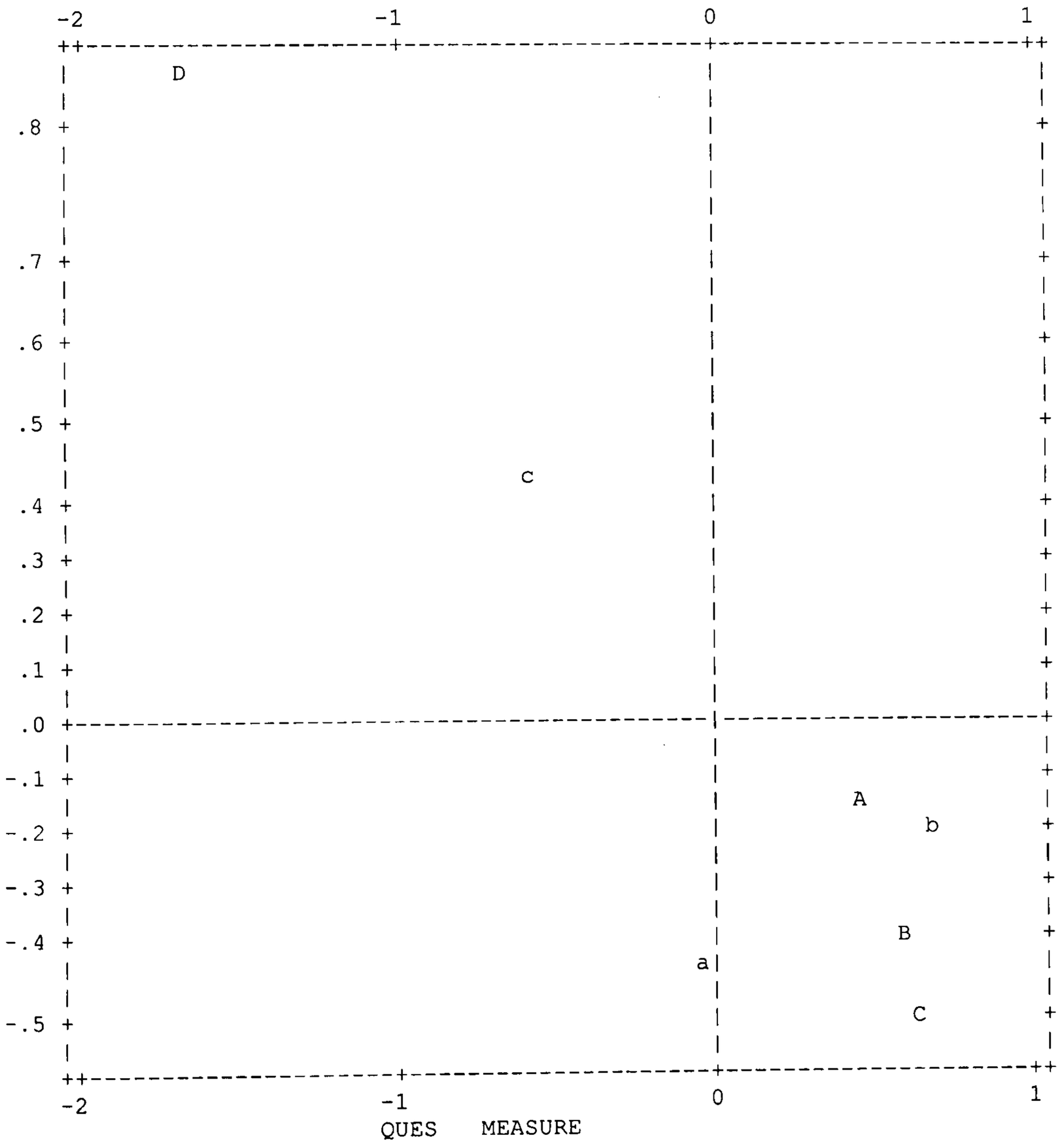
The results of the principal components analysis of the HADS-D (Table 5.4.2 and Figure 5.4.1) subscale resulted in 1.48 eigenvalues remaining in the analysis of the residuals. This figure is close enough to the criterion of 1.4 (Smith & Miao, 1994) to suggest that there is no remaining structure in the residuals once the Rasch factor has been extracted.

**Table 5.4.2. Factor 1 from principal component analysis of standardised residuals for the HADS-D (sorted by loading). Factor 1 explains 1.48 of 7**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | QUES |
|---|---|---|---|---|---|---|
| 1 | .83 | -1.68 | 1.01 | 1.07 | D 4 | DEP4 |
| 1 | .41 | -.61 | .98 | .93 | c 1 | DEP1 |
| 1 | -.50 | .62 | 1.27 | 1.12 | C 2 | DEP2 |
| 1 | -.42 | -.04 | .77 | .68 | a 6 | DEP6 |
| 1 | -.37 | .60 | 1.33 | 1.28 | B 7 | DEP7 |
| 1 | -.22 | .68 | .76 | .78 | b 3 | DEP3 |
| 1 | -.13 | .44 | 1.42 | 1.27 | A 5 | DEP5 |

Although there are clear differences in loading between the items two (items 1 and 4) and the remainder.

**Figure 5.4.1. Principal Components (Standardised Residual) Factor Plot of HADS-D**



*5.4.2. Analysis of Item Locations*

The location of the items from the HADS-D are shown in Figure 5.4.2. There is considerable overlap between items 2 ("I can laugh and see the funny side of things"), 3 ("I feel cheerful"), 5, and 7 with items 2 and 7 sharing the same location.

**Figure 5.4.2. Logit map of all items and patients for HADS-D**

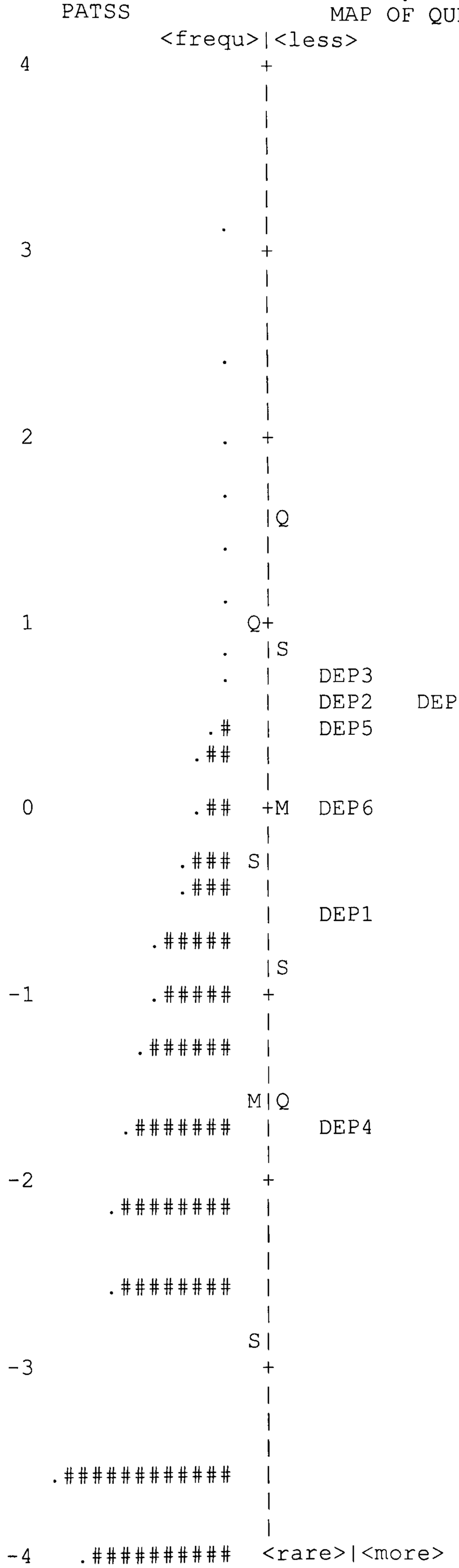


**Figure 5.4.2. Most Probable Response for HADS-D**

```
MOST PROBABLE RESPONSE: MODE  (BETWEEN "0" AND "1" IS "0", ETC.)
 -   -4    -3    -2    -1     0     1     2     3     4
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   QUES
0                          1          2 3               3    3   DEP3
0                          1          2 3               3    2   DEP2
0                          1          2 3               3    7   DEP7
0                         1          2 3                3    5   DEP5
|                                                       |
0                      1          2 3                   3    6   DEP6
|                                                       |
0                   1          2  3                     3    1   DEP1
|                                                       |
|                                                       |
0            1          2 3                            3    4   DEP4
|-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   QUES
 -   -4    -3    -2    -1     0     1     2     3     4

1        2    1  1  1 1
8        3    5  5  4 1 9 96 644 211 1
6        1    3  8  1 4 8 78 770 160 04 3 2 1    2        PATSS
     Q        S       M        S        Q
```

The most probable response table is shown in Figure 5.4.2. Items 2, 3, 5 and 7 confirm the overlap demonstrated in Figure 5.3.1. sharing virtually identical response probabilities for any given person ("ability") measure.

**Table 5.4.3. Item Summary for HADS-D**

SUMMARY OF 7 MEASURED QUESS

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 918.4 | 1283.0 | .00 | .05 | 1.08 | 1.1 | 1.02 | -.1 |
| S.D. | 465.0 | .0 | .81 | .01 | .25 | 5.0 | .22 | 3.9 |
| MAX. | 1931.0 | 1283.0 | .68 | .05 | 1.42 | 7.8 | 1.28 | 4.1 |
| MIN. | 566.0 | 1283.0 | -1.68 | .04 | .76 | -5.8 | .68 | -7.1 |
| REAL RMSE | .05 | ADJ.SD | .81 | SEPARATION | 16.03 | QUES | RELIABILITY | 1.00 |
| MODEL RMSE | .05 | ADJ.SD | .81 | SEPARATION | 17.34 | QUES | RELIABILITY | 1.00 |
| S.E. OF QUES | MEAN | .33 | | | | | | |

TABLE 3.2 Rasch analysis of DEPRESSION scores depress0.txt Apr 8 11:01 2003
INPUT: 1469 PATSS, 7 QUESS ANALYZED: 1283 PATSS, 7 QUESS, 4 CATS BIGSTEPS v2.82

The item summary for the HADS-D is given in Table 5.4.3. The mean infit and outfit statistics are close to 1.0 (1.08 and 1.02 respectively) indicating good overall fit. The separation index is approximately 16, indicating 16 distinct difficulty (D) strata.

**Table 5.4.4. Summary of Measured Steps for HADS-D**

SUMMARY OF MEASURED STEPS

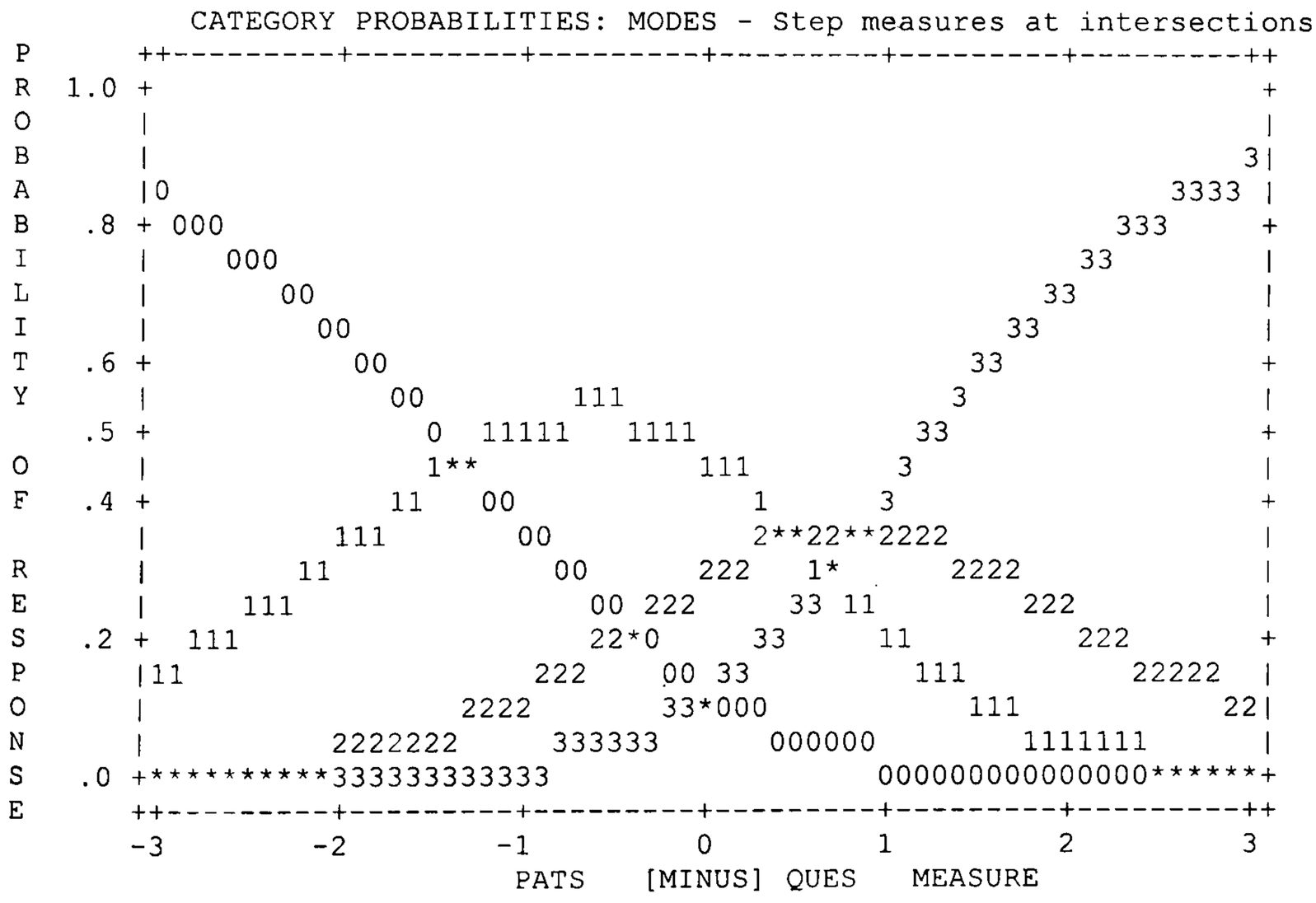
| CATEGORY LABEL | OBSERVED COUNT | AVERAGE MEASURE | EXP. | COHERENCE EXP% | COHERENCE OBS% | INFIT MNSQ | OUTFIT MNSQ | STEP CALIBRATN |
|---|---|---|---|---|---|---|---|---|
| 0 | 4624 | -2.56 | -2.48 | 80% | 77% | .91 | .96 | NONE |
| 1 | 2860 | -1.03 | -1.20 | 51% | 60% | .94 | .74 | -1.35 |
| 2 | 922 | -.02 | -.12 | 36% | 37% | .91 | .91 | .48 |
| 3 | 575 | .54 | .92 | 63% | 21% | 1.53 | 2.04 | .87 |

AVERAGE MEASURE is mean of (Bn-Di), EXP. is expected value.
EXP% = (expected & observed)/(all expected) [MEASURE->RATING?]
OBS% = (expected & observed)/(all observed) [RATING->MEASURE?]

The summary of measured steps is shown in Table 5.4.4, which shows that the average measure increases monotonically, although the outfit mean square is just greater than 2. Additionally, the difference between steps 1 and 2 is in excess of 1.4, whereas that between steps 2 and 3 is considerably less, i.e. more misinformation is being produced (Linacre, 1999a).

The category probability curve is shown in Figure 5.4.3, and shows that as a patient's "ability" increases, i.e. their psychological distress increases their probability of selecting a three increases, as the probability of other responses decreases.

**Figure 5.4.3. Category Probability Curve for HADS-D**



*5.4.3. Analysis of Person Locations*

The person measures from the HADS-D scale are shown in Table 5.4.5, and a summary of person measures is given in Table 5.4.6.

**Table 5.4.5. Person measures for the HADS-D**

| MEASURE | SCORE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| -4.29 | 0 | 1 | 0 | 1 | 0 |
| -3.5 | 1 | 1.33 | 0.31 | 1.97 | 0.62 |
| -2.64 | 2 | 0.69 | -0.47 | 1.04 | 0.04 |
| -2.09 | 3 | 0.94 | -0.09 | 0.92 | -0.13 |
| -1.67 | 4 | 0.44 | -1.23 | 0.51 | -1.03 |
| -1.32 | 5 | 0.96 | -0.07 | 0.81 | -0.37 |
| -1.01 | 6 | 1.06 | 0.12 | 1.16 | 0.28 |
| -0.73 | 7 | 2.58 | 2.14 | 2.66 | 2.18 |
| -0.47 | 8 | 0.15 | -2.72 | 0.14 | -2.74 |
| -0.22 | 9 | 1.56 | 0.94 | 1.61 | 0.99 |
| 0.01 | 10 | 0.7 | -0.67 | 0.69 | -0.67 |
| 0.24 | 11 | 1.02 | 0.03 | 0.98 | -0.04 |
| 0.46 | 12 | 0.84 | -0.35 | 0.79 | -0.43 |
| 0.68 | 13 | 0.72 | -0.68 | 0.65 | -0.75 |
| 0.91 | 14 | 0.39 | -1.76 | 0.73 | -0.55 |
| 1.14 | 15 | 1 | -0.01 | 0.91 | -0.16 |
| 1.38 | 16 | 1.92 | 1.53 | 1.58 | 0.76 |
| 1.65 | 17 | 0.46 | -1.32 | 0.55 | -0.73 |
| 1.98 | 18 | 1 | 0 | 0.74 | -0.32 |
| 2.4 | 19 | 0.61 | -0.58 | 0.48 | -0.62 |
| 3.09 | 20 | 0.75 | -0.24 | 0.53 | -0.38 |

## Table 5.4.6. Summary of Person Measures for HADS-D

SUMMARY OF 1283 MEASURED (NON-EXTREME) PATSS

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 5.0 | 7.0 | -1.61 | .69 | .98 | -.2 | 1.02 | -.2 |
| S.D. | 3.6 | .0 | 1.28 | .21 | .82 | 1.1 | 1.01 | 1.2 |
| MAX. | 20.0 | 7.0 | 3.09 | 1.09 | 5.38 | 4.4 | 6.93 | 5.2 |
| MIN. | 1.0 | 7.0 | -3.50 | .47 | .09 | -3.3 | .10 | -3.1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .79 | ADJ.SD | 1.00 | SEPARATION | 1.27 | PATS | RELIABILITY | .62 |
| MODEL RMSE | .72 | ADJ.SD | 1.05 | SEPARATION | 1.47 | PATS | RELIABILITY | .68 |
| S.E. OF PATS | MEAN | .04 | | | | | | |

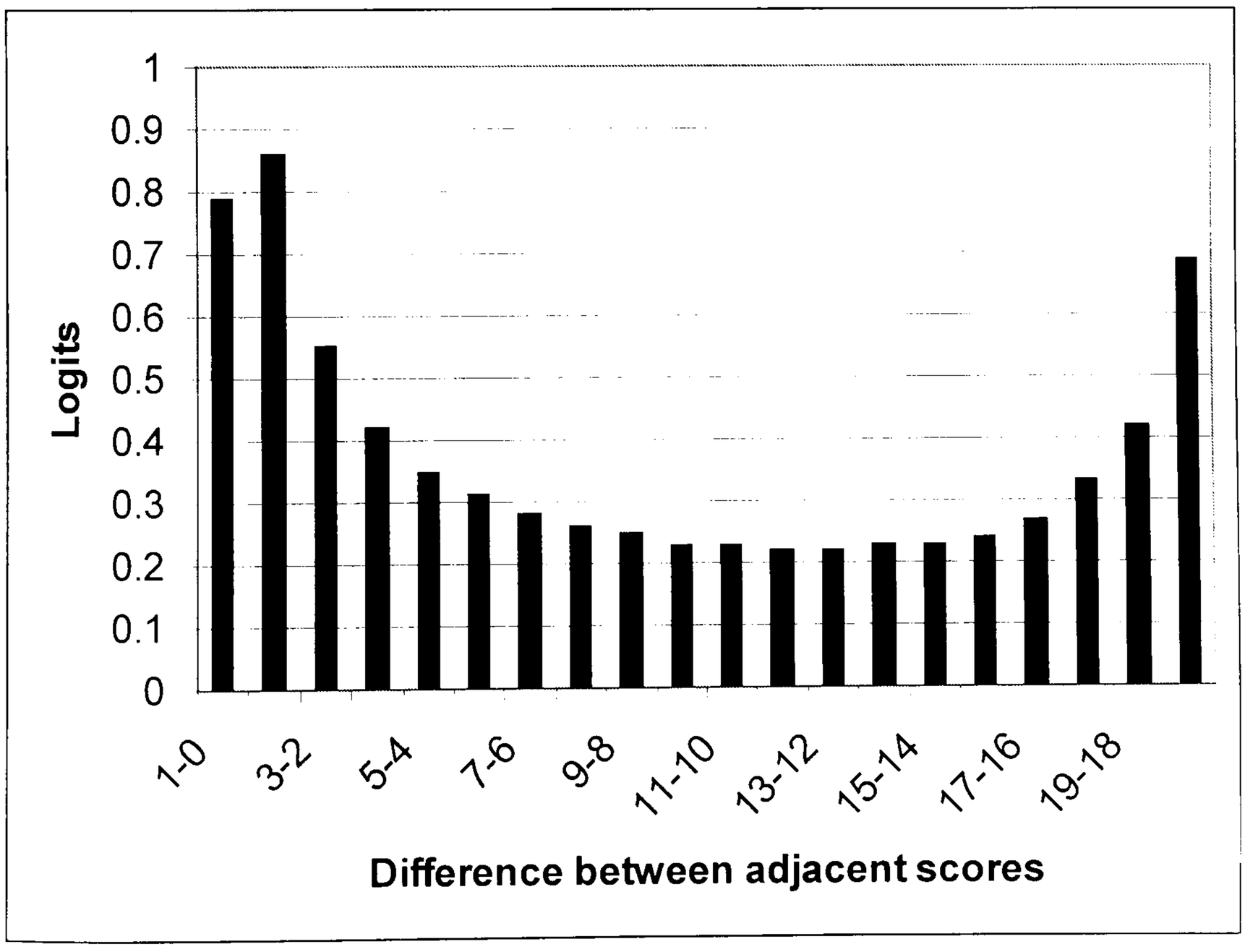
WITH 186 EXTREME PATSS = 1469 PATSS MEAN -1.95 S.D. 1.49

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .91 | ADJ.SD | 1.18 | SEPARATION | 1.30 | PATS | RELIABILITY | .63 |
| MODEL RMSE | .85 | ADJ.SD | 1.22 | SEPARATION | 1.43 | PATS | RELIABILITY | .67 |

MINIMUM EXTREME SCORE: 186 PATSS

Once again the mean fit statistics are close to 1.0 for both infit (.98) and outfit (1.02) indicating good overall fit. The separation index is 1.27 indicating approximately one distinct ability (B) stratum. The reliability index for the person measures was good (0.62).

## Figure 5.4.4. Difference in logits between adjacent scores of the HADS-D



The differences in logits between adjacent scores can be seen in Figure 5.4.4. Although the differences between scores in the mid-range (10 – 15) are roughly equal, beyond these scores differences increase sharply at both extremes.

**Figure 5.4.5. Test Information Curve for the HADS-D**

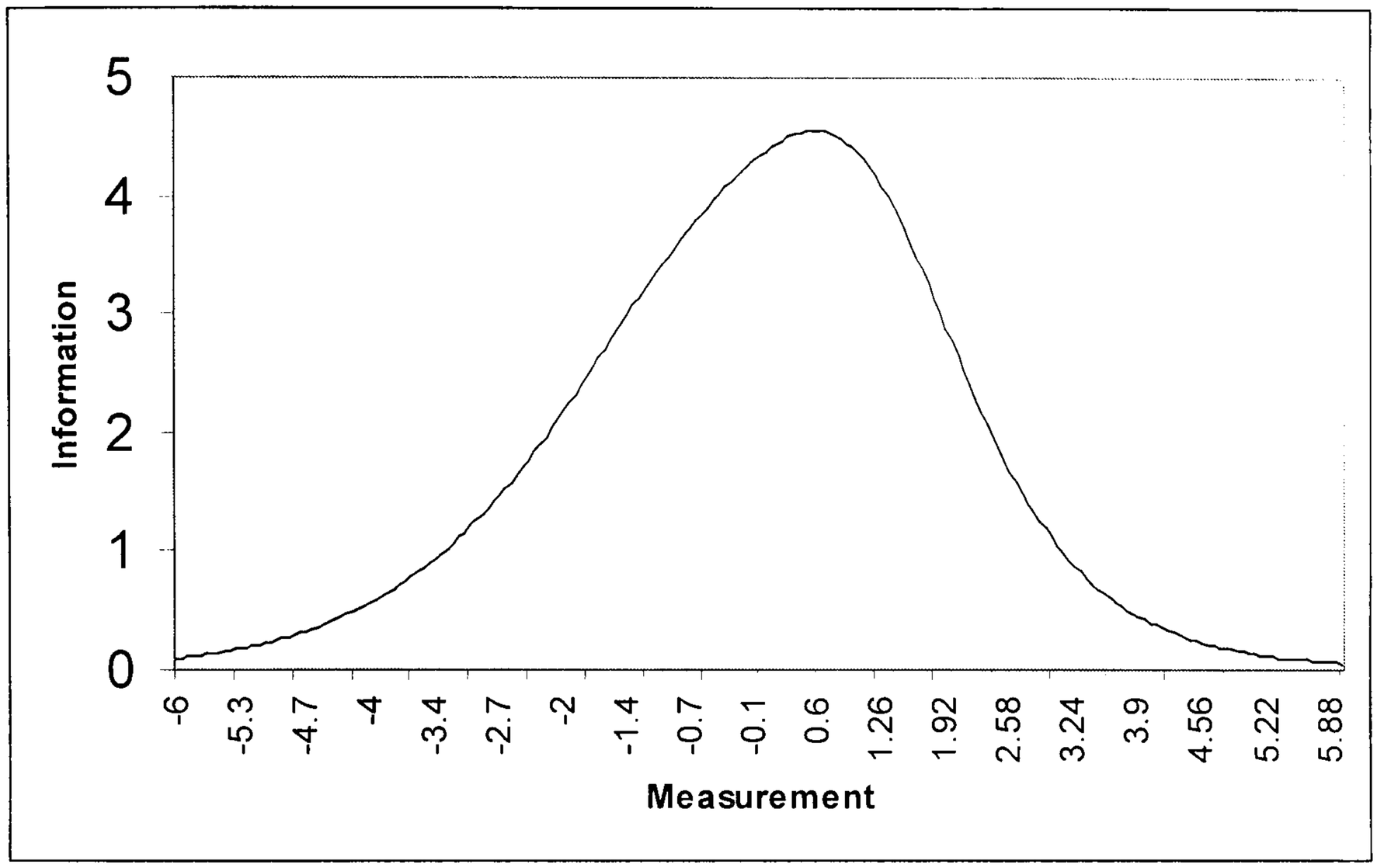



Figure 5.4.5 shows the test information curve for HADS-D. The most information is provided in the range between –0.73 and +1.38, or between scores of 7 and 16. This corresponds to 25.9% of all of patients' scores. The maximum occurs at around 0.60, which corresponds to a raw score of 13.

*5.4.4 Differential Item Functioning of HADS-D*

The HADS-D scale was investigated for differential item functioning, the results of which can be seen in Table 5.4.7.

**Table 5.4.7 Differential Item Functioning of HADS-D**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -.59 | .05 | 2 | -.64 | .06 | .05 | .08 | .59 | INF | 8 | D1 |
| 1 | .65 | .07 | 2 | .57 | .08 | .09 | .10 | .83 | INF | 9 | D2 |
| 1 | .73 | .07 | 2 | .60 | .08 | .13 | .11 | 1.22 | INF | 10 | D3 |
| 1 | -1.69 | .05 | 2 | -1.65 | .06 | -.04 | .08 | -.51 | INF | 11 | D4 |
| 1 | .39 | .06 | 2 | .52 | .08 | -.13 | .10 | -1.28 | INF | 12 | D5 |
| 1 | -.07 | .06 | 2 | .01 | .07 | -.09 | .09 | -.95 | INF | 13 | D6 |
| 1 | .61 | .07 | 2 | .58 | .08 | .03 | .10 | .33 | INF | 14 | D7 |

The sample was split into males (group 1) and females (group 2) for the differential item functioning analysis. As Table 5.4.7 suggests none of the items from the HADS-D exhibited item bias, i.e. all t-statistics are smaller than1.96 and none of the differences between items exceeds 0.50 logits.

## 5.5. Discussion

The Rasch analysis demonstrated good fit statistics for the majority of items from the HADS-Total and the subscales HADS-A and HADS-D, although there was a limited amount of overlap between some items from each scale. Reliability indices ranged from good to very good, and were in line with the results from the Cronbach's alpha statistics from Study 5.1. Step calibrations and step measures corresponded to published criteria (e.g. Linacre, 1999a). The analysis suggested that 3 of the items from the HADS-total and one from each of the subscales demonstrated a lack of fit with the model (infit mean squares greater than 1.30). None of the items from either the scales or subscales exhibited differential item functioning.

In addition, the overall fit and reliability for the person measures for the scales was also good. The Rasch analyses confirmed and extended the results from the traditional psychometrics undertaken in Study 5.1.

The results from the Principal Components Analysis (PCA) demonstrated an underlying factor, e.g. psychological distress, and two additional factors corresponding to Anxiety and Depression. These results are identical to those from the first and second order factor analysis carried in Study 5.1. Moreover, as with the

results from 5.1, item 4 from the HADS-A subscale ("I can sit at ease and feel relaxed") also loaded onto the HADS-D subscale, rather than the Anxiety subscale. The subsequent PCA of the subscales demonstrated two uniform structures corresponding to Anxiety and Depression.

The Rasch analysis extended results from traditional psychometrics as well by allowing differences between adjacent scores to be plotted. This analysis showed that the scales and subscales are interval based for mid-range scores, but that larger differences exist between scores either at one extreme (HADS-Scale) or both (HADS-A and HADS-D). Therefore caution needs to be exercised when interpreting changes, since changes at the extremes require a greater level of person measure than those occurring in the mid-range of scores.

Finally, the test information curves demonstrated the range for the greatest amount of information, or "test information", for each scale. The cutoff points (i.e. 8, Zigmond and Snaith, 1983) used for the subscales for identifying potential cases of either Anxiety or Depression fall within the range for these figures (i.e. scores of 7 to 16), although the Rasch analysis suggests that scores of 13 (HADS-D) or 14 (HADS-A) provide the greatest amount of information. The test information curve for the HADS-total suggest that scores around 24 provide the greatest amount of information. This threshold is considerably greater than limits suggested previously (Hopwood et al., 1991).

The next study will explore the factor analysis and Rasch analysis of the functional scales and some of the symptom scales of the EORTC-QLQ C30.

## 6. Factor and Rasch Analysis of the EORTC QLQ-C30

Quality-of-life (QOL) assessment is playing an increasingly important role in clinical cancer research, and more recently, in routine practice in oncology clinics (Detmar, & Aaronson, 1998; Taenzer, Bultz, Carlson et al., 2000; Velikova, Brown Smith et al., 2002). There are now several well-validated measurement tools in existence, prominent among which are the EORTC QLQ-C30 (EORTC QLQ-C30, Aaronson, Ahmedzai, Bergman et al., 1993) and the Functional Assessment of Cancer Therapy – General (FACT-G, Cella, Tulsky, Gray et al., 1993).

However, aside from the early development of these questionnaires (EORTC QLQ-c20, Aaronson et al., 1993; FACT-G, Cella et al., 1993) there has been little by way of investigation of the factorial structure and internal consistency of either instrument, other than the two studies by Ringdal and colleagues (Ringdal and Ringdal, 1993; Ringdal et al., 1999) which investigated the internal structure of the EORTC QLQ-C30 using Mokken scales, and similarly the single study by Kemmler et al. (2002) which explored the structure of the FACT-G, as noted in Chapter 2.

The following two studies (Chapters 6 and 7) have several aims, namely: 1). To explore the traditional psychometric properties of both the EORTC QLQ-C30 (Chapter 6) and the FACT-G (Chapter 7); 2). To apply a Rasch analysis to both instruments; and 3). To discuss the results in terms of the implications for interpretation of quality of life data in the context of clinical significance.

Therefore a factor analysis (principal components analysis) of the instruments was carried out, as well as an examination of the internal consistency and reliability (Cronbach's alpha and inter-item correlations). In addition to this the Rasch model for polytomous data (Andrich, 1978a, b) was applied to the instruments (i.e. the Physical Functioning, Emotional Functioning and Fatigue Scales of the EORTC QLQ-C30, and all four scales of the FACT-G: item (including, location and fit statistics) and person parameters (including differences between adjacent scores) were recorded.

## 6.1. Factor Analysis of the EORTC QLQ-C30

### *6.1.1. Aim*

This study investigated the factor structure of the EORTC QLQ C30 (version 3.1) in a large heterogeneous sample of 1625 cancer patients. Factor analysis (Principal Components Analysis) was carried out to investigate the factor structure of the instrument. In addition, the reliability of the functional subscales (Physical, Social, Role, Emotional and Cognitive Functioning), the Global Quality of Life Scale, and the Fatigue, Pain and Nausea and Vomiting subscales was assessed using Cronbach's alpha.

### *6.1.2. Method*

The patient data for the EORTC QLQ-C30 were collated from a total of six studies which have been carried out by the Cancer Research UK, Psychosocial and Clinical Practice Research Group (St. James's University Hospital, Leeds) and Cancer Research UK, Medical Oncology Unit (Western General Hospital, Edinburgh) over the past four years (Cull et al., 2001; Velikova et al., 1999; Velikova et al., 2002; Wright et al., 2003).

Patients completed an electronic version of the EORTC QLQ-C30 (version 3.1) on a standalone computer with touchscreen monitor. The raw scores from both questionnaires were recorded onto an MS-Access database and converted to the summated scales.

### *6.1.3 Participants*

A total of 1625 patients – 923 females (average age 56.4 years, s.d. 14.1) and 699 males (average age 54.1, s.d. 16.3) - completed the electronic questionnaire. Table 6.1.1 gives a breakdown of diagnosis by gender.

**Table 6.1.1 Diagnosis by gender and age**

| | | | | |
|---|---|---|---|---|
| EORTC QLQ-C30 | Female | | Males | |
| | | | | |
| | n = 923 | | n = 699 | |
| | | | | |
| Age, years (mean ± S.D.) | 56.4 ± 14.1 | | 54.1 ± 16.3 | |
| | | | | |
| | Count | % | Count | % |
| Diagnosis | | | | |
| | | | | |
| Breast | 351 | 38.0 | 5 | 0.7 |
| Colorectal | 75 | 8.1 | 108 | 15.5 |
| Gastrointestinal | 36 | 3.9 | 104 | 14.9 |
| Genitourinary/Gynae. | 353 | 38.2 | 319 | 45.6 |
| Lung | 17 | 1.8 | 47 | 6.7 |
| Melanoma | 13 | 1.4 | 33 | 4.7 |
| Renal | 14 | 1.5 | 21 | 3.0 |
| Sarcoma | 5 | 0.5 | 7 | 1.0 |
| Unknown | 22 | 2.4 | 14 | 2.0 |
| Other | 38 | 4.1 | 41 | 5.9 |

*6.1.4 Methodology*

A principal components analysis was carried out on the data. Factors were identified using a scree plot and Kaiser's criterion of eigenvalues greater than 1. Subsequently a factor analysis was carried out on the rotated data. An orthogonal rotation was carried out on the data rotated factors in order to minimise the level of correlation between the factors. The varimax procedure was selected since this is the optimal procedure for obtaining a simple structure rotation (Kline, 1997).

*6.1.5 Results*

Table 6.1.2 shows a breakdown of the scores from the EORTC QLQ-C30 scores for the functional and symptom scales.

**Table 6.1.2. Means and standard deviations of the EORTC QLQ-C30 scores**

| | Mean | Std. Deviation |
|---|---|---|
| Functional Scales | | |
| | | |
| Physical Functioning | 81.0 | 19.96 |
| Role Functioning | 67.3 | 32.34 |
| Emotional Functioning | 73.1 | 22.52 |
| Cognitive Functioning | 81.1 | 21.35 |
| Social Functioning | 71.3 | 30.10 |
| Global QL | 64.1 | 24.50 |
| | 34.4 | 26.29 |
| Symptoms | | |
| | | |
| Fatigue | | |
| Pain | 24.5 | 28.52 |
| Nausea & Vomiting | 11.2 | 20.42 |
| Dyspnoea | 22.4 | 29.09 |
| Appetite | 20.7 | 30.90 |
| Sleeplessness | 29.9 | 31.58 |
| Constipation | 16.7 | 27.67 |
| Diarrhoea | 10.3 | 21.44 |
| Finance | 14.2 | 25.85 |

*Abbreviations are as for Chapter 5

The correlation matrix for the scales from the EORTC QLQ-C30 is provided in Table 6.1.3. All correlations are significant at $p < 0.05$.

**Table 6.1.3 Correlation matrix for the EORTC QLQ-C30**

| | PF | RF | EF | CF | SF | QL | FA | NV | PA | DY | SL | AP | CO | DI | FI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PF | - | | | | | | | | | | | | | | |
| RF | 0.65 | - | | | | | | | | | | | | | |
| EF | 0.32 | 0.38 | - | | | | | | | | | | | | |
| CF | 0.35 | 0.38 | 0.48 | - | | | | | | | | | | | |
| SF | 0.54 | 0.74 | 0.44 | 0.42 | - | | | | | | | | | | |
| QL | 0.59 | 0.62 | 0.46 | 0.43 | 0.62 | - | | | | | | | | | |
| FA | -0.67 | -0.73 | -0.50 | -0.51 | -0.68 | -0.68 | - | | | | | | | | |
| NV | -0.31 | -0.41 | -0.31 | -0.31 | -0.42 | -0.41 | 0.49 | - | | | | | | | |
| PA | -0.49 | -0.58 | -0.41 | -0.39 | -0.55 | -0.53 | 0.58 | 0.41 | - | | | | | | |
| DY | -0.50 | -0.44 | -0.29 | -0.32 | -0.40 | -0.45 | 0.53 | 0.28 | 0.34 | - | | | | | |
| SL | -0.33 | -0.33 | -0.43 | -0.39 | -0.34 | -0.39 | 0.43 | 0.25 | 0.33 | 0.28 | - | | | | |
| AP | -0.38 | -0.47 | -0.36 | -0.34 | -0.47 | -0.50 | 0.55 | 0.54 | 0.44 | 0.32 | 0.27 | - | | | |
| CO | -0.30 | -0.35 | -0.26 | -0.29 | -0.34 | -0.32 | 0.39 | 0.32 | 0.36 | 0.25 | 0.23 | 0.32 | - | | |
| DI | -0.11 | -0.16 | -0.14 | -0.15 | -0.16 | -0.18 | 0.19 | 0.21 | 0.15 | 0.11 | 0.16 | 0.20 | 0.10 | - | |
| FI | -0.16 | -0.32 | -0.23 | -0.18 | -0.38 | -0.25 | 0.27 | 0.21 | 0.21 | 0.16 | 0.16 | 0.17 | 0.13 | 0.09 | - |

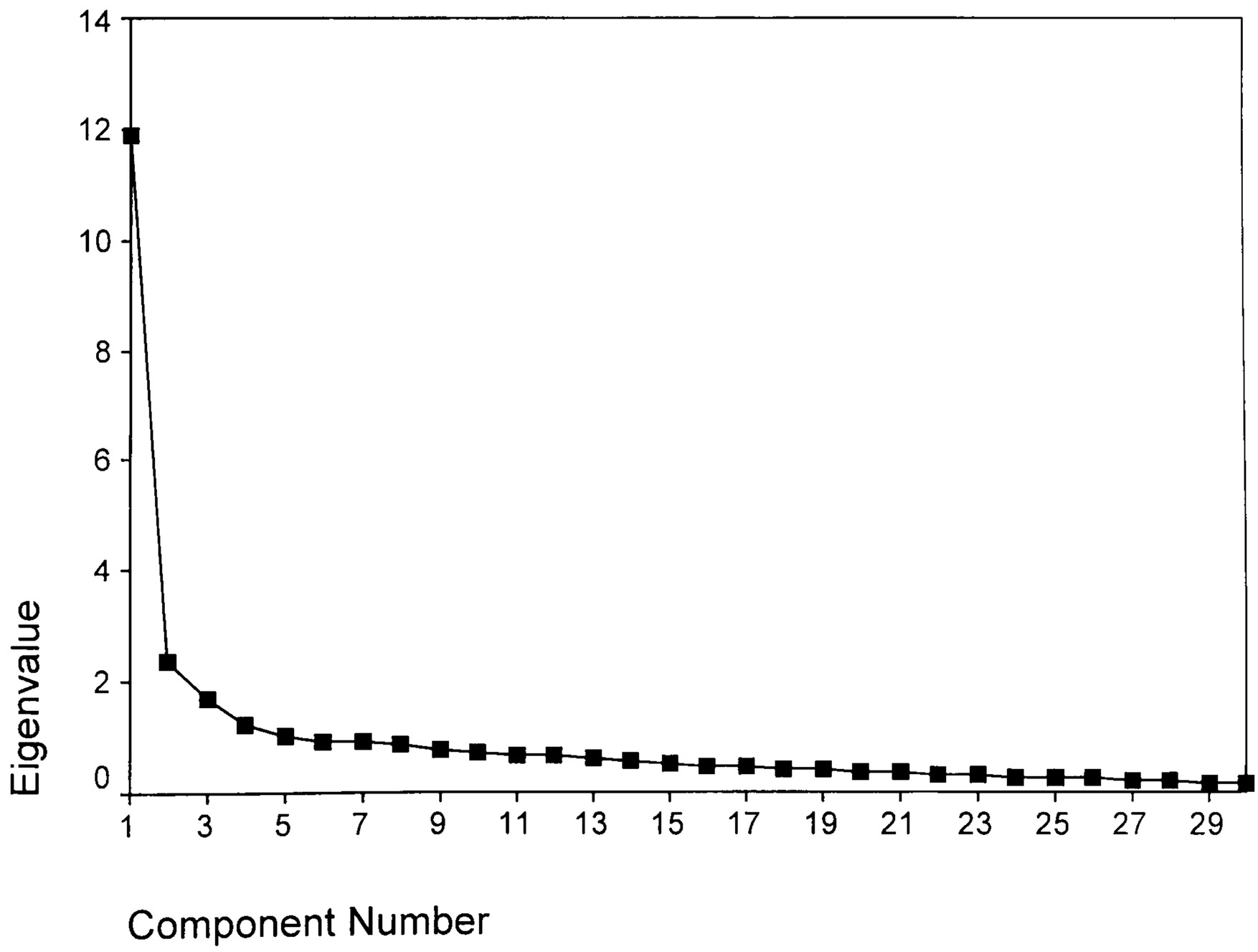
The eigenvalues from the principal components analysis is shown in Table 6.1.4. It can be seen that five factors are identified with eigenvalues greater than 1, collectively explaining just over 60% of the variance.

**Table 6.1.4 Eigenvalues from Factor Analysis of EORTC QLQ-C30**

| Component | Eigenvalues | % of Variance | Cumulative % |
|---|---|---|---|
| 1 | 11.894 | 39.645 | 39.645 |
| 2 | 2.369 | 7.895 | 47.540 |
| 3 | 1.677 | 5.590 | 53.131 |
| 4 | 1.248 | 4.162 | 57.292 |
| 5 | 1.031 | 3.438 | 60.730 |

The graphical representation of the eigenvalues is shown in the form of a scree plot in Figure 6.1.1.

**Figure 6.1.1. Scree plot from the Principal Components Analysis of the EORTC QLQ-C30**



The component matrix is shown in Table 6.1.5. Values below 0.30 have been suppressed. The rotated component matrix is shown in Table 6.1.6.

**Table 6.1.5 Component matrix from the Principal Components Analysis of the EORTC QLQ-C30**

| Component | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| PF1 | .57 | -.49 | | | |
| PF2 | .64 | -.51 | | | |
| PF3 | .55 | -.49 | | | |
| PF4 | .60 | | | | |
| PF5 | .41 | | | | .59 |
| RF1 | .77 | | | | |
| RF2 | .76 | | | | |
| DY1 | .58 | | | | |
| PA1 | .67 | | | | |
| FA1 | .81 | | | | |
| SL1 | .54 | | | | |
| FA2 | .81 | | | | |
| AP1 | .66 | | | | |
| NV1 | .61 | | -.47 | | |
| NV2 | .48 | | -.53 | | |
| CO1 | .49 | | | | |
| DI1 | | | | | |
| FA3 | .76 | | | | |
| PA2 | .71 | | | | |
| CF1 | .66 | | | | |
| EF1 | .56 | .50 | .41 | | |
| EF2 | .54 | .50 | | | |
| EF3 | .55 | .49 | | | |
| EF4 | .61 | .45 | | | |
| CF2 | .44 | | | | |
| SF1 | .73 | | | | |
| SF2 | .79 | | | | |
| FI1 | | | | -.53 | |
| QL1 | -.73 | | | | |
| QL2 | -.75 | | | | |

The rotated factor structure revealed four factors, Physical Functioning (Factor 2), Role, Social Functioning, Pain and Fatigue (Factor 1), Emotional and Cognitive Functioning (Factor 3), and a fourth factor consisting primarily of symptom scales, in particular appetite, nausea and vomiting, and diarrhoea (see Table 6.1.5).

Therefore with the exception of the Physical Functioning Scale, none of the other scales demonstrate a uniform structure predicted by the construction of the scales.

**Table 6.1.6 Rotated component matrix of the EORTC QLQ-C30**

| Component | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| PF1 | | .78 | | | |
| PF2 | | .86 | | | |
| PF3 | | .80 | | | |
| PF4 | | .69 | | | |
| PF5 | | | | | .62 |
| RF1 | .70 | .43 | | | |
| RF2 | .70 | .43 | | | |
| DY1 | | .53 | | | |
| PA1 | .47 | | | | |
| FA1 | .51 | .50 | | | |
| SL1 | | | .49 | | |
| FA2 | .48 | .44 | | | |
| AP1 | | | | .63 | |
| NV1 | | | | .77 | |
| NV2 | | | | .80 | |
| CO1 | | | | .47 | |
| DI1 | | | | .46 | |
| FA3 | .43 | .42 | | | |
| PA2 | .54 | | | | |
| CF1 | | | .53 | | |
| EF1 | | | .83 | | |
| EF2 | | | .79 | | |
| EF3 | | | .78 | | |
| EF4 | | | .77 | | |
| CF2 | | | .47 | | |
| SF1 | .72 | | | | |
| SF2 | .75 | | | | |
| FI1 | .62 | | | | |
| QL1 | -.46 | -.44 | | | |
| QL2 | -.50 | -.42 | | | |

The sample was split randomly into two samples comprising 50% of the overall sample each, and the factor analysis was performed again. Additionally, the factor analysis was also carried out on a sample consisting only of the breast cancer patients, as well as on the remaining sample consisting of patients who did not have breast cancer.

Although the primacy of the factors changed for the different samples, the factor structures remained the same, demonstrating stability in the factor structures.

In addition, the reliability and internal consistency of the Functioning scales, the Global Quality of Life scale, as well as the Fatigue scale were also assessed.

**Table 6.1.7 Item correlations and Cronbach's alpha**

| Items | Item-total Correlation | Cronbach's α |
|---|---|---|
| **PF1** | 0.71 | 0.78 |
| **PF2** | 0.82 | 0.74 |
| **PF3** | 0.72 | 0.78 |
| **PF4** | 0.64 | 0.80 |
| **PF5** | 0.39 | 0.86 |
| **RF** | 0.77 | 0.87 |
| **CF** | 0.44 | 0.61 |
| **EF1** | 0.75 | 0.82 |
| **EF2** | 0.74 | 0.83 |
| **EF3** | 0.69 | 0.85 |
| **EF3** | 0.72 | 0.84 |
| **SF** | 0.76 | 0.87 |
| **QL** | 0.85 | 0.92 |
| **FA1** | 0.75 | 0.81 |
| **FA2** | 0.75 | 0.82 |
| **FA3** | 0.75 | 0.82 |

*Since the cognitive-, social, and role functioning scales, as well as the global quality of life scales only consist of two items the correlation coefficients and Cronbach's alpha are only recorded once.

Table 6.1.7 shows the item-total correlations and revised Cronbach's alpha. Except for Physical Functioning item 5, and Cognitive Functioning scale all other items show high item-total correlations in excess of 0.60. Similarly all items with the exception of the Cognitive functioning items show a high revised Cronbach's alpha.

**Table 6.1.8. Cronbach's alpha for Individual Scales**

| | Total Cronbach's alpha |
|---|---|
| Physical Functioning | 0.83 |
| Role Functioning | 0.87 |
| Emotional Functioning | 0.87 |
| Cognitive Functioning | 0.61 |
| Social Functioning | 0.76 |
| Global QoL | 0.92 |
| Fatigue | 0.87 |

Cronbach's alpha for the scales is shown in Table 6.1.8, which demonstrates that all scales with the exception of the Social and Cognitive Functioning scales showed very high levels of internal consistency, although the levels of consistency for these two scales were still good.

*6.1.6 Conclusions*

This study investigated the factor structure of the EORTC QLQ-C30 using a principal components analysis. The results demonstrated that although the internal consistency and inter-item correlations were reasonably good for the scales a four factor structure emerged from the rotated component matrix which did not correspond to the functioning domains, namely: a Physical Functioning factor, a factor covering the Role and Social Functioning domains, as well as Pain and Fatigue, and a factor covering the Emotional and Cognitive Functioning domains, as well as fourth factor including the majority of the symptom scales.

In the next section the data are analysed using a Rasch model for polytomous data, i.e. the Rating Scale Model (Andrich, 1978a, b).

## 6.2 Rasch Analysis of Components of the EORTC QLQ-C30

### *6.2.1 Aim*

The aim of this study is to examine the Physical Functioning, Emotional Functioning and Fatigue scales of the EORTC QLQ-C30 using Rasch models.

The rule of thumb for the minimum number of categories required for the Rasch analysis is a minimum of 10, and is provided by the product of the number of response categories and the number of items for each scale. Since the minimum category requirement is only met by three scales of the EORTC QLQ-C30, namely the Physical and Emotional Functioning Scales, as well as the Fatigue scale, the item locations, as well as the unidimensionality and person measures will be explored for these scales only and compared to the results of the factor analysis.

### *6.2.2 Methodology*

The data employed in this section were the same as described in section 5.1.3. The scores from the Physical Functioning, Emotional Functioning and Fatigue scales of the EORTC QLQ-C30 were converted to interval-level logit (log-odds) scores using the *Winsteps* software (Linacre & Wright, 2000), and the Rating Scale Model for polytomous data (Andrich 1978a, b) as described in Chapter 1.2.

In this study item difficulty and person ability estimates were derived, as well as fit statistics (infit and outfit) for the items. A principal components analysis of the residuals was also performed for the scales, and test characteristic curves were calculated. Additionally, the differences between adjacent scores for person measures were plotted for each scale, and the unidimensionality of the scales was explored further using differential item analyses.

*6.2.3. Results for Physical Functioning*

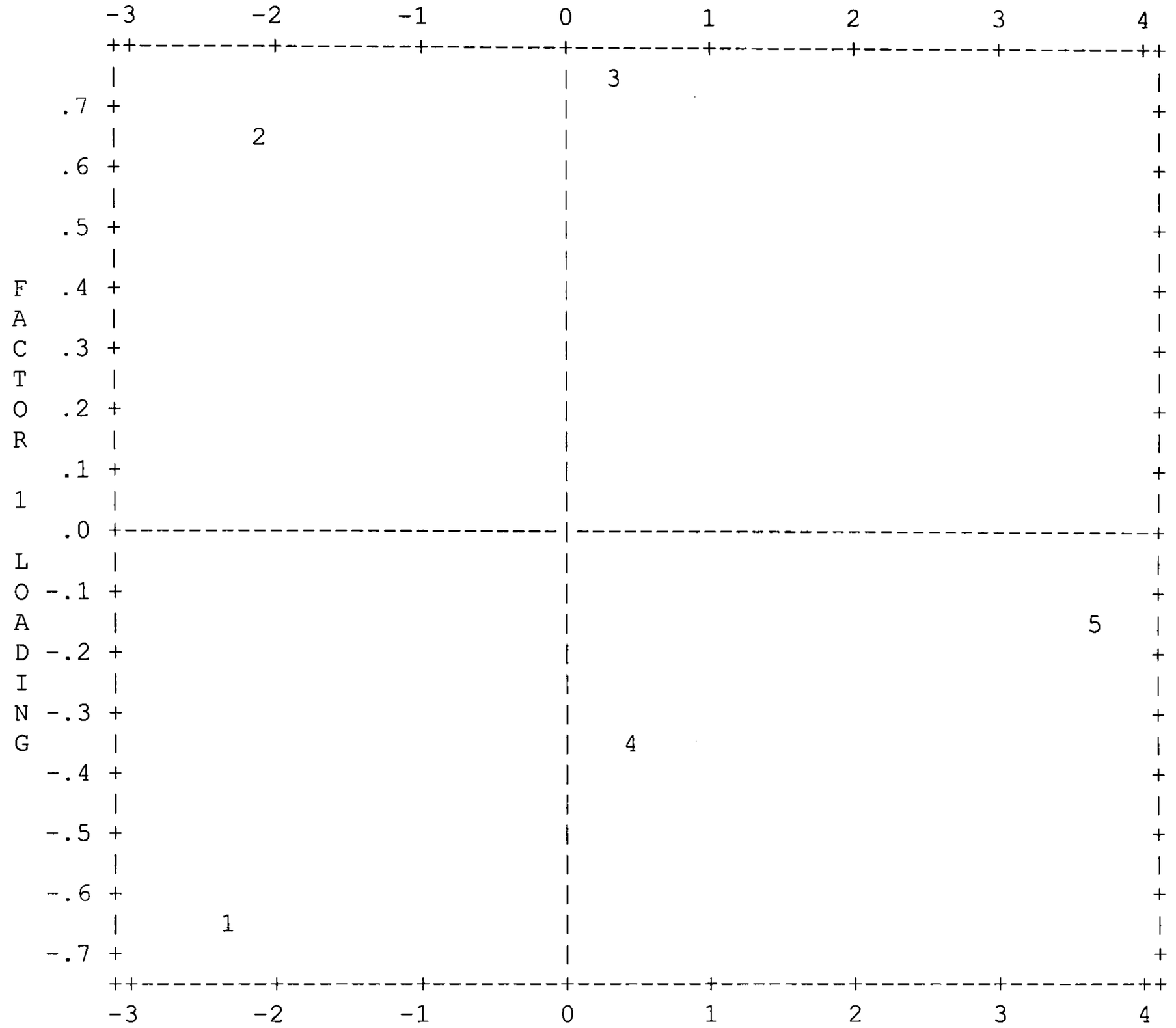
The location measures and fit statistics are given in Table 6.2.1 for the Physical Functioning scale. It can be seen from this table that two items from this scale exhibit poor fit, i.e. infit mean square statistics greater than 1.30 or smaller than 0.70 and standardised t-statistics greater than 1.96 (e.g. Wright et al., 1994), namely item 5 ("Do you need help with eating, dressing, washing yourself or using the toilet?"), which showed excessive "noise", and item 2 ("Do you have any trouble taking a long walk?") which demonstrated overfit with the model.

**Table 6.2.1 Unidimensionality measures for Physical Functioning**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | DISPLACE | QUESS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1332 | 1179 | 3.63 | .11 | 1.50 | 5.8 | 2.06 | 2.6 | .45 | .01 | PF5 |
| 4 | 1864 | 1179 | .43 | .06 | 1.19 | 3.9 | 1.41 | 5.6 | .73 | .00 | PF4 |
| 3 | 1890 | 1179 | .34 | .06 | 1.00 | .1 | .81 | -3.3 | .78 | .00 | PF3 |
| 2 | 2686 | 1179 | -2.07 | .05 | .64 | -9.9 | .63 | -9.2 | .91 | .00 | PF2 |
| 1 | 2782 | 1179 | -2.32 | .05 | .98 | -.4 | 1.03 | .6 | .86 | .00 | PF1 |
| MEAN | 2111. | 1179. | .00 | .07 | 1.06 | -.1 | 1.19 | -.7 | | | |
| S.D. | 547. | 0. | 2.15 | .02 | .28 | 5.4 | .51 | 5.1 | | | |

Figure 6.2.1. shows the factor plot of the principal components analysis of the standardised residuals for the Physical Functioning scale. A total of 1.5 eigenvalues were extracted from the residuals, indicating that there were no other factor structures in the residuals (Smith & Miao, 1994).

**Figure 6.2.1 Principal Components (Standardized Residual) Factor Plot of the Physical Functioning Scale**




The factor loadings from the principal components analysis and coding for Figure 6.2.1 can be seen in table 6.2.2.

**Table 6.2.2. Factor Loadings from the Principal Components Analysis of the Physical Functioning Scale**

| | | | INFIT | OUTFIT | ENTRY | | |
|---|---|---|---|---|---|---|---|
| FACTOR | LOADING | MEASURE | MNSQ | MNSQ | NUMBER | QUES | |
| 1 | .73 | .34 | 1.00 | .81 | 3 | 3 | PF3 |
| 1 | .63 | -2.07 | .64 | .63 | 2 | 2 | PF2 |
| 1 | -.67 | -2.32 | .98 | 1.03 | 1 | 1 | PF1 |
| 1 | -.34 | .43 | 1.19 | 1.41 | 4 | 4 | PF4 |
| 1 | -.15 | 3.63 | 1.50 | 2.06 | 5 | 5 | PF5 |

It is interesting to note that two items with the poorest fit load in an opposite direction to items with better fit.

The item map for the Physical Functioning Scale can be seen in Figure 6.2.2. This figure demonstrates the overlap between items PF3 and PF4, and between PF1 and PF2. In addition, the items do not cover the full range of person abilities (i.e. –6 to +6) with the majority of items falling between –2 and +2 on the ability scale.

The category probability curve for Physical Functioning is shown in Figure 6.2.3.

**Figure 6.2.2. Item Map for the Physical Functioning Scale**

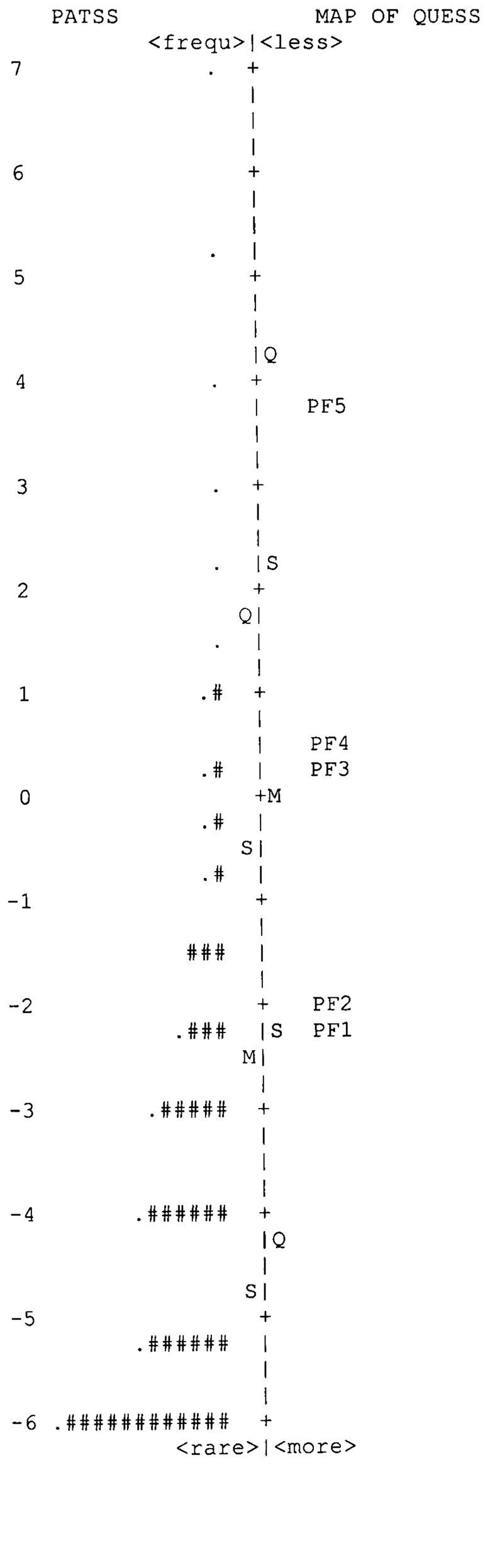
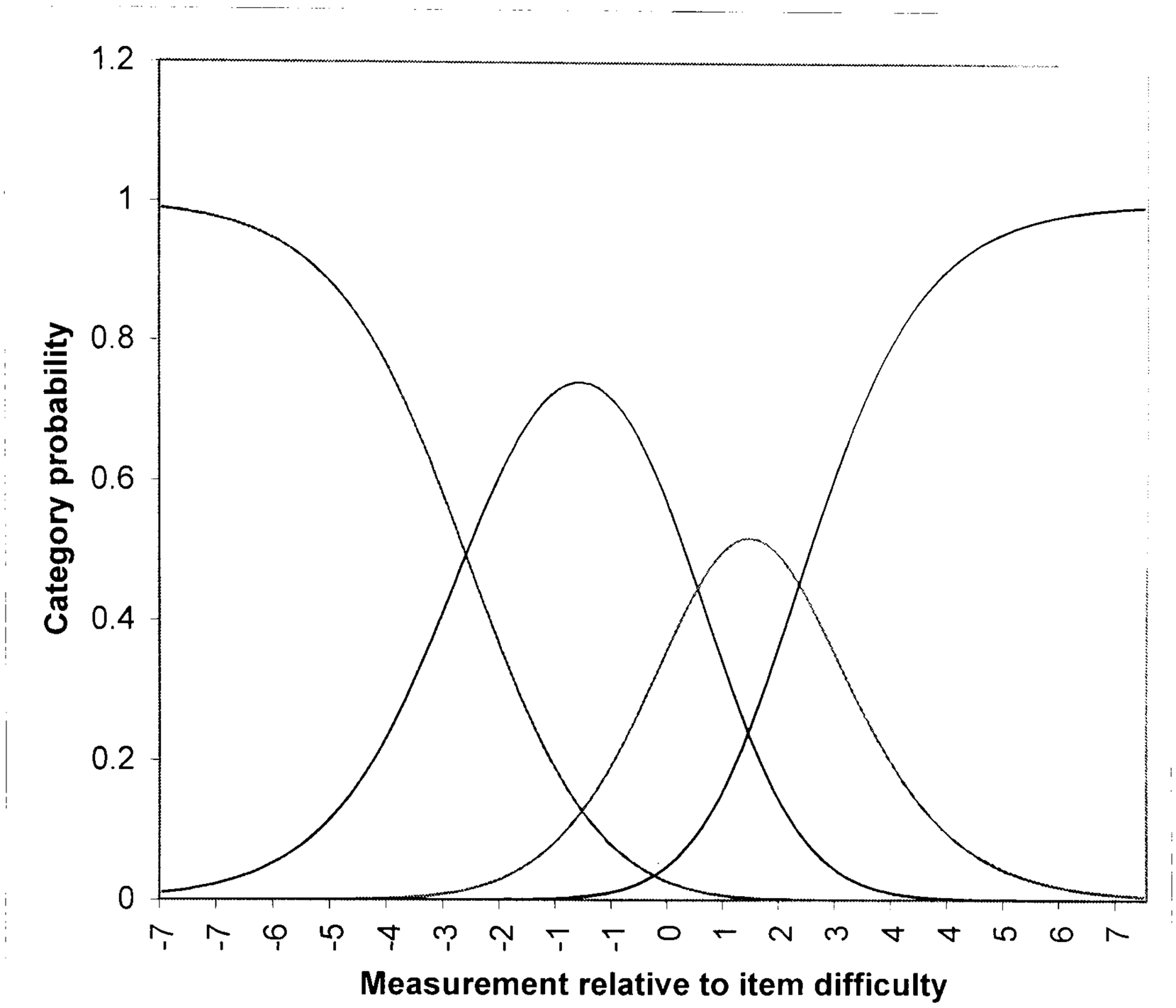


**Figure 6.2.3 Category Probability Curve for Physical Functioning**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Black = category 4.

It can be seen for instance, from figure 6.2.3 that as a patient's physical functioning declines, i.e. as the person measure increases in relation to the item difficulty, that the likelihood of the patient responding to items with "Not at all" (category 1) or "A little" (category 2 decreases, and the likelihood of responses such as "Quite a bit" (category 3) or "Very much" (category 4) increases.

Table 6.2.3 shows a summary of the items for the Physical Functioning Scale. The item separation index is approximately 28, demonstrating that the scale can distinguish between 28 levels of difficulty.

**Table 6.2.3. Summary of Items from the Physical Functioning Scale**

```
     SUMMARY OF 5 MEASURED (NON-EXTREME) QUESS
+-------------------------------------------------------------------------+
|           RAW                              MODEL         INFIT      OUTFIT |
|          SCORE      COUNT     MEASURE     ERROR       MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------|
| MEAN    2110.8     1179.0         .00       .07       1.06    -.1   1.19    -.7 |
| S.D.     547.3        .0        2.15       .02        .28    5.4    .51    5.1 |
| MAX.    2782.0     1179.0        3.63       .11       1.50    5.8   2.06    5.6 |
| MIN.    1332.0     1179.0       -2.32       .05        .64   -9.9    .63   -9.2 |
|-------------------------------------------------------------------------|
| REAL RMSE     .08  ADJ.SD    2.15  SEPARATION 27.48  QUES   RELIABILITY 1.00 |
|MODEL RMSE     .07  ADJ.SD    2.15  SEPARATION 30.92  QUES   RELIABILITY 1.00 |
| S.E. OF QUES MEAN = 1.08                                                |
+-------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
QUES RAW SCORE-TO-MEASURE CORRELATION = -.98
```

Table 6.2.4 shows a summary of the category measures for the Physical Functioning Scale. It can be seen that there is a good level of separation between the categories with a distance at least 1.4 logits (Linacre, 1999a) between each threshold (structure measure), and the distances between each category increase monotonically (category measure).

**Table 6.2.4. Summary of Category Measures for the Physical Functioning Scale**

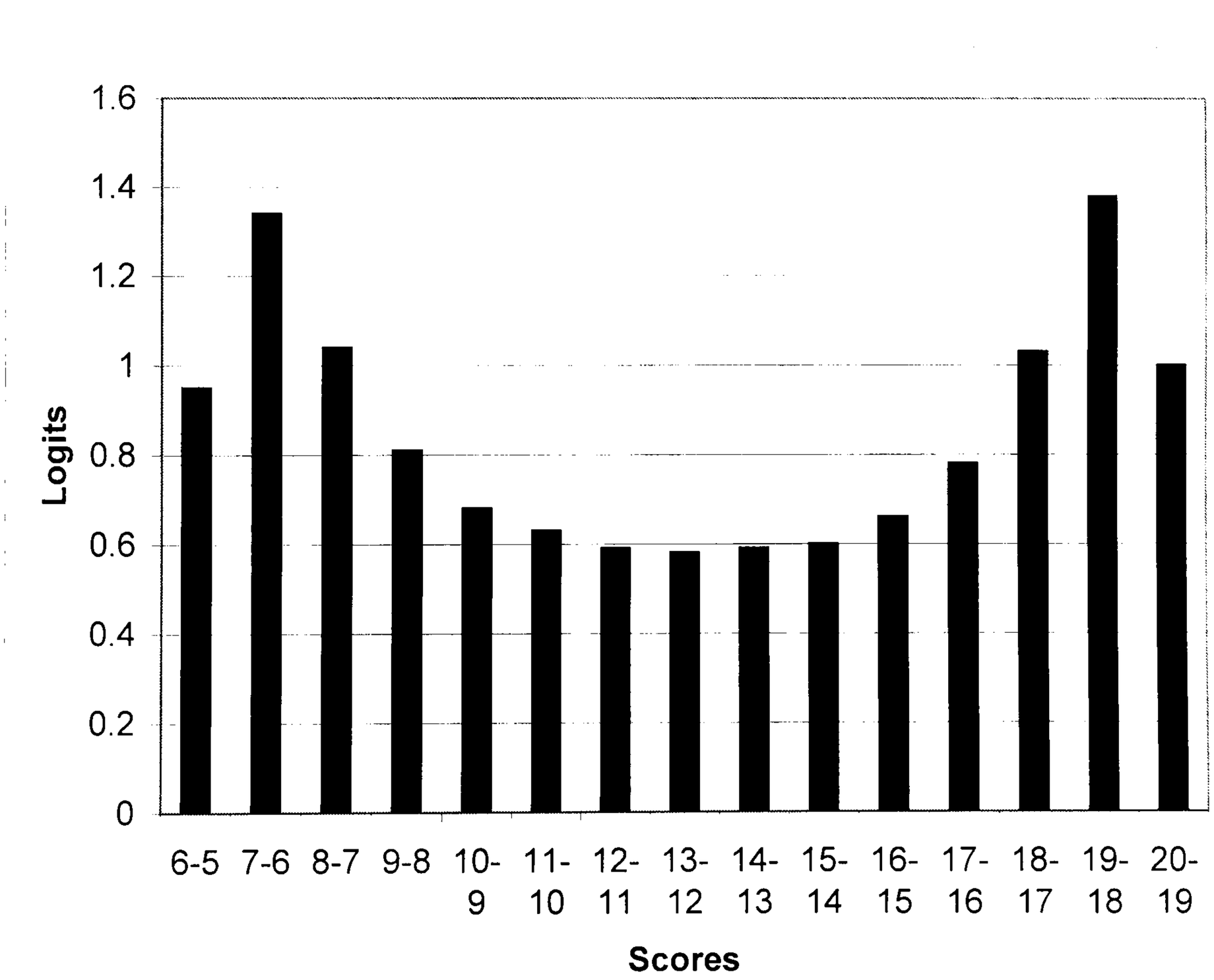
```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
+-------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT||STRUCTURE|CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ|| MEASURE | MEASURE|
|-------------------+------------+------------++---------+--------+
|  1    1    2699  46| -4.89 -4.91|  1.05  1.02||  NONE   |( -3.98)| 1
|  2    2    2120  36| -1.67 -1.63|   .92  1.35||   -2.86 |  -1.14 | 2
|  3    3     689  12|   .68   .63|   .92  1.06||     .66 |   1.44 | 3
|  4    4     387   7|  2.67  2.66|   .98  1.01||    2.20 |(  3.44)| 4
+-------------------------------------------------------------------
```

The person measures for the Physical Functioning scale are shown in Table 6.2.5 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 6.2.4.

**Table 6.2.5. Person measures for the Physical Functioning Scale**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| **5** | -6.31 | 1.00 | 0.00 | 1.00 | 0.00 |
| **6** | -5.36 | 0.91 | -0.13 | 0.46 | -0.16 |
| **7** | -4.02 | 0.16 | -1.67 | 0.14 | -0.64 |
| **8** | -2.98 | 0.52 | -0.79 | 0.46 | -0.49 |
| **9** | -2.17 | 0.43 | -1.13 | 0.37 | -0.87 |
| **10** | -1.49 | 1.13 | 0.19 | 2.31 | 1.23 |
| **11** | -0.86 | 0.03 | -3.29 | 0.05 | -2.70 |
| **12** | -0.27 | 1.58 | 0.79 | 1.72 | 0.92 |
| **13** | 0.31 | 0.68 | -0.59 | 0.72 | -0.52 |
| **14** | 0.90 | 0.59 | -0.80 | 0.60 | -0.78 |
| **15** | 1.50 | 0.53 | -0.92 | 0.97 | -0.04 |
| **16** | 2.16 | 9.10 | 4.81 | 9.90 | 5.14 |
| **17** | 2.94 | 0.49 | -0.79 | 0.32 | -0.71 |
| **18** | 3.97 | 0.40 | -0.84 | 0.22 | -0.55 |
| **19** | 5.35 | 2.70 | 1.14 | 3.15 | 0.33 |
| **20** | 6.35 | 1.00 | 0.00 | 1.00 | 0.00 |

**Figure 6.2.4 Difference between adjacent raw scores for the Physical Functioning Scale**



As can be seen from Figure 6.2.4 the differences between adjacent scores are roughly equal between the range of 10 to 15, however at either extremes the differences increase.

The summary of person measures is shown in Table 6.2.6. The person separation index is poor at 1.65, indicating that the scale is only able to detect fewer than 2 levels of person ability. The reliability measure however is good at 0.73.

**Table 6.2.6. Summary of Person Measures for the Physical Functioning Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 9.0 | 5.0 | -2.59 | .97 | .93 | -.4 | 1.02 | -.2 |
| S.D. | 2.8 | .0 | 2.12 | .18 | .88 | 1.1 | 1.38 | .9 |
| MAX. | 19.0 | 5.0 | 5.35 | 1.28 | 9.10 | 4.8 | 9.90 | 5.1 |
| MIN. | 6.0 | 5.0 | -5.36 | .76 | .03 | -3.3 | .05 | -2.7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | 1.11 | ADJ.SD | 1.81 | SEPARATION | 1.64 | PATS | RELIABILITY | .73 |
| MODEL RMSE | .99 | ADJ.SD | 1.88 | SEPARATION | 1.90 | PATS | RELIABILITY | .78 |
| S.E. OF PATS | MEAN | .06 | | | | | | |
| WITH 479 EXTREME PATSS | = | 1658 | PATSS | MEAN | -3.64 | S.D. | 2.50 | |
| REAL RMSE | 1.26 | ADJ.SD | 2.16 | SEPARATION | 1.72 | PATS | RELIABILITY | .75 |
| MODEL RMSE | 1.19 | ADJ.SD | 2.20 | SEPARATION | 1.86 | PATS | RELIABILITY | .77 |

MAXIMUM EXTREME SCORE: 4 PATSS
MINIMUM EXTREME SCORE: 475 PATSS

**Figure 6.2.5 Test Information Curve for the Physical Functioning Scale**

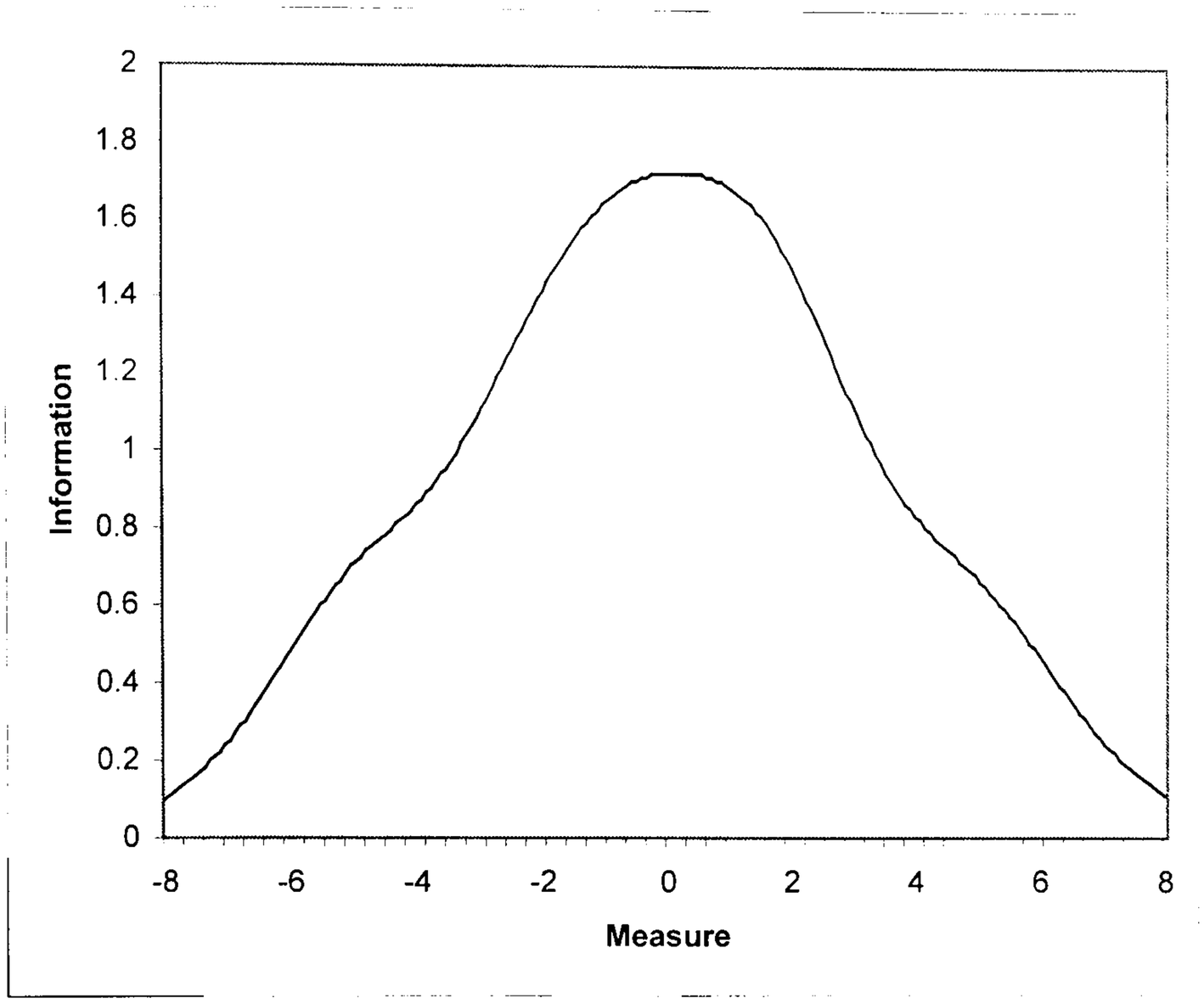



Figure 6.2.5 shows the test information curve for the Physical Functioning scale and demonstrates that the scale provides the most information at the centre of the scale (ability measures around zero).

Since one of the requirements of the Rasch model is that there should be item estimate invariance (e.g. Wright & Masters, 1982) across different samples of persons, in addition to the initial Rasch analysis of the scale, the unidimensionality of the Physical Functioning scales was further investigated using a differential item functioning analysis.

The results of the differential item analysis can be seen in Table 6.2.7, which shows the analysis between samples of male and female patients.

**Table 6.2.7 Differential Item Analysis of the Physical Functioning Scale**

| PATS GROUP | DIF MEASURE | DIF S.E. | PATS GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | QUES Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -1.96 | .11 | 2 | -2.41 | .08 | .45 | .14 | 3.31 | INF | 1 | PF1 |
| 1 | -1.80 | .11 | 2 | -2.05 | .08 | .25 | .14 | 1.84 | INF | 2 | PF2 |
| 1 | .35 | .13 | 2 | .43 | .10 | -.08 | .16 | -.49 | INF | 3 | PF3 |
| 1 | -.15 | .13 | 2 | .44 | .10 | -.59 | .16 | -3.74 | INF | 4 | PF4 |
| 1 | 2.72 | .21 | 2 | 3.65 | .17 | -.93 | .27 | -3.47 | INF | 5 | PF5 |

The criterion used for this analysis was a p-value of 0.05 (two-tailed), however in order to control for multiple testing a Bonferroni correction was applied (0.05/5), therefore the new statistical criterion was 0.01 significance evaluated against Student t value for infinity at 2.56. In addition to evaluating differential items functioning against statistical significance items were only considered to be exhibiting bias if the difference between the groups exceeded 0.5 logits (as suggested by Wright and Panchapakesan, 1969).

There were significant differences between male (group 1) and female patients (group 2) in response to question PF1, “Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?”, which females found easier to endorse. In addition, item 4 (“Do you have to stay in a bed or a chair for most of the day?”) also exhibited bias with men finding the item easier to endorse. However, despite the statistical significance neither difference between the groups for each item exceeded 0.5, although the contrast for item 4 is close to this criterion. The contrast between groups for item 5 was both statistically significant and greater than 0.50 logits. It can therefore be concluded that item 5, which had demonstrated poor fit also exhibited differential bias (both groups found this item hard to endorse).

In summary, a Rasch analysis of the Physical Functioning scale of the EORTC QLQ-C30 demonstrated poor fit for two of the five items from the scale (PF2, and PF5) with one of those items (PF5) not fitting the Rasch model. Furthermore, PF5 also exhibited differential item bias. It can be concluded from this that the Physical Functioning scale is a unidimensional scale (with the exception of PF5).

Additionally, the results from the person measures demonstrated that the scale was not interval based with equally spaced scores between the mid-range (10 – 15), but large, unequal differences at both extremes.

*6.2.4. Results for Emotional Functioning*

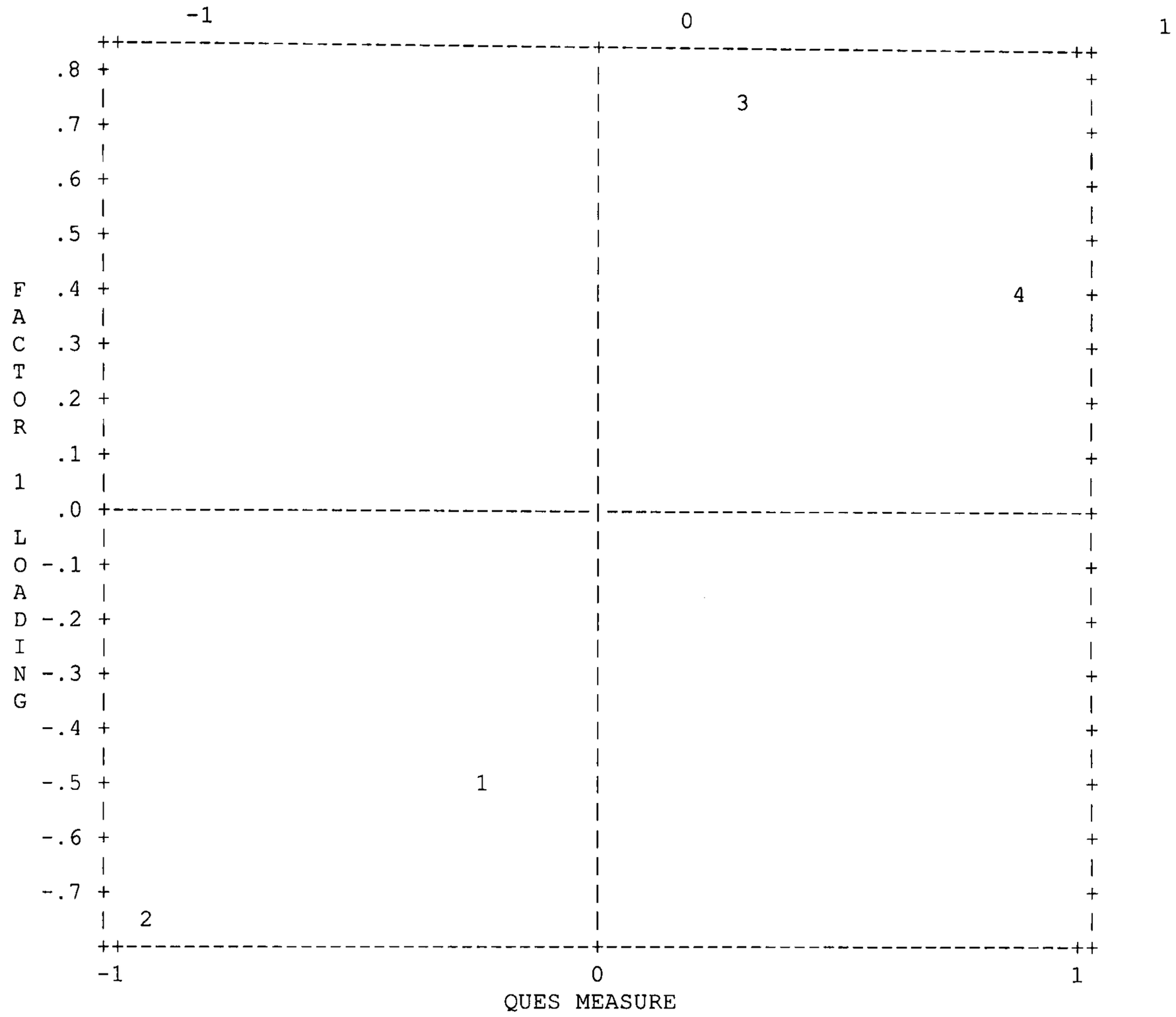
The location measures and fit statistics are given in Table 6.2.8 for the Emotional Functioning scale. The results of the Rasch analysis of this scale demonstrate a good level of fit for all items, i.e. infit mean square statistics within a range of 0.70 – 1.30 (e.g. Wright et al., 1994).

**Table 6.2.8 Unidimensionality measures for Emotional Functioning**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | | QUESS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2438 | 1306 | .29 | .06 | 1.11 | 2.8 | 1.11 | 2.4 | A | .82 | EF3 |
| 4 | 2283 | 1306 | .88 | .06 | 1.04 | 1.1 | 1.01 | .2 | B | .83 | EF4 |
| 2 | 2781 | 1306 | -.94 | .06 | .93 | -1.8 | .90 | -2.2 | b | .86 | EF2 |
| 1 | 2583 | 1306 | -.24 | .06 | .88 | -3.3 | .86 | -3.3 | a | .86 | EF1 |
| MEAN | 2521. | 1306. | .00 | .06 | .99 | -.3 | .97 | -.7 | | | |
| S.D. | 184. | 0. | .67 | .00 | .09 | 2.4 | .10 | 2.2 | | | |

Figure 6.2.6. shows the factor plot of the principal components analysis of the standardised residuals for the Emotional Functioning scale. A total of 1.5 eigenvalues were extracted from the residuals, indicating that there were no other factor structures in the residuals (Smith & Miao, 1994).

**Figure 6.2.6. Principal Components (Standardized Residual) Factor Plot of the Emotional Functioning Scale**




The factor loadings from the principal components analysis and coding for Figure 6.2.6 can be seen in Table 6.2.9.

**Table 6.2.9. Factor Loadings from the Principal Components Analysis of the Emotional Functioning Scale**

| FACTOR | LOADING | MEASURE | INFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | QUES | |
|---|---|---|---|---|---|---|---|
| 1 | .76 | .29 | 1.11 | 1.11 | 3 | 3 | EF3 |
| 1 | .41 | .88 | 1.04 | 1.01 | 4 | 4 | EF4 |
| 1 | -.74 | -.94 | .93 | .90 | 2 | 2 | EF2 |
| 1 | -.48 | -.24 | .88 | .86 | 1 | 1 | EF1 |

The item map for the Emotional Functioning Scale can be seen in Figure 6.2.7. This figure demonstrates that the four items on this scale are reasonably well spaced with a distance of approximately 0.50 logits between each item. However, the items are situated close the centre of the scale covering a small range of person abilities between –1.0 to +1.0.

The category probability curve for Emotional Functioning is shown in Figure 6.2.8.

**Figure 6.2.7. Item Map for the Emotional Functioning Scale**

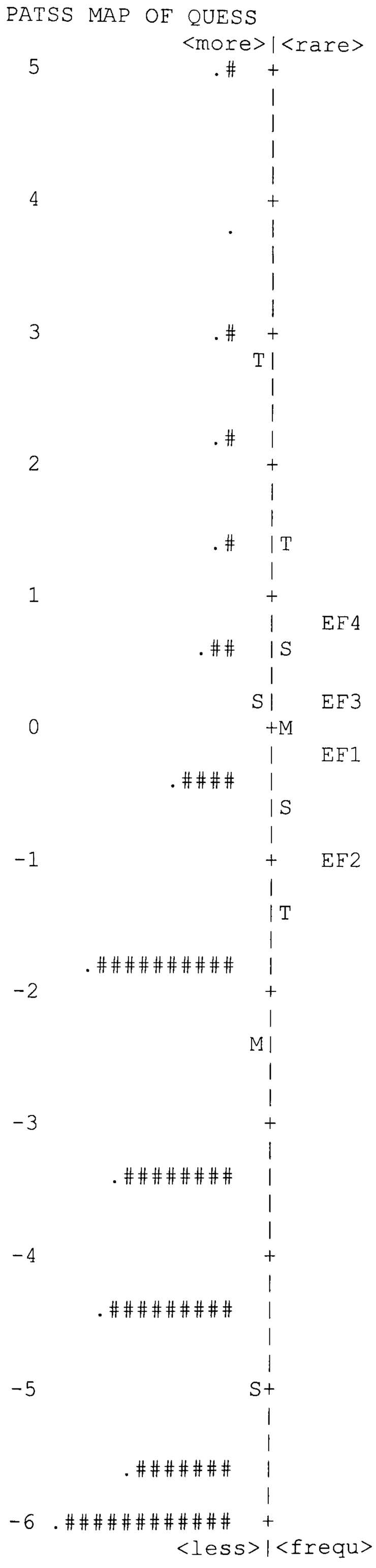
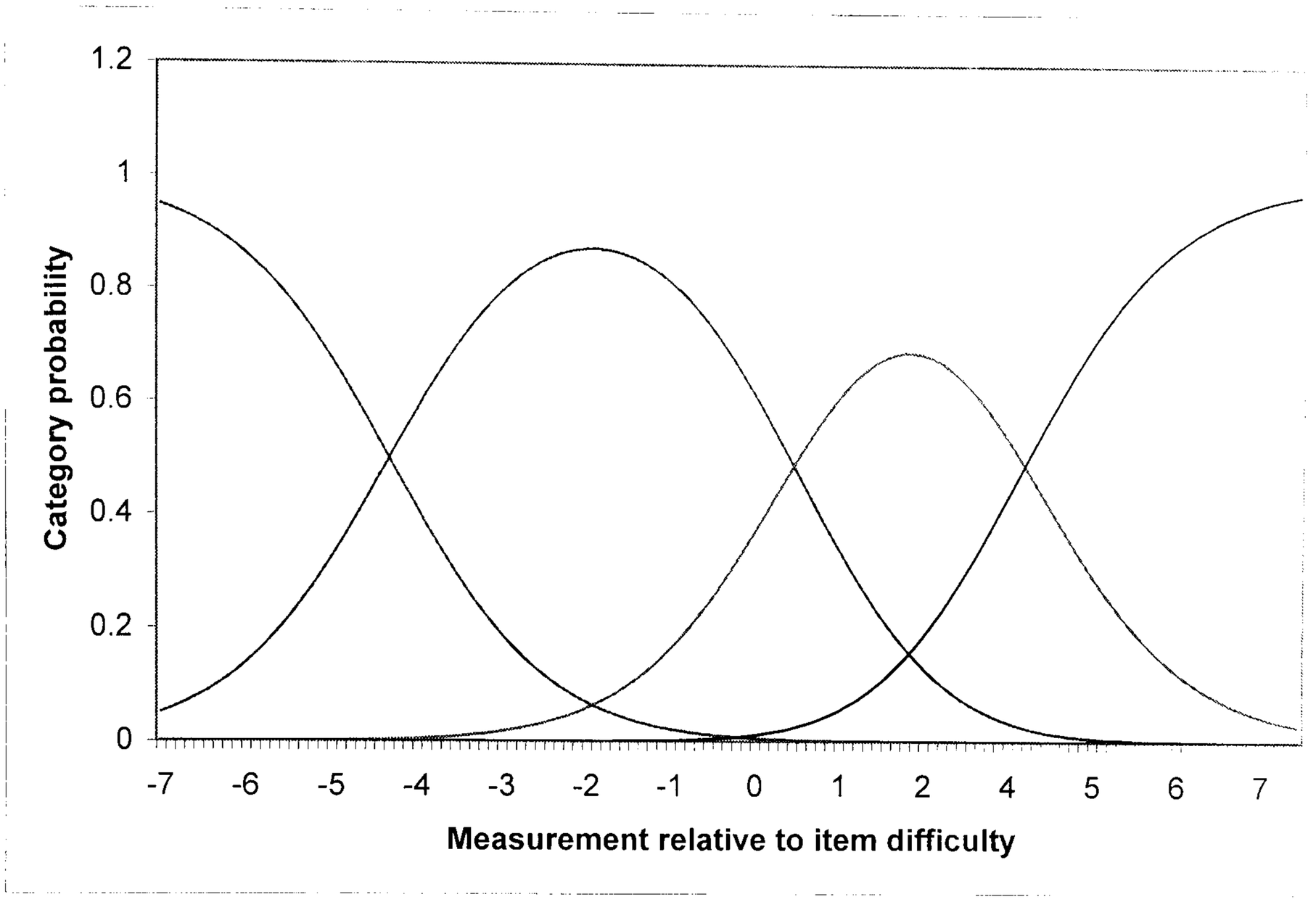


**Figure 6.2.8 Category Probability Curve for Emotional Functioning**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Black = category 4.

It can be seen for instance, from figure 6.2.8 that as a patient's emotional functioning declines, i.e. as the person measure increases in relation to the item difficulty, that the likelihood of the patient responding to items with "Not at all" (category 1) or "A little" (category 2 decreases, and the likelihood of responses such as "Quite a bit" (category 3) or "Very much" (category 4) increases.

Table 6.2.10 shows a summary of the items for the Emotional Functioning Scale. The item separation index is approximately 11, demonstrating that the scale can distinguish between 11 levels of difficulty.

**Table 6.2.10. Summary of Items from the Emotional Functioning Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 2521.3 | 1306.0 | .00 | .06 | .99 | -.3 | .97 | -.7 |
| S.D. | 183.7 | .0 | .67 | .00 | .09 | 2.4 | .10 | 2.2 |
| MAX. | 2781.0 | 1306.0 | .88 | .06 | 1.11 | 2.8 | 1.11 | 2.4 |
| MIN. | 2283.0 | 1306.0 | -.94 | .06 | .88 | -3.3 | .86 | -3.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| REAL RMSE | .06 | ADJ.SD | .67 | SEPARATION 10.79 | QUES | RELIABILITY | .99 |
| MODEL RMSE | .06 | ADJ.SD | .67 | SEPARATION 11.00 | QUES | RELIABILITY | .99 |
| S.E. OF QUES MEAN = .39 | | | | | | | |

Table 6.2.11 shows a summary of the category measures for the Emotional Functioning Scale. It can be seen that there is a good level of separation between the categories with a distance at least 1.4 logits (Linacre, 1999a) between each threshold (structure measure), and the average measures increase monotonically across the rating scale (category measure).

**Table 6.2.11. Summary of Category Measures for the Emotional Functioning Scale**

SUMMARY OF CATEGORY STRUCTURE. Model="R"

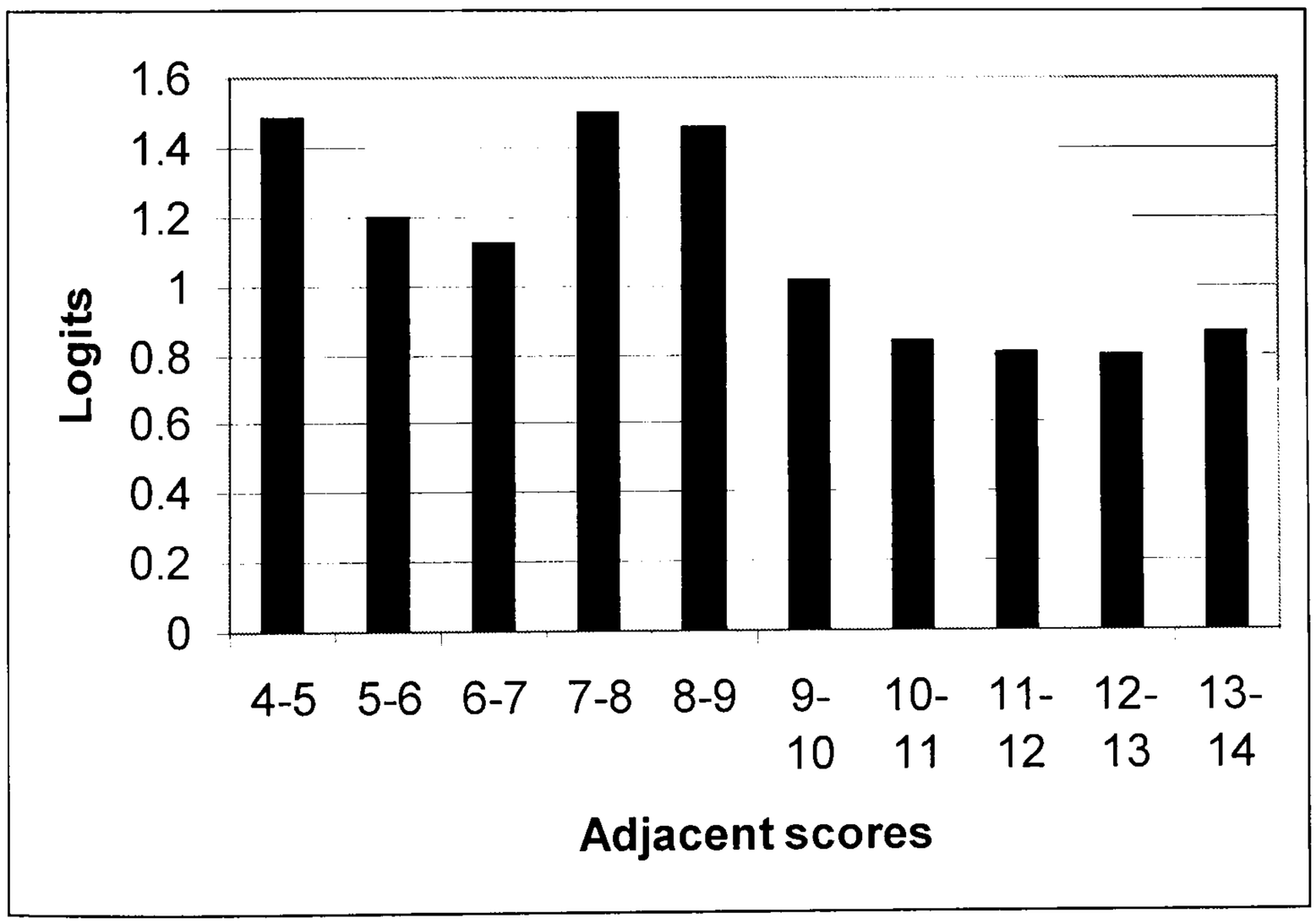
| CATEGORY LABEL | SCORE | OBSERVED COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1434 | 27 | -4.83 | -4.82 | 1.02 | .97 | NONE | ( -5.58) | 1 |
| 2 | 2 | 2903 | 56 | -2.43 | -2.42 | .95 | .95 | -4.48 | -1.87 | 2 |
| 3 | 3 | 703 | 13 | 1.15 | 1.05 | .92 | .88 | .75 | 2.24 | 3 |
| 4 | 4 | 184 | 4 | 3.40 | 3.57 | 1.19 | 1.39 | 3.73 | ( 4.86) | 4 |

The person measures for the Emotional Functioning scale are shown in Table 6.2.12 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 6.2.9.

**Table 6.2.12. Person measures for the Emotional Functioning Scale**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.Z |
|---|---|---|---|---|---|
| 4 | -7.17 | 1 | 0 | 1 | 0 |
| 5 | -5.69 | 1.09 | 0.14 | 0.93 | -0.09 |
| 6 | -4.49 | 1.16 | 0.39 | 1.18 | 0.42 |
| 7 | -3.37 | 0.55 | -0.84 | 0.47 | -0.88 |
| 8 | -1.87 | 2.61 | 0.99 | 2.37 | 0.86 |
| 9 | -0.41 | 1.16 | 0.2 | 1.22 | 0.23 |
| 10 | 0.6 | 1.09 | 0.15 | 1.04 | 0.06 |
| 11 | 1.44 | 0.31 | -1.53 | 0.31 | -1.51 |
| 12 | 2.24 | 0.14 | -2.13 | 0.14 | -2.14 |
| 13 | 3.04 | 0.9 | -0.16 | 0.9 | -0.16 |
| 14 | 3.9 | 2.62 | 1.93 | 2.38 | 1.68 |
| 16 | 6.43 | 1 | 0 | 1 | 0 |

**Figure 6.2.9 Difference between adjacent raw scores for the Emotional Functioning Scale**




*Scores beyond 14 were not estimated therefore not included

As can be seen from Figure 6.2.9 the differences between adjacent scores are roughly equal between the range of 10 to 14 (roughly 0.80 logits). However below this score this difference increases to between 1.20 and 1.40. It appears therefore

that there is a threshold for this scale around the raw score of 10, or a scale score of 50 with differences between raw scores above this (and consequently scale scores below this since the scale is scored negatively) being smaller than differences between raw scores below this score (and scale scores above this).

The summary of person measures is shown in Table 6.2.13. The person separation index is slightly better than that for the Physical Functioning scale, yet still poor at 1.74, indicating that this scale is also only able to detect fewer than 2 levels of person ability. The reliability measure however is good at 0.75.

**Table 6.2.13. Summary of Person Measures for the Emotional Functioning Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 7.7 | 4.0 | -2.40 | 1.12 | .97 | -.4 | .97 | -.4 |
| S.D. | 2.3 | .0 | 2.60 | .13 | 1.01 | 1.1 | 1.06 | 1.1 |
| MAX. | 15.0 | 4.0 | 4.98 | 1.31 | 9.90 | 4.3 | 9.90 | 4.2 |
| MIN. | 5.0 | 4.0 | -5.69 | .89 | .06 | -2.1 | .05 | -2.1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | 1.30 | ADJ.SD | 2.25 | SEPARATION | 1.74 | PATS | RELIABILITY | .75 |
| MODEL RMSE | 1.13 | ADJ.SD | 2.34 | SEPARATION | 2.08 | PATS | RELIABILITY | .81 |

S.E. OF PATS MEAN = .07

| | | |
|---|---|---|
| MAXIMUM EXTREME SCORE: | 13 | PATSS |
| MINIMUM EXTREME SCORE: | 336 | PATSS |
| LACKING RESPONSES: | 3 | PATSS |

**Figure 6.2.10 Test Information Curve for the Emotional Functioning Scale**

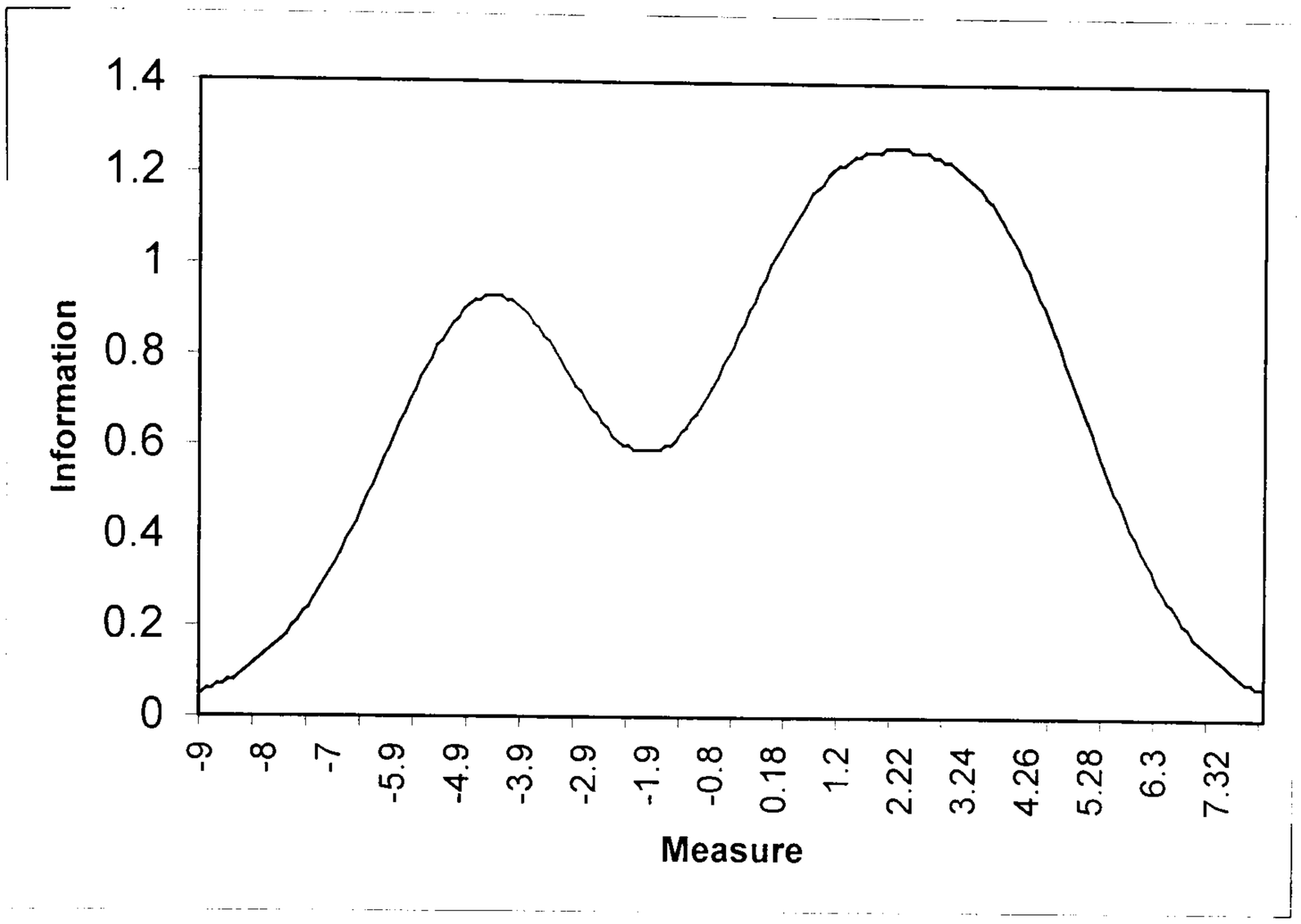



Figure 6.2.10 shows the test information curve for the Emotional Functioning scale and demonstrates that the scale provides the most information at two points, namely –4.60 and +2.05 relative to the ability measures.

In addition to the initial Rasch analysis of the scale, the unidimensionality of the Emotional Functioning scales was further investigated using a differential item functioning analysis.

The results of the differential item analysis can be seen in Table 6.3.7, which shows the analysis between samples of male (group 1) and female (group 2) patients.

**Table 6.2.14. Differential Item Analysis of the Emotional Functioning Scale**

| PATS GROUP | DIF MEASURE | DIF S.E. | PATS GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | QUES Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -.16 | .13 | 2 | -.15 | .10 | -.01 | .16 | -.06 | INF | 1 | EF1 |
| 1 | -.94 | .13 | 2 | -.91 | .09 | -.02 | .16 | -.15 | INF | 2 | EF2 |
| 1 | .28 | .13 | 2 | .35 | .10 | -.07 | .16 | -.45 | INF | 3 | EF3 |
| 1 | .83 | .13 | 2 | .71 | .10 | .12 | .17 | .70 | INF | 4 | EF4 |

As with the analysis of the Physical Functioning scale the criterion used for this analysis was modified in order to control for multiple testing using the Bonferroni correction (0.05/5). The new p-value was 0.01 evaluated against Student t value for infinity at 2.56. As well with the differential item analysis for the Physical Functioning scale the criterion of contrast between groups greater than 0.50 was also evaluated to identify item bias.

There were no significant differences between male (group 1) and female patients (group 2) for any of the items from the Emotional Functioning scale.

In summary, a Rasch analysis of the Emotional Functioning scale of the EORTC QLQ-C30 demonstrated good fit for all of the scale items, unlike the Physical Functioning scale. Furthermore, none of the items exhibited differential item bias. It can therefore be concluded that the Emotional Functioning scale is a unidimensional scale. In addition, the results from the person measures demonstrated that the scale was interval based with a threshold corresponding to the scale score of 50. The difference between adjacent scale scores below this point is smaller than the difference for scores above it. This demonstrates that relatively smaller differences in person ability are required to move between scale scores below this point than above it.

*6.2.5. Results for Fatigue*

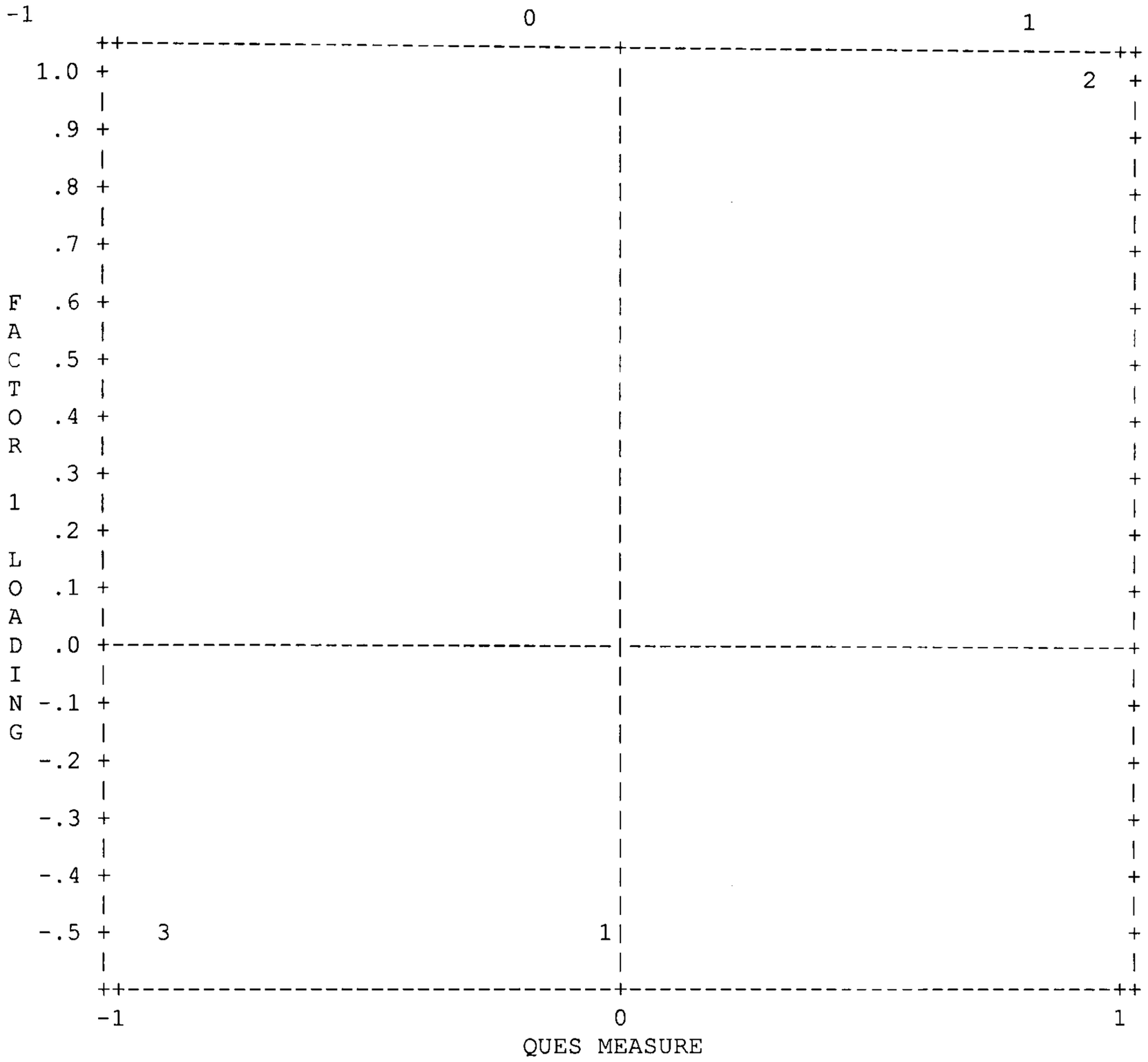
The location measures and fit statistics are given in Table 6.2.15 for the Fatigue scale. The results of the Rasch analysis of this scale demonstrate a good level of fit for all items, i.e. infit mean square statistics within a range of 0.70 – 1.30 (e.g. Wright et al., 1994).

**Table 6.2.15. Unidimensionality measures for Fatigue**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | | QUESS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 2645 | 1340 | .94 | .06 | 1.04 | 1.1 | 1.00 | .1 | A | .89 | FA2 |
| 1 | 2897 | 1340 | -.02 | .06 | .99 | -.2 | .94 | -1.2 | B | .89 | FA1 |
| 3 | 3136 | 1339 | -.92 | .06 | .93 | -1.9 | .89 | -2.6 | a | .88 | FA3 |
| MEAN | 2893. | 1340. | .00 | .06 | .99 | -.4 | .94 | -1.2 | | | |
| S.D. | 200. | 1. | .76 | .00 | .05 | 1.2 | .05 | 1.1 | | | |

Figure 6.2.11. shows the factor plot of the principal components analysis of the standardised residuals for the Fatigue scale. A total of 1.5 eigenvalues were extracted from the residuals, indicating that there were no other factor structures in the residuals (Smith & Miao, 1994).

**Figure 6.2.11. Principal Components (Standardized Residual) Factor Plot of the Emotional Functioning Scale**




The factor loadings from the principal components analysis and coding for Figure 6.2.11. can be seen in Table 6.2.16.

**Table 6.2.16. Factor Loadings from the Principal Components Analysis of the Fatigue Scale**

| FACTOR | LOADING | MEASURE | INFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | QUES | |
|---|---|---|---|---|---|---|---|
| 1 | 1.00 | .94 | 1.04 | 1.00 | 2 | 2 | FA2 |
| 1 | -.52 | -.02 | .99 | .94 | 1 | 1 | FA1 |
| 1 | -.50 | -.92 | .93 | .89 | 3 | 3 | FA3 |

The item map for the Fatigue Scale can be seen in Figure 6.2.12. This figure demonstrates that the three items on this scale are reasonably well spaced with a distance of approximately 1.00 logits between each item. However, the items are situated close the centre of the scale covering a small range of person abilities between –1.0 to +1.0.

The category probability curve for the Fatigue scale is shown in Figure 6.2.13.

**Figure 6.2.12. Item Map for the Fatigue Scale**

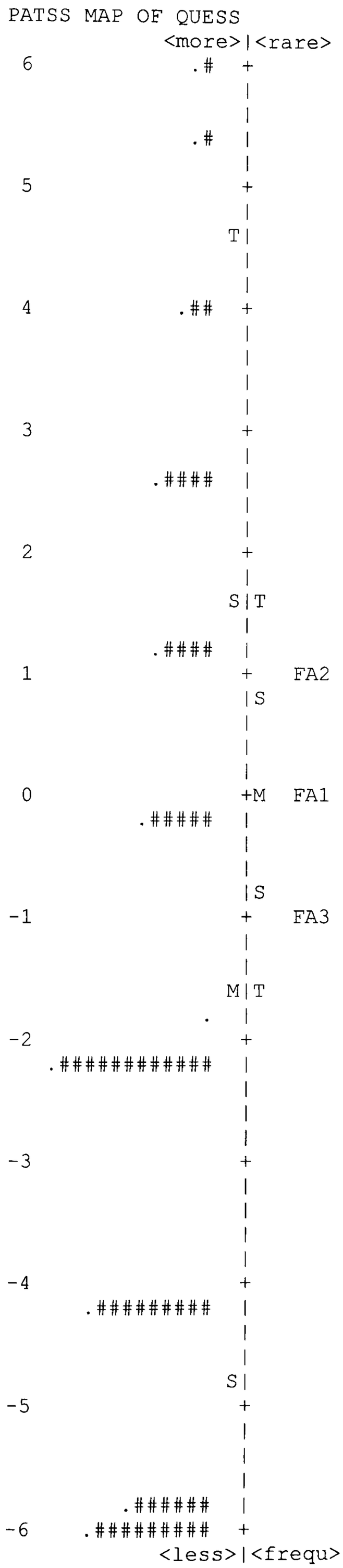
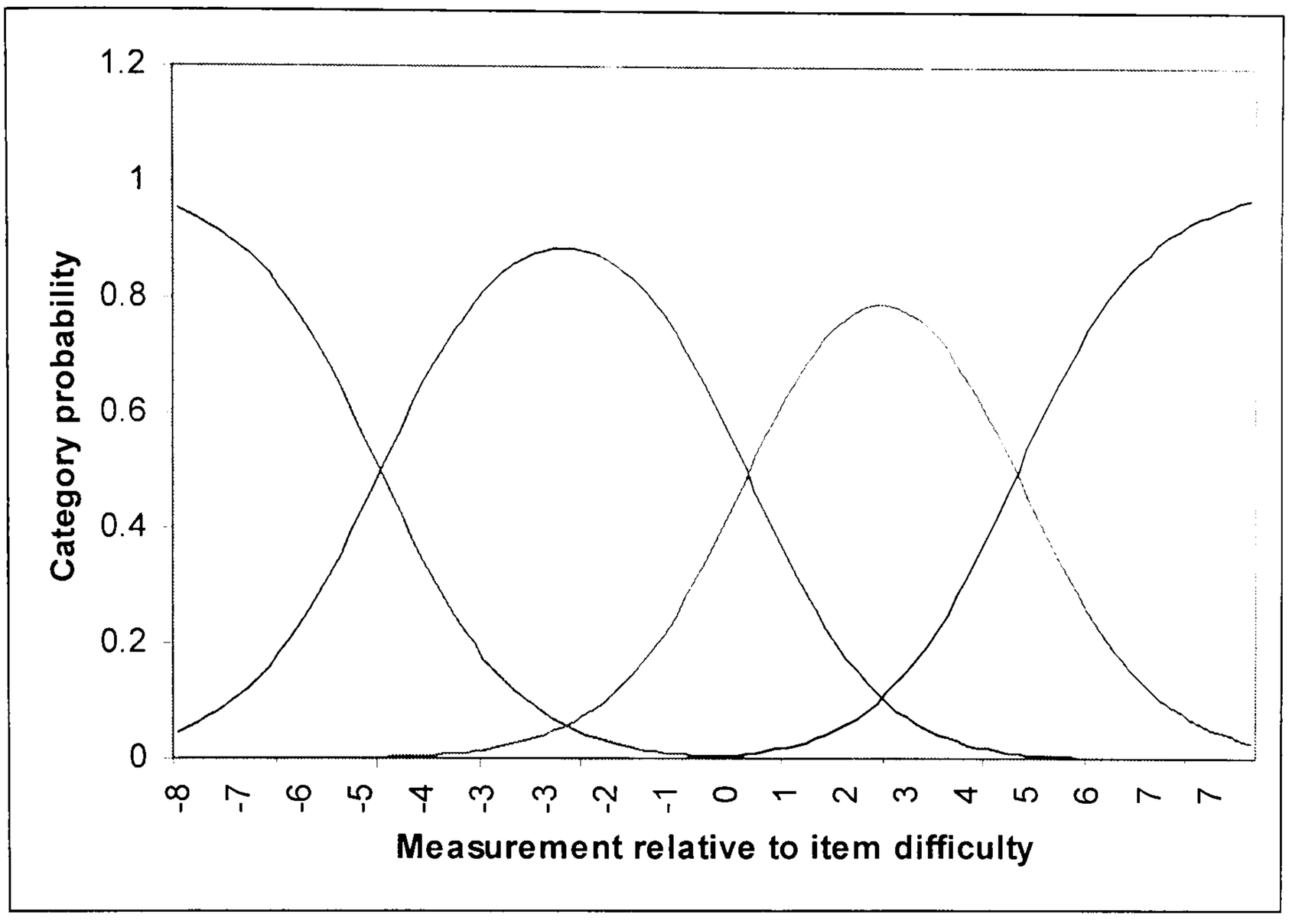


**Figure 6.2.13. Category Probability Curve for Emotional Functioning**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Black = category 4.

It can be seen for instance, from figure 6.2.13 that as a patient's fatigue declines, i.e. as the person measure increases in relation to the item difficulty, that the likelihood of the patient responding to items with "Not at all" (category 1) or "A little" (category 2 decreases, and the likelihood of responses such as "Quite a bit" (category 3) or "Very much" (category 4) increases.

Table 6.2.17 shows a summary of the items for the Fatigue Scale. The item separation index is approximately 12, demonstrating that the scale can distinguish between 12 levels of difficulty.

**Table 6.2.17. Summary of Items from the Fatigue Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 2892.7 | 1339.7 | .00 | .06 | .99 | -.4 | .94 | -1.2 |
| S.D. | 200.5 | .6 | .76 | .00 | .05 | 1.2 | .05 | 1.1 |
| MAX. | 3136.0 | 1340.0 | .94 | .06 | 1.04 | 1.1 | 1.00 | .1 |
| MIN. | 2645.0 | 1339.0 | -.92 | .06 | .93 | -1.9 | .89 | -2.6 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| REAL RMSE | .06 | ADJ.SD | .76 | SEPARATION | 12.20 | QUES | RELIABILITY .99 |
| MODEL RMSE | .06 | ADJ.SD | .76 | SEPARATION | 12.29 | QUES | RELIABILITY .99 |
| S.E. OF QUES MEAN = .54 | | | | | | | |

Table 6.2.18 shows a summary of the category measures for the Fatigue Scale. It can be seen that there is a good level of separation between the categories with a distance at least 1.4 logits (Linacre, 1999a) between each threshold (structure measure), and the average measures increase monotonically across the rating scale (category measure).

**Table 6.2.18. Summary of Category Measures for the Fatigue Scale**

| CATEGORY LABEL | SCORE | OBSERVED COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 732 | 18 | -5.20 | -5.13 | .99 | .91 | NONE | ( -6.12) | 1 |
| 2 | 2 | 2139 | 53 | -2.39 | -2.40 | .90 | .87 | -5.02 | -2.26 | 2 |
| 3 | 3 | 924 | 23 | 1.72 | 1.64 | 1.00 | .97 | .49 | 2.51 | 3 |
| 4 | 4 | 224 | 6 | 4.12 | 4.38 | 1.19 | 1.22 | 4.53 | ( 5.64) | 4 |
| MISSING | | 1 | 0 | -.88 | | | | | | |

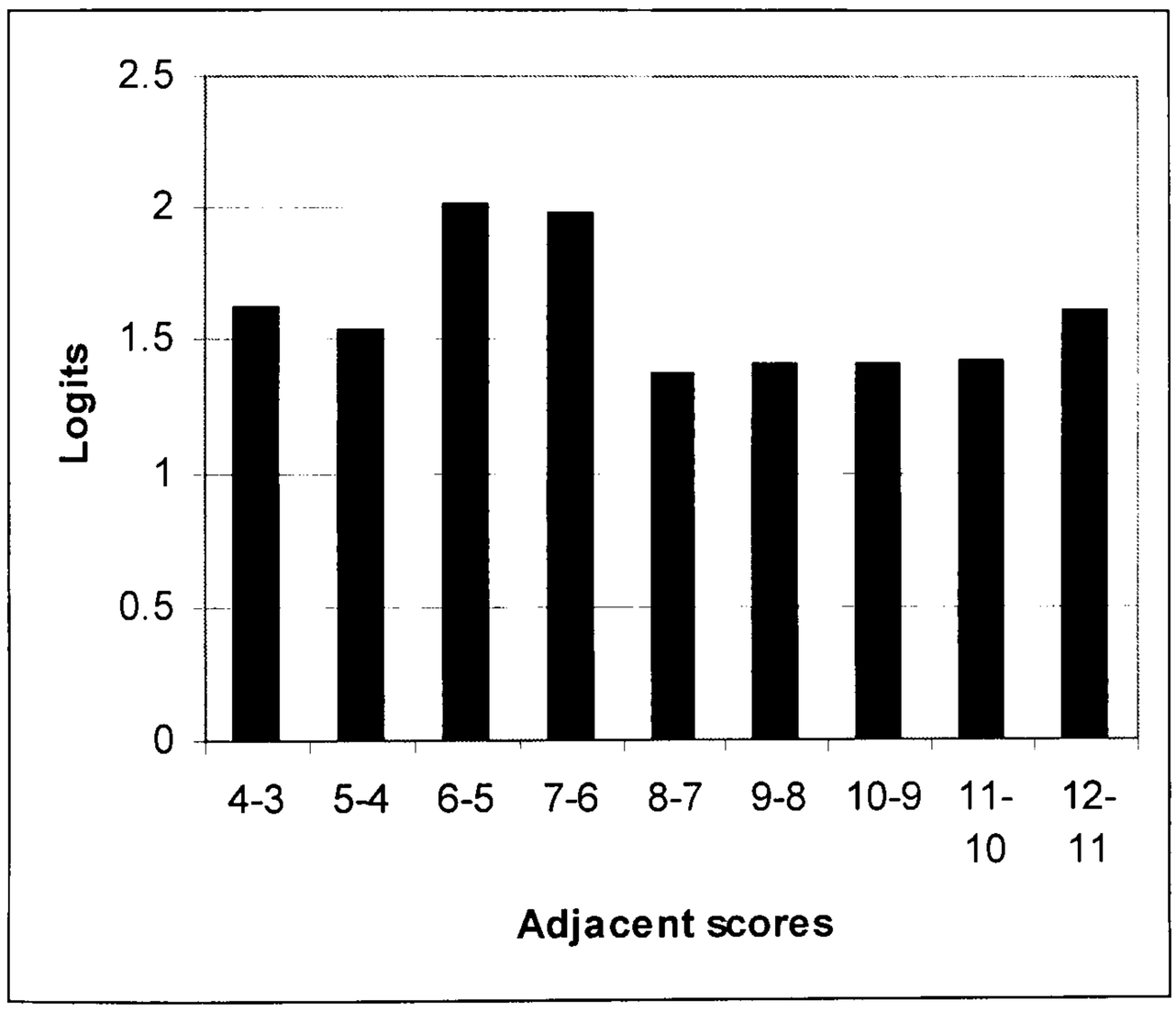
AVERAGE MEASURE is mean of measures in category.

The person measures for the Fatigue scale are shown in Table 6.2.19 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 6.2.14.

**Table 6.2.19. Person measures for the Fatigue Scale**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.Z |
|---|---|---|---|---|---|
| **3** | -7.43 | 1 | 0 | 1 | 0 |
| **4** | -5.81 | 0.56 | -0.96 | 0.5 | -0.86 |
| **5** | -4.27 | 1.26 | 0.39 | 1.14 | 0.18 |
| **6** | -2.26 | 0.07 | -1.32 | 0.06 | -1.31 |
| **7** | -0.28 | 1.76 | 0.9 | 2.15 | 1.08 |
| **8** | 1.1 | 0.39 | -1.13 | 0.37 | -1.12 |
| **9** | 2.51 | 0.12 | -1.62 | 0.12 | -1.62 |
| **10** | 3.92 | 0.41 | -1.11 | 0.4 | -1.09 |
| **11** | 5.34 | 1.3 | 0.44 | 1.17 | 0.2 |
| **12** | 6.95 | 1 | 0 | 1 | 0 |

**Figure 6.2.14. Difference between adjacent raw scores for the Fatigue Scale**




As can be seen from Figure 6.2.14 the differences between adjacent scores are roughly equal for all raw scores at roughly 1.5 logits, indicating that the Fatigue scale is interval based.

The summary of person measures is shown in Table 6.2.20. The person separation index is slightly better than that for the Physical Functioning scale, yet still poor at 1.70 similar to the level for the Emotional Functioning scale, indicating that

this scale is also only able to detect fewer than 2 levels of person ability. The reliability measure however is good at 0.74.

**Table 6.2.20. Summary of Person Measures for the Fatigue Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 6.5 | 3.0 | -1.59 | 1.33 | .94 | -.4 | .94 | -.4 |
| S.D. | 1.9 | .0 | 3.11 | .16 | 1.19 | 1.1 | 1.24 | 1.1 |
| MAX. | 11.0 | 3.0 | 5.34 | 2.03 | 9.90 | 4.7 | 9.90 | 4.4 |
| MIN. | 4.0 | 2.0 | -5.81 | 1.16 | .03 | -1.6 | .03 | -1.6 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | 1.57 | ADJ.SD | 2.68 | SEPARATION | 1.70 | PATS | RELIABILITY | .74 |
| MODEL RMSE | 1.34 | ADJ.SD | 2.81 | SEPARATION | 2.10 | PATS | RELIABILITY | .82 |

S.E. OF PATS MEAN = .09

MAXIMUM EXTREME SCORE: 52 PATSS
MINIMUM EXTREME SCORE: 264 PATSS
LACKING RESPONSES: 2 PATSS

**Figure 6.2.15. Test Information Curve for the Fatigue Scale**

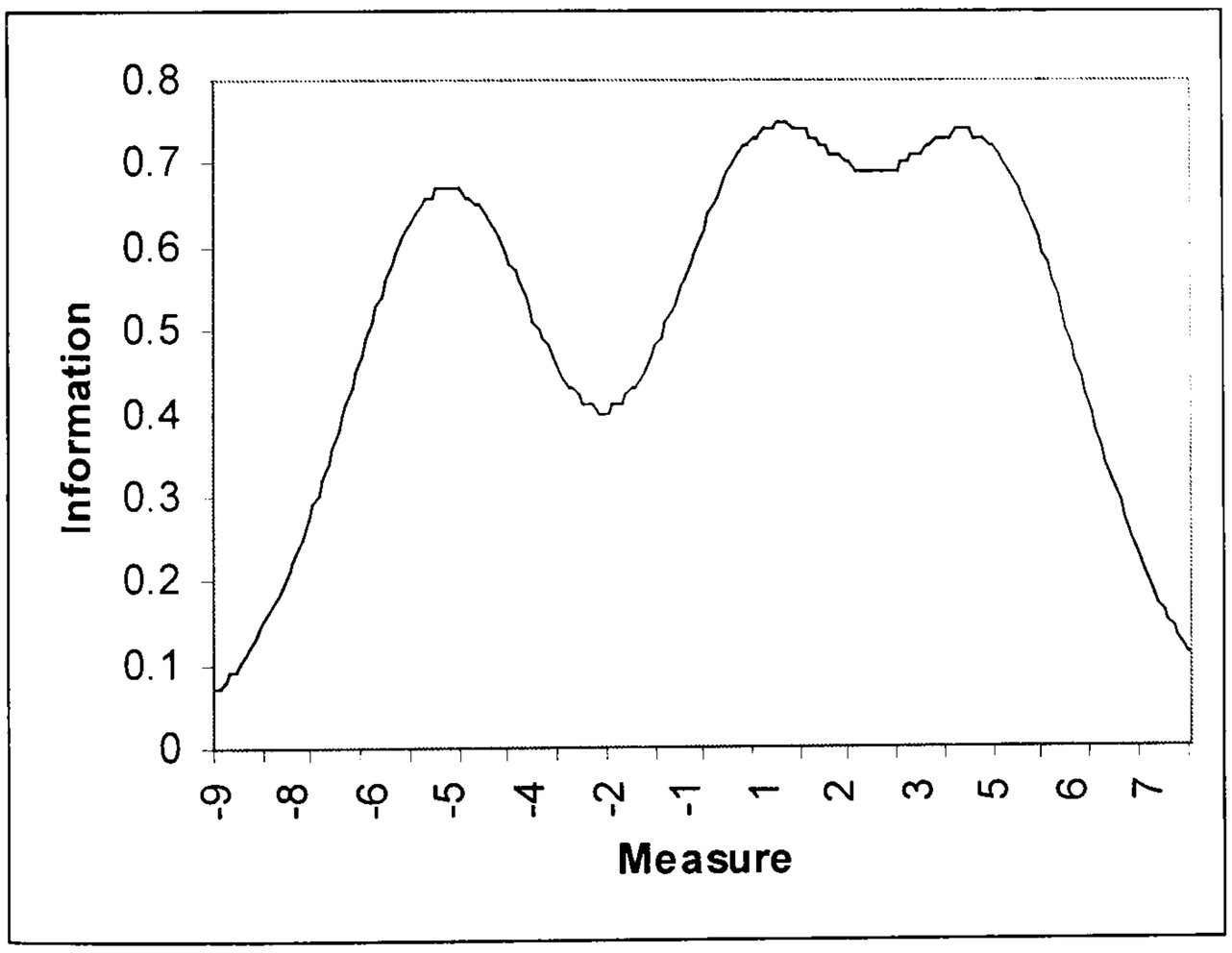



Figure 6.2.15 shows the test information curve for the Fatigue scale and demonstrates that the scale provides the most information at two points, namely – 5.50 and the range between +1.00 and +5.50 relative to the ability measures.

In addition to the initial Rasch analysis of the scale, the unidimensionality of the Fatigue scale was further investigated using a differential item functioning analysis.

The results of the differential item analysis can be seen in Table 6.4.7, which shows the analysis between samples of male (group 1) and female (group 2) patients.

**Table 6.2.21. Differential Item Analysis of the Fatigue Scale**

| PATS GROUP | DIF MEASURE | DIF S.E. | PATS GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | DIF S.E. | t | d.f. | JOINT Number | QUES Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .09 | .13 | 2 | -.18 | .10 | .28 | .16 | 1.72 | INF | 1 | FA1 |
| 1 | .52 | .13 | 2 | 1.06 | .10 | -.55 | .16 | -3.38 | INF | 2 | FA2 |
| 1 | -.61 | .13 | 2 | -.87 | .10 | .26 | .16 | 1.62 | INF | 3 | FA3 |

As with the analysis of the Physical Functioning scale the criterion used for this analysis was modified in order to control for multiple testing using the Bonferroni correction (0.05/5). The new p-value was 0.01 evaluated against Student t value for infinity at 2.56.

There were no significant differences between male (group 1) and female patients (group 2) for items FA1 and FA3, however a significant difference was observed for item FA2 (“Have you felt weak?“) with the males reporting fewer problems with Fatigue than females, although the contrast between groups was close enough to the criterion of 0.50 (e.g. Wright and Panchapakesan, 1969) to be ignored. It can therefore be concluded that none of the items from the Fatigue scale demonstrated differential item bias.

In summary, a Rasch analysis of the Fatigue scale of the EORTC QLQ-C30 demonstrated good fit for all of the scale items, unlike the Physical Functioning scale. However, one item (FA2) demonstrated differential item bias. It can therefore be concluded that the Fatigue scale is a unidimensional scale. In addition, the results from the person measures demonstrated that the scale was interval based with differences between the adjacent scores of around 1.50 logits. This demonstrates

that the steps between scores are equal and that the same level of ability is required to move between scores.

## 6.3. Discussion

This study described the factor analysis of the entire EORTC QLQ-C30 instrument and the Rasch analysis of three scales from the EORTC QLQ-C30. The results of the traditional psychometrics demonstrated levels of internal consistency for the scales equivalent to previous studies (e.g. Aaronson et al., 1993), although the factor analysis revealed a four factor structure differing from the conceptual structure, namely a physical functioning scale, a role and social functioning and pain and fatigue scale, a cognitive and emotional functioning scale, and a scale consisting of the symptom scales. There have been no other studies published to date on the factor analysis of the entire EORTC QLQ-C30. However, McLachlan et al (1999) did identify a two factor structure from an exploratory factor analysis (of scores from metastatic breast cancer patients) of four of the EORTC QLQ-C30 scales (Emotional, Role, Cognitive and Social Functioning) corresponding to "emotional distress", i.e. Emotional and Cognitive Functioning, and "functional ability", i.e. Role and Social Functioning. These results support the findings from the factor analysis in this study. However, these results contradict work by Ringdal and colleagues (Ringdal and Ringdal, 1993; Ringdal et al., 1999), using Mokken scales, which suggests that the Physical and Role functioning scales could be combined to form a unified scale.

The results of the Rasch analysis demonstrated that both the Emotional Functioning and Fatigue scales of the EORTC QLQ-C30 were unidimensional scales with all items exhibiting good fit. In addition, all of the items from the Emotional Functional scale and the majority of those from the Fatigue scale demonstrated no differential item bias. A series of recent studies (Groenvold, Petersen, and Bjorner, 2000; Petersen, Groenvold, & Bjorner, 2000) have used item-response theory analyses (e.g. Samejima, 1969) to identify items from both the Emotional Functioning

and Fatigue scales which provided the least information regarding the latent traits (i.e. emotional well-being and fatigue). On the basis of this analysis a single item was identified from each scale, namely item 3 from the Emotional Functioning scale ("Did you feel irritable?") and item 1 ("Did you need a rest?") from the Fatigue scale (Groenvold et al., 2000; Petersen et al., 2000). A further item (item 2, "Did you worry?") was removed in a subsequent analysis of the Emotional Functioning scale (Petersen et al., 2000). Both the 2- and 3-item Emotional Functioning scale and 2-item Fatigue performed well at predicting their respective full scale scores. In addition, Ringdal et al. (1999) also found that item 3 of the Emotional Functioning scale ("Did you feel irritable?") had the lowest corrected item-scale correlation score. The results of the item response theory analysis differ from the results of this study, but can however be explained by the different methodologies with the latter allowing for different item discrimination, whereas Rasch models assume item discrimination between items is equal, allowing for item and person estimates to be placed along the same metric (e.g. Wright and Masters, 1982). Clearly item-response based analyses allow for items to be removed on the basis of providing less information regarding the latent trait, although as Petersen et al (2000b) conceded, for the Fatigue scale at least there was "no obvious choice of items for a shortened scale" based on the information functions (Petersen et al., 2000, p. 9). Furthermore, as Ringdal et al (1999), have pointed out in order to measure "complex psychological phenomena" with just two items requires that the items are strongly correlated which could create problems in respect of content validity (Ringdal et al., 1999, p. 42).

The results of the Rasch analysis of the Physical Functioning scale differed slightly from the analysis of the other two scales. Two out of five of the items (PF2, and PF5) from this scale demonstrated poor fit statistics, although only one of these items did not fit the Rasch model (PF5, "Do you need help with eating, dressing, washing yourself or using the toilet?"). This item also demonstrated significant differential item bias. This result is similar to the findings by Ringdal et al (1999) who

also identified this item as having the lowest corrected item-scale correlation in comparison with the other items from this scale.

It can therefore be concluded that the Physical Functioning scale also reflects a unidimensional construct with the exception of item 5. This analysis confirmed the principal components analysis (the original factor analysis), which had also demonstrated a single factor corresponding to the Physical Functioning scale and had also shown high reliability (Cronbach's alpha 0.83), and is similar to the findings from other studies (Ringdal and Ringdal, 1993; Ringdal et al., 1999), which have suggested a strong structure (and high scalability) for this scale.

The analysis of the person measures from these scales showed differences between the Emotional Functioning and Fatigue scales, and the Physical Functioning scale. Both Emotional Functioning and Fatigue scales were interval-based, although the differences between the scale scores of the Emotional Functioning scale below 50 were closer than those above.

The interpretation of the scores from quality of life instruments has received considerable attention in recent years. This has been driven, in particular, by the need to determine how changes in scores can be interpreted, and what significance should be attached by patients and clinicians to changes in scores. A standard interpretation of change in QOL, namely the "minimal clinically important difference" (MCID) or "minimal important difference" (MID), has been defined by researchers at McMaster University as,

> "[T]he smallest difference in a score of a domain of interest that patients perceive to be beneficial and that would mandate, in the absence of troublesome side-effects and excessive costs, a change in the patient's management." (Jaeschke, Singer, & Guyatt, 1989).

There are two common approaches that have been adopted to explore the clinical significance of changes in QOL scores (e.g. Wyrwich & Wolinsky, 2000). These can be defined as the anchor-based approach, where changes in scores over time are noted, and compared to subjective significance ratings obtained from the patients. Attempts have been made to establish an MCID or MID which is unique to each QOL of life instrument. For example, the McMaster group has been able to establish from this that an average 0.5 per-item change represents a MCID for a number of questionnaires developed by this group [Jaeschke et al., 1989, Juniper, Guyatt, Willan et al., 1994). Similarly, others (Osoba, Rodrigues, Myles et al., 1998) have investigated how changes on the EORTC QLQ-C30 correspond with changes in the subjective rating made by patients. These researchers discovered that a change of one category in subjective ratings corresponded to a median change in QOL score of 8.75. This data corroborated a retrospective study (King, 1996) which found using data from previously published studies that QOL scores differed for groups of patients separated by performance status, weight loss, severity of disease and toxicity, and that that changes in the scale scores of the EORTC-QLQ of around 10 amount to negligible or low effects, changes of 15 are moderate effects, and a change of 20 or more indicates a clinically significant or high effect size.

Using a similar methodology, other research groups (Cella, Hahn, Dineen, 2002) have established that changes between 2 and 3 on the FACT-G scales (at least for Physical, and Functional Well-being) are associated with self-reported changes in patient well-being and can therefore be considered as clinically meaningful differences.

The second approach encompasses distribution-based methods, such as the standard error of measurement (SEM, McHorney & Tarlov, 1995). The SEM is sample-independent (Nunnally & Bernstein, 1994), i.e. it remains constant across the range of abilities of the population. Recent work has attempted to link the SEM and the MCID / MID, which has been achieved with some success (Wyrwich, Nienaber,

Tierney et al., 1999; Wyrwich, Tierney, & Wolinsky, 1999; Wyrwich, Tierney, & Wolinsky, 2000).

However, any statement that a given change in the score of a QOL instrument indicates a change in clinical states or subjective well-being, makes the assumption that the given change has the same implication at different parts of the scale, i.e. that items of questionnaires are equally spaced along a continuum corresponding to the underlying latent trait. This is known as an 'interval scale', and a change from a score of, for instance, 20 to 30 has the same clinical meaning, as a change from 40 to 50. If scores or items are not equally spaced then the interpretation of changes of scores becomes difficult, because a given numerical change will have a different meaning at different points along the scale.

A number of studies (Cook, Rabeneck, Campbell et al., 1999; Stucki, Daltroy, Katz et al., 1996) have addressed these issues using the Rasch model (Rasch, 1980). Using Rasch models, the raw scores generated by patients can be mapped onto an interval-level log-odds or logit scale, and distances between scores or items can be calculated. This can be used to establish whether the points on a scale are equally spaced, i.e. whether they are interval-scales. If distances between items are equal they will remain so when converted to logits, otherwise not (Cook et al., 1999).

In terms, of the "Minimal important difference" (Jaeschke, Singer, & Guyatt, 1989), from the results of this study, this means that changes in patients' level of fatigue are easily interpretable since differences in scores are equal across the scales. However, changes in Emotional Functioning are more problematical, since differences between scores are larger for higher levels of Emotional Functioning than for lower levels. This means that changes above the "threshold" of a scale score of 50 may be more significant (in terms of a minimal clinically important difference) than those below or around the threshold.

On the other hand, the results from the person measures demonstrated that the Physical Functioning scale was not interval based. Although the difference

between the range of scores from 10 to 15 were approximately equal, the differences at both extremes were not. This means that interpreting changes in physical functioning cannot be done with any confidence, since the impact of the change may differ on where it occurs along the scale.

In conclusion, the results from this study suggest that the three scales, Physical and Emotional Functioning, and Fatigue are unidimensional, which is broadly in agreement with previous studies (e.g. Ringdal and Ringdal, 1993; Ringdale et al., 1999). Future work could use Rasch analyses to explore whether the Emotional and Cognitive Functioning, and Role and Social Functioning scales could be combined to form unidimensional structures as suggested by the factor analysis in this study and the work by McLachlan et al (1999). In addition to this future work could also test the proposal by Ringdal and colleagues (Ringdal and Ringdal, 1993; Ringdal et al., 1999) that Physical and Role Functioning should combine, since their results have demonstrated that this combination shows both scalability and internal consistency (although their results also suggest that Physical Functioning works well by itself) and that this mediates the problems with low levels of consistency of the Role Functioning scale.

# 7.1. Factor and Rasch Analysis of the FACT-G

## 7.1. Factor Analysis of the FACT-G

### *7.1.1. Aim*

As previously discussed in Chapter 6, the aims of this study are as follows, namely: 1). To explore the traditional psychometric properties of the FACT-G; 2). To apply a Rasch analysis to the instrument; and 3). To discuss the results in terms of the implications for interpretation of quality of life data in the context of clinical significance.

This study investigated the factor structure of the FACT-G in a heterogeneous sample of 461 cancer patients. Factor analysis (Principal Components Analysis) was carried out to investigate the factor structure of the instrument. In addition, the reliability of the functional subscales (Physical (PWB), Social & Family (SFWB), Emotional (EWB) and Functional Well-being (FWB)) was assessed using Cronbach's alpha.

### *7.1.2. Method*

The patient data for the FACT-G were collated from a total of two studies which have been carried out by the Cancer Research UK, Psychosocial and Clinical Practice Research Group (St. James's University Hospital, Leeds, Velikova et al., 2002).

One group of patients completed an electronic version of the FACT-G on a standalone computer with touchscreen monitor. The raw scores from both questionnaires were recorded onto an MS-Access database and converted to the summated scales. The other group of patients (Velikova et al., 2004) completed the paper version of the FACT-G, the scores of which were then transferred onto an MS-Access database.

*7.1.3 Participants*

A total of 465 patients completed the questionnaires, however demographic details were only available for 461 patients: 323 females (average age 55.7 years, s.d. 12.4) and 138 males (average age 60.8, s.d. 13.0). Table 7.1.1 gives a breakdown of diagnosis by gender.

**Table 7.1.1 Diagnosis by gender and age for FACT-G**

| FACT-G | Female | | Males | |
|---|---|---|---|---|
| | n = 323 | | n = 138 | |
| Age, years (mean ± S.D.) | 55.7 ± 12.4 | | 60.8 ± 13.0 | |
| | Count | % | Count | % |
| Diagnosis | | | | |
| Breast | 99 | 30.7 | | |
| Colorectal | 35 | 10.8 | 37 | 26.8 |
| Gastrointestinal | 15 | 4.6 | 12 | 8.7 |
| Genitourinary | 111 | 34.4 | 21 | 15.2 |
| Lung | 9 | 2.8 | 13 | 9.4 |
| Melanoma | 10 | 3.1 | 11 | 8.0 |
| Renal | 18 | 5.6 | 26 | 18.8 |
| Sarcoma | 7 | 2.2 | 12 | 8.7 |
| Other | 17 | 5.3 | 6 | 4.4 |

*7.1.4 Methodology*

A principal components analysis was carried out on the data. Factors were identified using a scree plot and Kaiser's criterion of eigenvalues greater than 1. Subsequently a factor analysis was carried out on the rotated data (orthogonal rotation, e.g. varimax).

*7.1.5 Results*

Table 7.1.2 shows a breakdown of the scores from the FACT-G scales.

**Table 7.1.2. Means and standard deviations of the FACT-G scores**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| PWB | 465 | 1.00 | 33.00 | 17.65 | 5.99 |
| SFWB | 465 | 0.00 | 35.00 | 27.19 | 5.83 |
| EWB | 462 | 4.00 | 28.00 | 16.34 | 4.37 |
| FWB | 463 | 4.00 | 35.00 | 20.47 | 6.67 |
| TOTAL | 462 | 36.83 | 109.00 | 81.61 | 13.85 |

*PWB – Physical Well-being; SFWB – Social & Family Well-being; EWB – Emotional Well-being; FWB – Functional Well-being.

The correlation matrix for the scales from the FACT-G is provided in Table 7.1.3. All correlations are significant at $p < 0.01$.

**Table 7.1.3 Correlation matrix for scales from FACT-G**

| | PWB | SFWB | EWB | FWB | TOTAL |
|---|---|---|---|---|---|
| PWB | . | | | | |
| SFWB | .15 | . | | | |
| EWB | .44 | .26 | . | | |
| FWB | .64 | .42 | .44 | . | |
| TOTAL | .72 | .61 | .68 | .88 | . |

As can be seen from Table 7.1.3 although all individual scales correlate highly with the total score, however with the exception of the correlation between Physical Well-being and Functional Well-being, the other scales correlate poorly with each other.

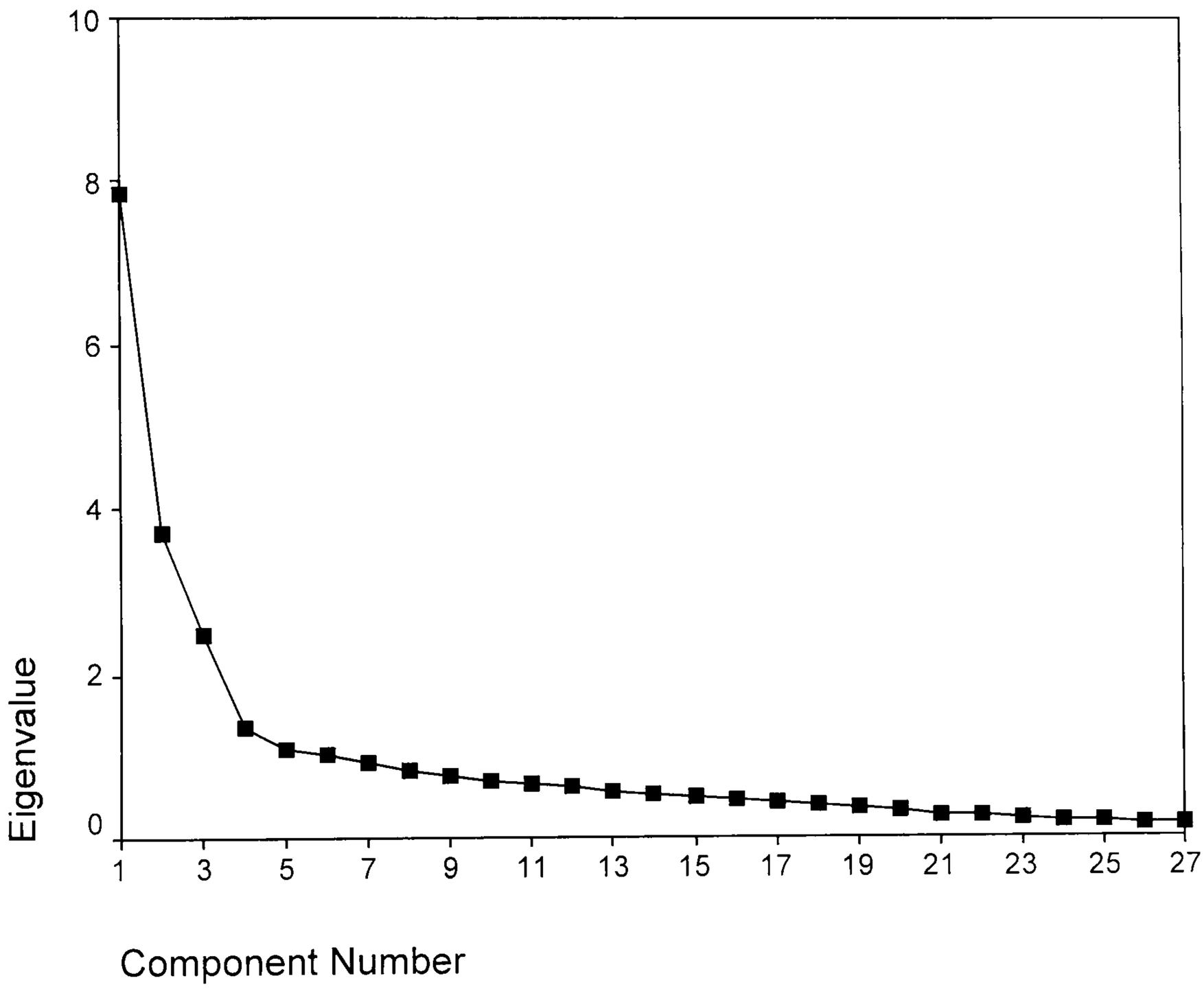
The eigenvalues extracted from the principal components analysis of the FACT-G can be seen in table 7.1.4. A total of 6 factors were extracted collectively representing around 65% of the variance.

**Table 7.1.4 Eigenvalues from Factor Analysis of FACT-G**

| Component | Total | % of Variance | Cumulative % |
|---|---|---|---|
| 1.00 | 7.85 | 29.08 | 29.08 |
| 2.00 | 3.71 | 13.73 | 42.82 |
| 3.00 | 2.48 | 9.19 | 52.00 |
| 4.00 | 1.34 | 4.97 | 56.97 |
| 5.00 | 1.09 | 4.05 | 61.03 |
| 6.00 | 1.02 | 3.77 | 64.80 |

The scree plot of the eigenvalues can be seen in Figure 7.1.1.

**Figure 7.1.1. Scree plot from the Principal Components Analysis of the FACT-G**




The component matrix is shown in Table 7.1.5. Values below 0.30 have been suppressed. The rotated component matrix is shown in Table 7.1.6.

**Table 7.1.5 Component matrix from the Principal Components Analysis of the FACT-G**

| Component | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **PWB1** | -0.60 | 0.37 | | | | |
| **PWB2** | -0.52 | | | | | |
| **PWB3** | -0.70 | 0.31 | | | | |
| **PWB4** | -0.47 | | | | 0.53 | |
| **PWB5** | -0.61 | 0.31 | | | | |
| **PWB6** | -0.73 | 0.34 | | | | |
| **PWB7** | -0.65 | | | | | |
| **SFWB1** | | 0.69 | | | | 0.39 |
| **SFWB2** | | 0.76 | | | | |
| **SFWB3** | | 0.72 | | | | 0.37 |
| **SFWB4** | | 0.63 | | | | -0.45 |
| **SFWB5** | 0.37 | 0.74 | | | | -0.36 |
| **SFWB6** | | 0.48 | | -0.34 | | |
| **SFWB7** | 0.53 | | | | 0.37 | 0.36 |
| **EWB1** | -0.60 | | 0.48 | | | |
| **EWB2** | 0.35 | | | 0.46 | | |
| **EWB3** | -0.46 | | 0.49 | | | |
| **EWB4** | -0.51 | | 0.64 | | | |
| **EWB5** | -0.50 | | 0.70 | | | |
| **EWB6** | -0.51 | | 0.61 | | | |
| **FWB1** | 0.61 | | 0.40 | | | |
| **FWB2** | 0.64 | | 0.34 | | | |
| **FWB3** | 0.80 | | | | | |
| **FWB4** | 0.40 | | | 0.51 | | |
| **FWB5** | 0.41 | | | | -0.57 | |
| **FWB6** | 0.78 | | | | | |
| **FWB7** | 0.79 | | | | | |

*PWB – items from PWB scale; SFWB – items from SFWB scale; EWB – items from EWB scale; FWBF – items from FWB scale.

In essence, the rotated factor structure revealed four factors, Physical Well-being (Factor 1), Emotional Well-being (Factor 2), Social and Family Well-being (Factor 3), and Functional Well-being (Factor 4). Therefore the rotated factor structure corresponded to the scale structures.

**Table 7.1.6 Rotated component matrix of the FACT-G**

| Component | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **PWB1** | 0.70 | | | | | |
| **PWB2** | 0.67 | | | | | |
| **PWB3** | 0.77 | | | | | |
| **PWB4** | 0.43 | | | | | 0.59 |
| **PWB5** | 0.61 | 0.31 | | | | |
| **PWB6** | 0.79 | | | | | |
| **PWB7** | 0.76 | | | | | |
| **SFWB1** | | | 0.33 | | 0.79 | |
| **SFWB2** | | | 0.74 | | 0.39 | |
| **SFWB3** | | | 0.40 | | 0.80 | |
| **SFWB4** | | | 0.75 | 0.39 | | |
| **SFWB5** | | | 0.85 | | | |
| **SFWB6** | | | 0.65 | | | |
| **SFWB7** | -0.41 | | | | 0.41 | 0.41 |
| **EWB1** | | 0.73 | | | | |
| **EWB2** | | | | 0.62 | | |
| **EWB3** | | 0.68 | | | | |
| **EWB4** | | 0.81 | | | | |
| **EWB5** | | 0.83 | | | | |
| **EWB6** | | 0.78 | | | | |
| **FWB1** | -0.67 | | | 0.36 | | |
| **FWB2** | -0.62 | | | 0.40 | | |
| **FWB3** | -0.63 | | | 0.47 | | |
| **FWB4** | | -0.33 | | 0.65 | | |
| **FWB5** | | | | 0.45 | 0.32 | -0.50 |
| **FWB6** | -0.65 | | | 0.45 | | |
| **FWB7** | -0.65 | | | 0.47 | | |

Since the full sample was too small to allow a split test reliability of the factor structure to be completes a random sample of 250 patients was drawn from the sample the factor analysis performed again. Although the primacy of the factors changed for the random sample, the factor structures remained the same, demonstrating stability in the factor structures.

In addition, the reliability and internal consistency of the FACT-G scales was also assessed. These are shown in Table 7.1.7 and 7.1.8.

**Table 7.1.7 Item correlations and Revised Cronbach's alpha for FACT-G**

| Items | Item-total correlation | Cronbach's α |
|---|---|---|
| **PWB1** | 0.63 | 0.83 |
| **PWB2** | 0.56 | 0.84 |
| **PWB3** | 0.63 | 0.83 |
| **PWB4** | 0.45 | 0.85 |
| **PWB5** | 0.61 | 0.83 |
| **PWB6** | 0.77 | 0.81 |
| **PWB7** | 0.66 | 0.83 |
| **SFWB1** | 0.63 | 0.75 |
| **SFWB2** | 0.56 | 0.73 |
| **SFWB3** | 0.63 | 0.73 |
| **SFWB4** | 0.45 | 0.75 |
| **SFWB5** | 0.61 | 0.72 |
| **SFWB6** | 0.77 | 0.77 |
| **SFWB7** | 0.66 | 0.84 |
| **EWB1** | 0.58 | 0.59 |
| **EWB2** | -0.26 | 0.83 |
| **EWB3** | 0.40 | 0.66 |
| **EWB4** | 0.64 | 0.58 |
| **EWB5** | 0.66 | 0.55 |
| **EWB6** | 0.68 | 0.54 |
| **FWB1** | 0.59 | 0.83 |
| **FWB2** | 0.64 | 0.82 |
| **FWB3** | 0.77 | 0.80 |
| **FWB4** | 0.37 | 0.85 |
| **FWB5** | 0.36 | 0.86 |
| **FWB6** | 0.76 | 0.80 |
| **FWB7** | 0.74 | 0.80 |

As can be seen from Table 7.1.7 item-total correlations are modest for most scales, with the exception of item 4 from the Social & Family Well-being scale, as well as item 2 from the Emotional Well-being scale, and items 4 and 5 from the Functional Well-being scale which demonstrated poor item-total correlations.

**Table 7.1.8. Cronbach's alpha for Individual Scales of FACT-G**

| | Total Cronbach's alpha |
|---|---|
| Physical Well-being | 0.85 |
| Social & Family Well-being | 0.82 |
| Emotional Well-being | 0.68 |
| Functional Well-being | 0.84 |

The Cronbach's alpha statistic was high for all scales with the exception of the Emotional Well-being scale (Table 7.1.8).

*7.1.6 Conclusions*

This study investigated the factor structure of the FACT-G using a principal components analysis. The results demonstrated that the internal consistency and item-total correlations were moderately good for the scales. Furthermore, a four factor structure emerged from the rotated component matrix which corresponded to the functioning domains, namely: a Physical, Social & Family, Emotional and Functional Well-being domains.

In the next section the data are analysed using a Rasch model for polytomous data, i.e. the Rating Scale Model (Andrich, 1978a, b). The item locations, as well as the unidimensionality and person measures will explored for these scales and compared to the results of the factor analysis.

## 7.2 Rasch Analysis of the FACT-G

*7.2.1 Aim*

The aim of this study is to examine each of the FACT-G subscales individually using a Rasch model for polytomous data, i.e. the Rating Scale Model (Andrich, 1978a, b).

*7.2.2 Methodology*

The data employed in this section were the same as described in section 5.1.3. The scores from the Physical Well-being, Emotional Well-being, Social and Family Well-being and Functional Well-being scales of the FACT-G were converted to interval-level logit (log-odds) scores using the *Winsteps* software (Linacre & Wright, 2000), and the Rating Scale Model for polytomous data (Andrich 1978a, b) as described in Chapter 1.2.

In this study item difficulty and person ability estimates were derived, as well as fit statistics (infit and outfit) for the items. A principal components analysis of the residuals was also performed for the scales, and test characteristic curves were calculated. Additionally, the differences between adjacent scores for person measures were plotted for each scale, and the unidimensionality of the scales was explored further using differential item analyses.

*7.2.3 Results for Physical Well-being*

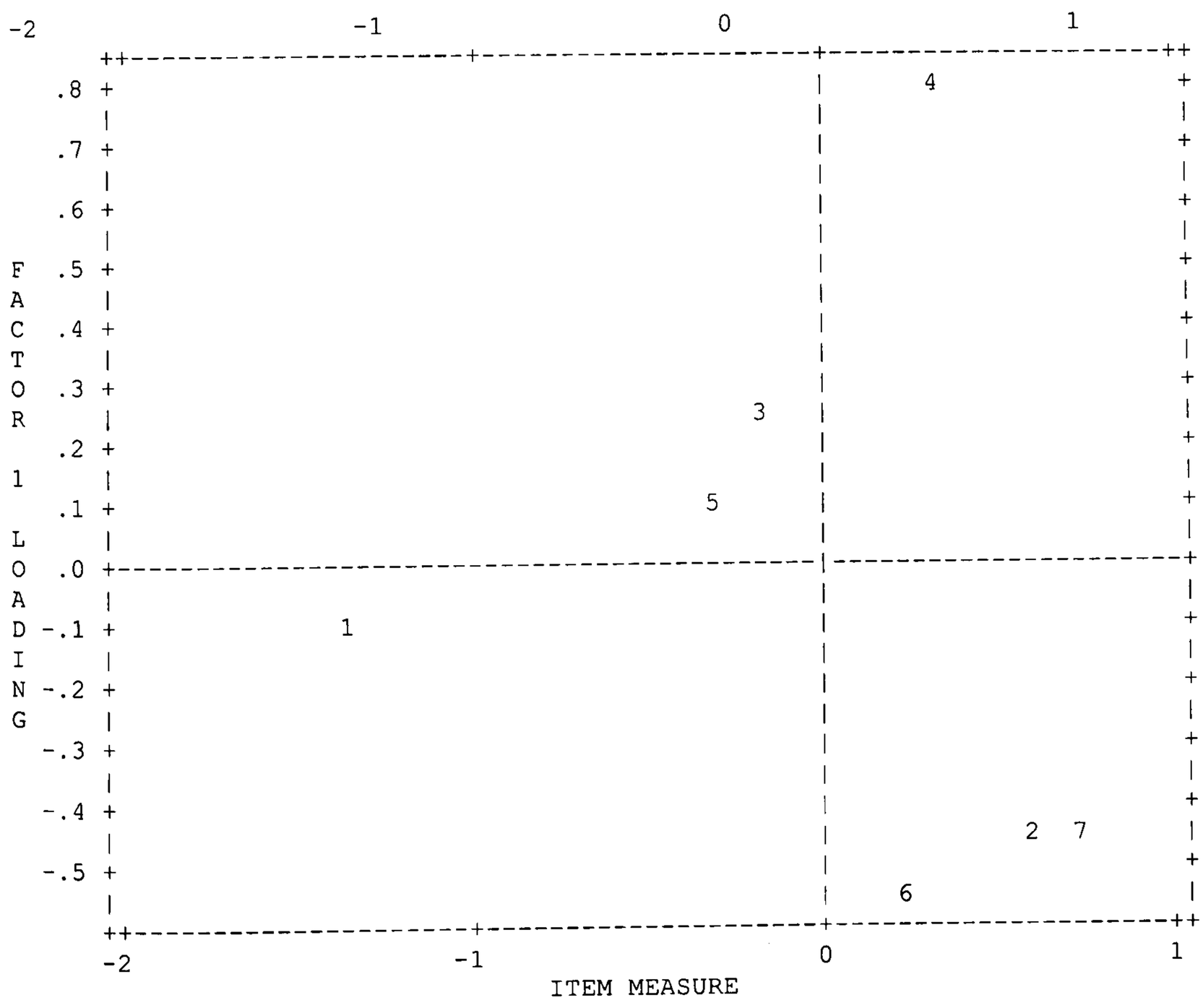
The location measures and fit statistics are given in Table 7.2.1 for the Physical Well-being scale. It can be seen from this table that two items from this scale exhibit poor fit, i.e. infit mean square statistics greater than 1.30 or smaller than 0.70 and standardised t-statistics greater than 1.96 (e.g. Wright et al., 1994), namely item 4 (" I have pain"), which added excessive "noise" to the model, and item 6 ("I feel ill"), which showed redundancy.

**Table 7.2.1 Unidimensionality measures for Physical Well-being**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | | ITEMS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 919 | 441 | .30 | .06 | 1.52 | 6.2 | 1.47 | 5.0 | A .58 | GP4 | PWB4 |
| 2 | 850 | 444 | .58 | .06 | 1.23 | 2.9 | 1.18 | 2.0 | B .61 | GP2 | PWB2 |
| 3 | 1062 | 442 | -.18 | .06 | 1.14 | 2.0 | 1.02 | .3 | C .71 | GP3 | PWB3 |
| 5 | 1112 | 443 | -.32 | .06 | .98 | -.2 | .99 | -.1 | D .70 | GP5 | PWB5 |
| 7 | 816 | 445 | .73 | .06 | .98 | -.2 | .81 | -2.2 | c .67 | GP7 | PWB7 |
| 1 | 1473 | 445 | -1.36 | .05 | .79 | -3.5 | .90 | -1.5 | b .78 | GP1 | PWB1 |
| 6 | 943 | 445 | .25 | .06 | .59 | -6.7 | .55 | -6.7 | a .77 | GP6 | PWB6 |
| MEAN | 1025. | 444. | .00 | .06 | 1.03 | .1 | .99 | -.5 | | | |
| S.D. | 208. | 1. | .65 | .00 | .28 | 3.9 | .27 | 3.4 | | | |

Figure 7.2.1. shows the factor plot of the principal components analysis of the standardised residuals for the Physical Well-being scale. A factor amounting to a total of 1.4 eigenvalues was extracted from the residuals, indicating that there were no other factor structures in the residuals (Smith & Miao, 1994).

**Figure 7.2.1 Principal Components (Standardized Residual) Factor Plot of the Physical Well-being Scale**



The factor loadings from the principal components analysis and coding for Figure 7.2.1 can be seen in Table 7.2.2.

**Table 7.2.2. Factor Loadings from the Principal Components Analysis of the Physical Functioning Scale**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | ITEM | |
|---|---|---|---|---|---|---|---|
| 1 | .79 | .30 | 1.52 | 1.47 | 4 | 4 GP4 | PWB4 |
| 1 | .23 | -.18 | 1.14 | 1.02 | 3 | 3 GP3 | PWB3 |
| 1 | .12 | -.32 | .98 | .99 | 5 | 5 GP5 | PWB5 |
| 1 | -.54 | .25 | .59 | .55 | 6 | 6 GP6 | PWB6 |
| 1 | -.47 | .73 | .98 | .81 | 7 | 7 GP7 | PWB7 |
| 1 | -.43 | .58 | 1.23 | 1.18 | 2 | 2 GP2 | PWB2 |
| 1 | -.12 | -1.36 | .79 | .90 | 1 | 1 GP1 | PWB1 |

The item map for the Physical Well-being Scale can be seen in Figure 7.2.2. This figure demonstrates the overlap between items PWB4 and PWB6. Furthermore, items PWB2 and PWB7, as well as PWB3 and PWB5 are close together. In addition, the items do not cover the full range of person abilities (i.e. –6 to +6) with the majority of items falling in a narrow range between –1.3 and +1 on the ability scale.

The category probability curve for Physical Well-being is shown in Figure 7.2.3.

**Figure 7.2.2. Logit map of all items (QUESS) and patients (PATSS) for Physical Well-being**

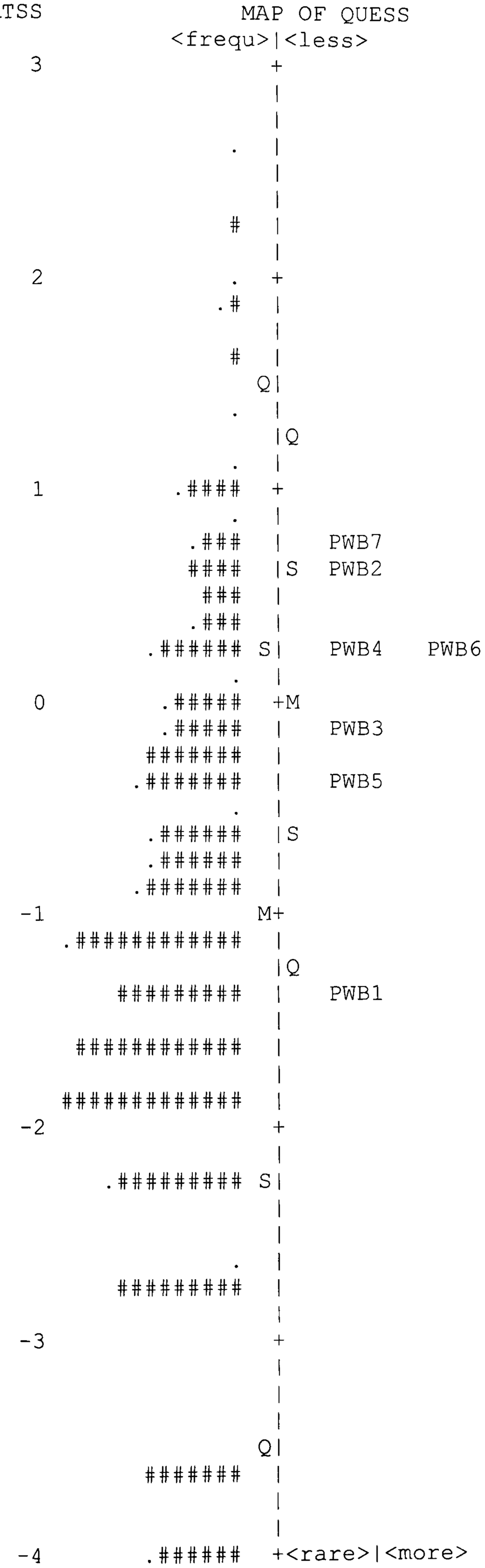
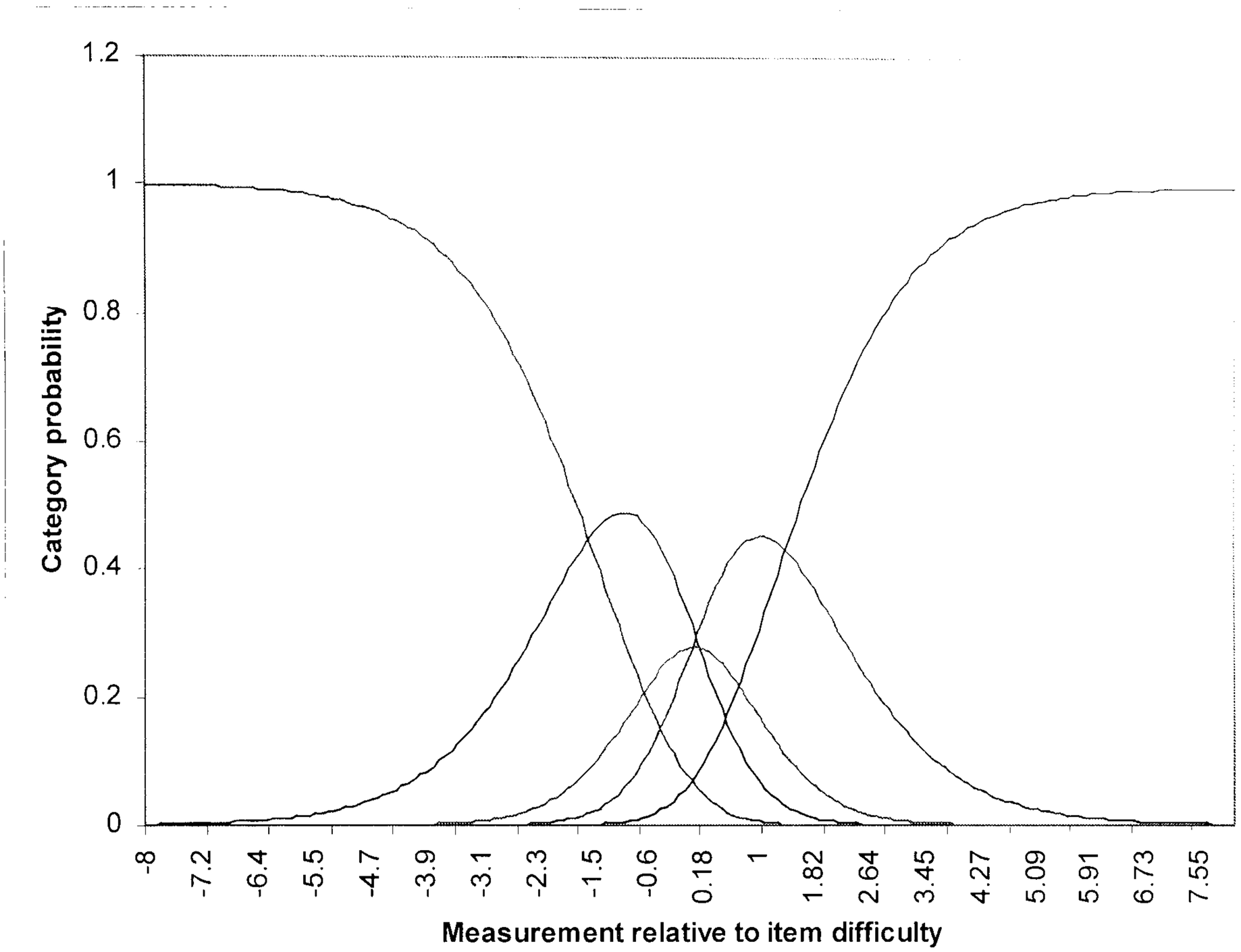


**Figure 7.2.3. Category Probability Curve for Physical Well-being**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Brown = category 4; 5). Black = category 5.

It can be seen for instance, from figure 7.2.3 that as a patient's physical functioning declines, i.e. as the person measure increases in relation to the item difficulty, that the likelihood of the patient responding to items with "Not at all" (category 1) or "A little" (category 2 decreases, and the likelihood of responses such as "Quite a bit" (category 3) or "Very much" (category 4) increases. However, the figure also shows that for person abilities around 0.18 logits, the likelihood of a person selecting categories 2, 3 or 4 in response to a question are equal (at approximately 30%).

Table 7.2.3 shows a summary of the items for the Physical Well-being Scale. The item separation index is approximately 10, demonstrating that the Physical Well-being scale can distinguish between 10 levels of item difficulty.

**Table 7.2.3. Summary of Items from the Physical Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1025.0 | 443.6 | .00 | .06 | 1.03 | .1 | .99 | -.5 |
| S.D. | 207.7 | 1.5 | .65 | .00 | .28 | 3.9 | .27 | 3.4 |
| MAX. | 1473.0 | 445.0 | .73 | .06 | 1.52 | 6.2 | 1.47 | 5.0 |
| MIN. | 816.0 | 441.0 | -1.36 | .05 | .59 | -6.7 | .55 | -6.7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .06 | ADJ.SD | .65 | SEPARATION | 10.36 | ITEM | RELIABILITY | .99 |
| MODEL RMSE | .06 | ADJ.SD | .65 | SEPARATION | 11.04 | ITEM | RELIABILITY | .99 |
| S.E. OF ITEM MEAN = .27 | | | | | | | | |

Table 7.2.4 shows a summary of the category measures for the Physical Well-being Scale. It can be seen that there is a good level of separation between the category 1 and 2, and category 2 and 3, and category 4 and 5 with a distance of around 1.4 logits (Linacre, 1999) between each threshold (structure measure). However, the distance between category 3 and 4 is less than 1.4, confirming the observations regarding the category probability as shown in Figure 7.2.3. In addition, the distances between each category increase monotonically (category measure).

**Table 7.2.4. Summary of Category Measures for the Physical Functioning Scale**

SUMMARY OF CATEGORY STRUCTURE. Model="R"

| CATEGORY LABEL | SCORE | OBSERVED COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1036 | 33 | -2.12 | -2.09 | 1.07 | 1.09 | NONE | ( -2.76) | 1 |
| 2 | 2 | 974 | 31 | -1.12 | -1.12 | .95 | .89 | -1.53 | -1.01 | 2 |
| 3 | 3 | 425 | 14 | -.23 | -.33 | .88 | .76 | .12 | .03 | 3 |
| 4 | 4 | 434 | 14 | .41 | .40 | .98 | .99 | .02 | 1.03 | 4 |
| 5 | 5 | 236 | 8 | 1.10 | 1.21 | 1.21 | 1.27 | 1.40 | ( 2.67) | 5 |
| MISSING | | 10 | 0 | -.52 | | | | | | |

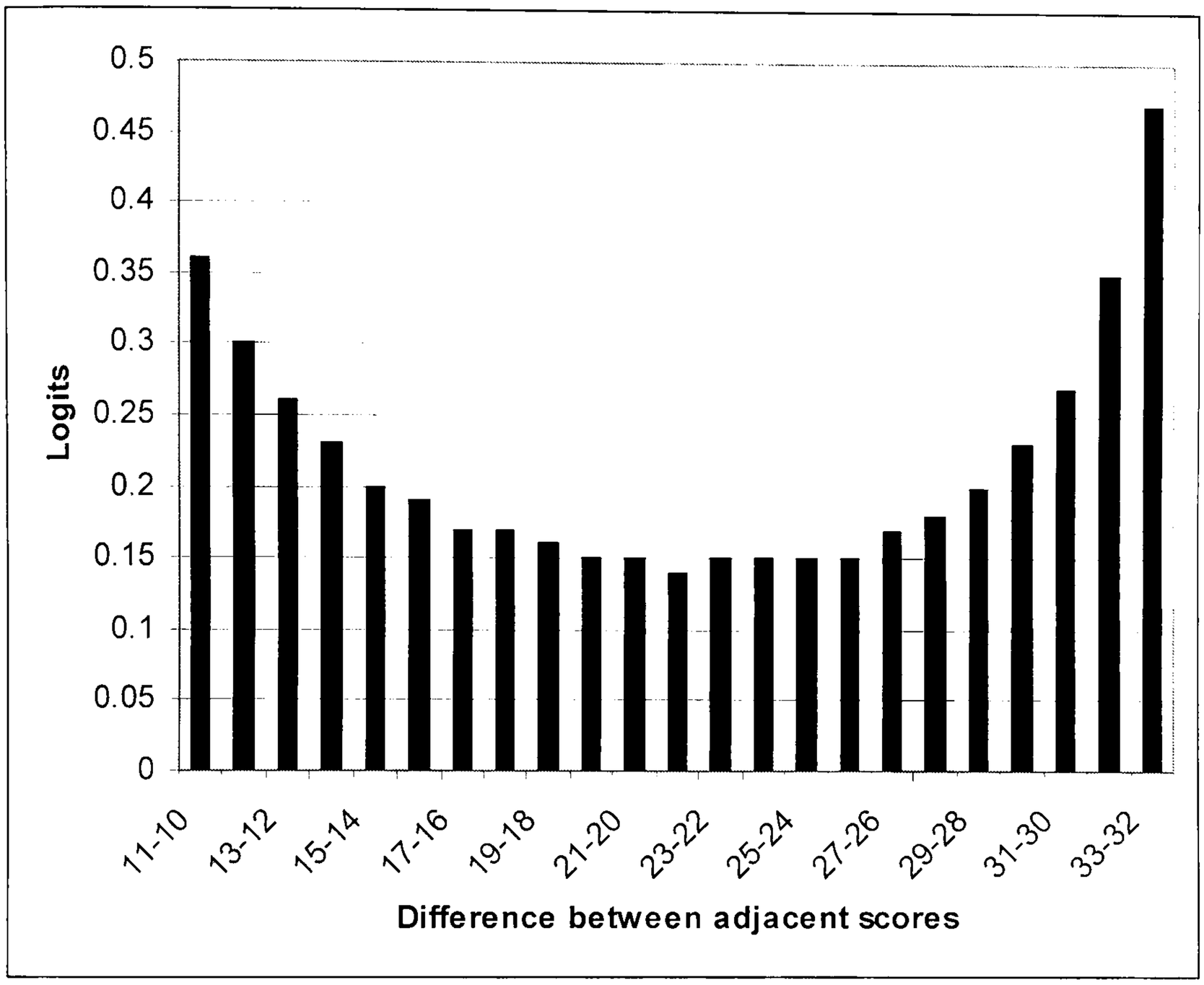
The person measures for the Physical Well-being scale are shown in Table 7.2.5 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 7.2.4.

**Table 7.2.5. Person measures for Physical Well-being**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 7 | -4.35 | 1 | 0 | 1 | 0 |
| 8 | -3.59 | 0.49 | -0.62 | 0.26 | -0.94 |
| 10 | -2.28 | 0.52 | -0.81 | 0.63 | -0.6 |
| 11 | -1.92 | 0.57 | -0.79 | 0.64 | -0.62 |
| 12 | -1.62 | 0.45 | -1.16 | 0.55 | -0.87 |
| 13 | -1.36 | 0.43 | -1.25 | 0.5 | -1.03 |
| 14 | -1.13 | 0.3 | -1.75 | 0.32 | -1.6 |
| 15 | -0.93 | 1.04 | 0.07 | 0.99 | -0.01 |
| 16 | -0.74 | 0.99 | -0.02 | 0.96 | -0.07 |
| 17 | -0.57 | 0.83 | -0.35 | 0.84 | -0.32 |
| 18 | -0.4 | 2.05 | 1.6 | 2.21 | 1.74 |
| 19 | -0.24 | 1.73 | 1.24 | 1.75 | 1.21 |
| 20 | -0.09 | 2.04 | 1.7 | 1.96 | 1.54 |
| 21 | 0.06 | 1.38 | 0.73 | 1.28 | 0.53 |
| 22 | 0.2 | 0.83 | -0.39 | 0.83 | -0.39 |
| 23 | 0.35 | 0.3 | -2.15 | 0.31 | -2.03 |
| 24 | 0.5 | 1.81 | 1.43 | 1.67 | 1.15 |
| 25 | 0.65 | 2.73 | 2.55 | 2.97 | 2.63 |
| 26 | 0.8 | 0.25 | -2.28 | 0.32 | -1.85 |
| 27 | 0.97 | 0.51 | -1.17 | 0.67 | -0.69 |
| 28 | 1.15 | 0.11 | -2.86 | 0.13 | -2.55 |
| 29 | 1.35 | 2.33 | 1.68 | 2.36 | 1.62 |
| 30 | 1.58 | 0.59 | -0.78 | 0.91 | -0.15 |
| 31 | 1.85 | 0.46 | -1 | 0.47 | -0.93 |
| 32 | 2.2 | 0.36 | -1.18 | 0.34 | -1.15 |
| 33 | 2.67 | 0.94 | -0.07 | 1.02 | 0.02 |

It can be seen from Figure 7.2.5 that the distance between adjacent scores is equal between 20 and 27, however that these differences increase at both extremes.

**Figure 7.2.4. Differences between adjacent scores of the Physical Well-being Scale**




The summary of person measures is shown in Table 7.2.6. The person separation index is close to 2.00 at 1.89, indicating that the scale is only able to detect fewer than 2 levels of person ability. The reliability measure however is good at 0.78.

**Table 7.2.6. Summary of Person Measures for the Physical Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 16.1 | 7.0 | -.95 | .51 | .97 | -.2 | .99 | -.2 |
| S.D. | 5.9 | .1 | 1.25 | .16 | .67 | 1.1 | .73 | 1.1 |
| MAX. | 33.0 | 7.0 | 2.67 | 1.06 | 4.07 | 3.5 | 7.05 | 5.3 |
| MIN. | 8.0 | 6.0 | -3.59 | .38 | .04 | -3.9 | .04 | -3.8 |
| REAL RMSE | .59 | ADJ.SD | 1.11 | SEPARATION | 1.89 | PERSON RELIABILITY | .78 | |
| MODEL RMSE | .54 | ADJ.SD | 1.13 | SEPARATION | 2.10 | PERSON RELIABILITY | .82 | |
| S.E. OF PERSON MEAN = .06 | | | | | | | | |

MINIMUM EXTREME SCORE: 20 PERSONS
LACKING RESPONSES: 1 PERSONS
VALID RESPONSES: 99.7%

**Figure 7.2.5. Test Information Curve for Physical Well-being Scale**

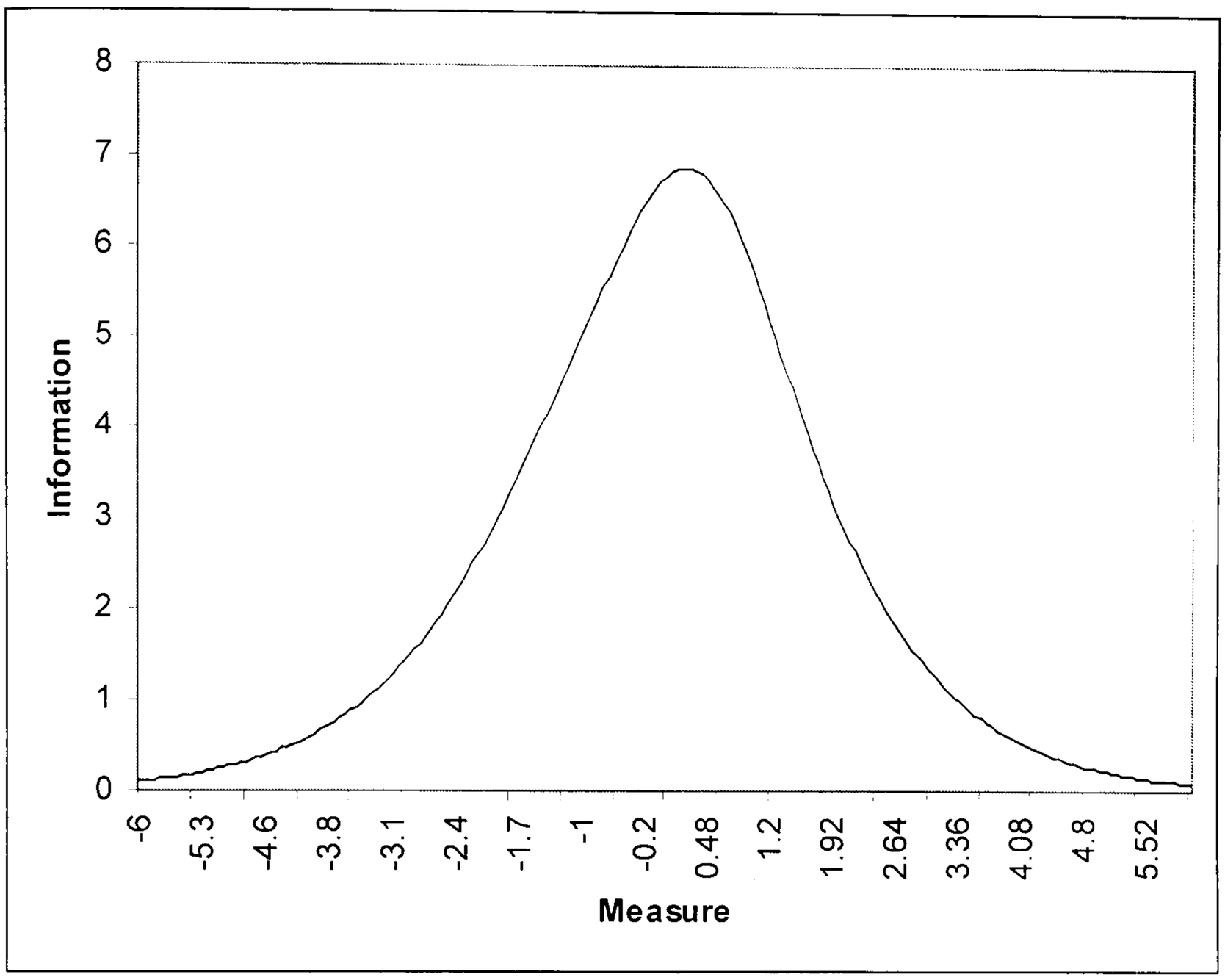



Figure 7.2.5 shows the test information curve for the Physical Well-being scale and demonstrates that the scale provides the most information over a narrow range at the centre of the scale (ability measures around zero).

In addition to the Rasch analysis described above the items from the Physical Well-being scale were also investigated for differential bias (as described in Chapter 6.2).

**Table 7.2.7 Differential Item Analysis of the Physical Well-being Scale**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | -1.31 | .06 | 1 | -1.47 | .10 | .16 | .12 | 1.40 | 443 | 1 | GP1 | pwb1 |
| 2 | .56 | .07 | 1 | .65 | .13 | -.10 | .15 | -.65 | 442 | 2 | GP2 | pwb2 |
| 2 | -.19 | .07 | 1 | -.15 | .11 | -.04 | .13 | -.31 | 440 | 3 | GP3 | pwb3 |
| 2 | .29 | .07 | 1 | .33 | .12 | -.04 | .14 | -.30 | 439 | 4 | GP4 | pwb4 |
| 2 | -.31 | .07 | 1 | -.35 | .10 | .04 | .12 | .36 | 441 | 5 | GP5 | pwb5 |
| 2 | .23 | .07 | 1 | .29 | .12 | -.06 | .14 | -.47 | 443 | 6 | GP6 | pwb6 |
| 2 | .72 | .07 | 1 | .77 | .13 | -.06 | .15 | -.37 | 443 | 7 | GP7 | pwb7 |

The sample was split into male and female groups for the differential item functioning analysis. The results of this can be seen in Table 7.2.7, which demonstrates that none of the items exhibited differential item bias.

In summary, a Rasch analysis of the Physical Well-being scale of the FACT-G demonstrated good fit for five of the seven items from the scale. However, two items (PWB4 and PWB6) exhibited poor fit statistics. Therefore with the exception of item PWB4 ("I have pain") it can be concluded that the Physical Well-being scale is a unidimensional structure. Additionally, none of the items demonstrated differential item bias. The results from the person measures demonstrated that the scale was not interval based with equally spaced scores for the range of scores between 20 and 27, but large, unequal differences at both extremes.

*7.2.4. Results for Social & Family Well-being*

The location measures and fit statistics are given in Table 7.2.8 for the Social & Family Well-being scale. It can be seen from this table that two items from this scale exhibited poor fit (0.70< Infit MNSQ > 1.30, and ZSTD > 2.00) namely items 6 ("I feel close to my partner (or the person who is my main support") and 7 ("I am satisfied with my sex life"), which both added excessive "noise" to the model.

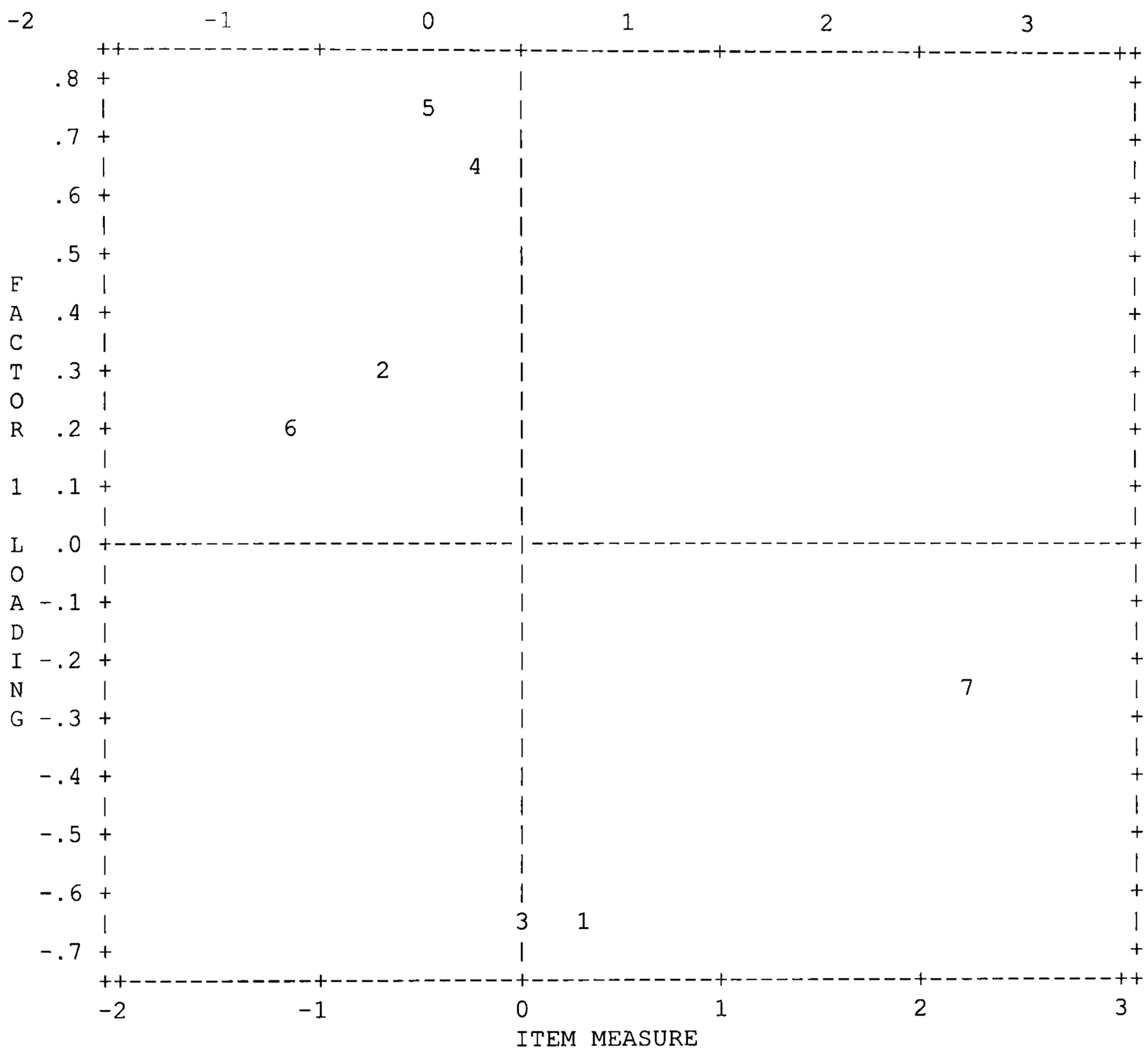
**Table 7.2.8. Unidimensionality measures for Social & Family Well-being**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | | ITEMS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 545 | 225 | 2.22 | .07 | 1.62 | 5.6 | 2.12 | 7.4 | A | .70 | GS7 | sfwb7 |
| 6 | 1652 | 355 | -1.14 | .10 | 1.63 | 4.4 | 1.18 | 1.1 | B | .47 | GS6 | sfwb6 |
| 4 | 1527 | 360 | -.19 | .07 | 1.00 | .0 | 1.20 | 1.7 | C | .62 | GS4 | sfwb4 |
| 1 | 1410 | 359 | .30 | .06 | .98 | -.2 | 1.06 | .6 | D | .68 | GS1 | sfwb1 |
| 2 | 1615 | 360 | -.69 | .08 | .93 | -.7 | .68 | -2.8 | c | .61 | GS2 | sfwb2 |
| 5 | 1568 | 357 | -.47 | .08 | .83 | -1.8 | .72 | -2.6 | b | .64 | GS5 | sfwb5 |
| 3 | 1491 | 360 | -.03 | .07 | .75 | -3.2 | .69 | -3.3 | a | .70 | GS3 | sfwb3 |
| MEAN | 1401. | 339. | .00 | .07 | 1.11 | .6 | 1.09 | .3 | | | | |
| S.D. | 357. | 47. | 1.01 | .01 | .34 | 3.0 | .47 | 3.5 | | | | |

Figure 7.2.6. shows the factor plot of the principal components analysis of the standardised residuals for the Social & Family Well-being scale. A factor with

eigenvalues of 2.1 was extracted from the residuals, indicating that there were other factor structures remaining in the residuals (Smith & Miao, 1994).

**Figure 7.2.6. Principal Components (Standardized Residual) Factor Plot of the Social & Family Well-being Scale**




The factor loadings from the principal components analysis and coding for Figure 7.2.6 can be seen in Table 7.2.9. This demonstrates that the two factors which remain after the principal components analysis of the residuals correspond broadly to items dealing with family concerns (SFWB2 “I get emotional support from my family”, SFWB4 “My family has accepted my illness” and SFWB5 “I am satisfied with family communication about my illness), and those that deal with friends (SFWB1 “I feel close to my friends”, and SFWB 3 “I get support from my friends”).

**Table 7.2.9. Factor Loadings from the Principal Components Analysis of the Physical Functioning Scale**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | ITEM | |
|---|---|---|---|---|---|---|---|
| 1 | .75 | -.47 | .83 | .72 | 5 | 5 GS5 | sfwb5 |
| 1 | .65 | -.19 | 1.00 | 1.20 | 4 | 4 GS4 | sfwb4 |
| 1 | .29 | -.69 | .93 | .68 | 2 | 2 GS2 | sfwb2 |
| 1 | .22 | -1.14 | 1.63 | 1.18 | 6 | 6 GS6 | swfb6 |
| 1 | -.67 | .30 | .98 | 1.06 | 1 | 1 GS1 | sfwb1 |
| 1 | -.65 | -.03 | .75 | .69 | 3 | 3 GS3 | sfwb3 |
| 1 | -.27 | 2.22 | 1.62 | 2.12 | 7 | 7 GS7 | sfwb7 |

The item map for the Social and Family Well-being Scale can be seen in Figure 7.2.7. This figure demonstrates no overlap between the items, although items GS1 to GS5 are close together. In addition, the items do not cover the full range of person abilities (i.e. –6 to +6) with the majority of items falling in a narrow range between -1.0 and +2 on the ability scale.

The category probability curve for Social and Family Well-being is shown in Figure 7.2.8.

**Figure 7.2.7. Logit map of all items and patients for Social & Family Well-being**

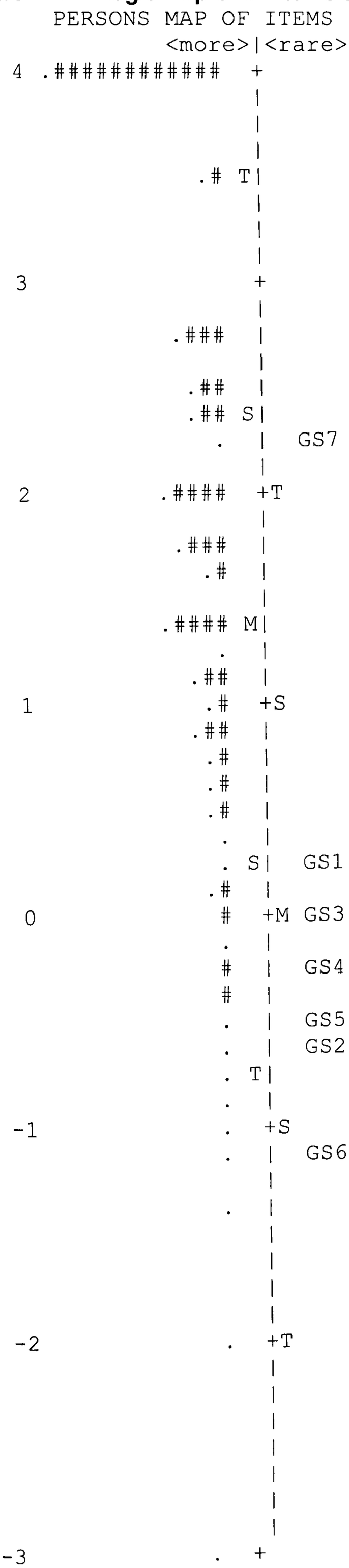
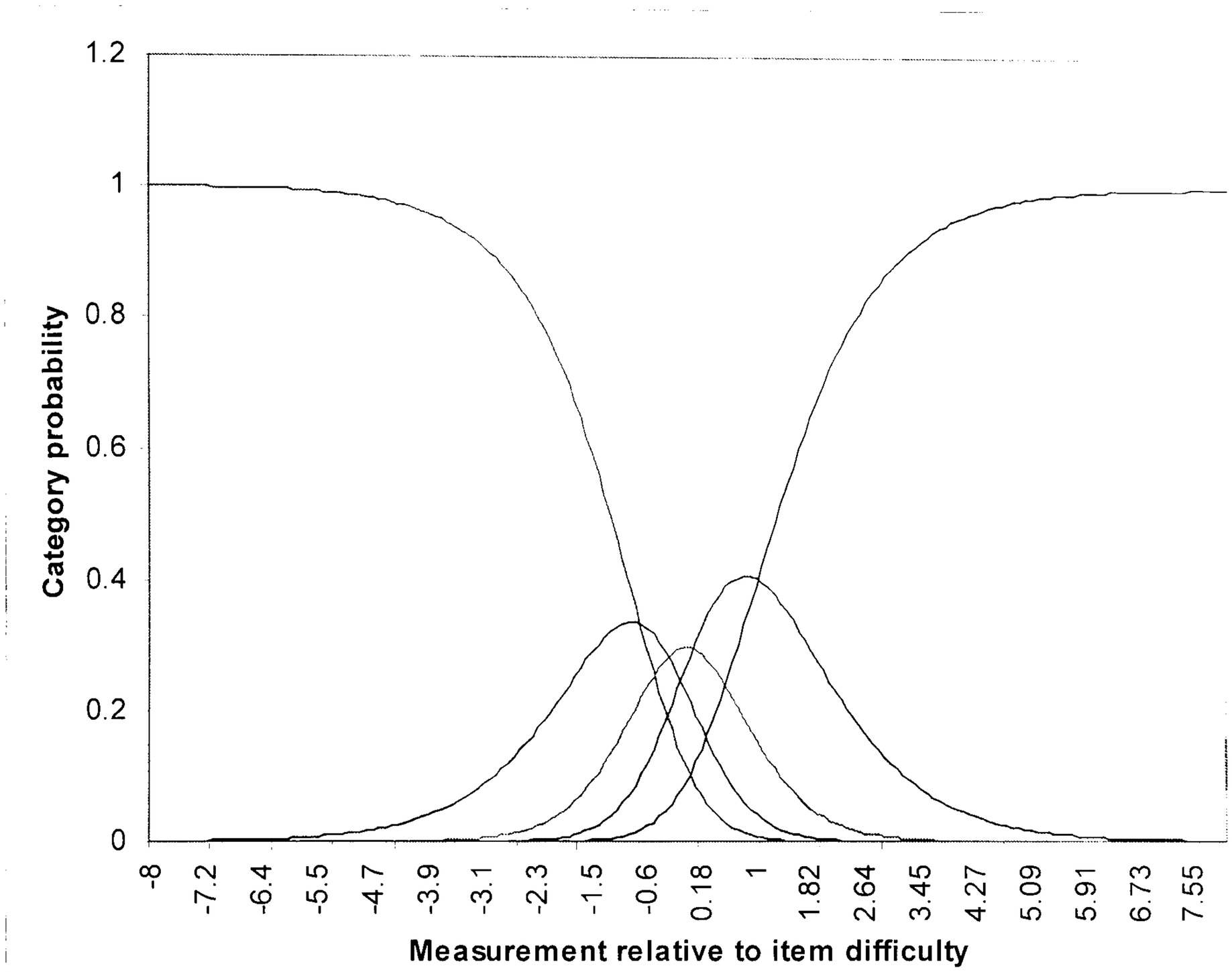


**Figure 7.2.8. Category Probability Curve for Social & Family Well-being**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Brown = category 4; 5). Black = category 5.

Although there is the expected change in probabilities of responding as a patient's social and family domain changes, i.e. as the person measure increases in relation to the item difficulty, that the likelihood of the patient responding to items with "Not at all" (category 1) or "A little" (category 2 decreases, and the likelihood of responses such as "Quite a bit" (category 3) or "Very much" (category 4) increases, there is a much greater overlap between category probabilities. For instance, the figure also shows that for person abilities around 0.18 logits, the likelihood of a person selecting categories, 3 or 4 in response to a question are equal (at approximately 30%).

Table 7.2.10 shows a summary of the items for the Social and Family Well-being Scale. The item separation index is approximately 12, demonstrating that the

Social and Family Well-being scale can distinguish between 12 levels of item difficulty.

**Table 7.2.10. Summary of Items from the Social & Family Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1401.1 | 339.4 | .00 | .07 | 1.11 | .6 | 1.09 | .3 |
| S.D. | 357.3 | 46.7 | 1.01 | .01 | .34 | 3.0 | .47 | 3.5 |
| MAX. | 1652.0 | 360.0 | 2.22 | .10 | 1.63 | 5.6 | 2.12 | 7.4 |
| MIN. | 545.0 | 225.0 | -1.14 | .06 | .75 | -3.2 | .68 | -3.3 |
| REAL RMSE | .08 | ADJ.SD | 1.00 | SEPARATION 11.97 | ITEM | RELIABILITY | .99 | |
| MODEL RMSE | .08 | ADJ.SD | 1.00 | SEPARATION 13.26 | ITEM | RELIABILITY | .99 | |
| S.E. OF ITEM MEAN = .41 | | | | | | | | |

Table 7.2.11 shows a summary of the category measures for the Social and Family Well-being Scale. It can be seen that there is a poor level of separation between all categories with less than a distance of 1.4, i.e. Linacre, 1999 between each threshold (structure measure). This confirms the observations regarding the category probability with the large overlap as shown in Figure 7.2.8. However, the distances between each category increase monotonically (category measure).

**Table 7.2.11. Summary of Category Measures for the Social & Family Well-being Scale**

| CATEGORY LABEL | SCORE | OBSERVED COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 135 | 5 | -.94 | -1.06 | 1.16 | 1.46 | NONE | ( -2.20) | 1 |
| 2 | 2 | 152 | 6 | -.11 | -.22 | 1.06 | 1.09 | -.75 | -.87 | 2 |
| 3 | 3 | 244 | 10 | .46 | .53 | .98 | 1.01 | -.32 | -.05 | 3 |
| 4 | 4 | 588 | 23 | 1.21 | 1.34 | 1.11 | .92 | .05 | .83 | 4 |
| 5 | 5 | 1257 | 50 | 2.28 | 2.23 | 1.03 | 1.02 | 1.02 | ( 2.35) | 5 |
| MISSING | | 151 | 6 | -.91 | | | | | | |

The person measures for the Social and Family Well-being scale are shown in Table 7.2.12 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 7.2.9. No estimates were provided for scores of 7,

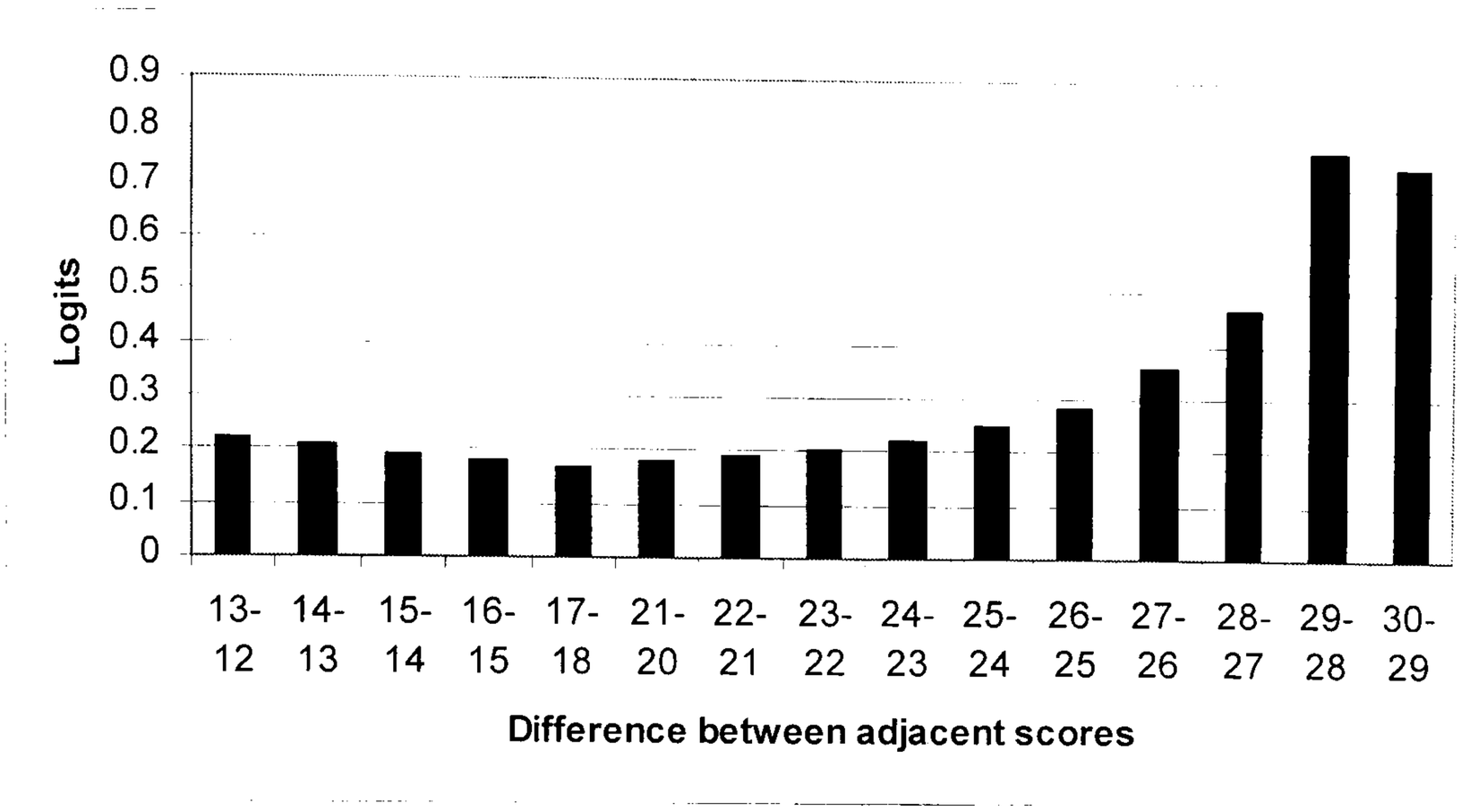
10, 11 or 19, which underlines the need to have large sample sizes for person ability estimates.

**Table 7.2.12. Person measures for Social & Family Well-being**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 6 | -4.03 | 1 | 0 | 1 | 0 |
| 9 | -2.04 | 2.14 | 1.19 | 3.28 | 1.97 |
| 12 | -1.14 | 2.27 | 1.6 | 2.18 | 1.48 |
| 13 | -0.92 | 1.15 | 0.25 | 0.99 | -0.02 |
| 14 | -0.71 | 0.86 | -0.27 | 0.77 | -0.46 |
| 15 | -0.52 | 0.29 | -1.92 | 0.27 | -1.98 |
| 16 | -0.34 | 1.37 | 0.64 | 1.38 | 0.65 |
| 17 | -0.17 | 1.93 | 1.46 | 1.9 | 1.4 |
| 18 | 0.01 | 0.3 | -1.97 | 0.33 | -1.82 |
| 20 | 0.35 | 1.68 | 1.12 | 1.68 | 1.1 |
| 21 | 0.53 | 0.82 | -0.36 | 0.85 | -0.3 |
| 22 | 0.72 | 2.82 | 2.33 | 2.71 | 2.17 |
| 23 | 0.92 | 1.63 | 0.96 | 1.41 | 0.64 |
| 24 | 1.14 | 0.37 | -1.45 | 0.37 | -1.42 |
| 25 | 1.39 | 0.97 | -0.05 | 1.08 | 0.12 |
| 26 | 1.67 | 0.63 | -0.65 | 0.6 | -0.69 |
| 27 | 2.03 | 1.52 | 0.62 | 1.36 | 0.43 |
| 28 | 2.5 | 0.44 | -0.88 | 0.39 | -0.94 |
| 29 | 3.27 | 0.65 | -0.38 | 0.42 | -0.65 |
| 30 | 4.01 | 1 | 0 | 1 | 0 |

It can be seen from Figure 7.2.9 that the distance between adjacent scores is equal between 12 and 25, however that beyond this score the difference increases.

**Figure 7.2.9. Differences between adjacent scores of the Social & Family Well-being Scale**




The summary of person measures is shown in Table 7.2.13. The person separation index is very poor at 1.20, indicating that the scale is able to detect fewer than 2 levels of person ability. The reliability measure however is also only moderately good at 0.59.

**Table 7.2.13. Summary of Person Measures for the Social & Family Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 27.2 | 6.6 | 1.35 | .59 | 1.00 | -.2 | 1.03 | -.2 |
| S.D. | 4.7 | .5 | 1.06 | .18 | .85 | 1.1 | 1.18 | 1.2 |
| MAX. | 34.0 | 7.0 | 3.56 | 1.04 | 5.77 | 3.8 | 9.14 | 5.3 |
| MIN. | 9.0 | 4.0 | -2.03 | .37 | .08 | -3.2 | .08 | -3.0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| REAL RMSE | .68 | ADJ.SD | .82 | SEPARATION | 1.20 | PERSON RELIABILITY .59 |
| MODEL RMSE | .62 | ADJ.SD | .87 | SEPARATION | 1.41 | PERSON RELIABILITY .67 |

S.E. OF PERSON MEAN = .06

MAXIMUM EXTREME SCORE: 103 PERSONS
MINIMUM EXTREME SCORE: 1 PERSONS
LACKING RESPONSES: 1 PERSONS
VALID RESPONSES: 94.0%

**Figure 7.2.10. Test Information Curve for Social & Family Well-being Scale**

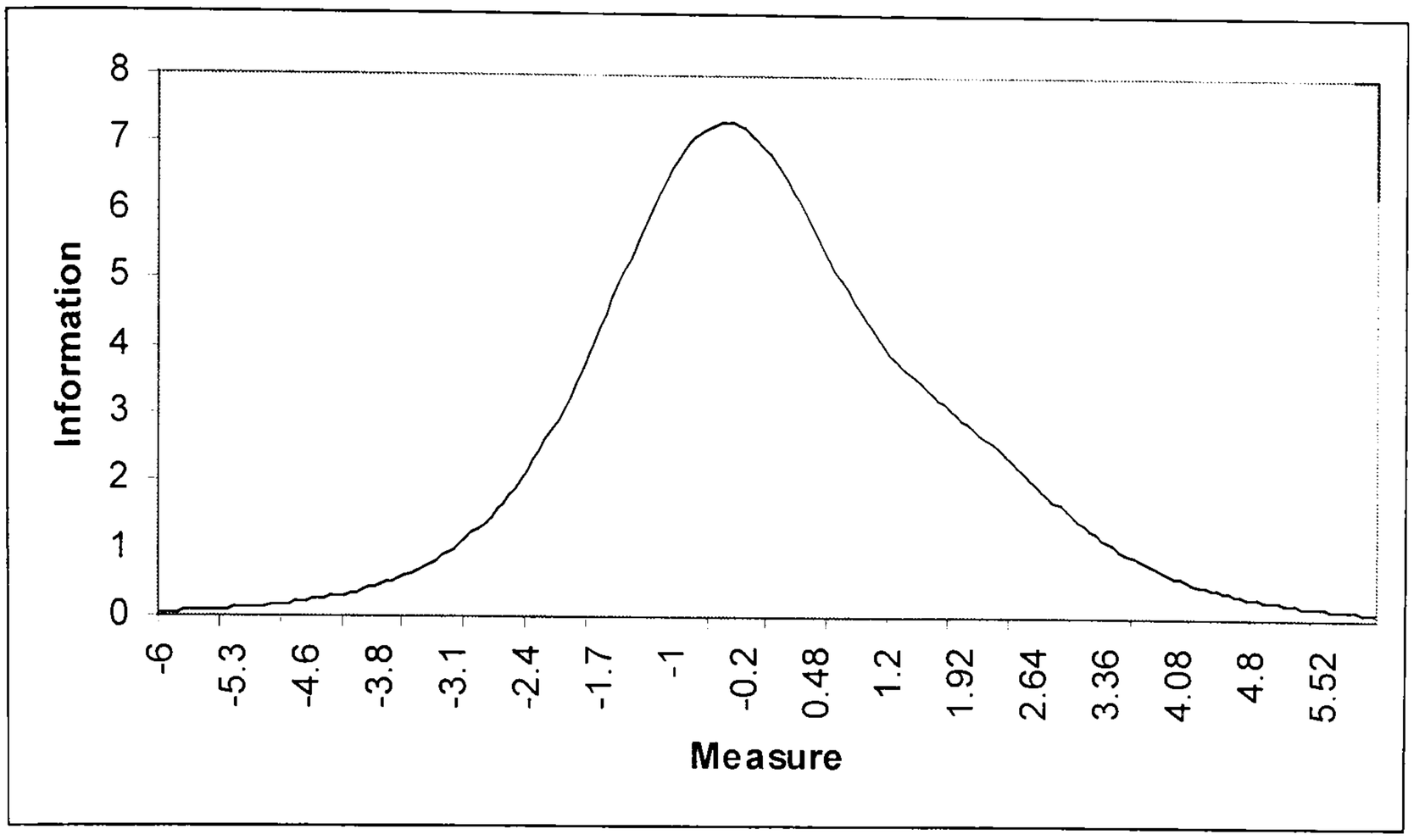



Figure 7.2.10 shows the test information curve for the Social & Family Well-being scale and demonstrates that the scale provides the most information over a narrow range at the centre of the scale (ability measures around –0.50).

In addition to the Rasch analysis described above the items from the Social & Family Well-being scale were also investigated for differential bias (as described in Chapter 6.2).

**Table 7.2.14. Differential Item Analysis of the Social & Family Well-being Scale**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | .19 | .08 | 1 | .53 | .11 | -.35 | .13 | -2.62 | 357 | 1 | GS1 sfwb1 |
| 2 | -.64 | .10 | 1 | -.80 | .15 | .16 | .18 | .89 | 358 | 2 | GS2 sfwb2 |
| 2 | -.13 | .08 | 1 | .19 | .12 | -.32 | .14 | -2.26 | 358 | 3 | GS3 sfwb3 |
| 2 | -.10 | .08 | 1 | -.42 | .14 | .32 | .16 | 2.04 | 358 | 4 | GS4 sfwb4 |
| 2 | -.43 | .09 | 1 | -.57 | .14 | .13 | .17 | .79 | 355 | 5 | GS5 sfwb5 |
| 2 | -.98 | .11 | 1 | -1.56 | .20 | .58 | .23 | 2.54 | 353 | 6 | GS6 sfwb6 |
| 2 | 2.22 | .09 | 1 | 2.23 | .12 | -.02 | .15 | -.11 | 223 | 7 | GS7 sfwb7 |

The sample was split into male and female groups for the differential item functioning analysis. The criterion used for this analysis was a p-value of 0.05 (two-tailed), however in order to control for multiple testing a Bonferoni correction was applied (0.05/5), therefore the new statistical criterion was 0.01 significance evaluated

against Student t value for infinity at 2.56. The results of this can be seen in Table 7.3.7, which demonstrates that only one of the items exhibited differential item bias, namely item 1 ("I feel close to my friends"), which female patients found easier to endorse than males, although this significance was marginal.

In summary, a Rasch analysis of the Social and Family Well-being scale of the FACT-G demonstrated good fit for five of the seven items from the scale. However, two items (SFWB5 and SFWB6) exhibited poor fit statistics. These two items deal with close relationships and patients' sex lives. In addition, the results of principal components analysis (PCA) of the residuals indicated that a factor structure remained in the residuals. Two factors emerged from the PCA, namely a factor corresponding to items dealing with "Friendship" (items 1 and 2), and factor corresponding to "Family" (items 2, 3 and 5). It can be concluded from these results that the Social & Family Well-being scale is not a unidimensional structure, but a scale that covers with three broad domains corresponding to patients' family, friendships and close relationships.

In addition, the results from the person measures demonstrated that the scale was interval based with equally spaced scores for the range of scores between 12 and 25, although beyond this score differences between adjacent scores increased. However, not all of the person ability measures could be estimated, and therefore these results must be interpreted with some caution.

*7.2.5. Results for Emotional Well-being*

The location measures and fit statistics are given in Table 7.2.15 for the Emotional Well-being scale. It can be seen from this table that four items from this scale exhibited poor fit (0.70< Infit MNSQ > 1.30, and ZSTD > 2.00) namely items 2 ("I am satisfied with how I am coping with my illness") and 3 ("I am losing hope in the fight against my illness"), which both added excessive "noise" to the model, and items 4 ("I

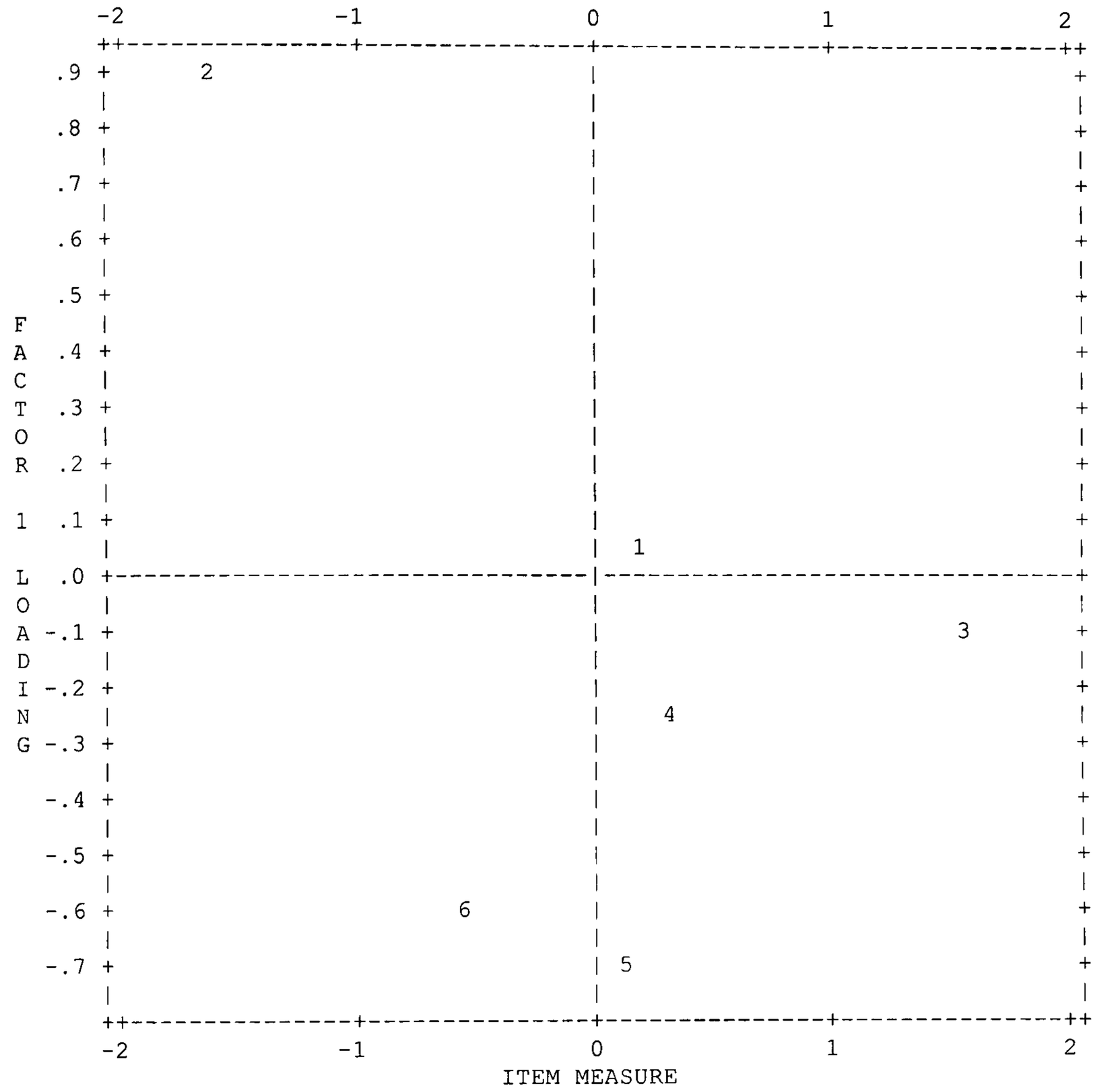
feel nervous") and 6 ("I worry that my condition will get worse") which both displayed some redundancy.

**Table 7.2.15. Unidimensionality measures for Emotional Well-being**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | | ITEMS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1796 | 457 | -1.65 | .05 | 1.93 | 9.9 | 3.71 | 9.9 | A .11 | | GE2 | ewb2 |
| 3 | 657 | 455 | 1.58 | .08 | 1.45 | 4.2 | 1.22 | 2.0 | B .53 | | GE3 | ewb3 |
| 5 | 1033 | 456 | .15 | .05 | .76 | -3.8 | .75 | -3.8 | C .74 | | GE5 | ewb5 |
| 1 | 1017 | 452 | .16 | .05 | .75 | -4.0 | .75 | -3.7 | c .70 | | GE1 | ewb1 |
| 4 | 976 | 455 | .30 | .05 | .63 | -6.2 | .62 | -5.8 | b .73 | | GE4 | ewb4 |
| 6 | 1316 | 457 | -.54 | .05 | .57 | -8.3 | .57 | -7.6 | a .78 | | GE6 | ewb6 |
| MEAN | 1133. | 455. | .00 | .06 | 1.01 | -1.4 | 1.27 | -1.5 | | | | |
| S.D. | 353. | 2. | .97 | .01 | .50 | 6.4 | 1.11 | 5.9 | | | | |

Figure 7.2.11. shows the factor plot of the principal components analysis of the standardised residuals for the Emotional Well-being scale. A factor with a total of 1.7 eigenvalues was extracted from the residuals, indicating that there were possibly other factor structures remaining in the residuals (Smith & Miao, 1994).

**Figure 7.2.11. Principal Components (Standardized Residual) Factor Plot of the Emotional Well-being Scale**




However, the factor loadings from the principal components analysis, which can be seen in Table 7.2.16, suggest that no other "meaningful" structure remains in the residuals.

**Table 7.2.16. Factor Loadings from the Principal Components Analysis of the Emotional Well-being Scale**

| | | INFIT | OUTFIT | ENTRY | | |
|---|---|---|---|---|---|---|
| FACTOR | LOADING | MEASURE | MNSQ | MNSQ | NUMBER | ITEM |
| 1 | .89 | -1.65 | 1.93 | 3.71 | 2 | 2 GE2 ewb2 |
| 1 | .03 | .16 | .75 | .75 | 1 | 1 GE1 ewb1 |
| 1 | -.72 | .15 | .76 | .75 | 5 | 5 GE5 ewb5 |
| 1 | -.60 | -.54 | .57 | .57 | 6 | 6 GE6 ewb6 |
| 1 | -.24 | .30 | .63 | .62 | 4 | 4 GE4 ewb4 |
| 1 | -.10 | 1.58 | 1.45 | 1.22 | 3 | 3 GE3 ewb3 |

The item map for the Emotional Well-being Scale can be seen in Figure 7.2.12. This figure demonstrates overlap between items 1 and 5, and item 4 is close to these items. In addition, the items do not cover the full range of person abilities (i.e. –6 to +6) with the majority of items falling in a range roughly between -2.0 and +2.0 on the ability scale.

The category probability curve for Emotional Well-being is shown in Figure 7.2.13.

**Figure 7.2.12. Logit map of all items and patients for Emotional Well-being**

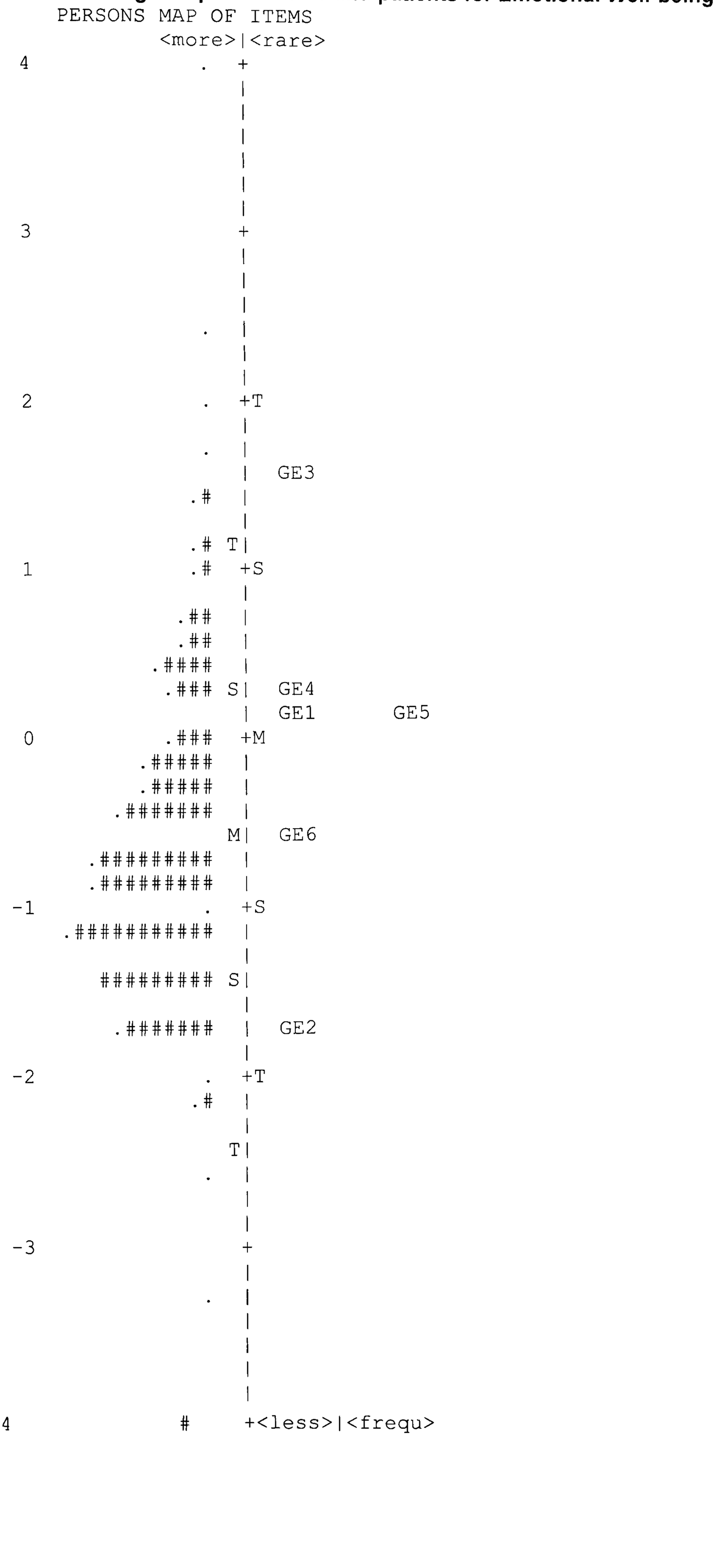
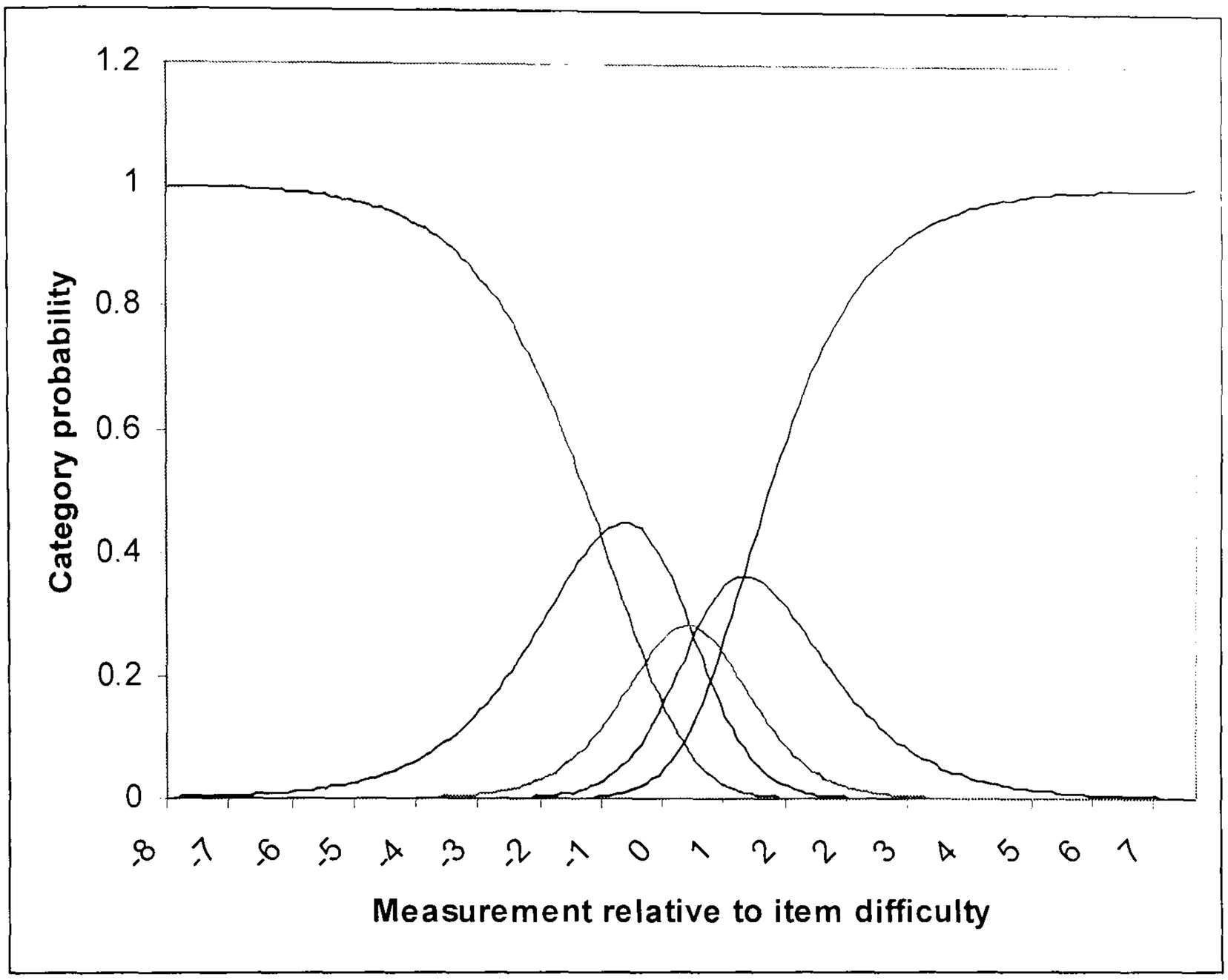


**Figure 7.2.13. Category Probability Curve for Emotional Well-being**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Brown = category 4; 5). Black = category 5.

As with the other two scales (Physical and Social & family Well-being) there is the predicted change in probabilities of responding as a patient's emotional functioning changes, however the figure also shows that for person abilities around 0, the likelihood of a person selecting specific categories overlaps considerably.

Table 7.2.17 shows a summary of the items for the Emotional Well-being Scale. The item separation index is approximately 15, demonstrating that the Emotional Well-being scale can distinguish between 15 levels of item difficulty.

**Table 7.2.17. Summary of Items from the Emotional Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1132.5 | 455.3 | .00 | .06 | 1.01 | -1.4 | 1.27 | -1.5 |
| S.D. | 353.0 | 1.7 | .97 | .01 | .50 | 6.4 | 1.11 | 5.9 |
| MAX. | 1796.0 | 457.0 | 1.58 | .08 | 1.93 | 9.9 | 3.71 | 9.9 |
| MIN. | 657.0 | 452.0 | -1.65 | .05 | .57 | -8.3 | .57 | -7.6 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| REAL RMSE | .06 | ADJ.SD | .97 | SEPARATION | 15.11 | ITEM | RELIABILITY 1.00 |
| MODEL RMSE | .06 | ADJ.SD | .97 | SEPARATION | 17.02 | ITEM | RELIABILITY 1.00 |
| S.E. OF ITEM MEAN = .43 | | | | | | | |

Table 7.2.18 shows a summary of the category measures for the Emotional Well-being Scale. It can be seen that there is a poor level of separation between categories 2 and 3, and 3 and 4 with less than a distance of 1.4, i.e. Linacre, 1999 between each of those thresholds (structure measure), as supported by the overlap between category probabilities shown in Figure 7.2.13. However, the distances between each category do increase monotonically (category measure).

**Table 7.2.18. Summary of Category Measures for the Emotional Well-being Scale**

| CATEGORY LABEL | SCORE | OBSERVED COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 833 | 30 | -1.82 | -1.69 | .78 | 1.15 | NONE | ( -2.50) | 1 |
| 2 | 2 | 785 | 29 | -.85 | -.91 | .84 | 1.28 | -1.23 | -.87 | 2 |
| 3 | 3 | 406 | 15 | .08 | -.22 | 1.13 | 1.71 | .10 | .07 | 3 |
| 4 | 4 | 366 | 13 | .61 | .47 | .77 | .82 | .23 | .92 | 4 |
| 5 | 5 | 342 | 12 | .89 | 1.22 | 1.33 | 1.42 | .91 | ( 2.30) | 5 |
| MISSING | | 10 | 0 | -.54 | | | | | | |

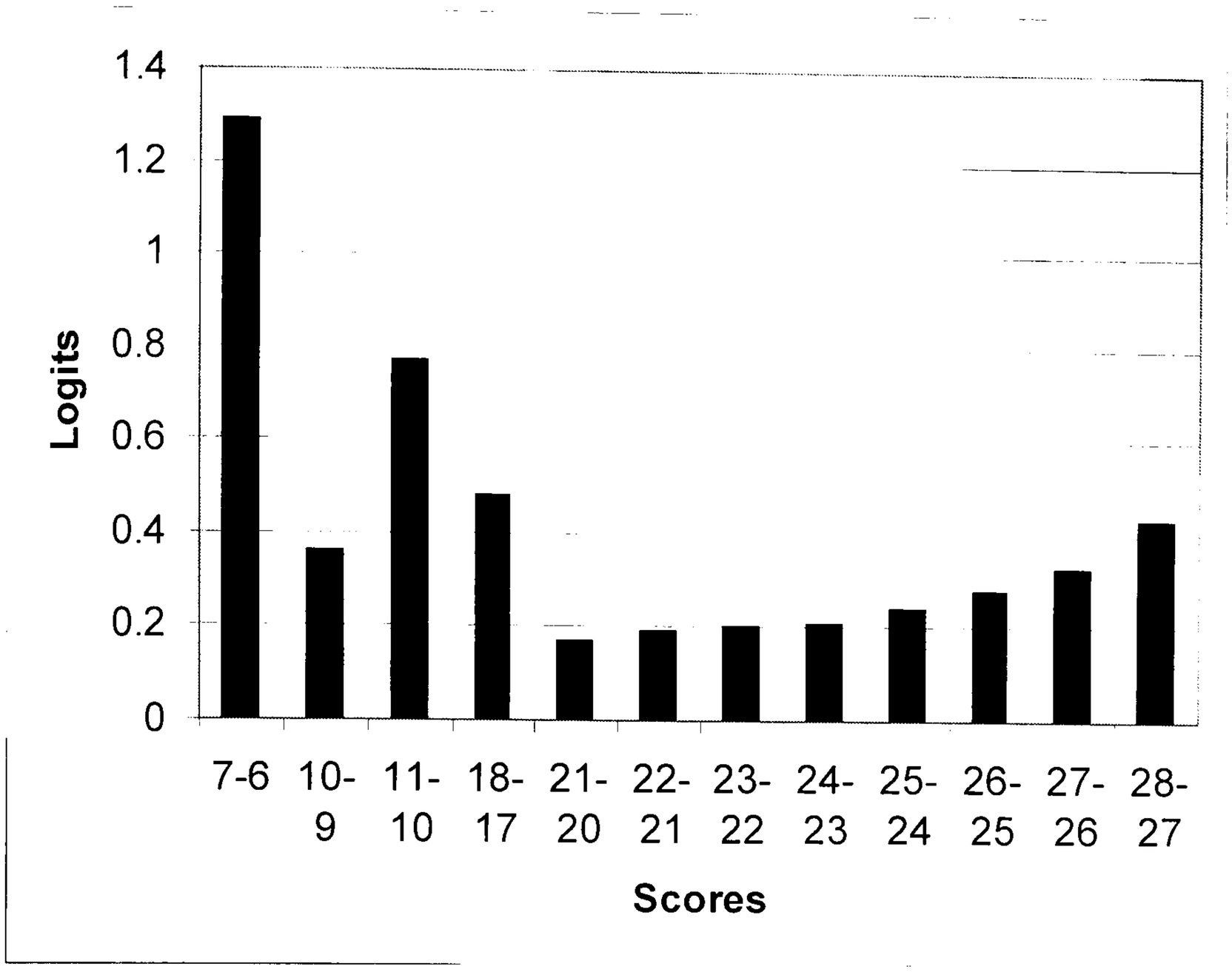
The person measures for the Emotional Well-being scale are shown in Table 7.2.19 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 7.2.14. No estimates were provided for person scores of 8, 12, 14, 16, 19 or 29.

**Table 7.2.19. Person measures for Emotional Well-being**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 6 | -4.65 | 1 | 0 | 1 | 0 |
| 7 | -3.36 | 1.05 | 0.1 | 0.79 | -0.1 |
| 8 | -1.94 | 0.51 | -0.7 | 0.62 | -0.5 |
| 9 | -2.08 | 1.28 | 0.3 | 0.69 | -0.4 |
| 10 | -1.72 | 1.9 | 1 | 1.05 | 0.1 |
| 11 | -0.95 | 0.63 | -0.6 | 0.55 | -0.7 |
| 12 | 0.43 | 0.07 | -2.2 | 0.12 | -1.6 |
| 13 | -0.92 | 0.1 | -2.6 | 0.11 | -2.3 |
| 14 | 0.49 | 0.89 | -0.1 | 0.94 | -0.1 |
| 15 | -0.49 | 0.95 | -0.1 | 1.07 | 0.1 |
| 16 | 0.23 | 1.2 | 0.3 | 1.9 | 1.1 |
| 17 | -0.12 | 0.26 | -2 | 0.3 | -1.7 |
| 18 | 0.36 | 1.59 | 0.9 | 2.53 | 1.7 |
| 19 | 0.23 | 0.51 | -1.2 | 0.45 | -1.2 |
| 20 | 0.41 | 0.76 | -0.5 | 0.85 | -0.3 |
| 21 | 0.58 | 1.26 | 0.5 | 1.33 | 0.5 |
| 22 | 0.77 | 1.08 | 0.1 | 1.4 | 0.5 |
| 23 | 0.97 | 0.23 | -1.9 | 0.73 | -0.4 |
| 24 | 1.18 | 0.44 | -1.1 | 1.09 | 0.1 |
| 25 | 1.42 | 0.89 | -0.2 | 0.79 | -0.3 |
| 26 | 1.7 | 1.55 | 0.7 | 2.3 | 1 |
| 27 | 2.03 | 1.68 | 0.7 | 9.9 | 3.3 |
| 28 | 2.46 | 2.65 | 1.3 | 9.9 | 3.6 |
| 30 | 4.35 | 1 | 0 | 1 | 0 |

It can be seen from Figure 7.2.19 that the distance between adjacent scores is equal between 20 and 25, however the difference increases on either side of this range.

**Figure 7.2.14. Differences between adjacent scores of the Emotional Well-being Scale**




The summary of person measures is shown in Table 7.2.20. The person separation index is very poor at 1.28, indicating that the scale is able to detect fewer than 2 levels of person ability. The reliability measure however is also only moderately good at 0.62.

**Table 7.2.20. Summary of Person Measures for the Emotional Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 14.9 | 6.0 | -.59 | .48 | .97 | -.3 | 1.22 | -.1 |
| S.D. | 4.2 | .2 | .89 | .08 | .73 | 1.2 | 1.62 | 1.4 |
| MAX. | 28.0 | 6.0 | 2.46 | 1.06 | 5.76 | 4.2 | 9.90 | 5.8 |
| MIN. | 7.0 | 3.0 | -3.36 | .42 | .04 | -3.3 | .07 | -2.8 |

| | | | | | |
|---|---|---|---|---|---|
| REAL RMSE | .55 | ADJ.SD | .70 | SEPARATION 1.28 | PERSON RELIABILITY .62 |
| MODEL RMSE | .49 | ADJ.SD | .75 | SEPARATION 1.52 | PERSON RELIABILITY .70 |

S.E. OF PERSON MEAN = .04

MAXIMUM EXTREME SCORE: 1 PERSONS
MINIMUM EXTREME SCORE: 5 PERSONS
LACKING RESPONSES: 3 PERSONS
VALID RESPONSES: 99.6%

**Figure 7.2.15. Test Information Curve for Emotional Well-being Scale**

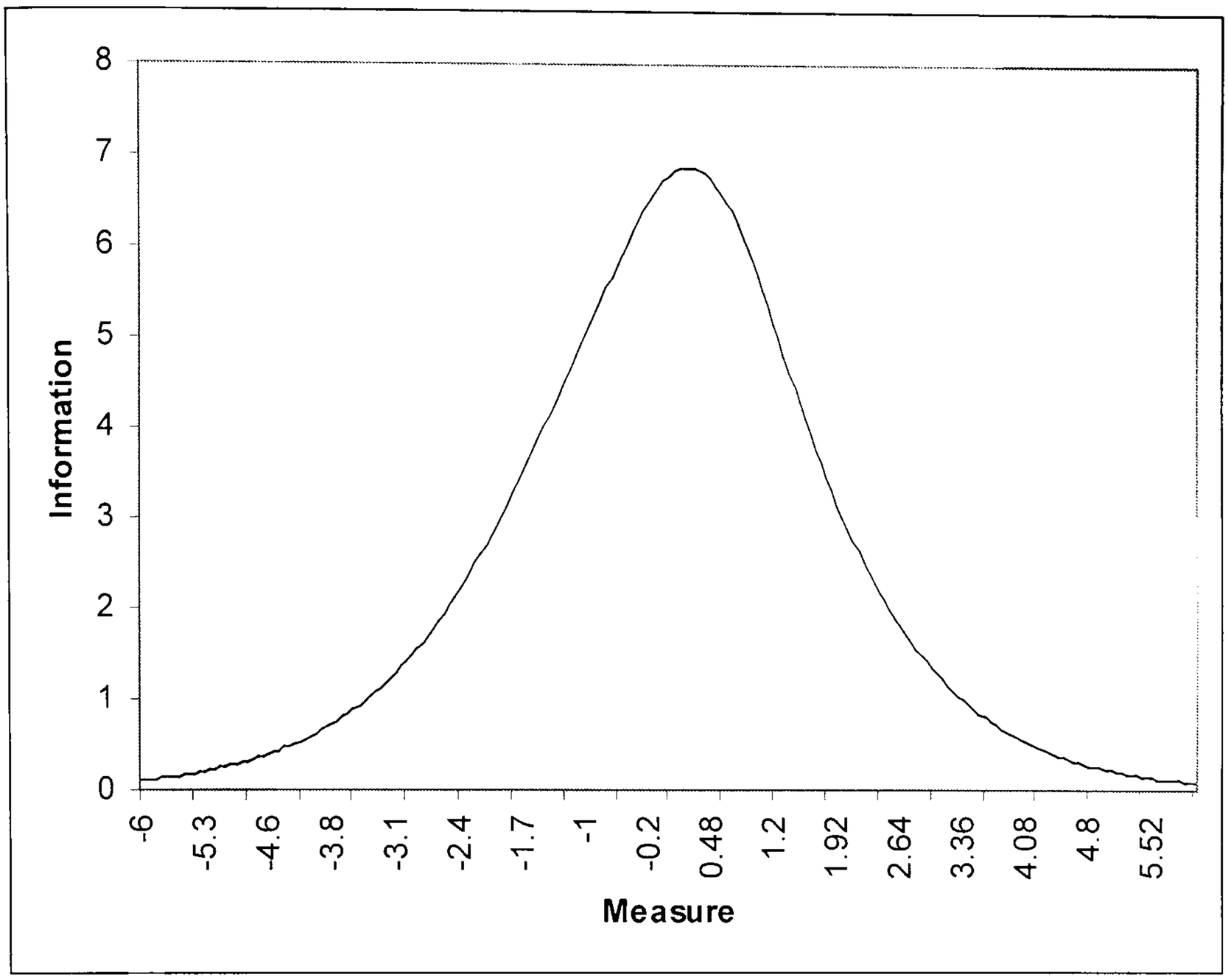



Figure 7.2.15 shows the test information curve for the Emotional Well-being scale and demonstrates that the scale provides the most information over a narrow range at the centre of the scale (ability measures around 0 logits).

In addition to the Rasch analysis described above the items from the Emotional Well-being scale were also investigated for differential bias (as described in Chapter 6.2).

**Table 7.2.21. Differential Item Analysis of the Emotional Well-being Scale**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | .13 | .06 | 1 | .26 | .10 | -.13 | .12 | -1.05 | 450 | 1 | GE1 | ewb1 |
| 2 | -1.55 | .06 | 1 | -1.85 | .09 | .30 | .11 | 2.71 | 455 | 2 | GE2 | ewb2 |
| 2 | 1.64 | .09 | 1 | 1.39 | .15 | .25 | .18 | 1.41 | 453 | 3 | GE3 | ewb3 |
| 2 | .23 | .06 | 1 | .49 | .11 | -.26 | .13 | -2.07 | 453 | 4 | GE4 | ewb4 |
| 2 | .11 | .06 | 1 | .25 | .10 | -.14 | .12 | -1.20 | 454 | 5 | GE5 | ewb5 |
| 2 | -.53 | .06 | 1 | -.57 | .09 | .04 | .10 | .39 | 455 | 6 | GE6 | ewb6 |

The sample was split into male and female groups for the differential item functioning analysis. The criterion used for this analysis was a p-value of 0.05 (two-tailed),

however in order to control for multiple testing a Bonferoni correction was applied (0.05/5), therefore the new statistical criterion was 0.01 significance evaluated against Student t value for infinity at 2.56. The results of this can be seen in Table 7.2.21, which demonstrates that only one of the items exhibited differential item bias, namely item 2 ("I am satisfied with how I am coping with my illness"), which female patients found slightly harder to endorse than males, although this significance was marginal.

In summary, a Rasch analysis of the Emotional Well-being scale of the FACT-G demonstrated poor fit for four of the six items from the scale. Two of these items (EWB2 and EWB3) deal with how patients are coping with their illness, and these two items did not fit the Rasch model. The remaining items, including the two redundant items (EWB4 and EWB6) deal with sadness, nervousness and worry. The results of principal components analysis (PCA) of the residuals indicated that no further factor structures remained in the residuals. It can be concluded from these results that the Emotional Well-being scale, aside from the two items dealing with coping is a unidimensional structure.

In addition, the results from the person measures were inconclusive given that 5 out a possible total of 30 person measures could not be estimated. However, it appeared that scores falling in the range between 20 and 25 were interval spaced, although beyond this score differences between adjacent scores increased at both extremes.

*7.2.6. Results for Functional Well-being*

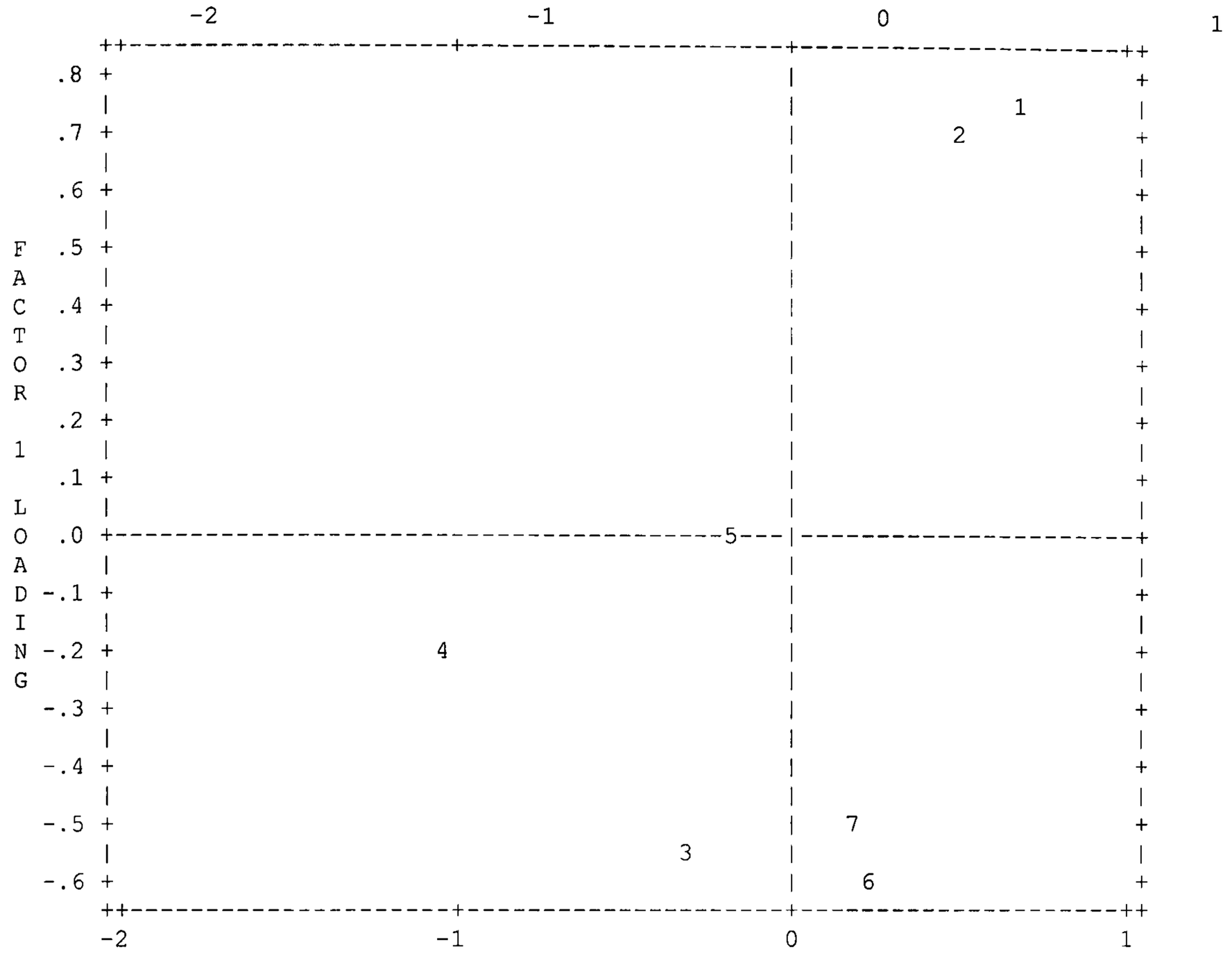
The location measures and fit statistics are given in Table 7.2.22 for the Functional Well-being scale. It can be seen from this table that two items from this scale exhibited poor fit (0.70< Infit MNSQ > 1.30, and ZSTD > 2.00) namely item 5 ("I am sleeping well"), which did not fit the Rasch model, and item 3 ("I am able to enjoy life") displayed some redundancy.

**Table 7.2.22. Unidimensionality measures for Functional Well-being**

| ENTRY NUMBER | RAW SCORE | COUNT | MEASURE | ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEA CORR. | ITEMS | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1514 | 439 | -.20 | .05 | 1.57 | 7.5 | 1.75 | 8.6 | A .53 | GF5 | fwb5 |
| 4 | 1784 | 440 | -1.05 | .06 | 1.36 | 4.5 | 1.56 | 5.4 | B .50 | GF4 | fwb4 |
| 1 | 1209 | 441 | .68 | .05 | 1.10 | 1.6 | 1.08 | 1.2 | C .72 | GF1 | fwb1 |
| 2 | 1259 | 433 | .50 | .05 | 1.02 | .4 | .98 | -.3 | D .73 | GF2 | fwb2 |
| 7 | 1391 | 441 | .17 | .05 | .78 | -3.7 | .80 | -3.1 | c .77 | GF7 | fwb7 |
| 6 | 1373 | 440 | .21 | .05 | .70 | -5.3 | .69 | -5.2 | b .79 | GF6 | fwb6 |
| 3 | 1566 | 442 | -.31 | .05 | .48 | -9.8 | .49 | -8.7 | a .78 | GF3 | fwb3 |
| MEAN | 1442. | 439. | .00 | .05 | 1.00 | -.7 | 1.05 | -.3 | | | |
| S.D. | 182. | 3. | .54 | .00 | .35 | 5.5 | .43 | 5.6 | | | |

Figure 7.2.16. shows the factor plot of the principal components analysis of the standardised residuals for the Functional Well-being scale. A factor with a total of 2.0 eigenvalues was extracted from the residuals, indicating that there were possibly other factor structures remaining in the residuals (Smith & Miao, 1994).

**Figure 7.2.16. Principal Components (Standardized Residual) Factor Plot of the Functional Well-being Scale**




The factor loadings from the principal components analysis, which can be seen in Table 7.2.23, suggest that two other structures remain in the residuals, namely a factor corresponding to work issues (i.e. item 1 “I am able to work (include work at home)” and item 2 “My work (include work at home) is fulfilling”), and a second factor corresponding to enjoyment of life (i.e. item 3, “I am able to enjoy life”, item 6 “I am enjoying the things I usually do for fun”, and item 7, “I am content with the quality of my life right now”).

**Table 7.2.23. Factor Loadings from the Principal Components Analysis of the Functional Well-being Scale**

| FACTOR | LOADING | INFIT MEASURE | OUTFIT MNSQ | OUTFIT MNSQ | ENTRY NUMBER | ITEM | |
|---|---|---|---|---|---|---|---|
| 1 | .77 | .68 | 1.10 | 1.08 | 1 | 1 GF1 | fwb1 |
| 1 | .68 | .50 | 1.02 | .98 | 2 | 2 GF2 | fwb2 |
| 1 | -.60 | .21 | .70 | .69 | 6 | 6 GF6 | fwb6 |
| 1 | -.54 | -.31 | .48 | .49 | 3 | 3 GF3 | fwb3 |
| 1 | -.52 | .17 | .78 | .80 | 7 | 7 GF7 | fwb7 |
| 1 | -.18 | -1.05 | 1.36 | 1.56 | 4 | 4 GF4 | fwb4 |
| 1 | .00 | -.20 | 1.57 | 1.75 | 5 | 5 GF5 | fwb5 |

The item map for the Functional Well-being Scale can be seen in Figure 7.2.17. This figure demonstrates overlap between items 6 and 7, which deal with enjoyment and quality of life. However, all items are very close together and do not cover the full range of person abilities (i.e –6 to +6) with the majority of items falling in a narrow range roughly between -1.0 and +1.0 on the ability scale.

The category probability curve for Functional Well-being is shown in Figure 7.2.18.

**Figure 7.2.17. Logit map of all items and patients for Functional Well-being**

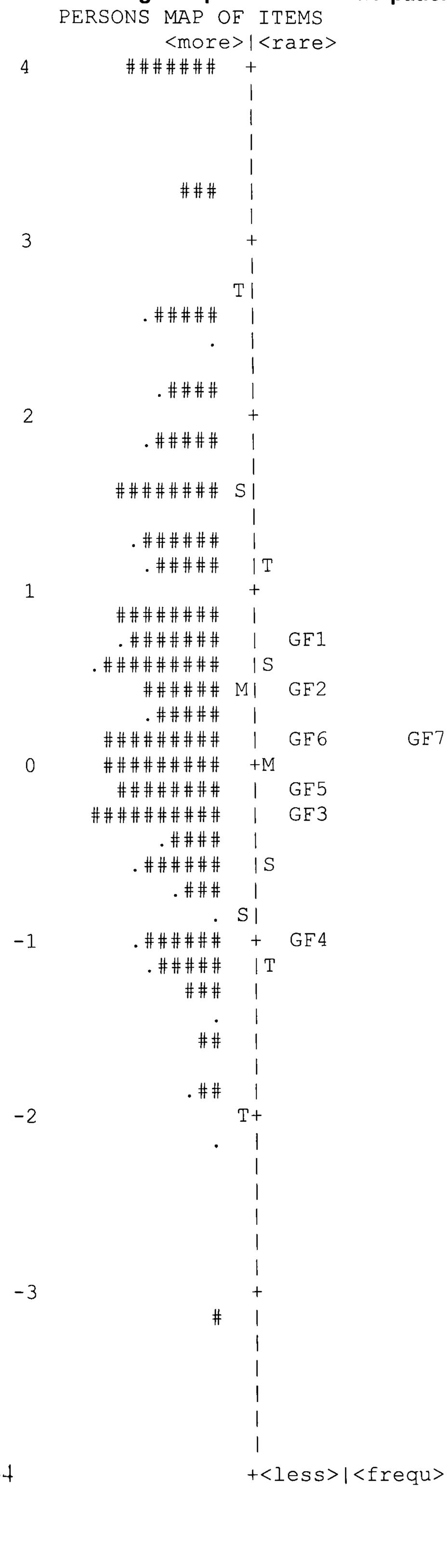
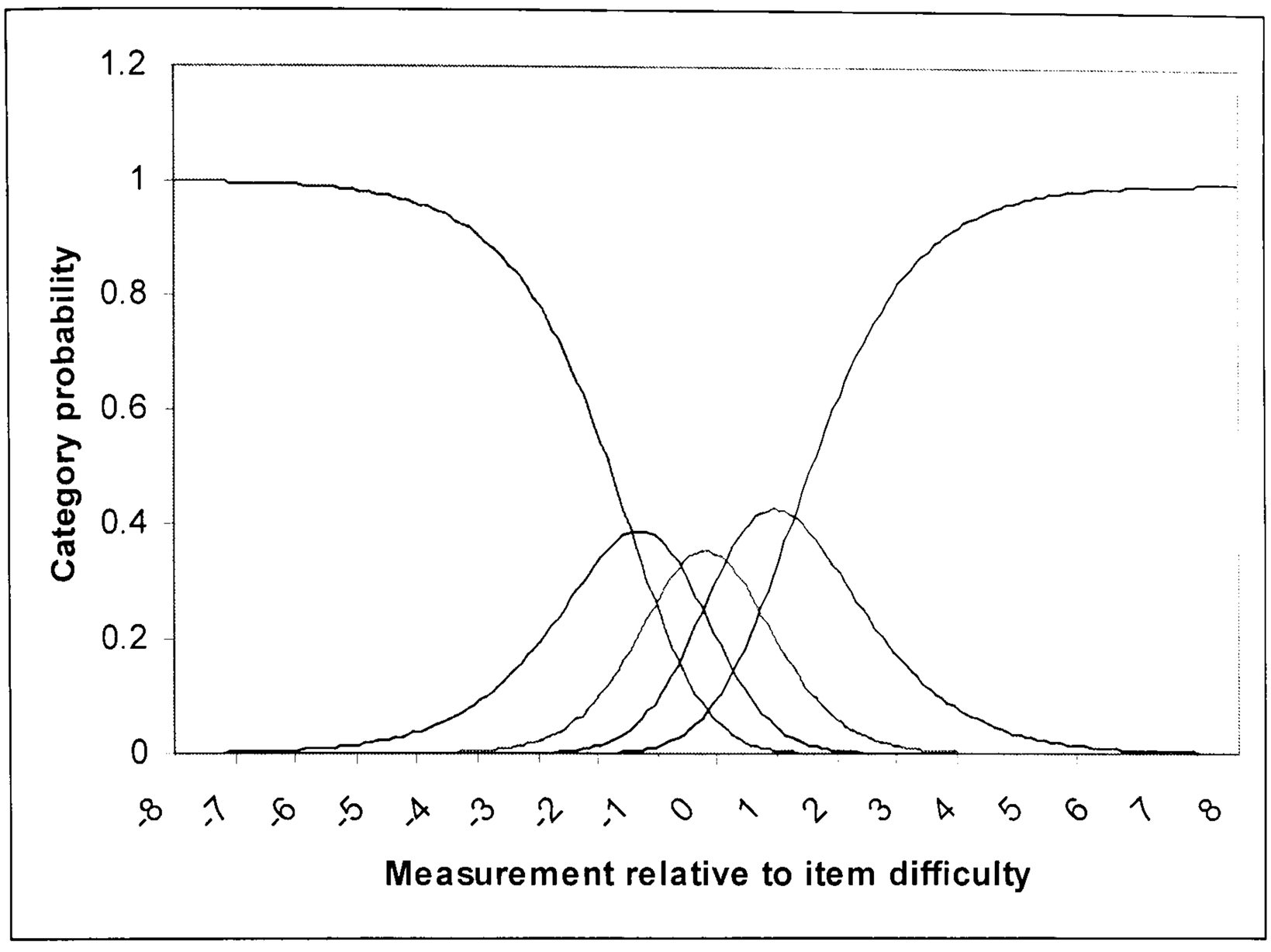


**Figure 7.2.18. Category Probability Curve for Functional Well-being**




*Key for the category probability curve: 1). Red = category 1;2). Blue = category 2; 3). Pink = category 3; 4). Brown = category 4; 5). Black = category 5.

As with the other scales from the FACT-G there is the predicted change in probabilities of responding as a patient's functional level changes, however the figure also shows that for person abilities around 0, the likelihood of a person selecting specific categories overlaps considerably.

Table 7.2.24 shows a summary of the items for the Functional Well-being Scale. The item separation index is approximately 9, demonstrating that the Functional Well-being scale can distinguish between 9 levels of item difficulty.

**Table 7.2.24. Summary of Items from the Functional Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 1442.3 | 439.4 | .00 | .05 | 1.00 | -.7 | 1.05 | -.3 |
| S.D. | 182.3 | 2.8 | .54 | .00 | .35 | 5.5 | .43 | 5.6 |
| MAX. | 1784.0 | 442.0 | .68 | .06 | 1.57 | 7.5 | 1.75 | 8.6 |
| MIN. | 1209.0 | 433.0 | -1.05 | .05 | .48 | -9.8 | .49 | -8.7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| REAL RMSE | .06 | ADJ.SD | .53 | SEPARATION | 9.12 | ITEM | RELIABILITY | .99 |
| MODEL RMSE | .05 | ADJ.SD | .53 | SEPARATION | 9.85 | ITEM | RELIABILITY | .99 |
| S.E. OF ITEM MEAN = .22 | | | | | | | | |

Table 7.2.25 shows a summary of the category measures for the Functional Well-being Scale. It can be seen that there is a poor level of separation between categories 1 and 2, 2 and 4 and 5 with less than a distance of 1.4, i.e. Linacre, 1999 between each of those thresholds (structure measure). However, the distances between each category do increase monotonically (category measure).

**Table 7.2.25. Summary of Category Measures for the Functional Well-being Scale**

| CATEGORY LABEL | OBSERVED SCORE | COUNT | % | OBSVD AVRGE | SAMPLE EXPECT | INFIT MNSQ | OUTFIT MNSQ | STRUCTURE MEASURE | CATEGORY MEASURE | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 380 | 12 | -1.12 | -1.11 | 1.02 | 1.05 | NONE | ( -2.53) | 1 |
| 2 | 2 | 547 | 18 | -.55 | -.50 | .93 | 1.08 | -1.17 | -1.03 | 2 |
| 3 | 3 | 673 | 22 | .12 | .09 | .91 | 1.00 | -.42 | -.03 | 3 |
| 4 | 4 | 777 | 25 | .84 | .78 | .81 | .90 | .27 | 1.01 | 4 |
| 5 | 5 | 699 | 23 | 1.61 | 1.67 | 1.19 | 1.19 | 1.31 | ( 2.63) | 5 |
| MISSING | | 18 | 1 | -.72 | | | | | | |

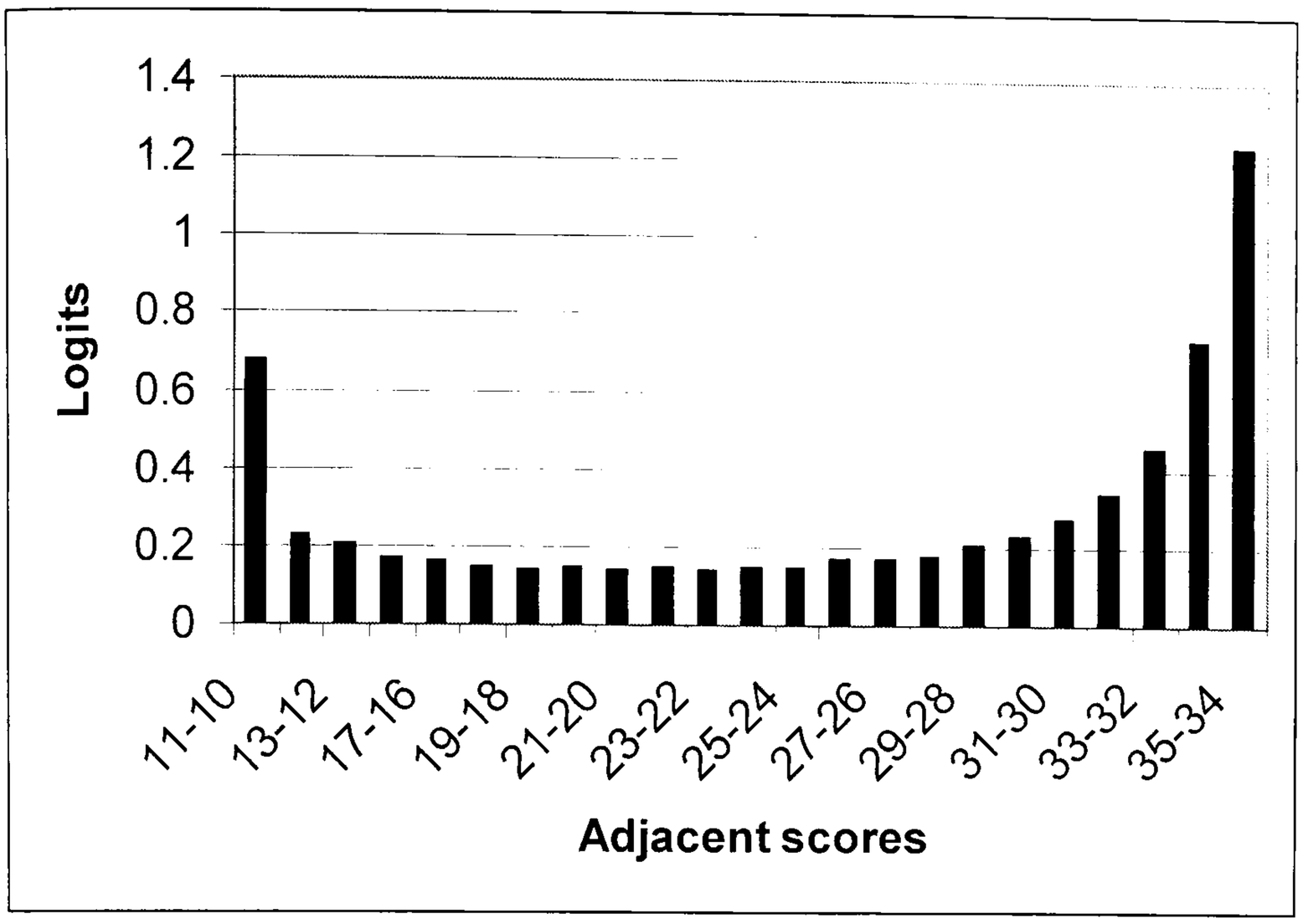
The person measures for the Functional Well-being scale are shown in Table 7.2.26 and the distances between adjacent raw scores (person measures) are represented graphically in Figure 7.2.19. No estimates were provided for person scores of 8, 9, or 14.

**Table 7.2.26. Person measures for Functional Well-being**

| SCORE | MEASURE | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD |
|---|---|---|---|---|---|
| 7 | -3.15 | 1.12 | 0.1 | 2.24 | 0.7 |
| 10 | -2.11 | 0.69 | -0.5 | 1.09 | 0.1 |
| 11 | -1.43 | 1.73 | 1 | 1.69 | 0.9 |
| 12 | -1.2 | 1.28 | 0.5 | 1.06 | 0.1 |
| 13 | -0.99 | 1.4 | 0.7 | 1.24 | 0.4 |
| 15 | -0.93 | 0.39 | -1.6 | 0.36 | -1.7 |
| 16 | -0.76 | 1.47 | 0.8 | 1.51 | 0.9 |
| 17 | -0.6 | 0.92 | -0.2 | 0.88 | -0.2 |
| 18 | -0.45 | 0.69 | -0.7 | 0.64 | -0.9 |
| 19 | -0.31 | 0.65 | -0.9 | 0.65 | -0.9 |
| 20 | -0.16 | 2.42 | 2.2 | 2.5 | 2.3 |
| 21 | -0.02 | 1.33 | 0.6 | 1.29 | 0.6 |
| 22 | 0.13 | 1.76 | 1.3 | 2 | 1.7 |
| 23 | 0.27 | 0.62 | -0.9 | 0.66 | -0.8 |
| 24 | 0.42 | 1.66 | 1.2 | 1.53 | 0.9 |
| 25 | 0.57 | 0.54 | -1.1 | 0.56 | -1 |
| 26 | 0.74 | 1.2 | 0.4 | 1.41 | 0.7 |
| 27 | 0.91 | 0.67 | -0.7 | 0.69 | -0.7 |
| 28 | 1.09 | 0.68 | -0.7 | 0.76 | -0.5 |
| 29 | 1.3 | 0.38 | -1.5 | 0.48 | -1.1 |
| 30 | 1.53 | 2.3 | 1.6 | 2.45 | 1.7 |
| 31 | 1.8 | 1.56 | 0.8 | 1.57 | 0.7 |
| 32 | 2.14 | 1.34 | 0.4 | 1.16 | 0.2 |
| 33 | 2.6 | 0.5 | -0.8 | 0.41 | -0.9 |
| 34 | 3.34 | 0.92 | -0.1 | 0.85 | -0.1 |
| 35 | 4.58 | 1 | 0 | 1 | 0 |

It can be seen from Figure 7.2.19 that the distance between adjacent scores is equal between 12 and 30, however the difference increases on either side of this range.

**Figure 7.2.19. Differences between adjacent scores of the Functional Well-being Scale**




The summary of person measures is shown in Table 7.2.27. The person separation index is close to the minimum person separation index requirement of 2 at 1.94. The reliability measure however is also good at 0.79.

**Table 7.2.27. Summary of Person Measures for the Functional Well-being Scale**

| | RAW SCORE | COUNT | MEASURE | MODEL ERROR | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD |
|---|---|---|---|---|---|---|---|---|
| MEAN | 22.8 | 7.0 | .36 | .46 | 1.01 | -.3 | 1.05 | -.2 |
| S.D. | 6.2 | .2 | 1.16 | .13 | .69 | 1.3 | .77 | 1.3 |
| MAX. | 34.0 | 7.0 | 3.34 | 1.03 | 4.42 | 3.3 | 4.99 | 3.5 |
| MIN. | 7.0 | 5.0 | -3.18 | .38 | .08 | -3.6 | .08 | -3.6 |

| | | | | | | |
|---|---|---|---|---|---|---|
| REAL RMSE | .53 | ADJ.SD | 1.03 | SEPARATION | 1.94 | PERSON RELIABILITY .79 |
| MODEL RMSE | .47 | ADJ.SD | 1.05 | SEPARATION | 2.22 | PERSON RELIABILITY .83 |
| S.E. OF PERSON MEAN = .06 | | | | | | |

MAXIMUM EXTREME SCORE: 21 PERSONS
LACKING RESPONSES: 3 PERSONS
VALID RESPONSES: 99.4%

**Figure 7.2.20. Test Information Curve for Functional Well-being Scale**

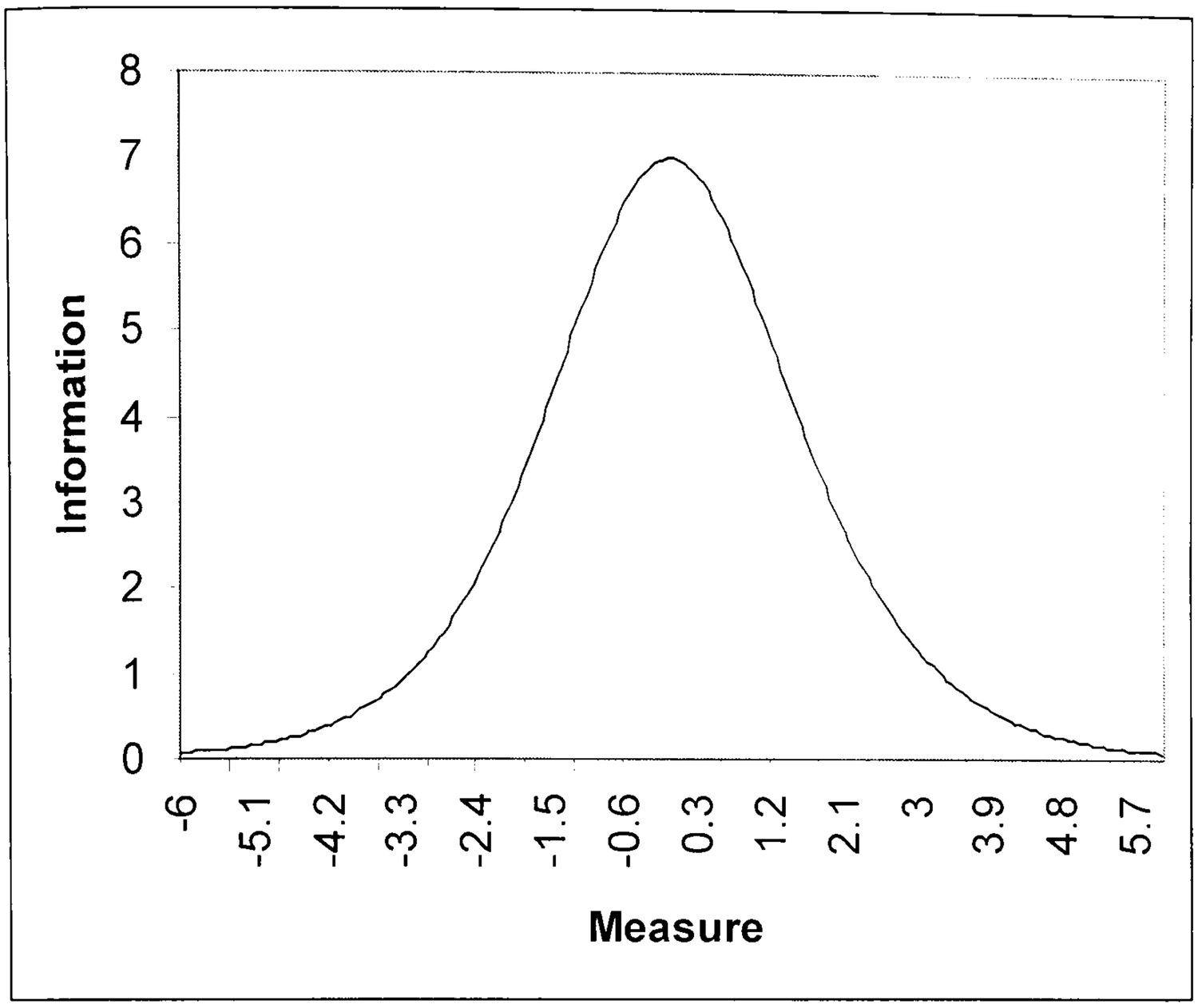



Figure 7.2.20 shows the test information curve for the Functional Well-being scale and demonstrates that the scale provides the most information over a narrow range at the centre of the scale (ability measures around –0.20 logits).

In addition to the Rasch analysis described above the items from the Functional Well-being scale were also investigated for differential bias (as described in Chapter 6.2).

**Table 7.2.28. Differential Item Analysis of the Functional Well-being Scale**

| PERSON GROUP | DIF MEASURE | DIF S.E. | PERSON GROUP | DIF MEASURE | DIF S.E. | DIF CONTRAST | JOINT S.E. | t | d.f. | ITEM Number | Name | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | .58 | .06 | 1 | .89 | .10 | -.31 | .12 | -2.69 | 439 | 1 | GF1 | fwb1 |
| 2 | .47 | .06 | 1 | .56 | .10 | -.09 | .11 | -.78 | 431 | 2 | GF2 | fwb2 |
| 2 | -.29 | .06 | 1 | -.36 | .10 | .07 | .12 | .58 | 440 | 3 | GF3 | fwb3 |
| 2 | -1.02 | .07 | 1 | -1.10 | .11 | .07 | .13 | .54 | 438 | 4 | GF4 | fwb4 |
| 2 | -.19 | .06 | 1 | -.21 | .10 | .02 | .12 | .20 | 437 | 5 | GF5 | fwb5 |
| 2 | .26 | .06 | 1 | .10 | .10 | .16 | .11 | 1.39 | 438 | 6 | GF6 | fwb6 |
| 2 | .20 | .06 | 1 | .10 | .10 | .10 | .11 | .85 | 439 | 7 | GF7 | fwb7 |

The sample was split into male and female groups for the differential item functioning analysis. The criterion used for this analysis was a p-value of 0.05 (two-tailed),

however in order to control for multiple testing a Bonferoni correction was applied (0.05/5), therefore the new statistical criterion was 0.01 significance evaluated against Student t value for infinity at 2.56. The results of this can be seen in Table 7.2.28, which demonstrates that only one of the items exhibited differential item bias, namely item 1 ("I am able to work (include work at home)"), which female patients found slightly harder to endorse than males, although this significance was marginal.

In summary, a Rasch analysis of the Functional Well-being scale of the FACT-G demonstrated poor fit for two of the seven items from the scale. One of these items (item 5, "I am sleeping well") did not fit the Rasch model. The results of principal components analysis (PCA) of the residuals indicated that two further factor structures remained in the residuals, relating to work (items 1 and 2), and enjoyment of life (items 3, 6 and 7). It can be concluded from these results that the Functional Well-being scale is therefore not unidimensional.

In addition, the results from the person measures demonstrated that the Functional Well-being scale was largely interval based within the range of scores between 13 – 30, although given that some person measures could not be estimated this should be interpreted with some caution.

## 7.3. Discussion

This chapter describes a series of analyses which were carried out on the subscales of the FACT-G (Cella et al., 1993). The results from the initial (rotated) factor analysis (principal components analysis) demonstrated a four factor structure corresponding to the FACT-G subscales, namely Physical, Social and Family, Emotional and Functional Well-being. The results of the reliability analysis of these scales demonstrated good levels for Cronbach's alpha. These results are similar to previous findings from studies using earlier versions of the FACT-G (version 2, e.g. Cella et al., 1993; Winstead-Fry and Schultz, 1997).

The results from the subsequent Rasch analysis demonstrated that only the factor structure of the Physical Well-being scale was unidimensional. The results of the Rasch analysis of the Social and Family Well-being scale demonstrated two factors corresponding to family concerns and friendships, as well as a factor relating to close relationships. The analysis of the Emotional Well-being scale suggested an additional two factors relating to coping and hope, and sadness, nervousness and worry. Two further factors were also revealed for the Functional Well-being scale relating to work and enjoyment of life.

The results of the Rasch analysis differ from those from the traditional psychometrics (with the exception of the Physical Well-being scale). This could be explained by the fact that highly correlated ordinal data can strongly influence factor analyses (Wright, 1996). It is interesting to note that the original development of the FACT-G included a Rasch analysis of potential candidate questions, and that the fit statistics generated by the analyses were used as exclusion criteria for ten of the original items (Cella et al., 1993). It is not reported whether Rasch analyses were carried out subsequently on the remaining items to assess fit.

These results could have important ramifications given the increasing use of the FACT-G and other quality-of-life measures in clinical practice (specifically in facilitating doctor-patient communication, e.g., Cella et al., 2002a, b; Detmar et al., 2003; Miller, Pittman & Strong, 2003; Taenzer et al., 2000; Velikova et al., 2002), as well as in clinical trials and in interpreting changes in quality of life (e.g. Osoba et al., 1998). Clearly caution needs to be exercised in interpreting the results from three of the four FACT-G subscales, particularly when employing a single score as an index of a clinically meaningful difference (e.g. Jaeschke et al., 1989), since if the subscales do not represent a single underlying construct it becomes difficult – if not almost meaningless – to draw valid conclusions from a shift in scores.

Unfortunately, the lack estimates for the person measures for most scales means that it is difficult to infer whether the FACT-G subscales are interval-based or

not. Most subscales demonstrated equal distances between adjacent scores for a range of person estimates, as well as larger distances at either extreme. However, this will have to be confirmed with an analysis of larger sample sizes (e.g. >1500, such as the analysis of the EORTC QLQ-c30 subscales).

Finally, all of the FACT-G subscales demonstrated overlap between the category threshold, particularly for category three, corresponding to the "somewhat" response category. Future work could explore the effect of removing this category (by re-categorising either category 2 or 4), and investigate the effect of this reclassification on item fit.

# 8. Item Reduction of HADS

## 8.1. Introduction

Anxiety and depression are common problems in patients diagnosed with cancer. Two recent reviews of the literature have estimated that prevalence of anxiety disorders in cancer patients ranges between 7% - 23% (Stark and House, 2000), whilst prevalence of depression has been reported to range between 7% and 47% (Sellick and Crooks, 1999). Clearly there is a need for oncologists to be able to identify those patients with clinically significant symptoms of psychological distress quickly and efficiently in the absence of a lengthy psychiatric interview.

The Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983) has been widely used as a self-report instrument for screening for psychiatric distress. The HADS (Zigmond and Snaith, 1983) was originally developed to identify psychiatric caseness in a general medical population, and has been used in screening in a variety of populations with varying degrees of success. Typically the screening efficacy parameters have demonstrated sensitivity of between 65%-90% at identifying cases of anxiety, and between 35%-90% for depression. Specificity has generally been lower for both disorders (Hermann, 1997). Factor analysis studies have revealed a stable two-factor structure of the HADS corresponding to its two subscales, although not all items load onto their respective subscales (Smith et al., 2002). In addition, the success of the HADS as a screening instrument to detect cases of psychiatric distress has been limited (Hall et al., 1999).

### *8.1.2. Aim*

Item parameter estimates have been derived for both the HADS-Scale and the two subscales, HADS-A and HADS-D, in Chapter 5. The Rasch analysis from this study (Chapter 5.2 and onwards) identified three items from the HADS-Scale and one from each of the subscales which demonstrated misfit. This study explored whether removing these items from the scales improved the psychiatric screening efficacy.

The misfitting items were removed from the analysis of the scales and new Rasch analyses were carried out to determine whether the remaining items still demonstrated good fit. Receiving Operating Characteristic curves (ROC) were plotted for the HADS-Scale and subscales containing all of the items, and sensitivity and specificity were measured for a sample of patients who had received a psychiatric assessment (either the Present State Examination (PSE) or the Schedule for Clinical Assessment in Neuropsychiatry (SCAN)). Subsequently, this process was repeated for the scales without the misfitting items to assess the impact on screening efficacy.

## 8.2 Method

### *8.2.1 Patients*

The sample of patients used for the Rasch analysis was reported in Chapter 5.1.3. The sample of patients used for the analysis of screening efficacy is reported in 8.2.2.

### *8.2.2 Psychiatric Interview*

A subset (n = 381) of the patients received a psychiatric interview based either on the Schedule for Clinical Assessment in Neuropsychiatry (SCAN) or the Present State Examination (PSE). The interviews were carried out by trained researchers and clinicians (Cull et al., 2001; Stark et al., 2002) in the patient's home within a fortnight of completion of the HADS.

The psychiatric interview data were re-scored using the Catego programme (Wing et al., 1974), which identifies cases of psychiatric disorder from the Index of Definition (ID). The scores range from ID1, where no symptoms are present, through borderline cases (ID5), to definite cases (ID scores from 6 to 8). Psychological distress was defined as a Catego score of 5 or more. Caseness of Anxiety or

Depression was defined as a Catego score of 5 or more in association with an ICD10 diagnosis.

In total 192 females and 189 males participated in these studies. The average age of the patients was 55.6 (s.d. = 12.41). The average age of females was 54.8 (s.d. = 12.19) and males was 56.5 (s.d. = 12.61).

### *8.2.3 Statistical analysis*

The original Rasch analyses which identified the misfitting items are described in Chapters 5.2.1, 5.3.1 and 5.4.1. The same analysis was carried out on the HADS-Scale and HADS-subscales with the misfitting items removed. In addition, t-tests were carried out to test for statistical differences in HADS scores between males and females. Furthermore, the HADS scores from the sub-sample were compared using one-way analyses of variance (ANOVA) to those from the larger sample described in Chapter 5.

Sensitivity and specificity analyses were also carried out. ROC curves were produced by plotting sensitivity against the false positive rate (1 – specificity), and the area-under-the-curve (AUC) was also recorded.

## 8.3 Results

### *8.3.1 HADS scores*

The summated scores of the HADS for the entire dataset were reported in detail in Chapter 5.1. Therefore the scores will only be reported here for the subset of patients who received the psychiatric interview.

The means for the HADS-scale and subscale scores for the subset of patients are given in Table 8.3.1 for males and females.

**Table 8.3.1 Mean HADS-Scale and Subscale scores by Gender**

| | HADS | s.d. | HADS-A | s.d. | HADS-D | s.d. |
|---|---|---|---|---|---|---|
| Females | 12.44 | 7.10 | 7.07 | 4.35 | 5.37 | 3.63 |
| Males | 11.15 | 6.59 | 5.58 | 4.05 | 5.57 | 3.42 |
| Total | 11.80 | 6.87 | 6.33 | 4.26 | 5.47 | 3.52 |

As can be seen from Table 8.3.1 females tended score higher on both the HADS-Scale and the HADS-A, but not the HADS-D subscale. However, only the difference between scores for the HADS-A subscale was statistically significant ($t = 3.45$, d.f. = 379, $p < 0.001$).

The mean scores for the sub-samples were 6.33 (s.d. 4.3) for HADS-A, 5.47 (s.d. 3.5) for HADS-D, and 11.80 (s.d. 6.9) for the HADS. The means for HADS-D and the HADS between the sample and sub-sample were significantly different ($F(1, 1848) = 26.41$, $p < 0.001$, and $F(1, 1848) = 12.16$, $p < 0.001$, respectively).

*8.3.2.Psychiatric caseness*

In total 77.2% (294/381) of the subset of patients scored lower than 5 on Catego, nearly 15% scored 5 (57/381) and 7.9% scored 6 or 7 (30/381). A breakdown of caseness revealed that 8.4% (32/381) of patients were experiencing anxiety disorders, 10.5% (40/381) depression, and 6.3% (24/381) of patients were experiencing both anxiety and depression. The scores from HADS-A demonstrated that 15.7% (60/381) of patients were experiencing anxiety disorders ('definite cases') according to this subscale using a cutoff of 11 (Zigmond and Snaith, 1983). Similarly, the results from HADS-D demonstrated that 8.9% (34/381) were experiencing depression using the same threshold.

A significantly greater proportion of females were categorized according to ICD-10 as having a psychiatric illness compared to men (females, 29.7%, 57/192 vs. males 20.6%, 39/189; $\chi^2 = 4.14$, d.f. =1, $p < 0.05$). A breakdown of caseness by gender indicated that for females the incidence of anxiety was 9.9%, depression

13.5%, anxiety and depression 6.3%, for males the incidence of anxiety was 6.9%, depression 7.4%, and anxiety and depression combined 6.3%.

*8.3.3 Rasch Analysis*

The Rasch analysis of the whole data set is described in Chapters 5.2.1, 5.3.1 and 5.4.1. To summarise the analysis revealed three misfitting items (i.e. items with infit mean square statistics greater than 1.30) from the HAD-Scale, namely Depression items 2 ("I can laugh and see the funny side of things"), 5 ("I have lost interest in my appearance"), and 7 ("I get sudden feelings of panics") as well as one from each of the subscales, i.e. Anxiety 6 ("I feel restless as if I have to be on the move") and Depression 6 ("I look forward with enjoyment to things"). The items were successively removed from each of the scales and the analysis repeated.

*8.3.4 HAD-Scale*

The item measures for the reduced HADS-total are given in Table 8.3.2.

**Table 8.3.2 Unidimensionality measures of the HADS with three items removed**

| ENTRY | MEASURE S | COUNT | IN.MSQ | IN.ZSTD | OUT.MS | OUT.ZSTD | NAME |
|---|---|---|---|---|---|---|---|
| 1 | -0.24 | 1416 | 0.63 | -9.9 | 0.74 | -7.04 | ANX1 |
| 2 | -0.17 | 1416 | 0.99 | -0.29 | 0.98 | -0.54 | ANX2 |
| 3 | -0.54 | 1416 | 0.78 | -6.46 | 0.8 | -5.6 | ANX3 |
| 4 | -0.28 | 1416 | 0.89 | -2.94 | 0.97 | -0.73 | ANX4 |
| 5 | 0.33 | 1416 | 0.96 | -0.97 | 0.97 | -0.58 | ANX5 |
| 6 | -0.42 | 1416 | 1.32 | 7.83 | 1.37 | 8.48 | ANX6 |
| 7 | 0.56 | 1416 | 0.93 | -1.79 | 0.86 | -2.96 | ANX7 |
| 8 | -0.03 | 1416 | 1.41 | 9.42 | 1.38 | 8.12 | DEP1 |
| 10 | 1.32 | 1416 | 0.84 | -3.65 | 0.73 | -4.61 | DEP3 |
| 11 | -1.11 | 1416 | 1.2 | 5.35 | 1.21 | 5.27 | DEP4 |
| 13 | 0.57 | 1416 | 1.19 | 4.42 | 1.1 | 1.91 | DEP6 |

The majority of items from the reduced HADS-Scale demonstrate infit statistics smaller than 1.30 demonstrating a good fit with the Rasch model. As with the Rasch analysis of the whole scale item 1 from the Anxiety subscale ("I feel tense or 'wound

up'") demonstrates a small amount of overfit or redundancy. Item 1 ("I still enjoy the things I used to") from the Depression subscale slightly exceeds the criterion of 1.3. However, this is marginal.

### *8.3.5 HADS-A*

Table 8.3.3 shows the item measures for the reduced HADS-A subscale.

**Table 8.3.3 Unidimensionality measures of HADS-A with one item removed**

| Entry | Measure | Count | In.MSQ | In.ZSTD | Out.MSQ | Out.ZSTD | Name |
|---|---|---|---|---|---|---|---|
| 1 | -0.25 | 1344 | 0.77 | -6.4 | 0.82 | -4.93 | ANX1 |
| 2 | -0.15 | 1344 | 1.02 | 0.46 | 1.01 | 0.19 | ANX2 |
| 3 | -0.68 | 1344 | 0.89 | -2.97 | 0.9 | -2.62 | ANX3 |
| 4 | -0.31 | 1344 | 1.37 | 8.6 | 1.4 | 9.25 | ANX4 |
| 5 | 0.54 | 1344 | 1.04 | 0.87 | 0.99 | -0.29 | ANX5 |
| 7 | 0.85 | 1344 | 0.95 | -1.36 | 0.88 | -2.69 | ANX7 |

Only item 4 ("I can sit at ease and feel relaxed") slightly exceeds 1.30, demonstrating good fit statistics for the reduced HADS-A subscale.

### *8.3.6 HADS-D*

Item measures for the reduced HADS-D subscale are shown in Table 8.3.4, which demonstrated good infit statistics for all of the remaining items in this subscale, which the exception of item 5.

**Table 8.3.4 Unidimensionality measures of HADS-D with one item removed**

| ;Entry | Measure | Count | In.MSQ | In.ZSTD | Out.MS | Out.ZSTD | Name |
|---|---|---|---|---|---|---|---|
| 1 | -0.61 | 1280 | 0.95 | -1.36 | 0.88 | -2.85 | DEP1 |
| 2 | 0.6 | 1280 | 1.27 | 5.01 | 1.08 | 1.29 | DEP2 |
| 3 | 0.66 | 1280 | 0.75 | -5.53 | 0.74 | -4.51 | DEP3 |
| 4 | -1.65 | 1280 | 0.91 | -2.6 | 0.94 | -1.56 | DEP4 |
| 5 | 0.43 | 1280 | 1.38 | 6.99 | 1.21 | 3.44 | DEP5 |
| 7 | 0.58 | 1280 | 1.29 | 5.36 | 1.21 | 3.17 | DEP7 |

### *8.3.7 Screening Efficacy of HADS Total*

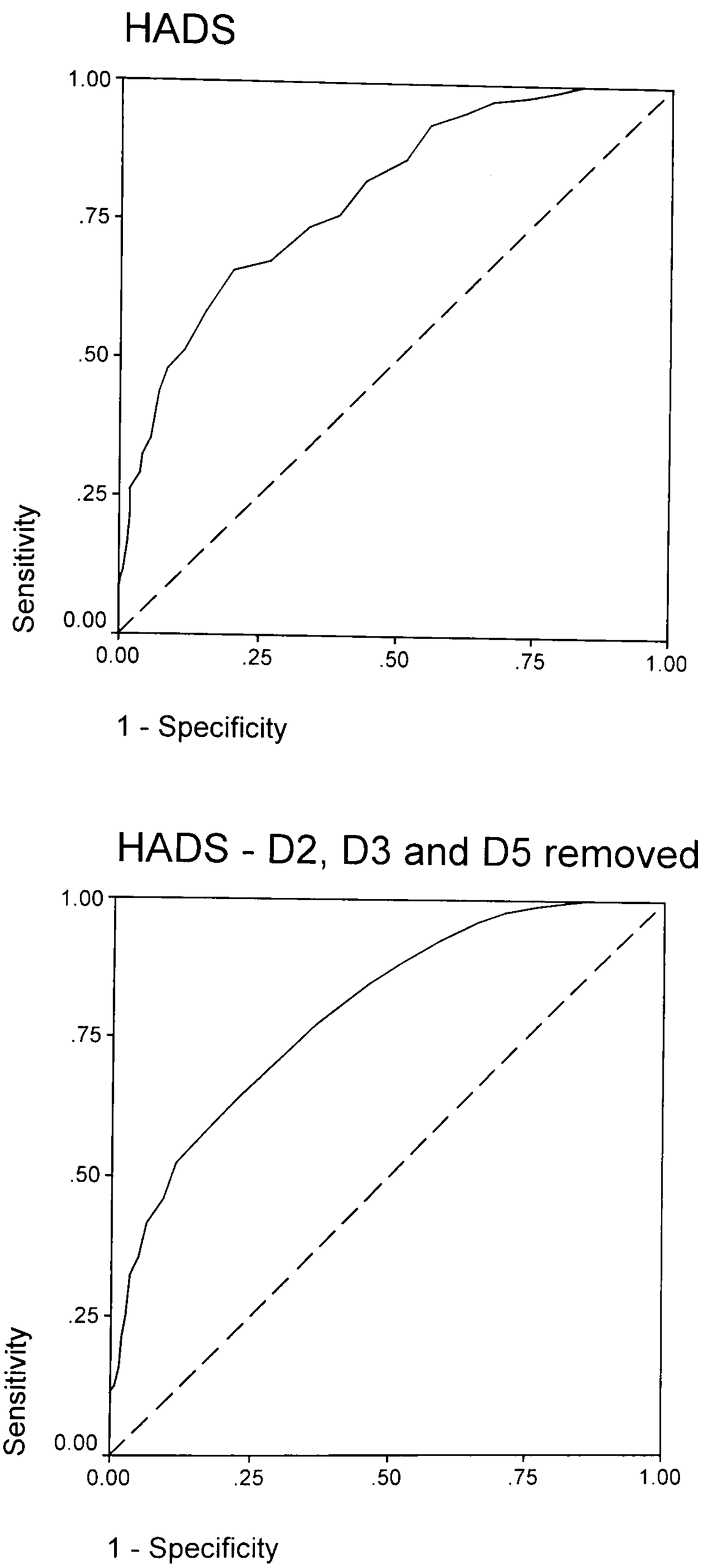
The HAD-Scale was evaluated for efficacy at detecting cases of psychological distress, i.e. Catego score of 5 or more. Figure 8.2.1 shows the ROCs for both the HAD-Scale and the reduced HAD-Scale.

**Table 8.3.5 Sensitivity, Specificity and Area under the Curve (AUC) for both HADS and Reduced HADS**

| | Sensitivity | Specificity | AUC |
|---|---|---|---|
| HADS | 0.71 | 0.70 | 0.80 |
| reduced HADS | 0.74 | 0.67 | 0.80 |

As can be seen from Table 8.3.5 the screening efficacy statistics for both scales are virtually identical, demonstrating that the removal of three items from the HADS total scale had no negative impact of its ability to detect psychological distress. Furthermore, a retest of the internal consistency of the HADS with the three 'misfitting' items removed demonstrated a high level of reliability (Cronbach's alpha 0.85).

**Figure 8.3.1 ROC Curves for the HADS and reduced HADS**

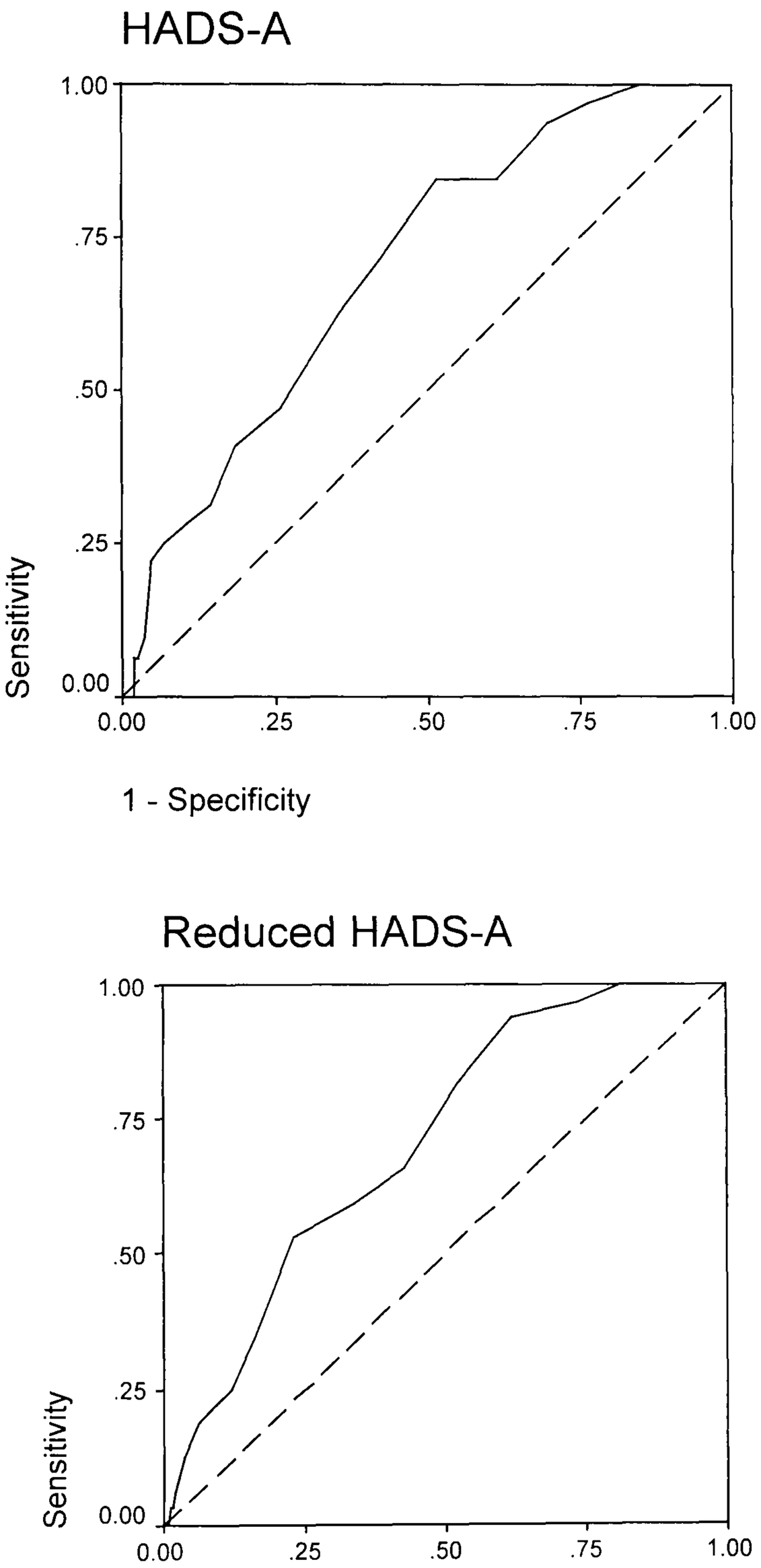



*8.3.8 HADS- A screening for Anxiety*

The HADS-A faired pretty poorly at detecting cases of Anxiety with levels of specificity and sensitivity of 0.67, and 0.61 respectively. Removing item 6 led to a deterioration in sensitivity (0.63), and only a marginal improvement in specificity. The

Area-under-the curve did not change (Figure 8.3.2). Cronbach's alpha for the reduced HADS-A scale was 0.83, demonstrating high levels of internal consistency.

**Figure 8.3.2 ROC curves for the HADS-A subscale and the Reduced HADS-A subscale**




Diagonal segments are produced by ties.

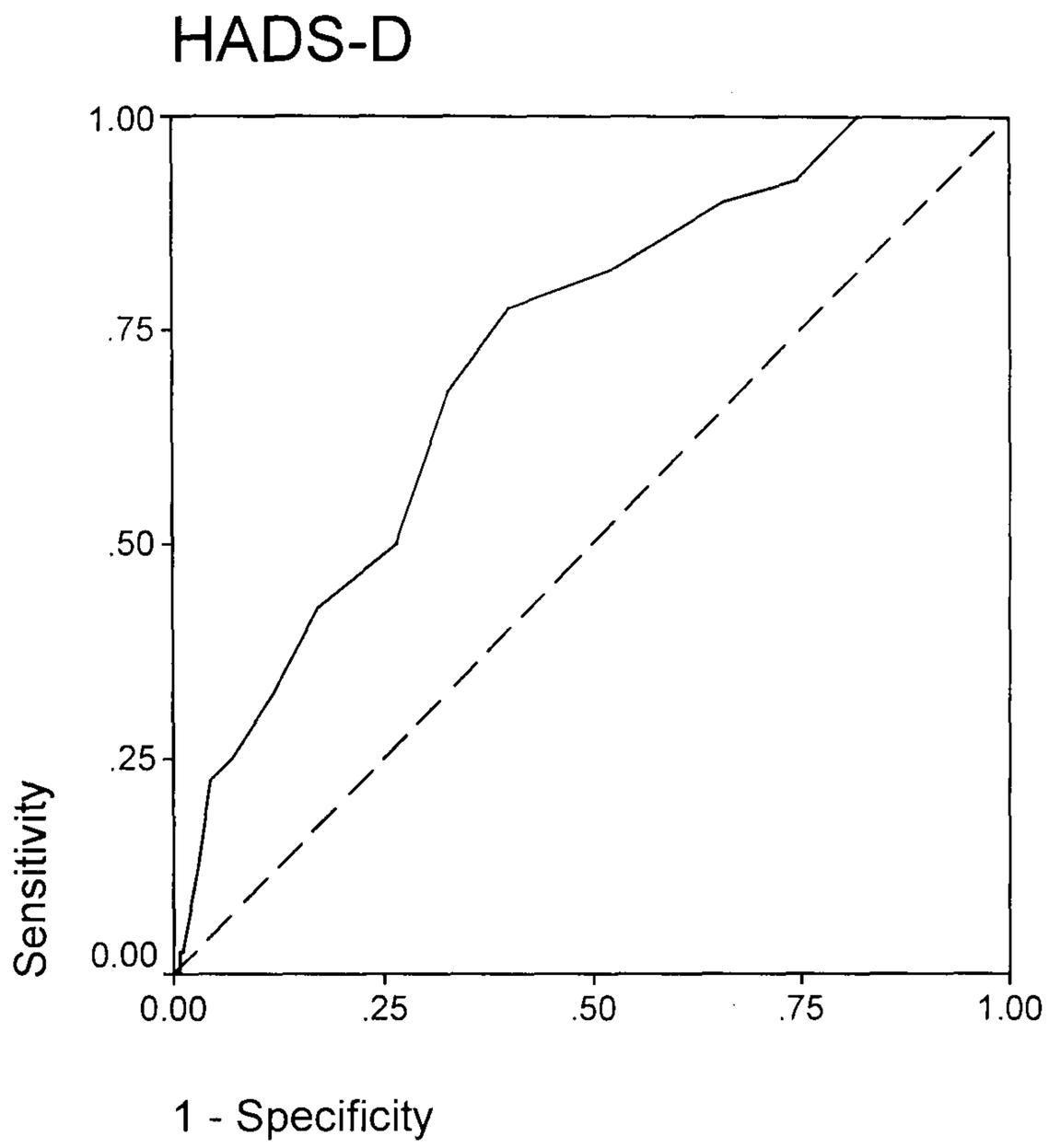
**Table 8.3.6 Sensitivity, Specificity and Area under the Curve (AUC) for both HADS-A and Reduced HADS-A**

| | Sensitivity | Specificity | AUC |
|---|---|---|---|
| HADS-A | 0.67 | 0.61 | 0.70 |
| reduced HADS-A | 0.63 | 0.62 | 0.71 |

*8.3.9 HADS-D screening for Depression*

HADS-D proved to be better at detecting cases of Depression compared to HADS-A and Anxiety. Sensitivity rates and the AUC were above .70. However, specificity was low at 0.64. Removing item 6 had little impact on sensitivity and the AUC, although specificity did decrease marginally (Figure 8.3.3 and Table 8.3.7). Cronbach's alpha for the reduced HADS-D was 0.73, which was slightly lower than for the total HADS-D, however this still demonstrated good levels of internal consistency.

**Figure 8.3.2 ROC curves for the HADS-D subscale and the Reduced HADS-D subscale**



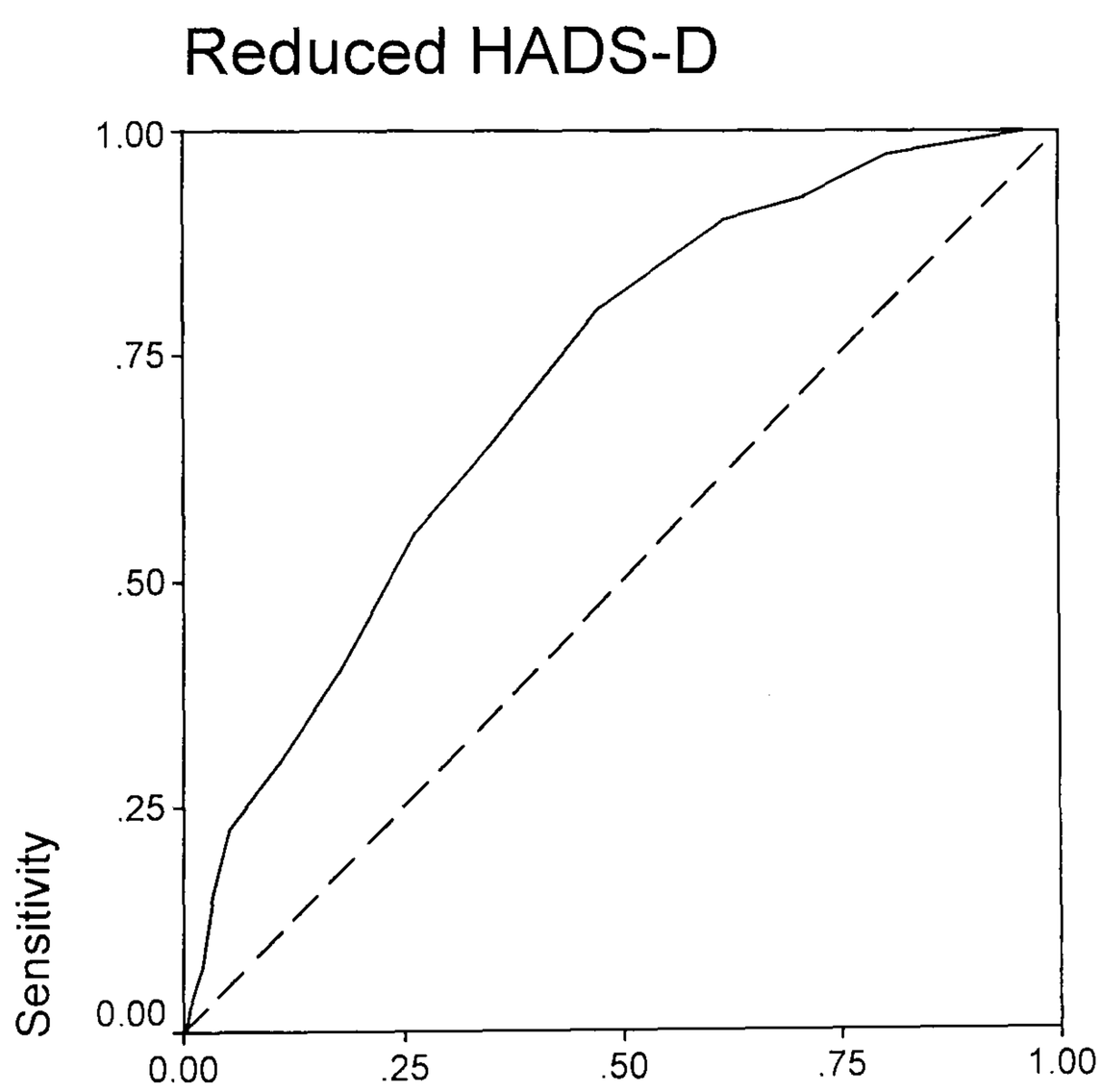



**Table 8.3.7 Sensitivity, Specificity and Area under the Curve (AUC) for both HADS-D and Reduced HADS-D**

| | Sensitivity | Specificity | AUC |
|---|---|---|---|
| HADS-D | 0.73 | 0.64 | 0.72 |
| reduced HADS-D | 0.73 | 0.59 | 0.72 |

## 8. 4 Discussion

Previous studies have found that the HADS has only moderate efficacy in screening for anxiety and depression (Hermann, 1997). Additionally, items from the HADS-A and HADS-D do not always load onto their respective subscales (Smith et al., 2002). This study investigated whether the screening efficacy of the HADS and the two subscales could be improved by removing items which had previously been identified as having poorer fit statistics (Chapter 5).

The Rasch analysis had demonstrated that three items from the HADS (Depression items 2, 5 and 7), and one item from each of the subscales (Anxiety 6 and Depression 6) had outfit statistics greater than 1.30. Removing these items had no significant effect on the fit statistics of the remaining items from the HADS and subscales, demonstrating that these items fitted the Rasch model. Furthermore, internal consistency for the total scale, as well as subscales remained good (>0.70). In addition, the screening efficacy of the HADS at detecting psychological distress, and that of the subscales at detecting cases of anxiety and depression respectively remained largely unaltered by the removal of the items: the area under the curve remained the same for all three scales, however the sensitivity of HADS-A and the HADS-D specificity decreased marginally.

Although the number of patients classified through either Catego and / or ICD-10 classification as having either psychological distress (23%), or anxiety (15%)

and depression (9%) was relatively low, which in turn may have had an effect on screening efficacy, this study demonstrated that Rasch analyses can be employed to identify misfitting items, and allow a shortened, and clinically useful version of the HADS questionnaire to be produced without detriment to the screening efficacy of the instrument. However, this conclusion has to be viewed with some caution given the fact that significant differences were found for the HADS-D and HADS-total between the larger sample and the sub-sample who also undertook a psychiatric interview. Nevertheless, these results are similar to those of Velozo et al's (2001) who identified items to be removed from the original 14-item version of the Visual Function scale by low or high mean squares or by having identical calibrations (measures). The authors were able to identify 4 items using these criteria. A subsequent retest of the reduced 10 item scale demonstrated good levels of internal consistency (Cronbach's alpha 0.89).

Rasch analyses can be utilised in this way to generate item banks in conjunction with computer-adaptive testing (McHorney, 1997) to produce adaptable screening tools for identifying psychological distress in cancer patients. Recent work using item response theory methods (Groenvold et al., 2000; Petersen et al., 2000a; Petersen et al., 2000b) has enabled the reduction of the Emotional Functioning scale of the EORTC QLQ-c30 from a 4- to a 2-item scale, and the Fatigue scale of the same instrument from 3 to 2 items. Scores from these reduced scales are able to predict scores derived from the full subscales. Finally, Lai et al (2003) have used Rasch models to identify and remove misfitting items from the FACIT-Fatigue scale in a process leading towards a computer-adaptive assessment of fatigue in cancer patients.

Future work in the domain of mental health could focus on Rasch analyses of other mental health measures along with the HADS, in order to identify items for an item bank and a possible computer-enabled assessment.

# 9. Comparison of Quality of Life Instruments

## 9.1. Introduction

As discussed in previous chapters the European Organization for Treatment and Research of Cancer Quality of Life Questionnaire C30 (EORTC QLQ C30, Aaronson, Ahmedzai, Bergman et al., 1993), and the Functional Assessment of Cancer Therapy questionnaire (FACT-G, Cella, Tulsky, Gray et al., 1993) are two widely used instruments for the assessment quality-of-life of cancer patients.

Although the use of both instruments is widespread there has been a tendency for questionnaires to be used independently, although a few studies have used both questionnaires in conjunction for Quality-of-Life (QOL) assessment (Doyle, Crump, Pintillie et al., 2001; Holzner, Kemmler, Kopp et al., 2001; Kopp, Schweigkofler, Holzner et al., 1998). In addition, recent studies have also attempted to compare and evaluate the functional scores of the EORTC QLQ-C30 and the scaled scores from the FACT-G (Kemmler, Holzner, Kopp et al., 1999; Kopp, Schweigkofler, Holzner et al., 2000; Sharp, Knight, Nadler et al., 1999). The results from these comparison studies in general have demonstrated good correlations between the Physical functioning scales of the two questionnaires (Physical Functioning - Physical Well-Being), slightly poorer correlations for the corresponding Emotional and Functional scales (Emotional Functioning - Emotional Well-Being, and Role Functioning – Functional Well-Being), and virtually no association between the corresponding Social Functioning scales (Social Functioning – Social & Family Well-Being).

Despite the pattern of association between the corresponding scales demonstrated by these studies, the conclusion has been proposed that direct comparisons between the two instruments is not possible (Kemmler et al., 1999; Kopp et al., 2000). This is an important issue because it implies that conclusions

drawn in studies using one of these instruments can only be extrapolated in a restricted way, which has important implications for clinical trials, and makes comparisons of quality of life results across studies difficult.

There are some methodological issues that arise when these important instruments are compared. The EORTC QLQ-C30 functional scales are scored on a scale of 0 to 100, whereas the individual FACT-G scales are scored out of a maximum of 28. Some of the comparison studies have converted the FACT-G scores by a simple linear transformation to the same scale as the EORTC QLQ-C30. However, comparisons of this kind between two different instruments that measure the same underlying construct (e.g. quality-of-life) can only be made if the questionnaires are converted to same scale. Unless this has been undertaken proper comparisons cannot be made, and definitive conclusions are problematical. Statistical techniques, such as the general Rasch model, exist which allow comparisons between different instruments. The Rasch model allows estimates of ability for a particular latent (i.e. unobserved) trait, e.g. physical functioning or QOL, to be made independently of the type of items presented to patients. It is therefore possible to compare scores from patients from their responses to different items, i.e. it allows comparisons to be made across different QOL measures (Gonin, Lloyd & Cella, 1996).

The second issue concerns the use of correlations as a measure of agreement. Previous comparisons between the EORTC QLQ-C30 and FACT-G have made use of correlation coefficients to compare the scores between the questionnaires. However, correlation is a measure of association, and not a measure of agreement. The magnitude of the correlation coefficients depends on the variation between individuals, and the measurement error or variation within individuals. Therefore, high variability of scores between individuals may give rise to high correlation coefficients, and conversely low variability may give rise to low correlation coefficients, both of which may lead to misleading results. Instrument comparison

can be best achieved through the measurement of agreement between scores. Techniques have been described for measuring agreement such as plotting the difference of two scores against the mean (Bland & Altman, 1986). These plots present a simple visual method of agreement, which allows the assessment of the size of disagreement (either error and bias), trend analysis and the identification of outliers.

### *9.1.1. Aim*

The present study compared the functional scales of the EORTC QLQ-C30 with the scales from the FACT-G, as well as the Global Quality-of-Life scores and FACT-G total. Measures of agreement of the two instruments were calculated from scores which had been converted using Rasch models. Statistical differences between corresponding scales were tested to ascertain whether the instruments, and their corresponding scales were functionally different.

## 9.2. Patients And Instruments

### *9.2.1. Study Sample*

The patient sample was collected from inpatients attending wards at Cookridge Hospital, a large cancer hospital in Leeds. The inclusion criteria for the study comprised ability to read and understand English, no visual or cognitive impairments, and no pre-existing psychological morbidity. Patients were approached on the wards and asked to participate in the study. A total of 245 were asked to participate in the study, and 200 patients agreed to take part (81.6%) and completed both versions of the questionnaire, 45 patients refused participation (19.4%). All patients who agreed to take part in the study were given an information sheet and asked to provide written consent.

Demographic and clinical details were available for 198 of the patients who participated in the study (99%). Patients not participating in the study were asked to give consent for their demographic and clinical details to be recorded as required by the 1999 Data Protection Act. These details were recorded for 28 (62%) of the non-participating patients.

Ethical approval for the study was provided by the local ethics committees of the Leeds Teaching Hospitals NHS Trust.

*9.2.2 Instruments*

Patients completed an electronic version of the EORTC QLQ-C30 (version 3.1) and the FACT-G (version 4) questionnaires on a portable touchscreen computer in a single sitting. The order of presentation of the two instruments was randomised by the computer programme. The raw scores were recorded into an MS-Access database and converted to the summated scales.

*9.2.3. Description of the EORTC QLQ-C30 and FACT-G*

Both instruments are described in detail in Chapter 2.

## 9.3. Data Analysis And Statistical Methods

The sample size calculations (Cohen, 1988) were derived from standard power and effect size tables, using a significance criterion of a = 0.05, and power = 0.80 for a two-tailed test. Assuming a correlation coefficient of r = 0.20 (rounded-up from 0.14, the correlation between SF and SFWB derived from two published studies, e.g. Kemmler et al., 1999; Sharp, Knight & Nadler, 1999), the number of patients needed to detect a significant association was calculated as 194. We therefore decided to recruit 200 patients into the study.

The comparison of age differences between participators and non-participators was carried out using non-parametric Mann-Whitney tests. Statistical

significance was evaluated against a p-value of less than 0.05. Spearman's correlations for ordinal data were used to assess associations between functional scales of EORTC QLQ-C30 and FACT-G. Internal reliabilities for the functional scales of the EORTC QLQ-C30 and the FACT-G scales were calculated using Cronbach's alpha. The raw scores from both instruments were converted to scale scores using the published algorithms (Cella, 1997; Fayers, Aaronson, Bjordal et al., 2001). The raw scores from the functional scales of the EORTC QLQ-C30 and the FACT-G scales were used to derive ability estimates and converted to logits using the *Winsteps* programme (Linacre & Wright, 2000).

The logits scores of the ability estimates from the scales were then converted to z-scores by subtracting the mean logit score from each scale, and dividing by the standard deviation for the logit scores for each scale (Gonin et al., 1996).

Agreement between the z-scores of the logits for corresponding scales was then calculated by plotting the difference between the scales (e.g. z-scores of logits for Physical Functional (zlPF) – z-scores of logits for Physical Well-being (zlPWB)) against the mean of the scores for corresponding scales (e.g. zlPF + zlPWB / 2) using scatter plots. Confidence intervals for the agreement plots were derived by calculating the number of difference measures that fell outside the 95% (i.e. $\pm$ 1.96) limit.

The difference between the z-scores of the logits of corresponding scales was assessed using paired t-tests and evaluated against a p-value of 0.05.

## 9.4 Results

### *9.4.1 Patients*

The diagnostic and clinical details for participators are provided in Table 9.4.1. The distribution of non-participators by diagnosis did not differ markedly from those patients who participated in the study. The mean age for females (participants and non-participants) was 67 years. However, the male participators were on average

slightly older than those not participating in the study (mean age of male participators: 63.6 ± 12.5 years, mean age of male non- participators: 59.8 ± 4.7 years). This was not statistically significant ($p > 0.05$).

**Table 9.4.1. Diagnoses and clinical details of patients**

| | Female | | Males | |
|---|---|---|---|---|
| | n = 127 | | n = 73 | |
| Age, years (mean ± S.D.) | 57.1 ± 12.2 | | 63.6 ± 12.5 | |
| | Count | % | Count | % |
| Diagnosis | | | | |
| Breast | 50 | 39.4 | | |
| Colorectal | 35 | 27.6 | 37 | 50.7 |
| Gastrointestinal | 15 | 11.8 | 12 | 16.4 |
| Genitourinary | 13 | 10.2 | 7 | 9.6 |
| Lung | 6 | 4.7 | 12 | 16.4 |
| Other | 6 | 4.7 | 5 | 6.9 |
| Months since diagnosis (mean ± S.D.): | 16.1 ± 19.5 | | 14.0 ± 14.6 | |
| Extent: | | | | |
| Disease free | 5 | 4.1% | | |
| Primary local | 48 | 39.0% | 30 | 42.3% |
| Local recurrent | 15 | 12.2% | 12 | 16.9% |
| Metastatic | 55 | 44.7% | 29 | 40.8% |
| Type of treatment: | | | | |
| None | 7 | 5.8% | 6 | 8.3% |
| Radiotherapy | 7 | 5.8% | 3 | 4.2% |
| Chemotherapy | 98 | 81.0% | 60 | 83.3% |
| Hormone therapy | 1 | .8% | | |
| Chemo/-radiotherapy | 8 | 6.6% | 3 | 4.2% |

*9.4.2. EORTC-QLQ C30 and FACT-G scores*

In total, due to the randomisation by the computer programme, 104 patients completed the EORTC QLQ-C30 first followed by the FACT-G, and 96 patients completed the two instruments in reverse order. Table 9.4.2 provides the means and standard deviations for the converted scores for both instruments.

The scales from both instruments demonstrated good internal reliability with Cronbach's alpha scores of around 0.81. The Global Quality-of-Life (QL) scale from

the EORTC QLQ-C30 demonstrated the highest reliability coefficient of 0.93, whereas the Emotional Well-being scale from the FACT-G had the lowest reliability coefficient of 0.63.

**Table 9.4.2. Means and standard deviations of EORTC QLQ c30 and FACT-G**

| QOL Domain | EORTC QLQ – c30 (Mean ± SD) | FACT-G (Mean ± SD) |
|---|---|---|
| Physical | 71.2 ± 21.7 | 21.5 ± 5.1 |
| Role/functional | 64.5 ± 31.3 | 18.2 ± 6.1 |
| Emotional | 80.3 ± 17.9 | 18.2 ± 4.3 |
| Social | 67.7 ± 28.9 | 23.4 ± 4.9 |
| Global QL/Total | 65.0 ± 19.2 | 81.2 ± 14.8 |
| Cognitive | 81.9 ± 23.4 | |
| Symptom scales | | |
| Fatigue | 39.9 ± 24.4 | |
| Nausea & vomiting | 11.3 ± 17.8 | |
| Pain | 19.4 ± 26.2 | |
| Dyspnoea | 25.8 ± 28.1 | |
| Insomnia | 27.0 ± 28.8 | |
| Appetite | 21.7 ± 28.9 | |
| Constipation | 15.8 ± 27.1 | |
| Diarrhoea | 11.5 ± 23.3 | |
| Finance | 10.2 ± 21.7 | |

The Physical Functioning (PF), Emotional Functioning (EF) and QL scales and their corresponding FACT-G scales (Physical Well-being, PWB; Emotional Well-being, EWB; and the Total FACT score) demonstrated high correlation coefficients, indicating good association between the scales from the two instruments. The Social Functioning (SF) and Social and Family Well-being (SFWB) scales showed the poorest association with a correlation coefficient of 0.18 (Table 9.4.3).

**Table 9.4.3. Spearman correlations between the Functional Scales of the EORTC-QLQ c30 and the FACT-G scales**

| | PF | RF | EF | CF | SF | QL | PWB | SFWB | EWB | FWB |
|---|---|---|---|---|---|---|---|---|---|---|
| PF | . | | | | | | | | | |
| RF | .68 | . | | | | | | | | |
| EF | .24 | .36 | . | | | | | | | |
| CF | .37 | .41 | .41 | . | | | | | | |
| SF | .54 | .64 | .38 | .50 | . | | | | | |
| QL | .58 | .52 | .39 | .41 | .54 | . | | | | |
| PWB | **.73** | .65 | .45 | .48 | .69 | .67 | . | | | |
| SFWB | .12 | .15 | .23 | .18 | **.18** | .35 | .15 | . | | |
| EWB | .37 | .32 | **.67** | .29 | .33 | .42 | .44 | .26 | . | |
| FWB | .54 | **.55** | .35 | .39 | .63 | .67 | .64 | .42 | .44 | . |
| TOTAL | .58 | .56 | .52 | .44 | .63 | **.70** | .72 | .61 | .68 | .88 |

*All correlations are significant at the .05 level (2-tailed) with the exception of SFWB.

The means and standard deviations of the z-score logits are shown in Table 9.4.4.

**Table 9.4.4. Means and standard deviations for the logit scores of the EORTC QLQ-c30 and the FACT-G**

| Scale | Mean ± SD | Scale | Mean ± SD |
|---|---|---|---|
| PF | -2.5 ± 2.5 | PWB | -1.7 ± 1.4 |
| SF | -2.2 ± 3.7 | SWB | 1.9 ± 1.4 |
| EF | -3.7 ± 2.3 | EWB | -0.9 ± 1.1 |
| RF | -1.4 ± 2.9 | FWB | 0.9 ± 1.4 |
| QL | 7.3 ± 8.8 | TOTAL | 0.1 ± 0.4 |

The agreement plots (Figures 9.4.1 to 9.4.5) showed very good agreement between three corresponding scales PF and PWB, EF and EWB, and between QL and the total FACT-G score. For the PF/PWB and EF/EWB plots less than 5% of the scores fell outside the 95% confidence interval[1]. Although the majority of plots for SF/SFWB and RF/FWB fell within the 95% confidence interval, the overall proportion was less than the other three scales at 74%.

[1] Confidence intervals were calculated for each scale, which then allowed the proportion of plots outside the limits to be determined.

**Figure 9.4.1. Difference against average of Physical Functioning and Physical Well-being**

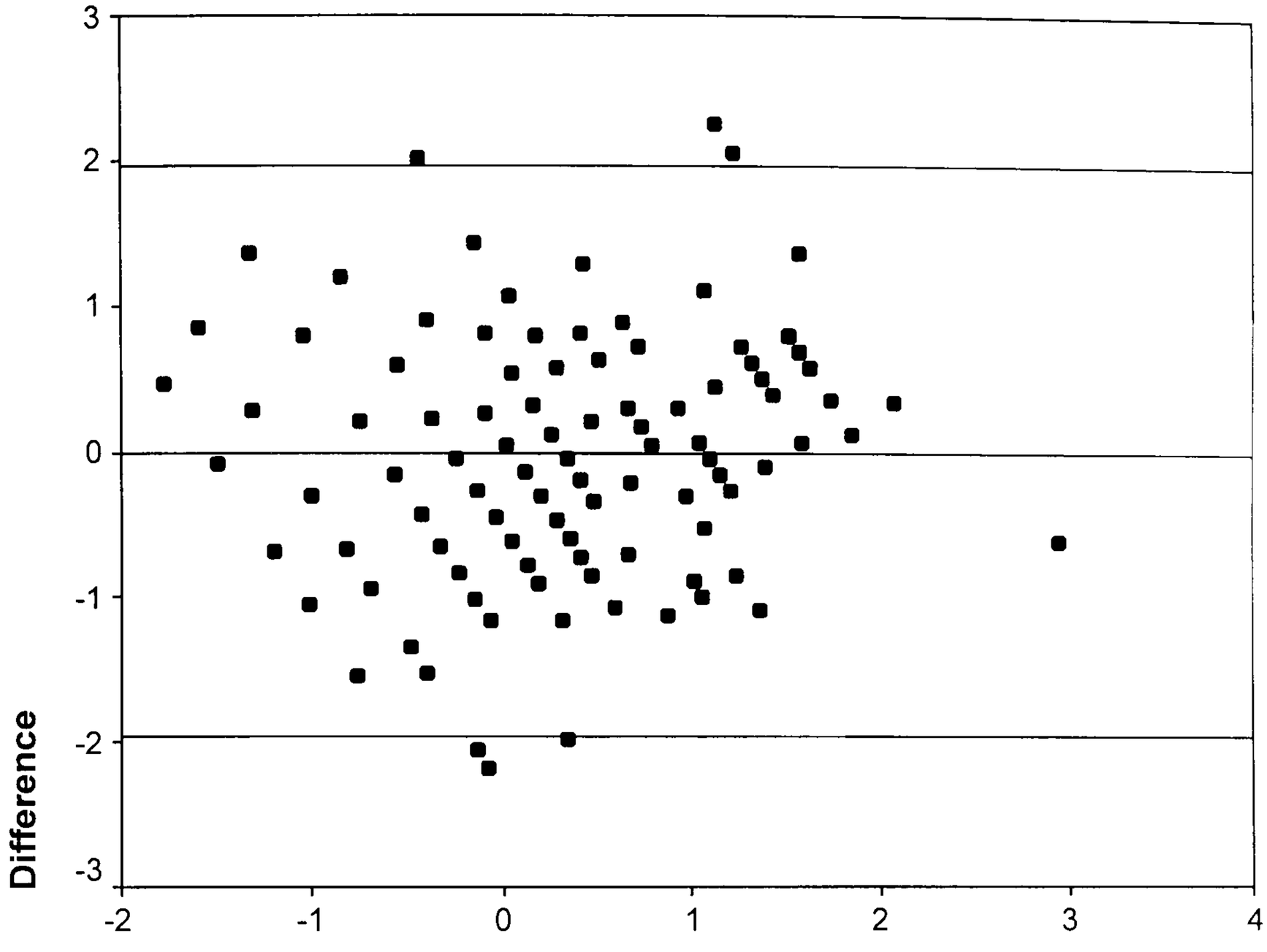


**Figure 9.4.2. Difference against average of Social Functioning and Social & Family Well-being**

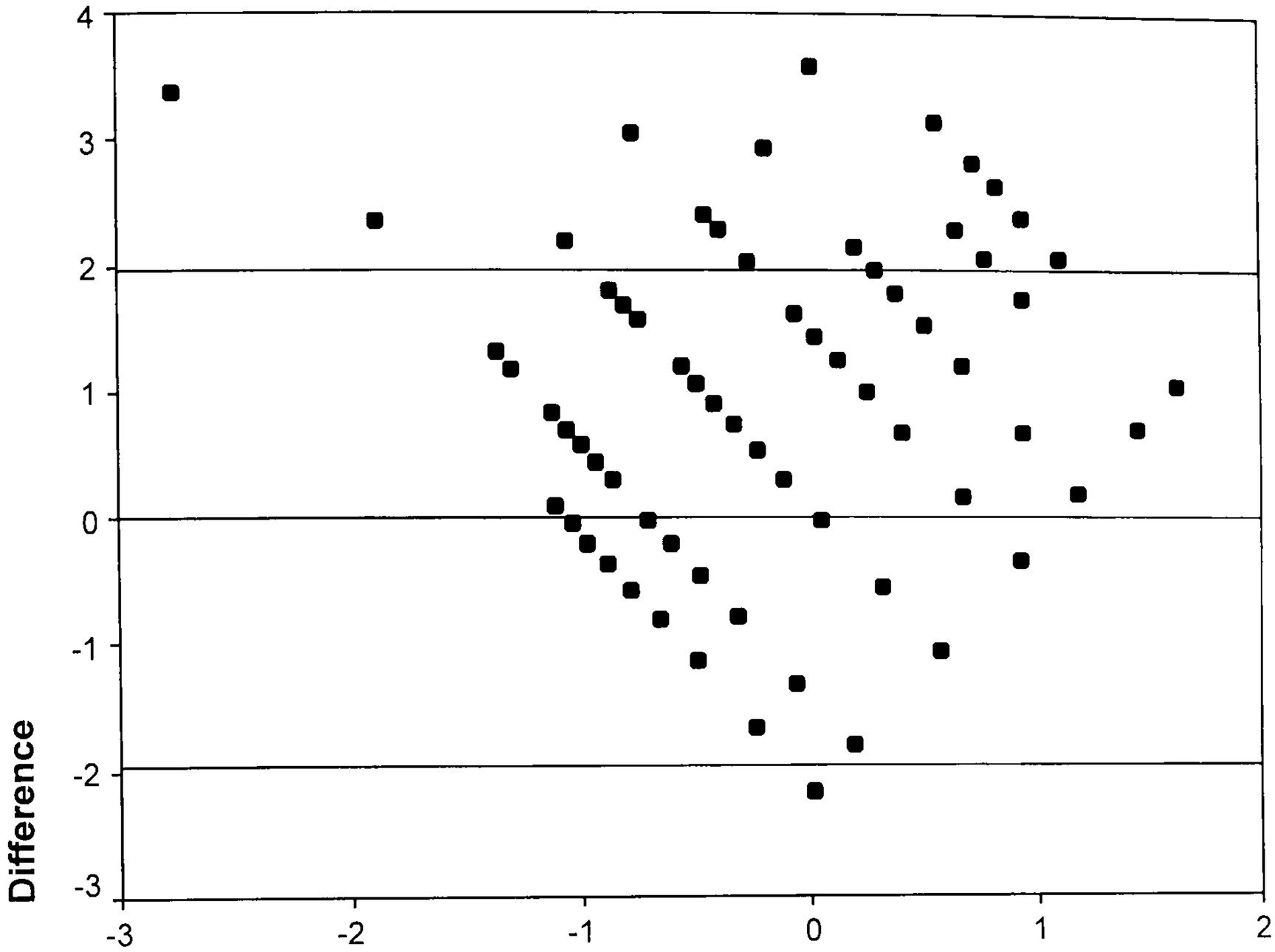


**Figure 9.4.3. Difference against average of Role Functioning and Functional Well-being**

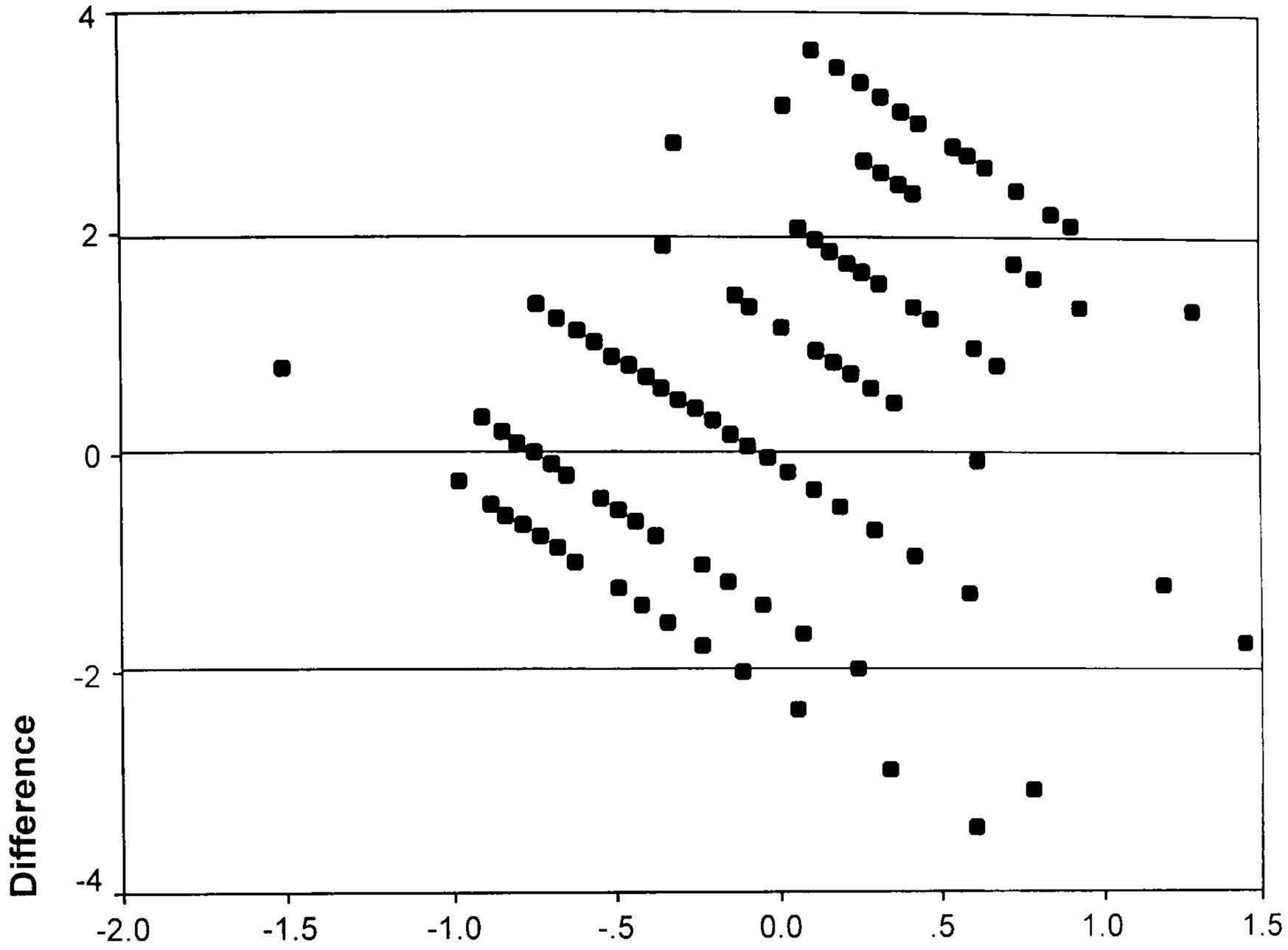


**Figure 9.4.4. Difference against average of Global QL and Total FACT-G**

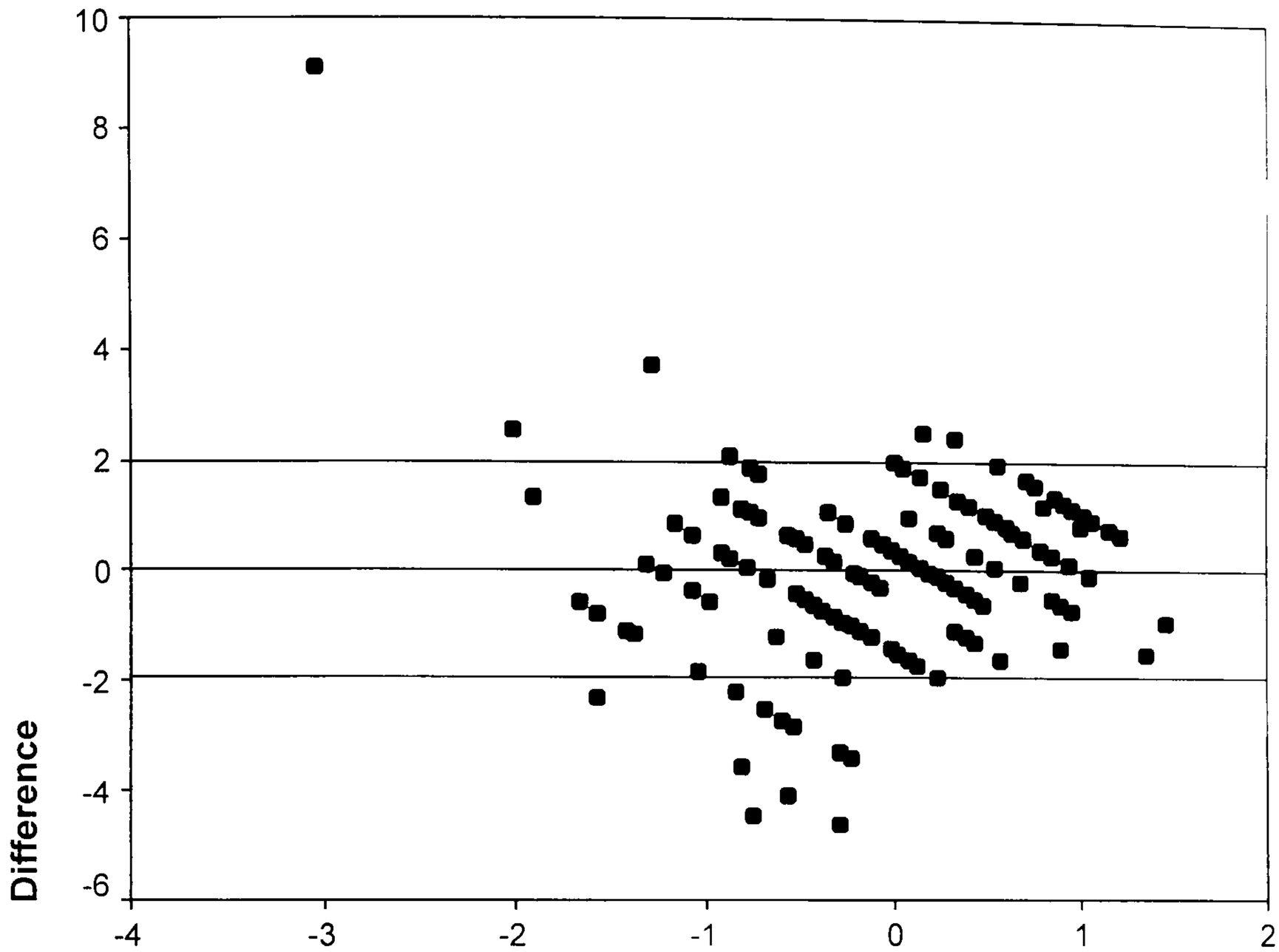
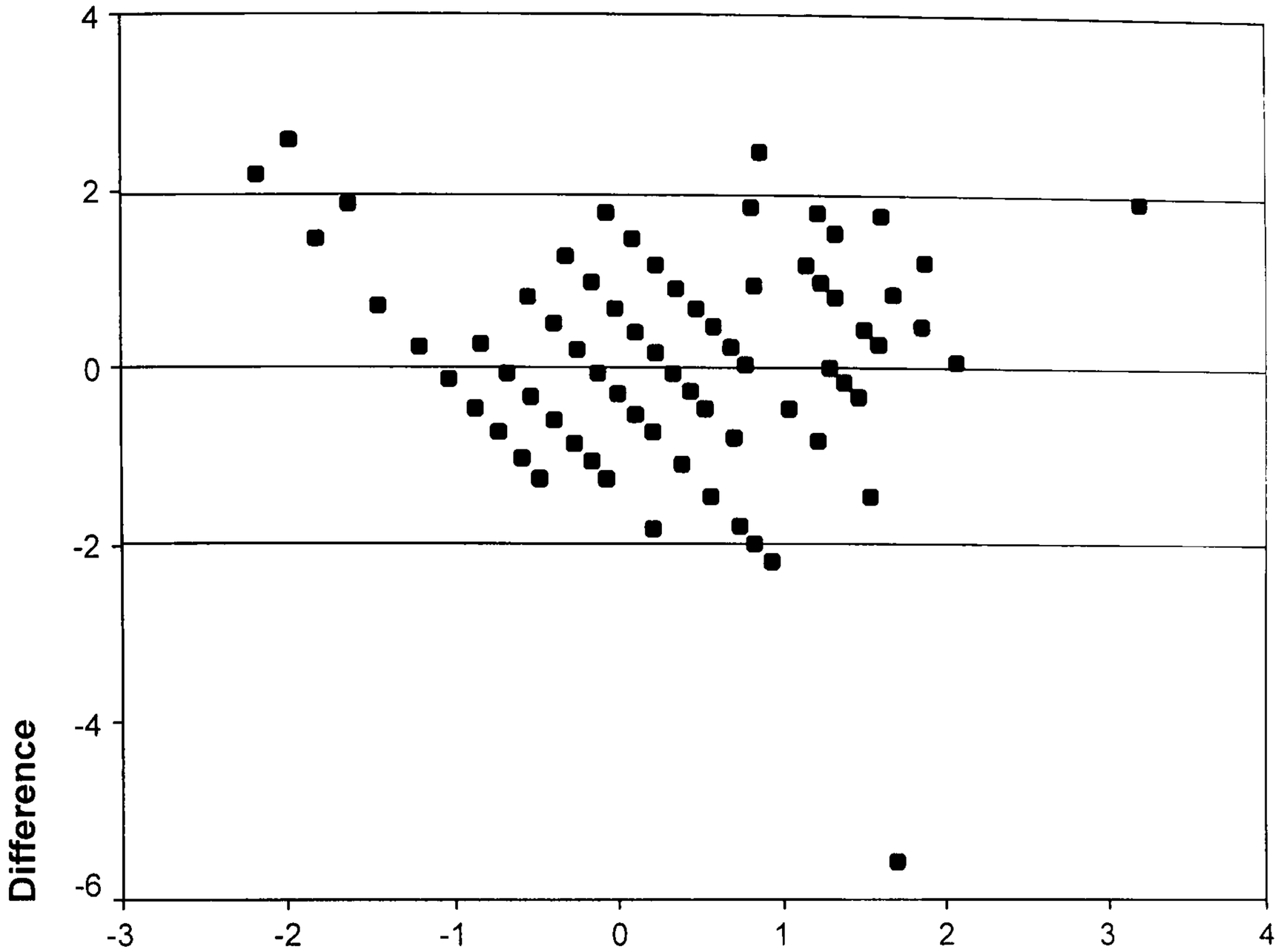


**Figure 9.4.5. Difference against average of Emotional Functioning and Emotional Well-being**




In addition, the paired t-tests demonstrated no significant differences between the corresponding scales from the two instruments ($t < 1$) for all comparisons.

## 9.5. Discussion

The results of this study demonstrated good levels of association between the corresponding scales of the EORTC QLQ-C30 and the FACT-G with the exception of Social Functioning and Social Well-being scale. Furthermore, when the scaled scores were converted to logits, to place scores from different instruments on the same metric, all corresponding scales showed high levels of agreement, and no

statistically significant differences were found between the scores from corresponding scales.

The agreement measures from this study therefore demonstrate that the scales from both quality of life instruments are comparable. The fact that no statistical differences were found between pairs of scales from the questionnaires indicated effectively that the corresponding scales from the EORTC QLQ-C30 and the FACT-G are equivalent.

Previous studies (Kemmler et al., 1999; Kopp et al., 2000; Sharp et al., 1999) have only focused on correlations between the scales from both instruments and have demonstrated good association between physical scales, but poorer associations between others, in particular social and emotional functioning. This has led to the conclusion that fundamental differences exist between the EORTC QLQ-C30 and FACT-G, and that the instruments cannot be used interchangeably (Kemmler et al., 1999; Kopp et al., 2000).

However, there are a number of methodological differences between this study and the other comparison studies. These comparison studies investigated associations, rather than agreement, between the two questionnaires, and as has been pointed out earlier, correlations are highly susceptible to variances in the data, and therefore are less reliable than agreement measures. In addition, the studies converted FACT-G scores to a scale equivalent to the EORTC QLQ-C30 scales using a simple linear algorithm, which fails to take the psychometric properties of the instrument into account, and does not convert scores from the scales of the instruments to an equivalent scale making conclusions hard to interpret.

Indeed, although the correlation coefficients in this study were similar to those of previous studies with the Social Functioning scale demonstrating the lowest correlation coefficients, when agreement plots were calculated from the scales with logits derived from the Rasch model, high agreement was found for all corresponding scales. Therefore, despite the differences in content of the questionnaires.

corresponding scales from each instrument are measuring the same underlying latent trait.

Additionally, the patient samples from these other comparison studies were limited to either breast and lymphoma patients, lymphoma and leukaemia patients or patients with metastatic prostate cancer. The greater diversity of patients in this study, in particular the wide range of diagnoses and numbers of patients on active treatment, from this study make the results more generalisable.

Finally, the patient responses in this study were collected using touchscreen computers. Touchscreen computers have been demonstrated to provide a reliable and efficient means of collecting quality of life data from oncology patients (Velikova, Wright, Smith et al., 1999). The use of touchscreen computers allowed true randomisation of the order of presentation of the two questionnaires, to avoid any order effects (Kemmler et al., 1999; Kopp et al., 2000).

These results have important implications for QOL assessment in clinical trials and in clinical practice in general. The agreement between the measures demonstrates that, in terms of functionality, there are no differences between the questionnaires, therefore the content of the two instruments may become more important as a selection criterion for use in quality of life assessment.

For example, the Emotional Functioning scale of the EORTC QLQ-C30 focuses more on general mental health issues, whereas the Emotional Well-being scale of the FACT-G focuses on mental health related to cancer specific concerns. Similarly, whereas the Social Functioning scale of the EORTC QLQ-C30 asks patients to rate whether their treatment or condition has interfered family and social life, the Social and Family Well-being scale of the FACT-G also includes questions regarding close relationships and sexual functioning.

In conclusion, both instruments focus on issues surrounding health-related quality of life, and in essence are assessing the same underlying traits. The precise content of the instruments is the main concern for investigators selecting

questionnaires and there are differences between the EORTC QLQ-C30 and FACT-G. However, both work well and are good alternatives. Therefore, with careful methodological attention, the results between studies using either instrument may be meaningfully compared.

# 10. Conclusions and Future Work

## 10.1 Conclusions

The hypothesis of this thesis was that computer-assisted questionnaires and modern psychometric techniques could be employed to improve the questionnaires presented to patients, by making the questionnaires shorter, and more relevant to patients. Furthermore, the aims of the thesis were to reduce patient burden by reducing the number of questions presented to patients, and secondly to increase the relevance to patients by developing systems that allow patients to select items or domains from questionnaires.

Traditional psychometrics and Rasch analyses were carried out on three important quality of life questionnaires, and additionally an experimental study was described comparing the results of a standard version of a QOL instrument to those from a computer-assisted questionnaire.

In essence the results of the studies described in the thesis demonstrated the following points:

1. The Rasch models can be employed to analyse quality of life instruments and that this analysis was informative in evaluating the questionnaires beyond the information provided by traditional psychometrics;
2. The results of the statistical analysis allow questionnaires to be evaluated which may help inform future selection of instruments by clinicians;
3. Items can be identified which may not fit the underlying models or assumptions;
4. Items can be removed on the basis of lack of fit to produce screening instruments which are as efficacious as the full instruments;
5. The use of computer-assisted questionnaires may shorten standard questionnaires, but at the cost of accuracy.

## 10.2 Future work

The thesis covered one broad theme, namely the potential reduction in patient burden through either the use of computer-assisted questionnaires in helping to evaluate patients' quality of life by allowing patients themselves to select domains, or the use of Rasch analysis in identifying items which can be removed from questionnaires.

There are several areas of future work which arise from the results in the thesis which are outlined below.

### *10.2.1. Methodological work*

The issue of unidimensionality is a central feature which is critical for the concept of fundamental measurement (e.g. Thurstone, 1925, 1926, 1928, 1931). It is a concept that underpins the identification of misfitting items and the selection of items for computer-adaptive testing.

However, the detection of items which do not "fit" the unidimensional construct relies in Rasch analysis on the fit statistics. As Smith and colleagues (Smith, 1988; Smith, 1991; Smith et al., 1998; Smith and Suh, 2003) have demonstrated with simulated datasets that for large samples, such as those in this thesis, the criteria used to identify fit should be more stringent. For instance, Smith et al. (1998) suggest that for sample sizes greater than 1000, the range for infit statistics should be 0.95 – 1.05, and outfit statistics should be 0.90 – 1.14, and not the commonly used range of 0.70 to 1.30 for both statistics (Wright et al., 1994). Karabatsos (2000) has recently added to this debate from a detailed critique of the fit statistics employed in Rasch analysis, and suggests that these statistics are not invariant across different samples and tests.

Given the central importance of unidimensionality to Rasch analysis, and the critical concomitant of fit statistics, it is apparent that future work needs to be carried

to explore the relationship between sample size, test properties and fit statistics. In particular since all of the work to date on this relationship has been carried out on simulated data with perfect fit (Karabatsos, 2000; Smith, 1988; Smith, 1991; Smith et al., 1998; Smith and Suh, 2003), there is a danger that given these stringent criteria and given the fact that perfect fit does not exist, that questionnaires could be reduced to very few items, which could concomitantly reduce face validity (Nunnally and Bernstein, 1994). Therefore, the relationship between these characteristics needs to be explored further on real data to ascertain whether this association holds.

There are two other important features of the Rasch analysis, that is the person ability estimate and the identification of categorical misfit (Linacre, 1995).

The person ability estimates are a critical factor in Rasch analysis in that they allow questionnaire scores to be converted to score estimates which are independent of item difficulties. A concomitant of this is that differences between adjacent scores can be plotted to assess whether the scale is an interval scale. It is evident that unless a scale can be shown to be an interval scale, then it is impossible to interpret any change in scores with either any confidence nor with a single "clinically meaningful difference" score (e.g. Cella, et al., 2002; King, 1996) since the impact of the change will differ depending on the location on the scale.

The results of the Rasch analysis of the EORTC QLQ-c30 scales suggests that whereas the Fatigue scale is largely an interval scale, the Physical Functioning scale follows an inverted "U" shape, and the Emotional Functioning scale appears to be bimodal. Future work could explore the relationship between score changes on these two scales, against other questionnaires (for instance, the Physical Well-being scale from the FACT-G, and the HADS respectively) and against a subjective evaluation of the impact of change by the patients (Osoba et al., 1998). One of the limitations identified from the Rasch analysis of the FACT was the that person measure estimates were limited due to a lack of data. Therefore, it is apparent that

additional data would need to be collected to accurately estimate the person ability measures for the FACT-G in future work.

The results demonstrated categorical misfit for 1 or 2 of the response categories of FACT-G for all subscales. Future methodological work should explore removing these categories, or collapsing categories, and whether this improves the categorical fit, and how this affects the overall item fit.

### *10.2.2. Computer-adaptive testing*

The unidimensional nature of many questionnaires which assess psychological distress clearly lends itself to the development of a computer-adaptive programme for identifying anxiety and depression in oncology patients. Furthermore, since the proportion of cancer patients experiencing psychological distress at some stage is estimated to be as high as 23% for anxiety (Stark and House, 2000), and 47% for depression (Sellick and Crooks, 1999), there is evidently a need for instruments which can identify vulnerable patients efficiently and effectively.

The development of such a programme requires additional methodological work, i.e. Rasch analysis on other questionnaires. This programme of work would require a substantial number (>1000) of additional data (Lai et al., 2003) from a number of other questionnaires measuring psychological distress *in order to develop an item bank (e.g. McHorney, 1997) consisting of a large number of questionnaires covering the psychological distress dimension*. This data would then need to be subjected to Rasch analyses to assess unidimensionality, and identify misfitting items and differential item bias. Subsequent work would then need to be carried out to establish clinical thresholds for the system (Embretson and Reise, 2000), for instance as demonstrated for the Emotional Functioning scale in Chapter 6, as well as to assess the screening efficacy of the computer-adaptive system.

This may have potential implications for patients involved in the initial development of the item bank in terms of the number of questionnaires which they

could potentially have to answer. The findings from the Rasch analysis of all three instruments demonstrated that most – if not all – questionnaires with the exception of the Physical Functioning scale of the EORTC QLQ-C30 and the Emotional Well-being scale of the FACT-G covered a narrow range of person abilities between –1 and +1. Furthermore, questionnaires for screening, for instance of psychological distress, whether they are static (i.e. a fixed number of items) or adaptive (flexible number of items) need to include more items around clinical threshold (Embretson and Reise, 2000). Items banks will need to be developed to include items covering a broader range of person abilities. Therefore it can be assumed that item banks would need to be developed to include items covering the range of abilities between, at least –2.5 and +2.5 separated by approximately 0.5 logit intervals (e.g. Lai et al., 2003). However, using techniques such as "common anchoring" (e.g. Bode et al., 2003) of common items to produce item banks may initially avoid the necessity of additional data collection, and certainly reduce the number of questionnaires which subsequent patients would have to answer. These techniques rely on item estimates derived from common items or questionnaires from separate samples to be used as "anchors" to enable item parameters to be calculated for the remaining items not shared by the separate datasets. This form of item bank development potentially reduces the need for patients to be presented with large numbers of questionnaires.

There is also an issue surrounding the number of questions which patients would need to answer as part of computer-adaptive test outside of the development process. Clearly there is a potential for those patients with poorer quality of life to have to answer more items, particularly so when screening for instance, for psychological distress. However, recent work by Ware and his colleagues (e.g. Ware et al., 2003) on computer-adaptive questionnaires for headaches and migraines has demonstrated that a reliable six-item (static) questionnaire can be developed using item-response theory to measure headache impact (Kosinski et al., 2003). Although it could be argued that there is not much difference (albeit in a different quality of life

domain) between this and for instance the seven-item HADS-A, or MHI-5 for instance, in terms of the number of questions, which in turn would argue against the development of new scales, however the efficacy of this reduced scale improved on the screening efficacy of existing measures (Kosinski et al., 2003). Therefore computer-adaptive testing or static questionnaires developed using Rasch models may have the potential to improve screening or assessment of quality of life domains utilising fewer or equal numbers of items as used in questionnaires developed with traditional psychometrics.

Other issues which would need to be explored include whether items behave in isolation, as they do when part of a test, and whether person abilities estimated by computer-adaptive questionnaires are equivalent not only to estimates derived from standard testing, but also that person estimates for similar levels of ability which have been estimated by different items are also equivalent.

If computer-adaptive systems are to be developed for quality of life assessment in general, rather than specific domains, such as psychological distress, then one of the issues for future work, which was touched upon in the methodology discussion (10.2.2), will be the issue of unidimensionality. Cancer patients often present at clinics with multiple problems in different domains, which clearly would be problematical for unidimensionality. This is reinforced by the research on computer-adaptive systems to-date in this field which has been limited to single domains, such as fatigue (Lai et al., 2003) and headaches (Ware et al., 2003).

However, the two technologies explored in this thesis, computer-assisted questionnaires and computer-adaptive questionnaires using Rasch models, could be combined to form "testlets" (e.g. Wainer & Kiely, 1987), which could overcome the potential problem of unidimensionality, for instance by allowing patients to select quality of life domains, which would then be explored and assessed further by linking these to item banks, specific to each domain.

The idea of computer-assisted questionnaires or "branching questionnaires", which presented questions to patients from pre-selected screens, was explored to a limited extent in Chapter 4. The results from this study, which were supportive of previous research (e.g. Boyes et al., 2002) in terms of the overall pattern of responses demonstrated that agreement between standard and computer-assisted versions of questionnaires may be influenced by the design of the computer-assisted questionnaire, in particular the selection screen. Although the results of this study may have been limited by the small sample size and the fact that the majority of patients were female, it is clear that future work would also need to explore how the two technologies could be combined and designed optimally. However, the findings also demonstrated that a minority of patients' responses corresponded closely between the two versions of the questionnaire. This was almost exclusively limited to those patients who had selected items from the CA-questionnaire. Therefore, in addition to further work on the design of the CA-questionnaire, cognitive models, such as the Cognitive Aspects of Survey Methodology models (e.g. Tourangeau et al., 2000) need to be explored to investigate whether these may be able to shed light on the reasons why, in some cases, a (small) majority of patients do not select items from the CA-questionnaire selection screen (and conversely other patients do). Future work could focus on whether the design of the CA-questionnaire interacted with one or more stages of processing (comprehension, retrieval, judgment and response) to differentially affect responses to the selection screen by some patients.

Clearly this is an important question which needs to be answered if technologies, such as computer-assisted questionnaires and computer-adaptive systems are going to be combined.

The final issue for the development of computer-adaptive tests surrounds the notion of copyright. Many quality of life instruments have copyright restrictions imposed on them. Current research has been able to circumvent this issue by using either instruments which have no copyright restrictions or by the authors of the

instruments themselves adapting their own questionnaire for computer-adaptive testing (e.g. Ware et al., 2003). This is an aspect affecting the development of item banks and computer-adaptive systems, which has not as yet been addressed in the literature. Furthermore, given the fact that the necessity of large item banks for the development of these systems will be furthered by multinational collaboration this in turn raises the issue of deriving items banks from multiple language sources. Recent research has demonstrated that differential item functioning can be identified through mistranslations or lack or correspondence between the original languages and the translations (Petersen, et al., 2003). This may necessitate the development of smaller computer-adaptive systems in the first instance before the logistical problems pertaining to copyright and multi-language sources can be overcome.

### *10.2.3. Implications for clinical practice*

The work described in this thesis and the future work highlighted above will help to inform health care professionals of the efficacy of different quality of life instruments. In addition, it will also facilitate the work of the Psychosocial Oncology Group in Leeds, particularly the aspects relating to the introduction of regular symptom and quality of life assessment into the care of individual patients.

The work also contributes to the interpretation of quality of life instruments by demonstrating that most of quality of life scales of EORTC QLQ-C30 and FACT-G are not interval based. This means that any interpretation of changes to quality of life scores must be undertaken with some caution, since the impact of the change may differ depending on the locality of change along the scale. For some of the scales, such as the Emotional Functioning scale of EORTC QLQ-C30, the results suggest that it may be possible to use a threshold level (e.g. below or above 50). This may be more beneficial to clinicians, who may be accustomed to employing and interpreting thresholds from, for instance, laboratory tests. This area has important clinical implications and is worth exploring in the future research.

In general, clinicians need to be confident of the meaning of scores derived from quality of life measures, irrespective of the form of the questionnaires, particularly since the measurement of quality of life is becoming more commonplace in oncology clinics and since emerging research is demonstrating the utility of measurement for facilitating the clinical consultation (Detmar et al., 2002; Velikova et al., 2004). Clinicians need to be informed of the meaning of QOL measures (e.g. Cella et al., 2002a; King, 1996; Osoba et al., 1998) and provided with guidelines to interpreting QOL scores (Velikova et al., 2002) in order to be confident that scores represent "true" quality of life of patient. This may be particularly so where clinicians are faced with QOL scores which are changing over time, and where interpretation of these changes may reflect a "response shift" in the patient's attitude, coping mechanisms, evaluation of life, etc. (Sprangers and Schwartz, 2000).

Similarly, systems such as those discussed in this thesis, like CA-questionnaires and computer-adaptive tests must also be developed so that they provide "accurate" reflection of an individual patient's quality of life (e.g. McGee et al., 1991).

Finally, some of the results from the research in this thesis can be directly applied in clinical practice. A shorter HADS questionnaire is planned to be used in the future studies of the Psychosocial Oncology Group, provided that the copyright issues can be overcome. The development of shorter forms of questionnaires and computer-adaptive systems in the future will overall reduce patient burden, whilst still providing reliable and relevant information to the health professionals. These systems could potentially enhance – but not replace - communication between healthcare professionals and patients, and may in the long-term improve patient care.

## Appendix 1 - Hospital Anxiety and Depression Scale

Name: Date:

Doctors are aware that emotions play an important part in most illnesses. If your doctor knows about these feelings he will be able to help you more. This questionnaire is designed to help your doctor know how you feel. Read each item and place a firm tick in the box opposite the reply which comes closest to how you have been feeling in the past week. Don't take too long over your replies: your immediate reaction to each item will probably be more accurate than a long thought-out response.

*Tick only one box in each section*

**I feel tense or 'wound up':**
Most of the time.....
A lot of the time.....
Time to time, occasionally.....
Not at all.....

**I still enjoy the things I used to enjoy:**
Definitely as much.....
Not quite so much.....
Only a little.....
Hardly at all.....

**I get a sort of frightened feeling as if Something awful is about to happen:**
Very definitely and quite badly.....
Yes, but not too badly.....
A little, but it doesn't worry me.....
Not at all.....

**I can laugh and see the funny side of things**:
As much as I always could.....
Not quite so much now.....
Definitely not so much now.....
Not at all.....

**Worrying thoughts go through my mind:**
A great deal of the time.....
A lot of the time.....
From time to time but not too often...
Only occasionally.....

**I feel cheerful:**
Not at all.....
Not often.....
Sometimes.....
Most of the time.....

**I can sit at ease and feel relaxed:**

Definitely.....
Usually.....
Not often.....
**Not at all.....**

**I feel as if I am slowed down:**
Nearly all the time.....
Very often.....
Sometimes.....
Not at all.....

**I get a sort of frightened feeling like 'butterflies' in the stomach:**
Not at all.....
Occasionally.....
Quite often.....
Very often.....

**I have lost interest in my appearance:**
Definitely.....
I don't take as much care as I should..
I may not take quite as much care.....
I take just as much care as ever.....

**I feel restless as if I have to be on the move:**
Very much indeed.....
Quite a lot.....
Not very much.....
Not at all.....

**I look forward with enjoyment to things:**
As much as I ever did.....
Rather less than I used to.....
Definitely less than I used to.....
Hardly at all.....

**I get sudden feelings of panic:**
Very often indeed.....
Quite often.....
Not very often.....
Not at all.....

**I can enjoy a good book or radio or TV programme:**
Often.....
Sometimes.....
Not often.....
**Very seldom.....**

# Appendix 2 – The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire

**EORTC QLQ-C30 (version 3.0)**

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

Please fill in your initials: |_____|

Please fill in your surname:
|_____|

Your birth date: |__|__|__|__|__|__|

Today's date: |__|__|__|__|__|__|

---

| | | Not at All | A Little | Quite a Bit | Very Much |
|---|---|---|---|---|---|
| 1. | Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase? | 1 | 2 | 3 | 4 |
| 2. | Do you have any trouble taking a long walk? | 1 | 2 | 3 | 4 |
| 3. | Do you have any trouble taking a short walk outside of the house? | 1 | 2 | 3 | 4 |
| 4. | Do you have to stay in a bed or a chair for most of the day? | 1 | 2 | 3 | 4 |
| 5. | Do you need help with eating, dressing, washing yourself or using the toilet? | 1 | 2 | 3 | 4 |

| **During the past week:** | | **Not at All** | **A Little** | **Quite a Bit** | **Very Much** |
|---|---|---|---|---|---|
| 6. | Were you limited in doing either your work or other daily activities? | 1 | 2 | 3 | 4 |
| 7. | Were you limited in pursuing your hobbies or other Leisure time activities? | 1 | 2 | 3 | 4 |
| 8. | Were you short of breath? | 1 | 2 | 3 | 4 |
| 9. | Have you had pain? | 1 | 2 | 3 | 4 |
| 10. | Did you need to rest? | 1 | 2 | 3 | 4 |
| 11. | Have you had trouble sleeping? | 1 | 2 | 3 | 4 |
| 12. | Have you felt weak? | 1 | 2 | 3 | 4 |
| 13. | Have you lacked appetite? | 1 | 2 | 3 | 4 |
| 14. | Have you felt nauseated? | 1 | 2 | 3 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| 15. | Have you vomited? | 1 | 2 | 3 | 4 |

| **During the past week:** | **Not at All** | **Not at Little** | **AQuite a Bit** | **Very Much** |
|---|---|---|---|---|
| 16. Have you been constipated? | 1 | 2 | 3 | 4 |
| 17. Have you had diarrhoea? | 1 | 2 | 3 | 4 |
| 18. Were you tired? | 1 | 2 | 3 | 4 |
| 19. Did pain interfere with your daily activities? | 1 | 2 | 3 | 4 |
| 20. Have you had difficulty in concentrating on things, like reading a newspaper or watching television? | 1 | 2 | 3 | 4 |
| 21. Did you feel tense? | 1 | 2 | 3 | 4 |
| 22. Did you worry? | 1 | 2 | 3 | 4 |
| 23. Did you feel irritable? | 1 | 2 | 3 | 4 |
| 24. Did you feel depressed? | 1 | 2 | 3 | 4 |
| 25. Have you had difficulty remembering things? | 1 | 2 | 3 | 4 |
| 26. Has your physical condition or medical treatment interfered with your family life? | 1 | 2 | 3 | 4 |
| 27. Has your physical condition or medical treatment interfered with your social activities? | 1 | 2 | 3 | 4 |
| 28. Has your physical condition or medical treatment caused you financial difficulties? | 1 | 2 | 3 | 4 |

**For the following questions please circle the number between 1 and 7 that best applies to you**

29. How would you rate your overall health during the past week?

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Very poor | | | | | | Excellent |

30. How would you rate your overall quality of life during the past week?

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| Very poor | | | | | | Excellent |

# Appendix 3 – Functional Assessment of Cancer – General

## FACT-G (Version 4)

Below is a list of statements that other people with your illness have said are important. By circling **one (1) number per line, please indicate how true each statement has been for you during the past 7 days.**

| | PHYSICAL WELL-BEING | Not at all | A little bit | Some-what | Quite a bit | Very much |
|---|---|---|---|---|---|---|
| GP1 | I have a lack of energy | 0 | 1 | 2 | 3 | 4 |
| GP2 | I have nausea | 0 | 1 | 2 | 3 | 4 |
| GP3 | Because of my physical condition, I have trouble meeting the needs of my family | 0 | 1 | 2 | 3 | 4 |
| GP4 | I have pain | 0 | 1 | 2 | 3 | 4 |
| GP5 | I am bothered by side effects of treatment | 0 | 1 | 2 | 3 | 4 |
| GP6 | I feel ill | 0 | 1 | 2 | 3 | 4 |
| GP7 | I am forced to spend time in bed | 0 | 1 | 2 | 3 | 4 |

| | SOCIAL/FAMILY WELL-BEING | Not at all | A little bit | Some-what | Quite a bit | Very much |
|---|---|---|---|---|---|---|
| GS1 | I feel close to my friends | 0 | 1 | 2 | 3 | 4 |
| GS2 | I get emotional support from my family | 0 | 1 | 2 | 3 | 4 |
| GS3 | I get support from my friends | 0 | 1 | 2 | 3 | 4 |
| GS4 | My family has accepted my illness | 0 | 1 | 2 | 3 | 4 |
| GS5 | I am satisfied with family communication about my illness | 0 | 1 | 2 | 3 | 4 |
| GS6 | I feel close to my partner (or the person who is my main support) | 0 | 1 | 2 | 3 | 4 |
| Q1 | *Regardless of your current level of sexual activity, please answer the following question. If you prefer not to answer it, please check this box* ☐ *and go to the next section.* | | | | | |
| GS7 | I am satisfied with my sex life | 0 | 1 | 2 | 3 | 4 |

FACT-G (Version 4)

**By circling one (1) number per line, please indicate how true each statement has been for you during the past 7 days.**

| | **EMOTIONAL WELL-BEING** | **Not at all** | **A little bit** | **Some-what** | **Quite a bit** | **Very much** |
|---|---|---|---|---|---|---|
| GB1 | I feel sad | 0 | 1 | 2 | 3 | 4 |
| GB2 | I am satisfied with how I am coping with my illness | 0 | 1 | 2 | 3 | 4 |
| GB3 | I am losing hope in the fight against my illness | 0 | 1 | 2 | 3 | 4 |
| GB4 | I feel nervous | 0 | 1 | 2 | 3 | 4 |
| GB5 | I worry about dying | 0 | 1 | 2 | 3 | 4 |
| GB6 | I worry that my condition will get worse | 0 | 1 | 2 | 3 | 4 |

| | **FUNCTIONAL WELL-BEING** | **Not at all** | **A little bit** | **Some-what** | **Quite a bit** | **Very much** |
|---|---|---|---|---|---|---|
| GF1 | I am able to work (include work at home) | 0 | 1 | 2 | 3 | 4 |
| GF2 | My work (include work at home) is fulfilling | 0 | 1 | 2 | 3 | 4 |
| GF3 | I am able to enjoy life | 0 | 1 | 2 | 3 | 4 |
| GF4 | I have accepted my illness | 0 | 1 | 2 | 3 | 4 |
| GF5 | I am sleeping well | 0 | 1 | 2 | 3 | 4 |
| GF6 | I am enjoying the things I usually do for fun | 0 | 1 | 2 | 3 | 4 |
| GF7 | I am content with the quality of my life right now | 0 | 1 | 2 | 3 | 4 |

# Appendix 4 – Coding for Computer-Assisted Version of the EORTC QLQ-C30

A total of five screens were designed to present the questionnaire to the patients and to collect the data using three forms.

## 1. Startup screen[1]

As the form for the Startup/barcode (**frmBarcode**) screen was loaded global variables were declared for the unique hospital identifier (hospitalid), and patient's initials (initials), and surname (surname). These variables were declared as variant .

The pulldown menu Program was used to control which questionnaires were presented to the patient, as well as containing the Quit command to end the programme. Both the *Click* events for the MBQ (Multiple Branching Questionnaire) and the QLQ (Quality-of-life Questionnaire) contained virtually identical code. **If-Else** statements were used to ascertain whether the *mnuMBQ/mnuQLQ Checked* properties were set to *True* or *False*. If the *Checked* property was set to *True* the programme simply reset it to *False* and vice versa. This property was used as a flag to determine which questionnaire would be presented to the patient later in the programme.

When patients scanned their barcode the number appeared in the textbox (*txtDemographics*) in the centre of the screen. At this stage patients had the opportunity to check the number appearing in the textbox and clear it if the barcode number was incorrect. The code for the 'Change' button simply reset the *Text* property of *txtDemographics* to contain no text.

---

[1] Bold lettering is used in this section to indicate Visual Basic forms or functions, whereas italic lettering is used for events or properties of objects on the forms.

The 'OK' button contained several functions. The **If-ElseIf-Else** structure initially simply checked whether the *txtDemographics* textbox was empty and presented an error message on the screen if the textbox was empty to ensure patients' details were recorded, the **Else-If** structure ensured that a six-digit code had been entered for the unique hospital identifier by using the **IsNumeric** and **Left** functions to extract the first six characters (**Left**(*txtDemographics*, 6) and checked whether this target is a numeral. The ElseIf structure presented an error message if this was not the case. Both these measures were included to minimize errors being introduced through unauthorized or incorrect use of the programme. When both If and ElseIf conditions had been met the code called the 'details' submodule.

The 'details' submodule employed two variables: 'Start' was a local variable that specified the position the cursor started from to derive the surname, and 'Target' which contained the current character being read by the programme.

Since all the barcodes had the same overall pattern, namely a six-digit number, a space, the surname, a dot and two initials, this structure could be used to greatly facilitate extracting the three variables from the single barcode. The 'hospitalid' variable was derived using the **Left** function to extract the first six characters from the text contained in the *txtDemographics* textbox e.g. (**Left**(*txtDemographics*, 6). Similarly the 'initials' were derived using the **Right** function to extract the last two characters from the *txtDemographics* textbox, since the last two characters are always the patient's initials.

A **Do-While** loop was employed to determine the 'surname' variable. The 'Target' was set using the **Mid** function and the 'Start' variable initialized to 8 (six characters allowed for the hospital identifier and a space before the surname). The **Mid**-function extracts a character or characters between two positions specified as parameters, e.g. **Mid**(txtDemographics, Start, 1) would set the first cursor at the position defined by the 'Start' variable, in this case position eight, and the last final cursor at the position defined by the second parameter, in this one character further

along from 'Start'. The intervening character(s) would then be extracted. In this programme the **Mid**-function was used to remove one character at a time in the loop. The end condition of the **Do**-loop was the full-stop character (the barcodes contained a full-stop between the surname and the initials). An **If-Else** statement checked whether the 'Target' variable met the end conditions, in which case the programme would exit from the loop, otherwise the 'Target' character was concatenated with the 'surname' variable, which had been initialized to contain nothing, and the 'Start' variable was incremented by one to move the **Mid**-function along by a single character. In this way the patient's surname was derived one character at a time until the programme encountered a full-stop at which point it would exit the **Do**-loop and the programme control would be returned to the code in the 'OK' button.

The next section of code in the 'OK' button set the *Text* property of four textboxes from the questionnaire screen (**frmQuestions**) to equal 'hospitalid', 'surname', 'initials' and the system (i.e. the current) date respectively.

The final section of code set the *Caption* property of the label from the introduction screen (**frmIntro**) to equal the introduction and instructions for the patient, before showing the introduction screen using the **Visible** function.

**2. Introduction screen**

The Introduction screen or form **frmIntro** only contained code for the *Click* event of the label. A series of **If-Nested If- ElseIf** statements controlled the flow of the programme. When the label (*Label1*) was clicked-on the first **If**-statement determined whether the instructions currently being presented related to the first set of instructions. If this condition was true then the **Nested-If** statement determined whether the programme should present the autoselection questionnaire or the standard questionnaire. If the *mnuMBQ* property had been set to *True* then the programme presented the instructions for the autoselection questionnaire on the screen (i.e. the *Label.Caption* property was set to equal the instructions), otherwise if

the *mnuQLQ* property had been set to *True* then the Questionnaire screen (**frmQuestions**) was shown (**frmQuestions**.*Visible* was set to *True*), and the Introduction screen was hidden (**frmIntro**.*Visible* = *False*). A textbox's *Text* property from the Questionnaire screen was set to equal the system (current) time, and the *lblQuestions*-label from this screen used for presenting the questions and its' *Caption* property was set to equal the first question from the standard questionnaire.

If none of the conditions from the **If-Nested-If** statements had been met then the next **ElseIf** statement determined whether the *Label1.Caption* was set to equal the instructions from the autoselection questionnaire. If this condition was satisfied then the selection screen (**frmSelect**) was made visible. Otherwise the final **ElseIf** statement determined whether the message thanking the patients had been presented, in which case the Startup/barcode screen was made visible and the Introduction screen was hidden.

**3. Selection screen**

Thirteen global variables were declared as integers for each of the scales presented to the patients. These variables were used as flags to indicate the scales the patients had selected. In the **Form_Load** section of the selection screen or **frmSelect**, the variables were initialized to zero.

Each of the labels used to represent the scales contained similar code on the **Click**-event. An **If-ElseIf** statement was used to check the status of the labels, i.e. whether they had been selected or not. The **If**-condition determined whether the label colour was 'White' (not selected), if this condition was met the Label.*BackColor* was changed to 'Yellow' and the flag variable for that scale was set to one to indicate that the label or scale had been selected. If that condition was not satisfied the **ElseIf**-condition checked whether the label had already been selected (i.e. it was 'Yellow'), and if this was the case the Label.*BackColor* was changed to 'White' and the appropriate flag variable set to zero to indicate that the scale had been de-selected.

Two submodules 'eortc1' and 'eortc2' were defined to be used in the Questionnaire screen to setup the response buttons when the Questionnaire screen was loaded. Module 'eortc1' was used to define the four response modes from the EORTC-QLQ c30 corresponding to 'Not at all', 'A little', 'Quite a bit', and 'Very much'. This code moved the frame, *Frame1* from the Questionnaire form over to the centre of the screen by adjusting the *Left*-property of the frame. *Frame1* contained seven buttons, numbered from zero to six. Buttons 0, 5, and 6 were hidden by setting the *Visible*-property of the buttons to *False*, i.e. **frmQuestions**.*Frame1.Command(0).Visible* = *False*, and the *Caption*-property for the remaining four buttons (1 to 4) was set to correspond to the appropriate captions, e.g. 'Not at all' (Command1(1)), 'A little' (Command1(2)), etc.

The code for module 'eortc2' was similar to the code for the 'eortc1' module. This module was used for the Global Health Status questions from the EORTC-QLQ c30, which contain seven response modes numbered one to seven. The extremes of the scale correspond to 'Very poor' (one) and 'Excellent' (seven). *Frame1* was moved over to the left to accommodate the extra response buttons by resetting the *Left*-property of the frame. The previously hidden command buttons (0, 5, and 6) were made visible by resetting their *Visible*-property to *True*, and a **For-Next** was used to change the command button captions to 1 to 7. The **For-Next** loop incremented a local variable 'j' through 0 to 6 and at each iteration the corresponding command button caption was set to equal j + 1.

Two labels from the Questionnaire screen whose captions read 'Very poor' and 'Excellent' were also made visible.

The *Click*-event for the *cmdContinue* command button contained the code to enable the programme to present the first question from the questionnaire to the patient on the Questionnaire screen. Initially the response buttons on the Questionnaire screen were set to correspond to the four buttons response mode by function call, calling the 'eortc1' module. A series of **If-ElseIf** statements was used to

determine which of the scales had been selected by the patient by checking which of the flag variables had been set to one.

The labels representing the scales were ordered into two columns as they were presented on the Selection screen. The programme moved through the **If-ElseIf**-statements checking each of the flag variables until it came across a variable that had been set to one. At this point the *lblQuestions*-label from the Questionnaire screen used to present the questions to the patients, *Caption*-property was set to equal the first question from the scale selected, e.g. if the Physical functioning scale had been selected then **frmQuestions**.*lblQuestions.Caption* = "Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?"

If none of the scales had been selected then the final **Else**-condition set the *lblQuestions.Caption* to equal the first question from the Global Health Status questions.

The Questionnaire was then presented to the patient and the Selection screen hidden, i.e. their *Visible*-properties were set to *True* and *False* respectively. Finally, the *Text*-property from the *txtTime1* textbox from the Questionnaire screen was set to equal the current (system) time to record the time the patient started questionnaire proper.

**4. Questionnaire screen**

A global variable 'counter' was used as marker to move through the questions to record the appropriate response to each question. The EORTC-QLQ c30 contains 30 questions, and some of these questions would potentially not be presented to patients as a consequence of the patients not selecting the scale from the Selection screen. In order to move through the appropriate questions 'counter' was used as a dynamic variable to indicate the start and end points of the questions relating to each scale so that the answers would be stored in the appropriate fields in the database.

Three modules were defined for the Questionnaire screen or **frmQuestions**: 'nextquestion' and 'nextquestion2' were used to present patients with the autoselection questionnaire and the standard questionnaire respectively, and 'check' which was used to determine which scales had been selected by the patients.

The 'nextquestion2' module consisted of a series of **If-ElseIf** statements which simply checked which question had been presented to the patient by reading the *Caption*-property of the *lblQuestions* label and presenting the next question in the list. The programme presented all thirty questions from the EORTC-QLQ c30 serially in this manner. After question 5 from the standard questionnaire the *lblDuring* label was made visible. The caption for this label read "During the last week", and is used on the EORTC-QLQ c30 for questions 6 to 28. Following question 28, the 'eortc2' module from the Selection form was invoked to alter the response frame and command buttons to include the seven response buttons for the Global Health Status questions. At the end of question 30 the Thank you-screen was presented to the patients, the Questionnaire screen was hidden, and the *Text*-property of the *Time2* textbox was set to the current time.

The 'nextquestion' module proceeded along similar lines to the 'nextquestion2' module with two important exceptions. Once the questions for a particular scale had been completed the corresponding flag variable (i.e. the global variable declared in the Selection screen) was set to zero to indicate to the programme that the questions had already been presented and to prevent the programme from repeating the same questions *ad infinitum*. Secondly, once the questions had been completed the code invoked the 'check' module.

The 'check' module consisted of a series of **If-Else** statements which determined the next scale to have been selected from the Selection screen and present the appropriate first question from the scale to the patient. The 'check' module also set the 'counter' variable. Once the selected questions had been presented to the patient they were also presented with the two questions from the Global Health Status scale,

and following this the Thank you-screen was presented, the Questionnaire screen was hidden, and the *Text*-property of the *Time2* textbox was set to the current time.

A *Data* object (*Data1*) was used to link the programme to an MS-Access database in order to store the patients' responses. The responses from the patients were stored in an array of textboxes named *txtScore*, whose *DataSource*-property had been set to *Data1* and whose *DataFields*-property had been set to the appropriate fields from the database.

The *Click*-event of the response buttons (*Command1*) contained **If-Else** statements which determined how the patients had responded by checking the *Caption*-property of the command button that had been pressed by the patients. The 'counter' variable was used as the index for the *txtScore* arrays to store the response in the correct textbox linked to the appropriate datafield. The *Text*-property of *txtScore* was set to equal either 1 to 4 (for the 'Not at all' to 'Very much' reponses) or 1 to 7 for the Global Health Status questions. After each response the 'counter' variable was incremented by one. An **If-ElseIf** statement checked which questionnaire was being presented to the patients by determining the status of *Checked* property of the *mnuMBQ*/*mnuQLQ* menu items.

Once the final questions had been presented on both forms of the questionnaire the temporary variables and textboxes were cleared using the **Unload** function.

**5. Final screen**

The final screen presented a message to the patients thanking them for completing the questionnaire and asking them to press the label containing the text to complete the task. The code for this screen has been described in the Introduction screen section (section 1).

**6. Data Storage**

An MS-Access 97 database was created to store the data generated from patients' responses to the questionnaires. The database, **Dq**, contained a single table, **ql**, for data storage. Two fields were created for storing the patients' surnames and initials. Both these fields were of datatype *Text* and had field sizes of 20 and 5 for surname and initials respectively. A field was created to store he unique hospital number. This was of datatype *Number* and had a field size *Double*. The date the patients completed the questionnaires, and the time they started and finished the tasks were stored in three fields of type *Date/Time*. The responses to the questionnaires were stored in thirty fields corresponding to the thirty questions from the EORTC-QLQ c30, and were of datatype *Number* and had a fieldsize of *Integer*. The default value for each of these fields was set to blank to allow for patients not selecting scales during the autoselection questionnaire.

The **Dq** database also contained a single query, **eortcq**, which comprised fields for patients' surnames, initials, hospital number and the date they completed the questionnaires, as well as fifteen fields containing the algorithms to convert the raw scores from the questionnaires to scaled scores. The algorithms were provided through the EORTC QLQ c30 Scoring Manual (Fayers, Aaronson, Bjordal, and Sullivan, 1995):

$$RawScore = (I_1 + I_2 + \ldots + I_n)/n$$

Functional scales:

$$Score = (1 - (RawScore - 1)/range)) * 100$$

Symptom scales and Global Health Status:

3. $Score = ((RawScore - 1)/range) * 100$

Range = (number of responses – 1), i.e. functional and symptom scales will have a range of three, and Global Health Status will have a range of six. These algorithms were used to convert the ordinal data from the questionnaires into continuous data ranging from 0 to 100. The algorithms depend on complete or near complete data for each scale and were therefore only used to convert the scores from the standard

questionnaire to avoid incorrect scores being produced through the missing data from the autoselection.

All the data generated from the questionnaires was stored temporarily on the Questionnaire screen (**frmQuestions**) as the patient proceeded through the programme. The Visual Basic *Database* object (*Data1*) was used to link the programme to the **Dq** database. The *Connect*-property of *Data1* was set to 'Access'. The *DatabaseName*-property of *Data1* was set to **Dq**, and this also contained information of path of the database, i.e. where the database was stored on the computer (c:\Dq\Dq.mdb). Similarly, the *RecordSource*-property of *Data1* was set to equal the **ql** table.

As described in section 5 the Questionnaire screen contained an array of thirty textboxes (*frmScores*) to store patients' responses to the questions, as well as six textboxes which were used to store patient details and time and date of completion of the questionnaires. The *DataSource*-property of each of these thirty-six textboxes was set to equal *Data1*, and their *DataField*-property was set to correspond to the appropriate field from the **ql** table.

## Bibliography

Aaronson, N.K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N.J., Filiberti, A., Flechtner, H., Fleishman, S.B., de Haes, J.C.J.M., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P.B., Schraub, S., Sneeuw, K., Sullivan, M., Takeda, F. for the EORTC Study Group on Quality of Life (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality of life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute* **85**, 365-376.

Aaronson, N.K., Ahmedzai, S. Bullinger, M., Crabeels, D., Estape, J., Filiberti, A., Flechtner, H., Frick, U., Hürny, C., Kaasa, S., Klee, M., Mastikica, M., Osoba, D., Pfansler, B., Razavi, D., Rofe, P.B.C., Schraub, S., Sullivan, S. & Taheda, F. for the EORTC Study Group on Quality of Life. (1991). The EORTC Core Quality of Life Questionnaire: Interim results of an international field study. In Osoba, D. (ed.) *Effect of Cancer on Quality of Life.* CRC Press Inc: Boca Raton, pp. 185-203.

Abiodun, O.A. (1994). A validity study of the Hospital Anxiety and Depression Scale in general hospital units and a community sample in Nigeria. *British Journal of Psychiatry* **165**, 669-672.

Allenby, A., Matthews, J., Beresford, J., & McLachlan, S.A. (2002). The application of computer touch-screen technology in screening for psychosocial distress in an ambulatory oncology setting. *European Journal of Cancer Care* **11**, 245-253.

Andersson, E. (1993). The Hospital Anxiety and Depression Scale. Homogeneity of the subscales. *Social Behavior and Personality* **21**, 197-204.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika* **43**, 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement* **38**, 665-680.

Angoff, W.H. (1993) Perspectives on differential item functioning methodology. Holland, P.W., & Wainer, H. (Eds.). *Differential item functioning*. Lawrence Erlbaum Associates: Hillsdale, NJ.

Baddeley, A. (1992). Working memory. *Science* **255**, 556-559.

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland: Maryland.

Beck, A.T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961) An inventory for measuring depression. *Archives of General Psychiatry* **4**, 561-571.

Bergman, B., Aaronson, N.K., Ahmedzai, S., Kaasa, S., & Sullivan, M. (1994). The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. *European Journal of Cancer* **30A**, 635-642.

Bergner, M., Bobbitt, R.A., Carter, W.B., & Gilson, B.S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care* **19**, 787-805.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M., & Novick, M.R. (Eds.). *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, Mass.

Bjordal K, Kaasa S. (1992). Psychometric validation of the EORTC Core Quality of Life Questionnaire, 30-item version and a diagnosis-specific module for head and neck cancer patients. *Acta Oncologica* **31**, 311-221.

Bjorner, J.B., Kosinski, M., & Ware JE. (2003). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research* **12**, 981-1002.

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical assessment. *Lancet* **8**, 307 – 310.

Blazeby, J.M., Alderson, D., Winstone, K., et al. (1996). Development of an EORTC questionnaire module to be used in quality of life assessment for patients with oesophageal cancer. The EORTC Quality of Life Study Group. *European Journal of Cancer* **32A**, 1912-1917.

Bliss, J.M., Robertson, B. & Selby, P.J. (1992). The impact of nausea and vomiting upon quality of life measures. *British Journal of Cancer* **66**(Supplement XIX), 14-23.

Bliss, J.M., Selby, P.J., Robertson, B., & Powles, T.J. (1992). A method for assessing the quality of life of cancer patients: Replication of the factor structure. *British Journal of Cancer* **65**, 961-966.

Bliven, B.D., Kaufman, S.E., & Spertus, J.A. (2001). Electronic collection of health-related quality of life data: validity, time benefits, and patient preference. *Quality of Life Research* **10**, 15-22.

Bond, T.G., & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Lawrence Erlbaum Associates: New Jersey.

Bode, R.K., Cella, D., Lai, J.S., Heineman, A.W. (2003). Developing an initial physical function item bank. *Journal of Applied Measurement* **4**, 124-136.

Boyes A, Newell S, Girgis A. (2002). Rapid assessment of psychosocial well-being: are computers the way forward in a clinical setting? *Quality of Life Research* **11**, 27-35.

Brady, M.J., Cella, D.F., Mo, F., Bonomi, A.E., Tulsky, D.S., Lloyd, S.R., Deasy, S., Cobleigh, M., & Shiomoto, G. (1997) Reliability and validity of the Functional Assessment of Cancer Therapy-Breast quality-of-life instrument. *Journal of Clinical Oncology* **15**, 974-86, 1997.

Broadbent, D.E. (1981).Selective and control processes. *Cognition* **10**, 53-58.

Browne, J.P., O'Boyle, C.A., McGee, H.M., Joyce, C.R.B., McDonald, N.J., O'Malley, K., & Hiltbrunner, B. (1994). Individual quality of life in the healthy elderly. *Quality of Life Research* **3**, 235-244.

Browne, J.P., O'Boyle, C.A., McGee, H.M., McDonald, N.J., Joyce, C.R.B. (1997). Development of a direct weighting procedure for quality of life domains. *Quality of Life Research* **6**, 301-309.

Buxton, J., White, M., & Osoba, D. (1998). Patients' experiences using a computerized program with a touch-sensitive video monitor for the assessment of health-related quality of life. *Quality of Life Research* **7**, 513-519.

Campbell, N.R. (1920). *Physics, the elements.* Cambridge: Cambridge University Press.

Calman, K.C. (1984). Quality of life in cancer patients--an hypothesis. *Journal of Medical Ethics* **10**, 124-127.

Carlson, L.E., Taenzer, P.A., Bultz, B.D., et al. (1999). Computerized quality of life screening in an outpatient lung cancer clinic. *Psycho-Oncology* **8**, 5.

Carlson, L.E., Speca, M., Hagen, N., & Taenzer P. (2001). Computerized quality-of-life screening in a cancer pain clinic. *Journal of Palliative Care* **17**, 46-52.

Cattell, J.M. (1890). Mental tests and measurement. *Mind* **15**, 373-380.

Cella, D.F. (1994). Quality of life: concepts and definition. *Journal of Pain and Symptom Management* **9**, 186-192.

Cella, D.F. (1997). *FACIT Manual.* Evanston, Centre on Outcomes, Research and Education.

Cella, D.F. (2002). The Functional Assessment of Cancer Therapy-Anemia (FACT-An) Scale: A new tool for the assessment of outcomes in cancer anemia and fatigue. *Seminars in Hematology* **34**, 13-19.

Cella, D.F., Bonomi, A.E., Lloyd, S.R., Tulsky, D.S., Kaplan, E., & Bonomi, P. (1995). Reliability and validity of the Functional Assessment of Cancer Therapy-Lung (FACT-L) quality of life instrument. *Lung Cancer* **2**, 199-220.

Cella, D.F., & Cherrin, E.A. (1988). Quality of life during and after cancer treatment. *Comprehensive Therapy* **14**, 69-75.

Cella, D., Eton, D.T., Lai, J.S., Peterman, A.H., & Merkel, D.E. (2002a). Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *Journal of Pain and Symptom Management* **24**, 547-561.

Cella, D., Hahn, E.A., & Dineen, K. (2002b). Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Quality of Life Research* **11**, 207-221.

Cella, D.F., Tulsky, D.S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S.B., Winicour, P., Brannon, J., Eckberg, K., Lloyd, S., Purl, S., Blendowski, C., Goodman, M., Barnicle, M., Stewart, I., McHale, M., Bonomi, P., Kaplan, E., Taylor, S., IV, Thomas, C.R., & Harris, J. (1993). The Functional Assessment of Cancer Therapy Scale: Development and validation of the general measure. *Journal of Clinical Oncology* **11**, 570-579.

Clark, L.A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology* **100**, 316-336.

Clark, L.A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. *Journal of Abnormal Psychology* **103** 103-116.

Coates, A., Dillenbeck, C.F., McNeil, D.R., et al. (1983). On the receiving end – II. Linear analogue self-assessment (LASA) in evaluation of aspects of the quality of life of cancer patients receiving therapy. *European Journal of Cancer and Clinical Oncology* **19**, 1633-1637.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1988.

Coombs, C.H. (1964). *A Theory of Data.* New York: John Wiley & Sons.

Cook, K.F., Rabeneck, L., Campbell, C.J.M. et al. (1999). Evaluation of a multidimensional measure of dyspepsia-related health for use in a randomized clinical trial. *Journal of Clinical Epidemiology* **52,** 381-392.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* **24**, 87-185.

Cull, A., Gould, A., Chappell, J., Wright, E.P., Smith, A.B. et al. (2001). Validating automated screening for psychological distress by means of computer touchscreens for use in routine oncology practice. *British Journal of Cancer* **85**, 1842-1849.

Cull, A., Stewart, M., & Altman, D.G. (1995). Assessment of and intervention for psychosocial problems in routine oncology practice. *British Journal of Cancer* **72**, 229-235.

Dagnan, D., Chadwick, P., & Trower, P. (2000). Psychometric properties of the Hospital Anxiety and Depression Scale with a population of members of a depression self-help group. *British Journal of Medical Psychology* **73**, 129 – 137.

De Haes, J.C. (1988). Quality of life: Conceptual and theoretical considerations. In Watson, M., Greer, S., Thomas, C. (Eds.). *Psychosocial Oncology*. Oxford: Pergamon Press, pp. 61-70.

De Haes, J.C., van Knippenberg, F.C., & Neijt, J.P. (1990). Measuring psychological and physical distress in cancer patients: Structure and application of the Rotterdam Symptom Checklist. *British Journal of Cancer* **62**, 1034-1038.

Detmar, S.B. & Aaronson, N.K. (1998). Quality of life assessment in daily clinical oncology practice: A feasibility study. *Quality of Life Research* **34**, 1181-1186.

Detmar, S.B., Muller, M.J., Schornagel, J.H., Wever, L.D., & Aaronson N.K. (2002). Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial. *Journal of the American Medical Association* **288**, 3027-3034.

Doward, L.C., Spoorenberg, A., Cook, S.A., Whalley, D., Helliwell, P.S., Kay, L.J., McKenna, S.P., Tennant, A., van der Heijde, D., & Chamberlain, M.A. (2003). Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Annals of the Rheumatic Diseases* **62**, 20-26.

Doyle, C., Crump, M., Pintilie, M., et al. (2001). Does palliative chemotherapy palliate? Evaluation of expectations, outcomes, and costs in women receiving chemotherapy for advanced ovarian cancer. *Journal of Clinical Oncology* **19**, 1266-1274.

Drummond, H.E., Ghosh, S., Ferguson, A., Brackenridge, D. & Tiplady, B. (1995). Electronic quality of life questionnaires: A comparison of pen-based electronic questionnaires with conventional paper in a gastrointestinal study. *Quality of Life Research* **4**, 21-26.

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). A confirmatory factor analysis of the Hospital Anxiety and Depression scale: Comparing empirically and theoretically derived structures. *British Journal of Clinical Psychology* **39**, 79 – 94.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Fayers, P., Aaronson, N., Bjordal, K., et al. (2001). *EORTC QLQ-c30 Scoring Manual*. EORTC Quality of Life Group, Brussels.

Fayers, P.M., Hopwood, P., Harvey, A., Girling, D.J., Machin, D., & Stephens R. (1997). Quality of life assessment in clinical trials guidelines and a checklist for protocol writers: The U.K. Medical Research Council experience. MRC Cancer Trials Office. *European Journal of Cancer* **33**, 20-28.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error. *Monthly Notices of the Royal Astronomical Society* **53**, 758-770.

Ford, S., Fallowfield, L., & Lewis, S. (1994). Can oncologists detect distress in their out-patients and how satisfied are they with their performance during bad news consultations? *British Journal of Cancer* **70**, 767-770.

Ganz, P.A. (1994). Long-range effect of clinical trial interventions on quality of life. *Cancer* **74**, 2620-2624.

Ganz, P.A., Moinpour, C.M., Cella, D.F., et al. (1992). Quality-of-life assessment in cancer clinical trials: A status report. *Journal of the National Cancer Institute* **84**, 994-995, 1992.

Gonin, R., Lloyd, S., Cella, D., & Gray, G. (1996). Establishing equivalence between scaled measures of quality of life. *Quality of Life Research* **5**, 20-26.

Groenvold, M., Bjorner, J.B., Klee, M.C., & Kreiner, S. (1995). Test for item bias in a quality of life questionnaire. *Journal of Clinical Epidemiology* **48**, 805 –816.

Groenvold, M., Fayers, P.M., Sprangers, M.A., Bjorner, J.B., Klee, M.C., Aaronson, N.K., Bech, P., & Mouridsen, H.T. (1999). Anxiety and depression in breast cancer patients at low risk of recurrence compared with the general population: a valid comparison? *Journal of Clinical Epidemiology* **52**, 523-530.

Groenvold, M., & Petersen, M.A. (2001). Using item response theory to shorten the length of scales in a questionnaire for palliative care. *Quality of Life Research* **10**, 196.

Groenvold, M., Petersen, M.Aa., & Bjorner, J. (2000). Use of item response theory (IRT) to develop a shortened version of the EORTC QLQ-C30 for palliative care patients: Emotional Functioning Scale. *Internal report to the EORTC Quality of Life Group*.

Guilford, J.P. (1954). *Psychometric methods*. New York: McGraw Hill.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star,S., & Clausen, J.A. (Eds.). *Measurement and Prediction* **4**, Princeton: Princeton University Press, pp. 60-90.

Hall, A., A'Hern, R., & Fallowfield, L. (1999). Are we using appropriate self-report questionnaires for detecting anxiety and depression in women with early breast cancer? *European Journal of Cancer* **35**, 79-85.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Sage: Newbury Park, CA.

Herrmann, C. (1997). International experiences with the Hospital Anxiety and Depression Scale: A review of validation data and clinical results. *Journal of Psychosomatic Research* **42**, 17-41.

Hickey, A.M., Bury, G., O'Boyle, C.A., Bradley, F., O'Kelly, F.D., & Shannon, W. (1996). A new short form individual quality of life measure (SEIQoL-DW): Application in a cohort of individuals with HIV/AIDS. *British Medical Journal* **313**, 29-33.

Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S. (1995). Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. *Journal of Clinical Oncology* **13**, 1249-1254.

Holzner, B., Kemmler, G., Kopp, M., et al. (2001). Quality of life in breast cancer patients – Not enough attention for long-term survivors? *Psychosomatics* **42**, 117-123.

Hopwood, P., Howell, A., & Maguire, P. (1991). Screening for psychiatric morbidity in patients with advanced breast cancer: Validation of two self-report questionnaires. *British Journal of Cancer* **64**, 353 – 356.

Hopwood, P., & Stephens, R.J. (2000). Depression in patients with lung cancer: prevalence and risk factors derived from quality-of-life data. *Journal of Clinical Oncology* **18**, 893-903.

Howell, D.C. (2002). *Statistical Methods for Psychology*. Belmont, CA: Duxbury Press.

Ibbotson, T., Maguire, P., Selby, P., Priestman, T., & Wallace, L. (1994). Screening for anxiety and depression in cancer patients: The effects of disease and treatment. *European Journal of Cancer* **30**, 37-40.

Jaeschke, R., Singer, J., & Guyatt, G.H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials* **10**, 407-415.

Jobe, J.B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research* **12**, 219-227.

Joyce, C.R., Hickey, A., McGee, H.M., O'Boyle, C.A. (2003). A theory-based method for the evaluation of individual quality of life: The SEIQoL. *Quality of Life Research* **12**, 275-280.

Juniper, E., Guyatt, G.H., Willan, A. et al. (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *Journal of Clinical Epidemiology* **47**, 81-87.

Kaasa, S., Bjordal, K., Aaronson, N., Moum, T., Wist, E., Hagen, S., & Kvikstad A. (1995). The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *European Journal of Cancer* **31A**, 2260-2263.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement* **1**, 152-176.

Kemmler, G., Holzner, B., Kopp, M., et al. (1999). Comparison of two quality-of-life instruments for cancer patients: The Functional Assessment of Cancer Therapy-General and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-c30. *Journal of Clinical Oncology* **17**, 2932-2940.

Kemmler, G., Holzner, B., Kopp, M., Dunser, M., Greil, R., Hahn, E., & Sperner-Unterweger, B. (2002). Multidimensional scaling as a tool for analysing quality of life data. *Quality of Life Research* **11**, 223-233.

King, M.T. (1996). The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Quality of Life Research* **5**, 555-567.

Kleinman L, Leidy NK, Crawley J, Bonomi A, Schoenfeld P. (2001). A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire. *Medical Care* **39**, 181-189.

Kline, P. (1992). *The handbook of psychological testing*. Routledge: London.

Kline, P. (1997). *An easy guide to factor analysis*. Routledge: London.

Kopp, M., Schweigkofler, H., Holzner, B., et al. (1998). Time after bone marrow transplantation as an important variable for quality of life: Results of a cross-sectional investigation using two different instruments for quality-of-life assessment. *Annals of Hematology* **77**, 27-32, 1998.

Kopp, M., Schweigkofler, H., Holzner, B., et al. (2000). EORTC QLQ-c30 and FACT-BMT for the measurement of quality of life in bone marrow transplant recipients: A comparison. *European Journal of Haematology* **65**, 97 –103.

Kosinski, M., Bayliss, M.S., Bjorner, J.B., et al. (2003). A six-item short-form survey for measuring headache impact: The HIT-6. *Quality of Life Research* **12**, 963-974.

Lai, J-S., Cella, D., Chang, C-H., Bode, R.K., & Heinemann, A.W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research* **12**, 485-501.

Leplege, A., & Hunt, S. The problem of quality of life in medicine. *Journal of the American Medical Association* **278**, 47-50.

Lewis, G. (1991). Observer bias in the assessment of anxiety and depression. *Social Psychiatry & Psychiatric Epidemiology* **26**, 265-272.

Lewis, G., & Wessely, S. (1990). Comparison of the General Health Questionnaire and the Hospital Anxiety and Depression Scale. *British Journal of Psychiatry* **157**, 860-864.

Linacre, J.M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions* **7**, 328.

Linacre, J.M. (1995). Prioritizing misfit indicators. *Rasch Measurement Transactions* **9**, 422-423.

Linacre, J.M. (1997) KR-20 or Rasch reliability: Which tells the "Truth"? *Rasch Measurement Transactions* **11**, 580-581.

Linacre, J.M. (1999a). Category disordering vs. step disordering. *Rasch Measurement Transactions* **13**, 675.

Linacre, J.M. (1999b). Understanding (or misunderstanding?) the Rasch model. *Rasch Measurement Transactions* **13**, 706.

Linacre, J.M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions* **16**, 878.

Linacre, J.M., & Wright, B.D. (2001). *User's guide to Winsteps*. Mesa Press: Chicago.

Lisspers, J., Nygren, A., & Söderman, E. (1997). Hospital Anxiety and Depression Scale (HAD): Some psychometric data for a Swedish sample. *Acta Psychiatrica Scandinavica* **96**, 281 – 286.

Loehlin, J.C. (1998).*Latent variable models* (3rd Ed.). Lawrence Erlbaum: Mahwah, NJ.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McColl, E., Meadows, K., & Barofsky, I. (2003). Cognitive aspects of survey methodology and quality of life measurement. *Quality of Life Research* **12**. 217-218.

Macduff, C. & Russell, E. (1998). The problem of measuring change in individual health-related quality of life by postal questionnaire: Use of the patient-generated index in a disabled population. *Quality of Life Research* **7**, 761-769.

McGee, H.M., O'Boyle, C.A., Hickey, A., O'Malley, K., & Joyce, C.R.B. (1991). Assessing the quality of life of the individual: The SEIQoL with a healthy and a gastroenterology unit population. *Psychological Medicine* **21**, 749-759.

McHorney, C.A. (1997). Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine* **127**, 743-750.

McHorney, C.A., Ware, J.E., Jr., & Raczek, A.E. (1993). The Mos 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care* **31**, 247-263.

McHorney, C., & Tarlov, A. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research* **4**, 293-307.

McLachlan, S.A., Allenby, A., Matthews, J., et al. (2001). Randomized trial of coordinated psychosocial interventions based on patient self-assessments versus standard care to improve the psychosocial functioning of patients with cancer. *Journal of Clinical Oncology* **19**, 4117-4125.

McLachlan, S.A., Devins, G.M., & Goodwin, P.J. (1999). Factor analysis of the psychosocial items of the EORTC QLQ-C30 in metastatic breast cancer patients participating in a psychosocial intervention study. *Quality of Life Research* **8**, 311-317.

McNamara, T. (1996). *Measuring second language performance*. Addison Wesley Longman Ltd.: New York.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Lawrence Erlbaum Associates. Hillsdale: New Jersey.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge University Press: Cambridge.

Miller, B.E., Pittman, B., & Strong, C. (2003). Gynecologic cancer patients' psychosocial needs and their views on the physician's role in meeting those needs. *International Journal of Gynecological Cancer* **13**. 111-119.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**,81-97.

Mokken, R.J. & Lewis, C. (1982) A non-parametric approach to the analysis of dichotomous responses. *Applied Psychological Measurement* 6, 417-430.

Moorey, S., Greer, S., Watson, M., Gorman, C., Rowden, L., Tunmore, R., Robertson, B., & Bliss, J. (1991). The factor structure and factor stability of the Hospital Anxiety and Depression Scale in patients with cancer. *British Journal of Psychiatry* **158**, 255 – 259.

Newell, S., Girgis, A., Sanson-Fisher, R.W. et al. (1997). Are touchscreen computer surveys acceptable to medical oncology patients? *Journal of Psychosocial Oncology* **15**, 37-46.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory*. McGraw-Hill: New York.

O'Boyle, C.A., McGee, H., Hickey, A., O'Malley, K., & Joyce, C.R.B. (1992). Individual quality of life in patients undergoing hip replacement. *Lancet* **339**, 1088-1091.

Osoba, D. (1999). What has been learned from measuring health-related quality of life in clinical oncology. *European Journal of Cancer* **35**, 1565-1570.

Osoba, D., Aaronson, N.K., Muller, M., et al. (1996). The development and psychometric validation of a brain cancer quality of life questionnaire for use in combination with general cancer-specific questionnaires. *Quality of Life Research* **5**, 139-150.

Osoba, D., Aaronson, N., Zee, B., Sprangers, M., & te Velde, A. (1997). Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. The Study Group on Quality of Life of the EORTC and the Symptom Control and Quality of Life Committees of the NCI of Canada Clinical Trials Group. *Quality of Life Research* **6**, 103-108.

Osoba, D., Rodrigues, G., Myles, J. et al. (1998). Interpreting the significance of changes in health-related quality-of-life scores. *Journal of Clinical Oncology* **16**, 139-144.

Petersen, M. A., Groenvold, M., & Bjorner, J. (2000). Use of item response theory (IRT) to develop a shortened version of the EORTC QLQ-C30 for palliative care patients. *Internal report to the EORTC Quality of Life Group*.

Petersen, M.A., Groenvold, M., Bjorner, J.B., Aaronson, N., Conroy, T., Cull, A., Fayers, P., Hjermstad, M., Sprangers, M., & Sullivan, M. The European Organisation for Research and Treatment of Cancer Quality of Life Group. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research* **12**, 373-385.

Pouwer, F., Snoek, F.J., van der Ploeg, H.M., Heine, R.J., & Brand, A.N. (1998). A comparison of the standard and the computerized versions of the Well-being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research* **7**, 33-38.

Priestman, T.J., & Baum, M. (1976). Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet* **24**, 899–900.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes* **1**.

Raczek, A.E, Ware, J.E., Bjorner, J.B., et al. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in

seven countries: results from the IQOLA Project. *Journal of Clinical Epidemiology* **51**, 1203-1214.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* London: University of Chicago Press.

Razavi, D., Delvaux, N., Farvacques, C., & Robaye, E. (1990). Screening for adjustment disorders and major depressive disorders in cancer in-patients. *British Journal of Psychiatry* **156**, 79 - 83.

Reese, T.W. (1943). The application of the theory of physical measurement to the measurement of psychological magnitudes with three experimental examples. *Psychological Monographs* **55**, 1-89.

Revicki, D.A., & Cella, D.F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research* **6**, 595-600.

Ringdal, G.I., & Ringdal, K. (1993). Testing the EORTC Quality of Life Questionnaire on cancer patients with heterogeneous diagnoses. *Quality of Life Research* **2**, 129-140.

Ringdal K, Ringdal GI, Kaasa S, Bjordal K, Wisloff F, Sundstrom S, Hjermstad MJ. (1999). Assessing the consistency of psychometric properties of the HRQoL scales within the EORTC QLQ-C30 across populations by means of the Mokken Scaling Model. *Quality of Life Research* **8**, 25-43.

Ryan, J.M., Corry, J.R., Attewell, R., & Smithson, M.J. (2002). A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Quality of Life Research* **11**, 19-26.

Ryser, L., Wright, B.D., Aeschlimann, A., Mariacher-Gehler, S., Stucki, G. (1999). A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. *Arthritis Care And Research* **12**, 331-335.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* **17**.

Schag, C.A., Ganz, P.A., & Heinrich, R.L. (1991). CAncer Rehabilitation Evaluation System - Short Form (CARES - SF): A cancer specific rehabilitation and quality of life instrument. *Cancer* **66**, 1406-1413.

Schag, C.A., & Heinrich, R.L. (1990). Development of a comprehensive quality of life measurement tool: CARES. *Oncology* **4**, 135-138.

Schipper, H., Clinch, J., McMurray, A., & Levitt, M. (1984). Measuring the quality of life of cancer patients: The Functional Living Index-Cancer: Development and validation. *Journal of Clinical Oncology* **2**, 472-483.

Schmitz, N., Hartkamp, N., Brinschwitz, C., Michalek, S., & Tress W. (2000). Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. *Acta Psychiatrica Scandinavica* **102**, 147-152.

Sprangers, M.A.G., & Schwartz, C.E. (2000). Integrating response shift into health-related quality-of-life research: A theoretical model. In (Eds. Schwartz, C.E. & Sprangers, M.A.G.) *Adaptation to changing health*. American Psychological Association: Washington, DC.

Selby, P.J., Chapman, J.A., Etazadi-Amoli, J., Dalley, D., & Boyd, N.F. (1984). The development of a method for assessing the quality of life of cancer patients. *British Journal of Cancer* **50**, 13-22.

Sellick, S.M., & Crooks, D.L. (1999). Depression and cancer: an appraisal of the literature for prevalence, detection, and practice guideline development for psychological interventions. *Psycho-Oncology* **8**, 315-333.

Sharp, L.K., Knight, S.J. Nadler, R., et al. (1999). Quality of life in low-income patients with metastatic prostate cancer: Divergent and convergent validity of three instruments. *Quality of Life Research* **8**, 461-470.

Sigle, J. & Porzsolt, F. (1996). Practical aspects of quality-of-life measurement: Design and feasibility study of the quality-of-life recorder and the standardized

measurement of quality of life in an outpatient clinic. *Cancer Treatment Reviews* **22** (Supplement A), 75-89.

Sigurdardottir, V., Bolund, C., Brandberg, Y., Sullivan, M. (1993). The impact of generalized malignant melanoma on quality of life evaluated by the EORTC questionnaire technique. *Quality of Life Research* **2**, 193-203.

Silverstone, P.H. (1994). Poor efficacy of the Hospital Anxiety and Depression Scale in the diagnosis of major depressive disorder in both medical and psychiatric patients. *Journal of Psychosomatic Research* **38**, 441 – 450.

Smith, A.B., Selby, P.J., Velikova, G., et al. (2002). The Factor structure of the HADS Questionnaire from a large cancer population. *Psychology and Psychotherapy* **75**, 165 – 176.

Smith, R.M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement* **48**, 657-667.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement* **51**, 541-565.

Smith, R.M. (1992). *Applications of Rasch measurement.* JAM Press: Maple Grove, MN.

Smith, R.M., & Miao, C.Y. (1994). Assessing unidimensionality for Rasch measurement. In Wilson, M. *Objective measurement* (vol 2). Ablex Publishing Corporation: Norwood, New Jersey, pp. 316-327.

Smith, R.M., Schumacker, R.E., & Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement* **2**, 66-78.

Smith, R.M., & Suh, K.K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement* **4**, 153-163.

Spielberger, C.D., Gorsuch, R.L., & Luchene, R.E. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto: Consulting Psychologists Press.

Spinhoven, Ph., Ormel, J., Sloekers, P.P.A., Kempen, G.I.J.M., Speckens, A.E.M., & van Hemert, A.M. (1997). A validation study of the Hospital Anxiety and

Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine* **27**, 363 – 370.

Spitzer, W.O., Dobson, A.J., Hall, J., et al. (1981). Measuring the quality of life of cancer patients: A concise QL-index for use by physicians. *Journal of Chronic Disease* **34**, 585-597.

Sprangers, M.A., Cull, A., Bjordal, K., Groenvold, M, & Aaronson, N.K. (1993). The European Organization for Research and Treatment of Cancer approach to quality of life assessment: Guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. *Quality of Life Research* **2**, 287-295.

Sprangers, M.A., Cull, A., Groenvold, M, Bjordal, K., Blazeby, J., & Aaronson, N.K. (1998). The European Organization for Research and Treatment of Cancer approach to quality of life assessment: An update and overview. EORTC Study Group on Quality of Life. *Quality of Life Research* **7**, 291-300.

Staquet, M., Berzon, R., Osoba, D., & Machin D. (1996). Guidelines for reporting results of quality of life assessments in clinical trials. *Quality of Life Research* **5**, 496-502.

Stark, D.P., & House, A. (2000). Anxiety in cancer patients. *British Journal of Cancer* **83**, 1261-1267.

Stark, D., Kiely, M., Smith, A., Velikova, G., House, A., & Selby, P. (2002). Anxiety in cancer patients: Their nature, associations and relation to quality of life. *Journal of Clinical Oncology* **20**, 3137-3148.

Stead, M.L., Brown, J.M., Velikova, G., et al. (1999). Development of an EORTC questionnaire moduel to be used in health-related quality-of-life assessment for patients with multiple myeloma. *British Journal of Haematology* **104**, 605-611.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science* **103**, 667-680.

Stucki, G., Daltroy, L., Katz, J.N. et al. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology* **49**, 711-717.

Suen, H.K. (1990). *Principles of Test Theories*, Lawrence Erlbaum Associates, Hillsdale, N.J.

Taenzer, P., Bultz, B.D., Carlson, L.E. et al. (2000). Impact of computerized quality of life screening on physician behaviour and patient satisfaction in lung cancer outpatients. *Psycho-oncology* **9**: 203-213.

Taenzer, P.A., Speca, M., Atkinson, M.J., Bultz, B.D., Page, S., Harasym, P., Davis, J.L, (1997). Computerized quality-of-life screening in an oncology clinic. *Cancer Practice* **5**, 168-175.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* **16**, 433-451.

Thurstone, L.L. (1926). The scoring of individual performance. *Journal of Educational Psychology* **17**, 446-457.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review* **34**, 278-286.

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology* **23**, 529-554.

Thurstone, L.L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology* **26**, 249-269.

Thurstone, L.L., & Chave, E.J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Tourangeau, R., Rips., L.J., & Rasinski, K. (2000). *The psychology of survey responses.* New York: Cambridge University Press.

Velikova, G., Booth, L., Smith, A.B., Brown, P.M., Lynch, P., Brown, J.M., & Selby, P.J. (2004). Measuring quality of life in routine oncology practice improves

communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology* **22**, 714-724.

Velikova, G., Brown, J.M., Smith, A.B., Selby, P.J. (2002). Computer-based quality of life questionnaires may contribute to doctor-patient interactions in oncology. *British Journal of Cancer* **86**, 51-59.

Velikova, G., Stark, D., & Selby, P.J. (1999). Quality of life instruments in oncology. *European Journal of Cancer* **35**, 1571-1580.

Velikova, G., Wright, E.P., Smith, A.B., Cull, A., Gould, A., Forman, D., Perren, T., Stead, M., Brown, J., & Selby, P.J. (1999). Automated collection and recording of quality of life data: A comparison of paper and computer touchscreen questionnaires. *Journal of Clinical Oncology* **17**, 998-1007.

Velikova, G., Wright, E.P., Smith, A.B et al. (2001). Self-reported quality of life of individual cancer patients: Concordance of results with clinical course and medical records. *Journal of Clinical Oncology* **19**, 2064 - 2073.

Velozo, C.A., Lai, J.S., Mallinson, T., & Hauselman, E. (2001). Maintaining instrument quality while reducing items: application of Rasch analysis to a self-report of visual function. *Journal of Outcome Measurement* **4**, 667-680.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale, N.J: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. (1987) Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational Measurement* **24**, 185-201.

Waldron, D., O'Boyle, C.A., Kearney, M., Moriarty, M., & Carney, D. (1999). Quality-of-life measurement in advanced cancer: Assessing the individual. *Journal of Clinical Oncology* **17**, 3603-3611.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology* **57**, 1051-1058.

Ward, W.L., Hahn, E.A., Mo, F., Hernandez, L., Tulsky, D.S., Cella, D. (1999). Reliability and validity of the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) quality of life instrument. *Quality of Life Research* **8**, 181-195.

Ware, J.E. Jr., Bjorner, J. & Kosinski, M. (1999). Dynamic health assessments: The search for more practical and more precise outcomes measures. *Quality of Life Newsletter* **21**, 11-13.

Ware, J.E, Bjorner, J.B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care* **38**, 43-59.

Ware, J.E, Kosinski, M. , Bjorner, J.B., Bayliss, M.S., Batenhorst, A., Dahloef, C.G.H., Tepper, S., & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research* **12**, 935-952.

Watson, C.G., Thomas, D., & Anderson, P.E. (1992). Do computer-administered Minnesota Multiphasic Personality Inventories underestimate booklet-based scores? *Journal of Clinical Psychology* **48**, 744-748.

White, D., Leach, C., Sims, R., Atkinson, M., & Cottrell, D. (1999). Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *British Journal of Psychiatry* **175**, 452 – 454.

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America* **17**, 684-688.

Wing, J., Cooper, J., & Sartorius, N. (1974). *Measurement and classification of psychiatric symptoms*. Cambridge: Cambridge University Press.

Winstead-Fry, P., & Schultz, A. (1997). Psychometric analysis of the Functional Assessment of Cancer Therapy-General (FACT-G) scale in a rural sample. *Cancer* **79**, 2446-2452.

World Health Organisation (1946). *Constitution of the World Health Organisation.* WHO, New York.

Wright, B.D. (1967). Sample-free test calibration and person measurement. *MESA Research Memorandum* **1**.

Wright, B.D. (1992). IRT in the 1990's: Which models work best? *Rasch Measurement Transactions* **6**, 196-200.

Wright, B.D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modelling* **3**, 3-24.

Wright, B.D. (1997). History of Social Science Measurement. *MESA Research Memorandum* **62**.

Wright, B.D., Linacre, J.M., Gustafson, J-E., & Martin-Loef, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions* **8**, 370.

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. MESA Press, Chicago.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement* **29**, 23-48.

Wright, E.P., Selby, P., Gillibrand, A., Smith, A., Velikova, G., Crawford, M., et al. (2003). Feasibility and compliance of automated measurement of quality of life in oncology practice. *Journal of Clinical Oncology* **21**, 374 – 382.

Wyrwich, K.W., Nienaber, N.A., Tierney, W.M. et al. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes for health-related quality of life measures. *Medical Care* **37**, 469-478.

Wyrwich, K.W., Tierney, W.M., & Wolinsky, F.D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality-of-life. *Journal of Clinical Epidemiology* **52**, 861-873.

Wyrwich, K.W., Tierney, W.M., & Wolinsky, F.D. (2002). Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. Quality of Life Research **11**, 1-7.

Wyrwich, K.W., & Wolinsky, F.D. (2000). Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice* **6**, 39-49.

Zabora, J., Brintzenhofeszoc, K., Curbow, B., Hooker, C., & Piantadosi, S. (2001). The prevalence of psychological distress by cancer site. *Psycho-oncology*, **10**, 19-28.

Zigmond, A.S., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica* **67**, 361-370.

Zung, W.W.K. (1965). A self-rating depression scale. *Archives of General Psychiatry* **12**, 63-70.

Psychology and Psychotherapy: Theory, Research and Practice (2002), 75, 165–176


www.bps.org.uk

# Factor analysis of the Hospital Anxiety and Depression Scale from a large cancer population

Adam B. Smith[1]*, Peter J. Selby[1], Galina Velikova[1], Dan Stark[1], E. Penny Wright[1], Ann Gould[2] and Ann Cull[3]

[1]ICRF Cancer Medicine Research Unit, St. James's University Hospital, UK
[2]Information & Statistics Division, NHS Scotland, UK
[3]ICRF Department of Clinical Psychology, Western General Hospital, UK

The Hospital Anxiety and Depression Scale (HADS) is widely used as a tool for assessing psychological distress in patients and non-clinical groups. Previous studies have demonstrated conflicting results regarding the factor structure of the questionnaire for different groups of patients, and the general population. This study investigated the factor structure of the HADS in a large heterogeneous cancer population of 1474 patients. It also sought to investigate emerging evidence that the HADS conforms to the tripartite model of anxiety and depression (Clark & Watson, 1993), and to test the proposal that detection rates for clinical cases of anxiety and depression could be enhanced by partialling out the effects of higher order factors from the HADS (Dunbar *et al.*, 2000). The results demonstrated a two-factor structure corresponding to the Anxiety and Depression subscales of the questionnaire. The factor structure remained stable for different subgroups of the sample, for males and females, as well as for different age groups, and a subgroup of metastatic cancer patients. The two factors were highly correlated ($r = .52$) and subsequent secondary factor analyses demonstrated a single higher order factor corresponding to psychological distress or negative affectivity. We concluded that the HADS comprises two factors corresponding to anhedonia and autonomic anxiety, which share a common variance with a primary factor namely psychological distress, and that the subscales of the HADS, rather than the residual scores (e.g. Dunbar *et al.*, 2000) were more effective at detecting clinical cases of anxiety and depression.

Originally designed to assess psychological distress of patients in medical and surgical settings, the Hospital Anxiety and Depression Scale (HADS, Zigmond & Snaith, 1983) has now been evaluated and validated for different medical and psychiatric patient

*Requests for reprints should be addressed to Adam Smith, ICRF Cancer Medicine Research Unit, St. James's University Hospital, Leeds LS9 7TF, UK (e-mail: medabs@cancermed.leeds.ac.uk)

populations (Spinhoven *et al.*, 1997; White, Leach, Sims, Atkinson, & Cottrell, 1999), and non-medical populations (Dagnan, Chadwick, & Trower, 2000; Lisspers, Nygren, & Söderman, 1997).

The HADS is a 14-item scale that requires respondents to endorse a verbal response which is scored as an index of the severity of anxiety or depression. The scores are then summed to produce two subscales corresponding to Anxiety (HADS-A), and Depression (HADS-D). As well the subscale totals, an overall total can be derived to indicate the level of psychological distress. Zigmond and Snaith (1983) advocated cut-offs between 8 and 10 for 'possible cases', and scores of 11 or more for 'definite cases'. The rates of prevalence reported using the HADS differ markedly. For instance, Hall, A'Hern, and Fallowfield (1999) report a rate of 13.5% for anxiety and 7.5% for depression in breast cancer patients using a cut-off of 11. These rates increased to 39.4 and 16.5% for anxiety and depression, respectively, using a threshold of 7. Hopwood, Howell, and Maguire (1991) reported that 27% of their sample of women with breast cancer had a probable case of affective disorder using a HADS threshold of 11. Similarly, Hopwood and Stephens (2000) reported levels of depression at 33% and anxiety at 34% in a sample of patients with lung cancer.

A number of studies have reported the efficacy of the HADS as a screening instrument for mental health problems (Abiodun, 1994; Hall *et al.*, 1999; Hopwood *et al.*, 1991; Ibbotson, Maguire, Selby, Priestman, & Wallace, 1994; Lewis & Wessely, 1990; Razavi, Delvaux, Farracques, & Robaye, 1990; Silverstone, 1994; Spinhoven *et al.*, 1997). These studies demonstrate that the HADS is a more consistent measure for detecting generalized anxiety disorders (sensitivity ranging from 59 to 93%, and specificity ranging from 73 to 90%), compared with depressive disorders (sensitivity ranging from 14 to 90%, and specificity from 73 to 100%). The combined HADS scores perform similarly in detecting either depressive or anxiety disorders with sensitivity ranging from 20 to 92%, and specificity from 74 to 95%

The original two-factor structure of the HADS, corresponding to the Anxiety and Depression subscales, has been confirmed by a number of subsequent studies (e.g. Dagnan *et al.*, 2000; Lisspers *et al.*, 1997; Moorey *et al.*, 1991; Spinhoven *et al.*, 1997; White *et al.*, 1999). Although two studies have demonstrated different factor structures (Andersson, 1993; Lewis, 1991), both studies involved small sample sizes which may have contributed to a distorted factor structure.

More recent interest in the HADS has centred on its relationship to the tripartite theory of anxiety and depression (Clark & Watson, 1991), and whether the factor structure corresponds to this model. In a recent study, Dunbar, Ford, Hunt, and Der (2000) proposed a three-level factor structure for the HADS corresponding to Clark and Watson's (1991) tripartite theory of anxiety and depression. Clark and Watson's model comprises three factors; psychological distress or negative affectivity (NA), autonomic anxiety, and depression as anhedonia. Dunbar *et al.* suggest that four items from the Anxiety subscale (items 1, 5, 7 and 11) represent negative affectivity, and the other three items correspond to autonomic anxiety, and consider the Depression subscale to correspond to anhedonia. The study collected HADS data from a large community sample. Although confirmatory factor analyses suggested that there was evidence supporting the two-factor structure, the tripartite model provided a better fit of their data. In fact, Dunbar *et al.*'s (2000) data suggested that the HADS conformed to a hierarchical model, as proposed by Clark, Watson, and Mineka (1994), in which the two secondary factors, anhedonia and autonomic anxiety, are subordinate to the higher factor, psychological distress or NA.

The model of the HADS proposed by Dunbar *et al.* (2000) also led to the prediction that the residual scores resulting from the regression of Autonomic Anxiety and Anhedonic Depression onto Negative Affectivity would be more effective at detecting clinical cases of anxiety and depression than the HADS subscales.

This study investigated the factor structure of the HADS in a large heterogeneous sample of 1474 cancer patients. Factor analyses were carried out to investigate the factor structure of the instrument across gender and different age groups, and with a subgroup of patients with metastatic cancer. A second-order factor analysis was also performed on the total dataset to explore whether HADS conforms in general to the tripartite model, and specifically to a hierarchical model. In addition, the predictions of improved screening efficacy for the residual scores of Autonomic Anxiety and Anhedonic Depression made by Dunbar *et al.* (2000) were also tested on a subset of patients who had also received a psychiatric interview.

## Method

Patient data were collated from a total of five studies which have been carried out by the ICRF Psychosocial Oncology Groups, at St. James's University Hospital, Leeds and Western General Hospital, Edinburgh over the past 4 years. The data from two studies have been previously reported (Cull *et al.*, 2001; Velikova *et al.*, 1999). All patients completed the HADS on a computer with a touchscreen monitor. A subset of 381 patients also received a psychiatric interview (either the Present State Examination or the Schedule for Clinical Assessment in Neuropsychiatry) by a trained interviewer (EPW and DS) at home within 2 weeks of completing the HADS. On the basis of the interview, patients were categorized into one of four categories: no psychiatric distress, anxiety disorder, depression, anxiety and depressive disorder.

Ethical approval for the studies had been given by the local hospital ethics committees in Leeds and Edinburgh.

### *Participants*

A total of 1474 patients participated in the studies. The majority of patients were recruited from outpatient oncology clinics (609 patients from St. James's Hospital, Leeds, and 785 patients from the Western General Hospital, Edinburgh). Eighty patients were recruited from a general oncology ward (Cookridge and St. James's Hospitals).

The average age of the sample was 55.9 years (range = 15.3–99.6). The total number of males taking part was 632, with an average age of 54.5 years (range = 15.3–91.7), and the number of females was 842, average age = 57.0 years (range = 17.8–99.6). Some data from five patients was missing, the remaining data from these patients has been included in the analyses. Table 1 gives the diagnoses.

## Results

### *HADS scores by age and gender*

The sample was split into three groups of approximately the same size based on age (group 1, <50 years, $N = 492$; group 2, ≥50 and <65, $N = 519$; and group 3, ≥65, $N = 454$) to allow comparisons of scores across different ages.

**Table 1.** Diagnoses of patients – number of patients (%)

| | |
|---|---|
| Breast | 379 (25.7) |
| Gastro-intestinal | 152 (10.3) |
| Lung | 46 (3.1) |
| Male genito-urinary | 262 (17.8) |
| Female genito-urinary | 207 (20.8) |
| Lymphoma | 45 (3.1) |
| Other | 283 (19.2) |

The mean score on the Anxiety subscale for all patients was 6.05 (SD 4.03) with a range between 0 and 21. The mean score for the Depression subscale was lower at 4.38 (SD 3.73) with a range between 0 and 20.

A breakdown of the scores by gender demonstrated higher scores for both Anxiety (mean 6.65, SD 3.99) and Depression (mean 4.61, SD 3.93) for women compared with men (mean 5.22, SD 3.73, and mean 4.06, SD 3.73, respectively). This pattern was also observed for each age group. Table 2 shows the mean HADS scores by age and gender.

**Table 2.** Mean HADS scores by age and gender (SD)

| Age (years) | <50 | ≥50 & <65 | ≥65 | Max. score | Total across age |
|---|---|---|---|---|---|
| HADS-A (Anxiety) | | | | | |
| Females | 6.74 (4.01) | 6.97 (4.01) | 6.07 (3.90) | 18 | 6.65 (3.99) |
| Males | 5.56 (4.14) | 5.31 (3.79) | 4.75 (3.77) | 21 | 5.22 (3.94) |
| HADS-D (Depression) | | | | | |
| Females | 4.08 (3.75) | 4.69 (3.81) | 4.99 (3.52) | 20 | 4.61 (3.73) |
| Males | 3.48 (3.75) | 4.33 (3.66) | 4.54 (3.68) | 20 | 4.06 (3.73) |

Although the summated anxiety and depression scales were not normally distributed, and were also highly skewed (skewness .63 (SE 0.064) for Anxiety, and .62 (SE 0.064) for Depression), Levene's test of equality of error variances revealed no significant differences for either subscale ($F < 1$). Therefore, a multivariate analysis of variance was performed on the subscale scores with two 'between group' factors (gender with two levels, and age with the three levels), which showed significant effects for Anxiety and Depression by gender ($F(1,1464) = 42.66$, $p < .0001$) and ($F(1,1464) = 5.54$, $p < .05$) respectively, and by age group ($F(2,1459) = 5.19$, $p < .05$) and ($F(2,1459) = 9.05$, $p < .001$), respectively. The interaction between gender and age group for the two subscales was not statistically significant ($F < 1$).

*Post hoc* Bonferroni contrasts demonstrated significant differences ($p < .05$) between the Anxiety subscales for groups 1 and 3, and groups 2 and 3, indicating that the oldest patients were experiencing significantly lower levels of anxiety than the youngest group of patients. The contrasts for the Depression subscale indicated significant differences between groups 1 and 2, and groups 1 and 3. The contrast between group 2 and 3 was not statistically different. These results demonstrated that levels of depression increased with age.

### Depression and anxiety

Employing a threshold score of 8, as recommended by Zigmond and Snaith (1983) resulted in 33.3%(491/1469) of patients being identified as at least 'possible cases' of anxiety, and 19.8%(291/1469) as 'possible cases' for depression. For each subscale the proportion of patients scoring more than 8 was higher among women than men (women: anxiety 334/842; depression 80/842, vs. men: anxiety 156/627, depression 110/627). These gender differences were statistically significant (anxiety: $\chi^2 = 64.66$, d.f. = 1, $p < .005$, and depression, $\chi^2 = 16.9$, d.f. = 1, $p < .005$).

### Factor analysis

The correlation matrix for the HADS items is shown in Table 3. Table 4 shows the inter-item reliability for the HADS. The various subgroups show the same Cronbach's alpha, which is around .83 for the Anxiety subscale and .79 for Depression. Both these are within the acceptable limits. Table 5 demonstrates the item-total correlations and the changes to the reliability coefficient when items are removed from the subscales.

The literature suggests a correlation between the anxiety and depression factors (Clark & Watson, 1991), therefore a principal components analysis with an oblique rotation was used for the factor analysis. The factor analysis of the entire dataset revealed a two-factor structure (Table 6). The first factor explained 37.69% of the variance (eigenvalue of 5.27) and the second factor accounted for 11.49%of the variance (eigenvalue of 1.61). The remaining factors had eigenvalues <1, and were therefore not selected for subsequent analysis (Figure 1). The rotated factor structure revealed two factor structures corresponding to the Anxiety and Depression subscale. Only one item, (item 7, "I can sit at ease and feel relaxed") from the Anxiety subscale loaded more strongly on the Depression subscale. The two factors were significantly correlated ($r = .52$).

Factor analyses of the data by gender, age-group and a split-half reliability sample, in which the dataset was divided into two subsets and the factor structure analysed for both sets, revealed similar factor structures. The factor loadings for the first two age groups and females are reversed compared with the entire dataset and the other factor analyses. However, the factor structures remain the same. In addition, extent of disease had also been recorded for a subset of the patients. The data from 197 patients with metastatic disease were also analysed and the same factor structure was again demonstrated.

Because the factor analyses revealed that the two subscales were strongly correlated, a second-order factor analysis and a Schmid-Leiman transformation were carried out on the data. The first-order factors had factor loadings of 0.59 on a second-order factor, and around 70% of the variance of these factors was explained by the second-order factor. As can be seen from Table 7, psychological distress accounts for about a third of the common variance.

### Screening for caseness of clinical anxiety and depression

The residuals for the Autonomic Anxiety and Anhedonic Depression scores were calculated by regressing scores from these scales onto NA. These scores were then used to measure screening efficacy by plotting the area-under-the curve (AUC) from ROC curves, and recording sensitivity (proportion of true positives) and specificity (proportion of true negatives).

**Table 3.** Interitem correlation matrix for HADS

| | | HADS-A | | | | | | | HADS-D | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Item 1 | Item 3 | Item 5 | Item 7 | Item 9 | Item 11 | Item 13 | Item 2 | Item 4 | Item 6 | Item 8 | Item 10 | Item 12 | Item 14 |
| HADS-A | Item 1 | – | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Item 3 | .50 | – | – | – | – | – | – | – | – | – | – | – | – | – |
| | Item 5 | .55 | .61 | – | – | – | – | – | – | – | – | – | – | – | – |
| | Item 7 | .36 | .32 | .36 | – | – | – | – | – | – | – | – | – | – | – |
| | Item 9 | .41 | .49 | .44 | .33 | – | – | – | – | – | – | – | – | – | – |
| | Item 11 | .33 | .32 | .36 | .29 | .25 | – | – | – | – | – | – | – | – | – |
| | Item 13 | .52 | .58 | .57 | .30 | .50 | .37 | – | – | – | – | – | – | – | – |
| HADS-D | Item 2 | .20 | .22 | .27 | .38 | .21 | .14 | .20 | – | – | – | – | – | – | – |
| | Item 4 | .23 | .21 | .24 | .36 | .23 | .09 | .20 | .36 | – | – | – | – | – | – |
| | Item 6 | .43 | .39 | .47 | .39 | .27 | .26 | .37 | .38 | .39 | – | – | – | – | – |
| | Item 8 | .30 | .26 | .35 | .29 | .16 | .24 | .26 | .47 | .22 | .39 | – | – | – | – |
| | Item 10 | .26 | .21 | .25 | .23 | .11 | .21 | .25 | .26 | .21 | .35 | .29 | – | – | – |
| | Item 12 | .37 | .32 | .37 | .42 | .27 | .18 | .28 | .51 | .46 | .49 | .41 | .40 | – | – |
| | Item 14 | .27 | .25 | .27 | .35 | .24 | .21 | .24 | .26 | .31 | .35 | .25 | .29 | .40 | – |

*All correlations are significant at $p < .001$.

**Table 4.** Inter-item reliability coefficients (Cronbach's alpha) of the HADS

| | HADS-A | HADS-D |
|---|---|---|
| Total | .83 | .79 |
| Age group 1 | .85 | .82 |
| Age group 2 | .83 | .79 |
| Age group 3 | .81 | .74 |
| Females | .82 | .78 |
| Males | .83 | .79 |
| Split-half 1 | .84 | .80 |
| Split-half 2 | .82 | .78 |

**Table 5.** Item-total correlations and the revised Cronbach's alpha for HADS-A and HADS-D when items are removed from subscales

| Items from HADS-A subscale | | | Items from HADS-D subscale | | |
|---|---|---|---|---|---|
| | Item-total correlation | Revised Cronbach's $\alpha$ | | Item-total correlation | Revised Cronbach's $\alpha$ |
| Item 1 | .63 | .80 | Item 2 | .56 | .75 |
| Item 3 | .67 | .79 | Item 4 | .46 | .77 |
| Item 5 | .68 | .79 | Item 6 | .58 | .75 |
| Item 7 | .44 | .83 | Item 8 | .50 | .77 |
| Item 9 | .56 | .81 | Item 10 | .43 | .78 |
| Item 11 | .43 | .83 | Item 12 | .67 | .73 |
| Item 13 | .67 | .79 | Item 14 | .44 | .77 |

The AUC for Autonomic Anxiety was low at .62 (confidence interval, CI, .53–.70). At sensitivity levels of .70, the specificity for AA was .41, and at sensitivity levels of .80, the specificity for AA decreased to .28. Similarly, the AUC for Anhedonic Depression was also low at .64 (CI .57–.72). At levels of .70 sensitivity, specificity was .48 for Anhedonic Depression, and at sensitivity levels of .85, specificity for AD was only .24.

## Discussion

Previous studies have reported a single factor, as well as two-, three- and four-factor structures for the HADS. The results from this study demonstrated a two-factor structure for HADS in a very large heterogeneous sample of cancer patients. The factor structure remained when the sample was divided into three age groups, and comparisons between males and females, as well as patients with metastatic cancer, revealed the same two factors approximately corresponding to anxiety and depression.

Although the factor structure remained constant across age groups and between males and females, age effects and gender differences were observed for the subscales

**Table 6.** Rotated factor structure for the entire dataset ($N = 1474$)

| | Factor 1 | Factor 2 |
|---|---|---|
| Items from HADS-A subscale | | |
| Item 1–Tense | .70 | .09 |
| Item 3–Frightened | .82 | – .04 |
| Item 5–Worrying | .76 | .09 |
| Item 7–Relaxed | .20 | .52 |
| Item 9–Butterflies | .72 | – .06 |
| Item 11–Restless | .57 | – .02 |
| Item 13–Panic | .85 | – .08 |
| Items from HADS-D subscale | | |
| Item 2–Enjoy things | – .17 | .81 |
| Item 4–Laugh | – .12 | .70 |
| Item 6–Cheerful | .23 | .57 |
| Item 8–Slowed down | .05 | .60 |
| Item 10–Appearance | .02 | .54 |
| Item 12–Enjoyment | – .02 | .81 |
| Item 14–Enjoy book | .06 | .55 |

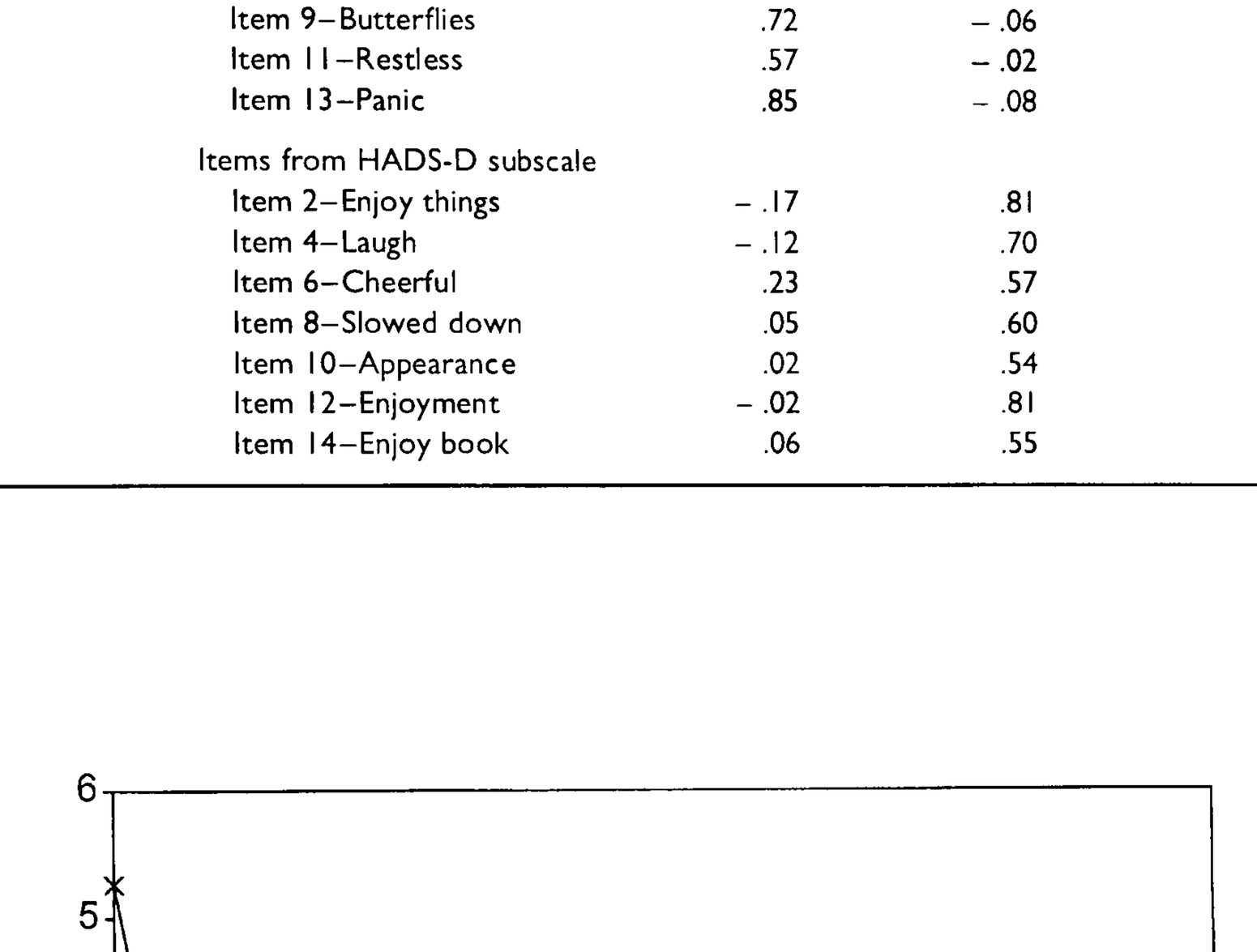



**Figure 1.** The scree-plot from the unrotated factor analysis of the entire dataset ($N = 1474$).

**Table 7.** Factor structure following second-order factor analysis and transformation

| Items | Psychological distress | Anxiety | Depression |
|---|---|---|---|
| Anxiety | | | |
| Item 1 | .46 | .57 | .06 |
| Item 3 | .47 | .66 | −.02 |
| Item 5 | .50 | .62 | .06 |
| Item 7 | .42 | .16 | .42 |
| Item 9 | .39 | .58 | −.05 |
| Item 11 | .33 | .46 | −.01 |
| Item 13 | .45 | .69 | −.06 |
| Depression | | | |
| Item 2 | .38 | −.14 | .66 |
| Item 4 | .31 | −.14 | .57 |
| Item 6 | .47 | .19 | .46 |
| Item 8 | .38 | .03 | .49 |
| Item 10 | .33 | .02 | .44 |
| Item 12 | .47 | −.02 | .66 |
| Item 14 | .36 | .05 | .45 |
| Eigenvalue | 2.38 | 2.27 | 2.21 |
| % common variance | 34.70 | 33.07 | 32.23 |

of HADS. Older patients generally reported higher levels of depression, but lower levels of anxiety than younger patients, and females in the second age group (between 50 and 65 years of age) reported the highest levels of anxiety.

There may be a number of explanations for the observed differences in factor structures found in other studies (e.g. Andersson, 1993; Lewis, 1991), and those reported here. For instance, the sample size used by Andersson (1993) was relatively small for factor analysis ($N = 163$). Also differences in reported factor structures have been found with groups from non-medical community samples. The HADS was designed originally for use in hospital clinics and wards, and the results from studies using patient groups find similar factor structures (Spinhoven *et al.*, 1997; White *et al.*, 1999). It may, therefore, be that non-patient groups respond differently to the HADS than patient groups, and that different measures may be needed to identify non-hospital-based cases of psychological distress (Groenvold *et al.*, 1999).

The two factors were strongly correlated (.52), which taken together with a second-order factor analysis confirmed a single higher-order factor, corresponding to psychological distress or NA in the tripartite model, and two subordinate factors, Anhedonia and Autonomic Anxiety. Although Clark and Watson's (1991) original model placed the three factors on equal footing, more recent formulations of the model (e.g., Clark *et al.*, 1994) have suggested a hierarchical structure. Therefore, these results provide evidence that the HADS corresponds to the tripartite structure, but that this structure differs from the model demonstrated by Dunbar *et al.* (2000).

The difference between this study and that by Dunbar *et al.* (2000) lies in the statistical models used. Dunbar *et al.* (2000) employed confirmatory factor analysis, whereas the method chosen for this study was the same as the method used by Clark *et al.* (1994), i.e. hierarchical or second-order factor analysis.

Both confirmatory factor analysis (CFA) and hierarchical factor analysis (HFA) are special forms of structural equation modelling. These models depend on correlations between variables to define or surmise relationships between latent (i.e. unobserved) and observed variables. In many ways the two methods are similar, the main difference between the models being that in CFA the latent model is fitted to a correlation matrix, and the 'goodness of fit' of the model is then evaluated, whereas HFA relies on the rotation of the correlation matrix to produce first and subsequent order factors (Loehlin, 1998). Because both statistical techniques depend on correlation between the observed and latent variables, neither model is truly causal: both models may posit causality, but it remains for subsequent empirical work to either confirm or dispute this.

The difference between Dunbar *et al.*'s (2000) study and this one in terms of the results, is subtle. The results from both studies demonstrate that a hierarchical tripartite structure probably underlies the HADS. However, whereas Dunbar *et al.* (2000) suggest that the HADS-A scale is split into NA and AA items, the results from this study suggest that the NA is a common factor whose variance is shared by both AA and Anhedonic Depression, and which is not specific to either subscale.

The results of these two studies have different implications in terms of clinical utility. Dunbar *et al.* (2000) have suggested that Autonomic Anxiety and Anhedonic Depression scores would have improved detection rates for clinical cases of anxiety and depression with NA effectively partialled out. However, the results from this study suggest that individual subscales would be better at detecting clinical cases of anxiety and depression.

We have carried out an analysis testing the efficacy of the two HADS subscales at detecting caseness of clinical anxiety and depression (Smith *et al.*, 2002). The AUC values for both HADS-A and HADS-D were higher than those derived from the Dunbar *et al.* model (.78 compared with .62 for anxiety, and .77 compared with .64 for depression, respectively). Similarly, both subscales maintained higher levels of specificity for differing levels of sensitivity.

These results suggest that the HADS subscales, rather than the residual scores proposed by Dunbar *et al.* (2000), have a greater screening efficacy for clinical cases of anxiety and depression.

In summary, our results confirm a two-factor structure of the HADS and an additional common factor, psychological distress, and extend this model to a large cancer patient population. In addition, the results provide further evidence that the model underlying the HADS corresponds to a tripartite structure.

The hierarchical model outlined here provided a different and more efficacious method for detecting clinical cases of anxiety and depression than the structure proposed by Dunbar *et al.* (2000). Therefore, these findings support a clinical strategy for screening, in which both the total scores (e.g. Razavi *et al.*, 1990) and the scores on the anxiety and depression scales may be used to identify potential cases. This information may be helpful to clinicians to identify patients requiring further clinical assessment and to focus their enquiry.

The factor structures proved to be robust, inter-item correlations were high, and the results demonstrated a reasonable fit of all the items with the exception of one item from the Anxiety subscale, item 7, "I can sit at ease and feel relaxed", which loaded heavily onto the Depression subscale. This result has been reported in a number of other studies (e.g. Moorey *et al.*, 1991) and taking that together with the generally mixed results at screening efficacy suggests that the HADS could benefit from additional work to investigate item fit. Future work could explore using other psychometric

techniques, such as item-response theory, to analyse item fit and remove items from HADS in order to improve its ability to detect cases of psychological distress.

## Acknowledgements

We are very grateful to Dr Phillip Snaith for the very helpful discussions at the beginning of this study. The work was supported by a grant from the National Health Service Research and Development Programme (ABS, EPW, AG, AC, PJS), the Imperial Cancer Research Fund (ABS, EPW, AC, PJS), the National Lottery Charities Board (GV), and St. James's and Seacroft National Health Service Trust Special Trustees (DS).

## References

Abiodun, O. A. (1994). A validity study of the Hospital Anxiety and Depression Scale in general hospital units and a community sample in Nigeria. *British Journal of Psychiatry, 165*, 669–672.

Andersson, E. (1993). The Hospital Anxiety and Depression Scale: Homogeneity of the subscales. *Social Behavior and Personality, 21*(3), 197–204.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*(3) 316–336.

Clark, L. A., Watson, D., & Mineka, S. (1994). Temperament, personality, and the mood and anxiety disorders. *Journal of Abnormal Psychology, 103*(1) 103–116.

Cull, A., Gould, A., House, A., Smith, A., Strong, V., Velikova, G., Wright, P., & Selby, P. (2001). Validating automated screening for psychological distress by means of computer touchscreens for use in routine oncology practice. *British Journal of Cancer, 85*, 1842–1849.

Dagnan, D., Chadwick, P., & Trower, P. (2000). Psychometric properties of the Hospital Anxiety and Depression Scale with a population of members of a depression self-help group. *British Journal of Medical Psychology, 73*, 129–137.

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). A confirmatory factor analysis of the Hospital Anxiety and Depression scale: Comparing empirically and theoretically derived structures. *British Journal of Clinical Psychology, 39*, 79–94.

Gorsuch, R. L. (1983). *Factor Analysis*, London: Erlbaum.

Groenvold, M., Fayers, P. M., Sprangers, M. A., Bjorner, J. B, Klee, M. C., Aaronson, N. K., Bech, P., & Mouridson, H. T. (1999). Anxiety and depression in breast cancer patients at low risk of recurrence compared with the general population: A valid comparison? *Journal of Clinical Epidemiology, 52*(6), 523–530.

Hall, A., A'Hern, R., & Fallowfield, L. (1999). Are we using appropriate self-report questionnaires for detecting anxiety and depression in women with early breast cancer. *European Journal of Cancer, 35*(1), 79–85.

Hopwood, P., Howell, A., & Maguire, P. (1991). Screening for psychiatric morbidity in patients with advanced breast cancer: Validation of two self-report questionnaires. *British Journal of Cancer, 64*, 353–356.

Hopwood, P., & Stephens, R. J. (2000). Depression in patients with lung cancer: Prevalence and risk factors derived from quality-of-life data. *Journal of Clinical Oncology, 18*, 893–903.

Ibbotson, T., Maguire, P., Selby, P., Priestman, T., & Wallace, L. (1994). Screening for anxiety and depression in cancer patients: The effects of disease and treatment. *European Journal of Cancer, 30*(1), 37–40.

Lewis, G. (1991). Observer bias in the assessment of anxiety and depression. *Social Psychiatry & Psychiatric Epidemiology, 26*(6), 265–272.

Lewis, G. & Wessely, S. (1990). Comparison of the General Health Questionnaire and the Hospital Anxiety and Depression Scale. *British Journal of Psychiatry, 157*, 860–864.

Lisspers, J., Nygren, A., & Söderman, E. (1997). Hospital Anxiety and Depression Scale (HAD): Some psychometric data for a Swedish sample. *Acta Psychiatrica Scandinavica*, *96*(4), 281–286.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

Moorey, S., Greer, S., Watson, M., Gorman, C., Rowden, L., Tunmore, R., Robertson, B., & Bliss, J. (1991). The factor structure and factor stability of the Hospital Anxiety and Depression Scale in patients with cancer. *British Journal of Psychiatry, 158*, 255–259.

Razavi, D., Delvaux, N, Farvacques, C., & Robaye, E. (1990). Screening for adjustment disorders and major depressive disorders in cancer in-patients. *British Journal of Psychiatry, 156*, 79–83.

Silverstone, P. H. (1994). Poor efficacy of the Hospital Anxiety and Depression Scale in the diagnosis of major depressive disorder in both medical and psychiatric patients. *Journal of Psychosomatic Research*, *38*(5), 441–450.

Smith, A. B., Velikova, G., Stark, D., Wright, E. P., Strong, V., Gould, A., Cull. A., & Selby, P. J. (2002). Screening efficacy of the HADS: An item-response theory analysis. Manuscript submitted for publication.

Spinhoven, Ph., Ormel, J., Sloekers, P. P. A., Kempen, G. I. J. M., Speckens, A. E. M., & van Hemert, A. M. (1997). A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychological Medicine*, *27*, 363–370.

Velikova G., Wright, E. P., Smith, A., Cull, A., Gould, A., Forman, D., Perren, T., Stead, M., Brown, J., & Selby, P. J. (1999). Automated collection of quality of life data: A comparison of paper and computer touchscreen questionnaires. *Journal of Clinical Oncology*, *17*, 998–1007.

White, D., Leach, C., Sims, R., Atkinson, M., & Cottrell, D. (1999). Validation of the Hospital Anxiety and Depression Scale for use with adolescents. *British Journal of Psychiatry, 175*, 452–454.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*, 361–370.

*Received 1 August 2001; revised version received 29 November 2001*