Item response theory in clinical outcome measurement

Rebecca Holman

Item response theory in clinical outcome measurement
Thesis, University of Amsterdam, the Netherlands

This thesis was prepared in the Department of Clinical Epidemiology and Biostatistics and the Department of Neurology at the University of Amsterdam, the Netherlands with the Department of Educational Measurement and Data Analysis, University of Twente, Enschede, the Netherlands. The project was supported by a grant from the Anton Meelmeijer fonds of the Academic Medical Center, Amsterdam, The Netherlands.

ISBN: 90-9019123-2

Cover: Andrea Graftdijk, Beunderreclame, Hoorn, the Netherlands.
Printed by Febodruk BV, Enschede, the Netherlands, (www.febodruk.nl).
The printing of this thesis was financially supported by Stichting tot bevordering van de Klinische Epidemiologie, University of Amsterdam, the Netherlands.

Item response theory in clinical outcome measurement

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op dinsdag 22 maart 2005, te 12.00 uur

door
Rebecca Holman
geboren te Bristol, Engeland

Promotiecommissie

Promotores	Prof. dr. R.J. de Haan
	Prof. dr. C.A.W. Glas
Co-promotor	Prof. dr. M. Vermeulen
Overige leden	Prof. dr. M.P.F. Berger
	Prof. dr. G.J. Bonsel
	Prof. dr. M.H. Prins
	Prof. dr. M.A.G. Sprangers
	Prof. dr. J.G.P. Tijssen

Faculteit der Geneeskunde

Contents

Chapter 1

Introduction

Measurement in clinical research

In well-known early medical study, the success of the various types of treatment was judged on how quickly sailors with scurvy were fit to resume normal duties onboard[1]. Similarly, the majority of research into acute, life threatening illnesses compares the proportion of patients who recover, rather than die, following various treatment regimes. Medications for illnesses, which are less immediately life threatening, can be compared using the time that patients remain alive, or the time to a specified worsening of their condition. However, a large part of the current burden of disease in Europe stems from chronic conditions. Chronic conditions are generally not fatal in the short term, but do present a substantial barrier to full health and participation in society and economic activities. Examples of common chronic illnesses are asthma, arthritis, heart failure and stroke.

The severity of many chronic illnesses can be measured using a wide range of physiological parameters. Examples include blood tests, imaging techniques and lung volume. These parameters can often be measured very accurately and experienced clinicians often find them easy to interpret. However, these parameters often do not tell the whole story about how the disease process affects patients and their whole life. In addition, physiological parameters often do not take 'side effects' of chronic disease, such as mild depression, reduced social contacts and lowered economic participation, into account. As a result of these limitations of expressing the severity of chronic illness in terms of physiological parameters, interest has moved towards patient centred outcomes. Examples of patient centred outcomes are health related quality of life, cognitive functioning, mobility and disability. Each of these constructs reflects an aspect of the impact of disease on the patient as a whole.

An important aspect of quality of life is the 'disability' status of patients. This is often described in terms of their ability to carry out 'activities of daily life' and measured using multi-item questionnaires, which grade each patient on whether they are able to perform certain activities. A disadvantage of this method is that all items have to be presented to all patients. This has lead to the bandwidth-fidelity problem, where detailed estimates of the status of patients spread across the whole range of functional levels can only be obtained with long questionnaires (broad bandwidth, high fidelity), such as the physical dimension of the Sickness Impact Profile with 65 items[2]. It can cost patients, clinicians and researchers an excessive amount of time to complete such instruments. Shorter instruments either cover a wide range of possible functional status (broad bandwidth, low fidelity), such as the Health Assessment Questionnaire[3] or remain detailed, but cover a much smaller range of levels of functional status (narrow bandwidth, high fidelity), such as the Barthel index[4].

Recently, interest in item response theory (IRT) techniques has grown. These form an alternative measurement paradigm to the sum score and correlation methods, which are becoming popular in research into quality of life and functional status. IRT measures at the item level, in contrast to sum score methods, which are based on a whole instrument[5]. This means that functional status can be assessed in a much more flexible way and that each patient can be presented with a smaller selection of items than is possible using sum score based methods. If these items are carefully selected from a properly constructed item bank, then the estimates of functional status will be detailed and completely comparable, even if each patient is offered a different selection of items. This means that adaptive testing procedures can be implemented resulting in a broad bandwidth, high fidelity instrument to assess functional status, which can be tailored to the functional status of the individual patient[6].

The AMC Linear Disability Score project

The AMC Linear Disability Score (ALDS) project aims to construct an item bank to measure the functional status of patients with a broad range of stable, chronic diseases[7, 8]. Functional status was defined as the ability to perform the activities of daily life required to live independently or in an appropriate care setting[9]. Once the ALDS item bank has been calibrated, it will be used as a basis for using computerised adaptive and other innovative testing procedures to assess the functional status of patients in a wide variety of clinical studies. In addition, the item bank will be used to compare the burden of disability in a wide range of more precisely defined patient groups and to allocate patients to appropriate care settings.

Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific functional health instruments[10] and supplemented by diaries of activities performed by healthy adults. A total of 190 items were identified and then described in detail. For example, ‘shopping’ was expanded to ‘travelling to the shopping centre, on foot or by car, bike or public transport, walking around the shopping centre, going into a number of shops, trying on clothes or shoes, buying a number of articles including paying for them, and returning home’. Two response categories were used: ‘I could carry out the activity’ and ‘I could not carry out the activity’. If patients had never had the opportunity to experience an activity a ‘not applicable’ response was recorded. For example, responses from patients who had never held a full driving licence to the item ‘driving a car’ were recorded in this category. Patients were asked, by trained nurse interviewers, whether they could, rather than did, carry out the activities given. Phrasing questions in terms of capacity may overestimate functional status, but phrasing them in terms of actual performance, may underestimate functional status, since actual performance also depends on personal characteristics and interests[11]. Even though asking patients what they could do is seen as more subject to bias than direct observation[12], it was chosen in the ALDS project as it is practical in both inpatient and outpatient settings and does not place patients in an unnatural ‘laboratory’ situation[13]. The item bank has been calibrated by using the responses from over 4000 patients with a broad range of stable chronic conditions. The patients were interviewed in nursing or care homes, sheltered accommodation or during a visit to one of a range of outpatients’ clinics at one general and two teaching hospitals in Amsterdam, The Netherlands. Each patient was presented with between 32 and 80 items.

Outline of this thesis

This thesis examines some of the statistical and methodological issues, which arise when calibrating and implementing an item bank to quantify functional status as expressed by the ability to perform activities of daily life. These issues have not been previously examined in the light of health status assessment in enough depth to provide a solid foundation for the AMC Linear Disability Score project.

The first part of this thesis examines problems encountered during the calibration phase of the AMC Linear Disability Score project item bank. As in any type of research, when constructing an item bank, it is essential to have a clear plan for carrying out the research and analysing the data. This is described in Chapter 2[8]. A number of problems highlighted are discussed in more depth in Chapters 3, 4 and 5. In the majority of unrushed educational examinations, it is acceptable to assume that pupils, who did not answer given questions, were unable to provide the correct answer to the problem. However, in clinical research, when patients have never experienced an activity, it is less logical to assume that they are unable to perform the activity. Chapter 3 examines four practical procedures for dealing with responses in a 'not applicable' category[14]. One of these methods is examined in more mathematical detail in Chapter 4[15]. When constructing an item bank, it is essential to consider the measurement properties of the items in subgroups in the patient population. In Chapter 5 the measurement properties for men and women and for patients aged 84 or under and patients aged over 84 are compared[16]. Finally, in Chapter 6, the measurement properties of the ALDS item bank are examined in a group of respondents requiring residential care[17].

The second part of this thesis considers two issues influencing the number of patients required to demonstrate the effectiveness of a novel treatment when item response theory based methods are used. Item response theory provides a framework, in which it is fairly easy to adjust the number of items offered to patients. Chapter 7[18] considers how varying the *number* of items used to assess the functional status of patients affects the number of patients required in a study. In item response theory, as with other methods of analysis, different selections of items provide varying degrees of information on patients. Chapter 8 examines the effect of *types* of items from an item bank on the power to detect differences between treatment groups.

Chapter 2

Constructing and calibrating the AMC Linear Disability Score project item bank

This chapter has been published as:

> Holman R, Lindeboom R, Glas CAW, Vermeulen R, de Haan RJ. Constructing an item bank using item response theory: the AMC linear disability score project. *Health Services and Outcomes Research Methodology* 2003: **4**; 19–33.

and is available from `http://dx.doi.org/doi:10.1023/A:1025824810390`

Introduction

Recently, interest in the the use of patient relevant outcomes, such as cognitive functioning, disability, functional status and quality of life, measured using questionnaires, as endpoints in medical research has increased. In spite of the popularity of sum score based approaches[19], some problems are associated with their use. Firstly, responses to all items on a scale are required to calculate a sum score, leading researchers to shorten health instruments to make them more practical but less detailed[20]. Secondly, since sum scores are dependent on the items included in the instrument, it is difficult to compare scores obtained on different instruments, even if they measure the same health concept[10]. Thirdly, the ordinal nature of sum scores makes it difficult to analyse them properly using parametric statistical techniques[21].

Item response theory (IRT) was proposed as an alternative to sum score based approaches for analysing data resulting from the responses of pupils to examination questions and is gaining acceptance in many areas of medical research, including cognitive screening[22, 23], psychiatric research[24], physical functioning[25], mobility[26], disability[27, 28] and quality of life[29]. IRT techniques have proved particularly useful in complex aspects of questionnaire development such as cross-cultural adaptation[30] and multidimensionality[31]. There are a number of advantages to the use of IRT in clinical measurement. One of the most exciting, is the implementation of computerised adaptive testing methods, in which more difficult items are presented to less disabled patients and easier items to more severely impaired patients, whilst ensuring that estimates of health status remain completely comparable[6]. Computerised adaptive testing methods can only be applied, if a calibrated item bank is available. This is a collection of items, which have calibrated by obtaining information on the measurement properties of the items from large groups of appropriate patients. The algorithms involved in computerised adaptive testing require prior knowledge about the measurement properties of the individual items, meaning that it would only be possible to calibrate an item bank using computerised adaptive testing in the second or subsequent stage of a multistage calibration procedure[32]. Other advantages of IRT include proper ways of dealing with ceiling and floor effects, some useful solutions to the problem of missing data and straightforward ways of dealing with heteroscedasticity between treatment or other groups[5]. In addition, there have been suggestions that IRT results in more accurate assessments of health status at patient level and hence in greater power to detect treatment or longitudinal effects[33, 34, 35].

Many publications describe and illustrate IRT techniques in clinical[22, 24, 25, 26, 28, 29, 30, 31] or mathematical terms[32, 36]. However, only a few have examined the practical aspects of the methodological processes involved in calibrating an item bank to measure a clinical construct[27, 23]. This article describes the methodology used during the AMC Linear Disability Score (ALDS) project, which was primarily set up to construct and calibrate an item bank to measure functional health status as expressed by the ability to perform activities of daily life[7], which is assumed to form a unidimensional construct.

Item response theory

A multitude of IRT models have been proposed for a wide variety of types of data[37]. The IRT model, which is most suitable for a particular data set depends on an interplay between the number of response categories for each item, the amount of data available and the reason for carrying out the analysis. The model to be used should be chosen in conjunction with someone with considerable experience of applying IRT models[38] and, often, after the the data have been collected and examined.

In this paper, models based on a logistic function of a single latent trait, θ, for the responses of patients to individual items scored in two categories and forming a unidimensional construct will be examined. In the two-parameter logistic model[39], the probability, p_{ik}, that patient k will respond in category '1' of item i is modelled using

$$p_{ik} = \frac{\exp(\alpha_i(\theta_k - \beta_i))}{1 + \exp(\alpha_i(\theta_k - \beta_i))} \tag{2.1}$$

where θ_k is the value of θ associated with patient k and α_i and β_i describe the behaviour of item i in relation to θ and are known as the slope and location parameters, respectively. If each item has been presented to a relatively small number of patients, say less than 200, then good estimates of both the α and β parameters are difficult to obtain[40] without using Bayes' modal methods. In this type of situation, more stable estimates of item parameters can be obtained if simpler models, such as the one-parameter logistic, or Rasch, model[41], in which α_i is constrained to be equal to 1 for all items, are considered[42]. The one-parameter logistic model has enjoyed widespread popularity in health status outcomes, but rarely fits a given data set satisfactorily[43]. This model can be extended by including a slope parameter, a_i. The parameter a_i plays a similar role to α_i in the two-parameter logistic model, but can only take integer values and is imputed given the results of fit statistics for the one-parameter logistic model. Although it has been shown that constraining the a_i to integer values places very little restriction on the fit of the model to data[44], the extended one-parameter logistic model may be an unsatisfactory final model for a data set. However, its flexibility and computational advantages result in an extremely useful exploratory model.

Item parameters are usually estimated using maximum likelihood methods[45], but it is not possible to maximise the likelihood without making further assumptions about the values of θ_k. Assuming that the values of θ_k are drawn from a Normal distribution and integrating θ out of the likelihood, results in marginal maximum likelihood estimates[46]. For the one-parameter logistic model and its extension, it is possible to assume that the sum of the individual item scores is a sufficient statistic for θ, leading to conditional maximum likelihood estimates of β_i[47]. Once the item parameters have been obtained, θ_k can be estimated using maximum likelihood or empirical Bayesian procedures[48]. The overall fit of IRT models to a data set can be tested by comparing the likelihoods of two hierarchical models[49], or using a Lagrange multiplier approach[50]. However, when calibrating an item bank, interest is often primarily in testing the fit of the model to individual items, by examining whether the proportion of responses predicted by the model to be in each of the response categories is close enough to the observed proportions across the range of θ [51, 45, 52]. The S_i statistic uses the fact that the sum of the scores on the individual items is a sufficient statistic for θ and compares the expected and observed numbers of patients, with given ranges of sum scores, responding in each item response category using a χ^2 based procedure[44]. In practice, parameters are estimated and fit statistics calculated using one of a range of specially developed software packages, such as Bilog[40] to fit the two-parameter logistic model and the package OPLM[53] to fit the one-parameter logistic model and its extension.

Constructing an item bank

The construction of an item bank can be split into four phases: (1) definition of content; (2) choice of calibration design; (3) data collection; and (4) fitting the IRT model. The first and third phases are also an important part of the construction of an instrument using sum score based methods. The second and fourth phases are unique to the use of IRT and will form the focus of this description.

Definition of content

It is important to define the concept to be measured and the patient population of interest carefully. When defining the concept, it can be useful to examine definitions given in previous studies, to study theoretical models for illness and health outcomes and consider whether the definitions given are likely to result in a unidimensional construct. Similarly, a useful starting point for identifying items is a review of existing instruments, to gain insight into how others have seen the construct[10]. A large number of potential items should be identified, since it will not be possible to model the response pattern to a proportion of the items using an IRT model. The number of potential items can be increased by asking patients or healthy volunteers to keep dairies of health related activities, symptoms or moods as appropriate. It is also important to consider how the data will be gathered from the patients, the number of scoring categories per item and how those categories are to be assigned to the responses made by patients. Using two, or at most three, response categories results in stability of scoring across time and researchers and increases clinical interpretability[54].

Choice of calibration design

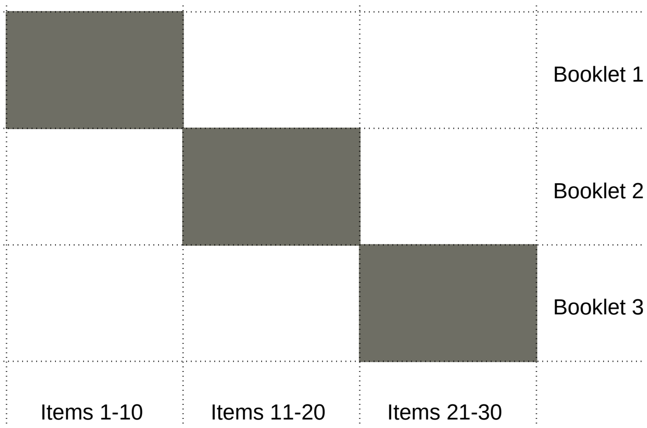

Figure 2.1: an incomplete unanchored calibration design

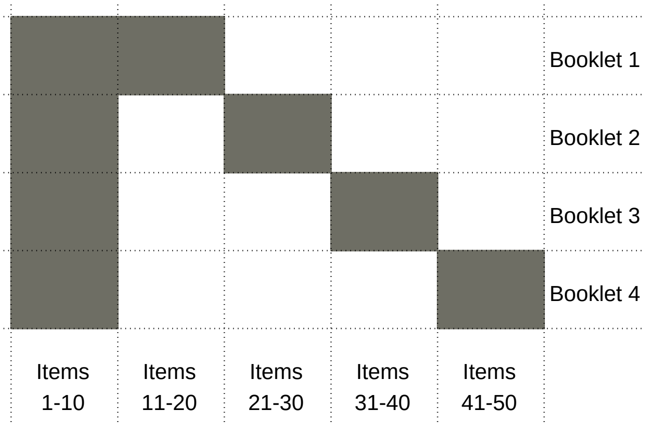

Figure 2.2: an incomplete calibration design with a common anchor

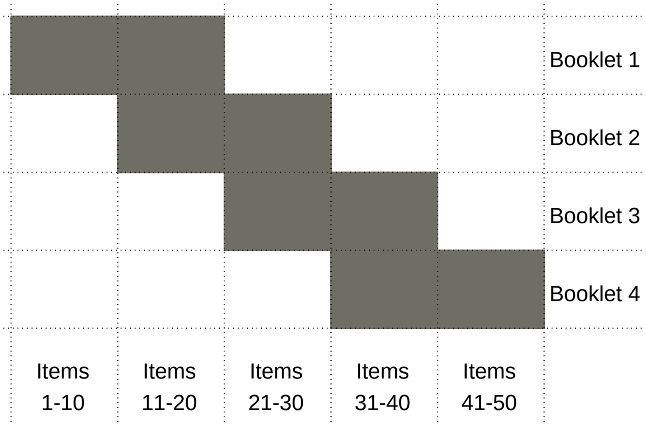

Figure 2.3: an incomplete calibration design with stepped anchors

The calibration design used in the construction of an item bank describes which items are presented to which patients in the data collection phase. The most natural choice may seem to be a ‘complete’ design, where each item is presented to every patient. However, a complete design is inefficient since particular items will be either too easy or too difficult for many patients, meaning that patients will provide very little statistical information on the item parameters. In contrast, an incomplete design presents different subsets of items, often called booklets, to different subgroups of patients. If an unanchored, incomplete calibration design, illustrated in Figure 2.1, were to be used, it would only be possible to place all items on a single scale if patients are randomised to booklets and thus it were reasonable to assume that the distributions of the values of the latent trait, associated with the patients to whom each booklet were administered, were identical.

An incomplete anchored design combines aspects of complete and incomplete calibration designs. The booklets are linked using common items meaning that it is possible to place all items on a single scale, without making any assumptions about the relationships between the the distributions of the values of the latent trait, associated with the patients to whom each booklet were administered[55]. An incomplete anchored calibration design can be constructed in three ways. Firstly, a single set of items can be used as an anchor, resulting in a common item design. This is illustrated in Figure 2.2, where items 1 to 10 form the anchor. A common item design could be used if a number of instruments were to be compared to an existing 'gold standard' instrument. Secondly, each booklet can have a number of items in common with, say, two other, but not all, booklets, resulting in a stepped design, illustrated in Figure 2.3. A stepped design is useful if the patients to be assessed differ greatly in ability and the booklets are ranked by difficulty, meaning that 'healthier' patients are administered completely different items to the most sick. Thirdly, it is also possible to combine these two types of incomplete anchored calibration design. This often occurs, when pre-existing data is used to calibrate a number of related instruments or to compare patient populations measured using different, but related instruments[27].

Data collection

The aim of calibrating an item bank is to obtain information on the measurement properties of the items and on the fit of the IRT model chosen to analyse the responses given by patients to the items. When the item bank is implemented, perhaps in conjunction with computerised adaptive testing[6], the emphasis shifts to estimating the health status of the individual patient. When using the logistic IRT models described in this paper, little statistical information on item parameters is obtained from patients, whose ability level is a lot different from the overall difficulty of the items. The precise point at which most information can be obtained varies according to the IRT model used and the value of the item parameters themselves[56]. However, in general, an item bank can be efficiently calibrated if the difficulty of items in a booklet are roughly matched to the ability of the patients.

Fitting the IRT model

As with the majority of statistical models, fitting an IRT model to a data set can be more of an art, performed using experience and intuition, than a science with exact rules. The process of fitting a model consists of a number of, perhaps iterative, steps including sum score and IRT based techniques. High values of sum score based statistics indicate that the item bank has good measurement qualities and, hence, can be used to construct instruments, which can discriminate between respondents appropriately. IRT based techniques are used to link the item bank and model the measurement properties of the items at given levels of the common latent trait, θ, so that results obtained using instruments assembled from the item bank are interpretable via θ. It should be noted that sum score based statistics give no indication on whether the main assumptions of IRT, unidimesionality[57] or local independence[32], are met and that a perfectly fitting IRT model does not automatically imply good measurement properties. However, if both of these are fulfilled, then a reasonable item bank should result. In this section, some guidance, resulting from the authors' experience, on when to exclude items from an item bank, will be given.

Before fitting an IRT model and examining its quality, it is useful to carry out a number of preliminary analyses. An overall impression of the data can be obtained by counting the number of patients who responded in each of the response categories to each item. It is difficult to obtain accurate estimates of parameters for items, to which the vast majority of responses, say over 90%, are in a single response category[40]. In addition, items with these characteristics do not contribute to the quality of measurements obtained using the item bank. Furthermore, in the authors' opinion, items, to which a substantial proportion, say over 10%, were in categories such as 'not applicable' or 'don't know' may not be suitable for the patient population being used to calibrate the item bank. If responses to any items, actually presented to patients, were omitted or in categories such as 'not applicable' or 'don't know', it can be useful to apply an imputation technique to replace these values. The choice of imputation technique depends on the number of data points to be replaced and the acceptability of certain contextual assumptions[58, 14]. For instance, if items are designed to measure cognitive status, it may be reasonable to assume that if patients do not complete a given item, then they are unable to do so. This assumption may be less valid when measuring functional health status.

The global measurement properties of the items can be investigated using sum score based methods. In order to examine whether particular items measure the same construct as the other items in the same booklet, the correlation, r_{is}, between the scores on a particular item and the total score in a booklet will be examined. The values of r_{is} are often classified as[59]: very good if $r_{is} \geq 0.40$; good if $0.30 \leq r_{is} < 0.40$; moderate if $0.20 \leq r_{is} < 0.30$; and poor if $r_{is} < 0.20$. Often items are removed from further analysis if r_{is} is smaller than a given value. The internal consistency can be assessed, within each booklet, using Cronbach's α[60], and is regarded as acceptable if $\alpha > 0.70$[19].

If the anchor between two booklets does not consist of at least two items, to which the IRT model can be fitted, then it is not possible to place the items from both booklets on a single scale, without assuming that the distributions of the latent trait of patients assessed using the two booklets are identical, as the common value of the standard deviation of θ cannot be estimated. However, the authors feel that at least four items are required in each anchor to enable the parameters of items in different booklets to be compared with sufficient precision to allow the coherence, between booklets, of the functional status construct to be examined satisfactorily. The anchors between booklets consisting of items scored in two categories can be examined using the following graphical method. Booklet h is represented by a line, q_h, ranging from 0 to 1. The parameter q_{hi} for item i in booklet h is

$$q_{hi} = \frac{x_{hi1}}{x_{hi0} + x_{hi1}} \tag{2.2}$$

where x_{hi0} and x_{hi1} are the number of respondents, to booklet h, who responded in the categories '0' and '1' of item i, respectively. The values of q_{hi} for a given item appearing in different booklets are linked to enable the anchors to be visualised. It is unlikely that the values of q_{hi} for a given item in different booklets will be the same. This method is illustrated in Figure 2.4.

Usually, a preliminary choice of IRT model will have been made before the data are collected. However, it can often be useful to examine additional models with more appropriate fit statistics or estimation methods. As described previously, the extended one-parameter logistic model, used together with the associated S_i statistics, is a very suitable exploratory model, as it uses estimation techniques, which make few assumptions about the distribution of the values of the latent variable amongst the patients. For example, consider a data set consisting of responses to items, which have been scored in two categories and examined using the extended one-parameter logistic model. Items can be removed from the analysis in an iterative process, in which the item with the lowest p-value of the fit statistic are removed first, until the p-value of S_i is more than a given value for all items. When choosing this value, it should be borne in mind that it is equal to the type I error rate, that is the proportion of items, for which the model is true but are rejected for not fulfilling the assumptions of the IRT model chosen anyway. For this reason, the p-value of 0.05, which is almost universal in clinical studies, may not be appropriate, particularly when calibrating large item banks. In this situation, a p-value of 0.02 or even 0.01 may be preferred, as this will mean that fewer items will be incorrectly removed from the item bank.

The extension of the one-parameter logistic model has two disadvantages. These are that the values of a_i can only take integer values and the estimates of the item parameters can only be obtained from a data set resulting from the application of an imputation procedure, rather than the original data set. Hence, the two-parameter logistic model should be used as a final IRT model. It is usually not necessary to examine the fit of items to the two-parameter logistic model if the extended one-parameter logistic model has been fitted, as any item which fits this model reasonably, will fit the two-parameter logistic model as well if not better. However, item fit can be checked using suitable statistics[61].

Finally, the quality of the final model should be examined. The anchors between the booklets may have been eroded during model fitting process and should be re-examined. In the process of fitting an IRT model, a substantial number of items may be removed from the analysis, meaning that the anchors may have been weakened.

In addition, the density of the fitted items can be examined by plotting all items on a single scale. The items should be spread across all levels of health status and there should be no large gaps between item difficulties across the range of health status, for which the item bank is to be used. In addition, if the aim of calibrating the item bank was to provide a basis for computerised adaptive testing methods, then it is useful for the majority of items to have a relatively high value of α_i, since such items provide the most information on whether a patient has health status above or below the value of β_i.

The AMC Linear Disability Score project

Definition of content

The AMC Linear Disability Score (ALDS) project aims to construct an item bank to measure the functional status of patients with a broad range of stable, chronic diseases[7]. Functional status was defined as the ability to perform the activities of daily life required to live independently or in an appropriate care setting[9]. Once the ALDS item bank has been calibrated, it will be used as a basis for using computerised adaptive and other innovative testing procedures to assess the functional status of patients in a wide variety of clinical studies. In addition, the item bank will be used to compare the burden of disability in a wide range of more precisely defined patient groups and to allocate patients to appropriate care settings.

Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific functional health instruments[10] and supplemented by diaries of activities performed by healthy adults. A total of 190 items were identified and then described in detail. For example, ‘shopping’ was expanded to ‘travelling to the shopping centre, on foot or by car, bike or public transport, walking around the shopping centre, going into a number of shops, trying on clothes or shoes, buying a number of articles including paying for them, and returning home’. Two response categories were used: ‘I could carry out the activity’ and ‘I could not carry out the activity’. If patients had never had the opportunity to experience an activity a ‘not applicable’ response was recorded. For example, responses from patients who had never held a full driving licence to the item ‘driving a car’ were recorded in this category. Patients were asked, by trained nurse interviewers, whether they could, rather than did, carry out the activities given. Phrasing questions in terms of capacity may overestimate functional status, but phrasing them in terms of actual performance, may underestimate functional status, since actual performance also depends on personal characteristics and interests[11]. Even though asking patients what they could do is seen as more subject to bias than direct observation[12], it was chosen in the ALDS project as it is practical in both inpatient and outpatient settings and does not place patients in an unnatural ‘laboratory’ situation[13].

Choice of calibration design

In the ALDS project, data was collected using an incomplete anchored calibration design similar to the one presented in Figure 3, but using 10 booklets, ranging from difficult (booklet 1) to very easy (booklet 10). Booklet 1 contains activities, which can only be carried out by those who are relatively healthy, and booklet 10 those, which can be carried out by all but the most severely disabled patients. Half of all the items in a given booklet are common with the booklet above and the other half with the booklet below, meaning that each item is in two booklets and the whole design is anchored. This design was chosen because it allowed a lot of statistical information to obtained, whilst keeping the burden on patients as low as possible. It should be emphasised that the ‘booklet’ structure described in this paper was designed to be used in the calibration process only. The authors do not advocate that these booklets should be used in future studies, but it would have been difficult to calibrate the item bank using computerised adaptive testing, as the algorithms involved require more detailed knowledge about the measurement properties of the individual items than it was possible to obtain before the calibration process began.

Data collection

The data described in this article were collected from 730 moderately disabled patients with a broad range of stable chronic conditions. The patients were interviewed during a visit to one of the neurology, rheumatology, pulmonology, internal medicine, vascular surgery, cardiology, rehabilitation medicine and gerontology outpatients' clinics at one general and two teaching hospitals in Amsterdam, The Netherlands. Each patient was presented with one of the four most difficult booklets in the calibration design, which encompass a total of 75 distinct items. Data to calibrate the easier items, more suited to patients with a lower level of functional status, in the remaining 6 booklets are currently being collected in institutions providing a variety of types of residential care. In order to increase the statistical efficiency of the design, the nurses interviewing the patients roughly matched the 'difficulty' of a booklet to the ability level of each patient, using their clinical experience. Hence, the easiest booklet was only presented to patients with substantial disabilities and the most difficult to those with minimal impairments. In practice, if a patient was able to carry out fewer than ten or more than twenty of the 32 activities described in each booklet to which they were allocated, the patient was re-assessed using an easier or more difficult booklet as appropriate.

Fitting the IRT model

Firstly, the number of responses to each item in each category was examined. More than 10% of the responses to 12 items were in the category 'not applicable', meaning that 63 of the original 75 items proceeded to a hot deck imputation procedure, based on logistic regression[16]. This was implemented and the resulting data set used to evaluate the correlations between the scores on the individual items and the sum scores within each of the four booklets. Fifteen of the remaining 63 items were removed because the correlation between their score and the total scores were less than 0.3 for all booklets in which the item appeared, leaving a total of 48 items. At this point, 22 items remained in booklet 1, 25 in booklet 2, 21 in booklet 3 and 17 in booklet 4 and gave values of Cronbach's α for the booklets were 0.79, 0.64, 0.71 and 0.66, respectively. Two of these values are below the recommended minimum of 0.70, but removing more items may have weakened the anchors between the booklets, making it impossible to fit the IRT model.

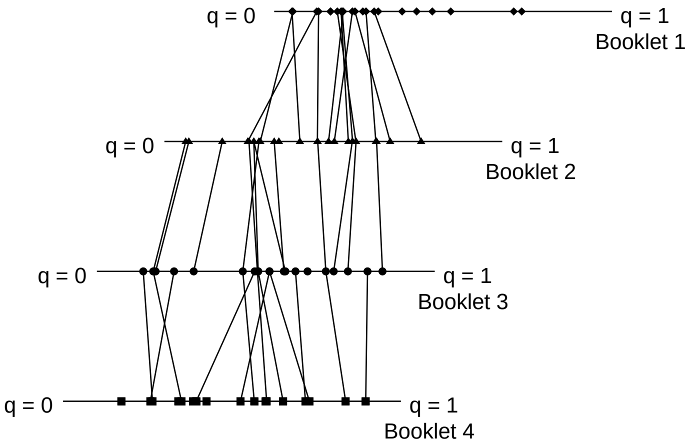

Figure 2.4: the anchors between the four booklets examined in this paper

The strength and structure of the anchors, which link a booklet to the one above and the one below it, have been examined using a graphical method and are illustrated in Figure 2.4. It can be seen that the proportion of patients responding in the category ‘I could carry out the activity’, q_{hi}, are well spread across the axes representing each of the four booklets, indicating that there are no strong ceiling or floor effects. In addition, there is a reasonable number of items in each of the three anchors and the items in the anchors are spread across the axes representing the booklets, suggesting that the anchors are ‘strong’ enough to enable the common value of the standard deviation of functional status to be estimated and, thus, comparable values of all item parameters obtained. Furthermore, the original design, in which the difficulty level of the booklets was matched to the functional status of the patients is partly reflected in Figure 4. It can be seen that the ‘easier’ items in booklet 1 form the anchor with booklet 2, whereas this is less clear for the anchors between booklets 2 and 3 and booklets 3 and 4.

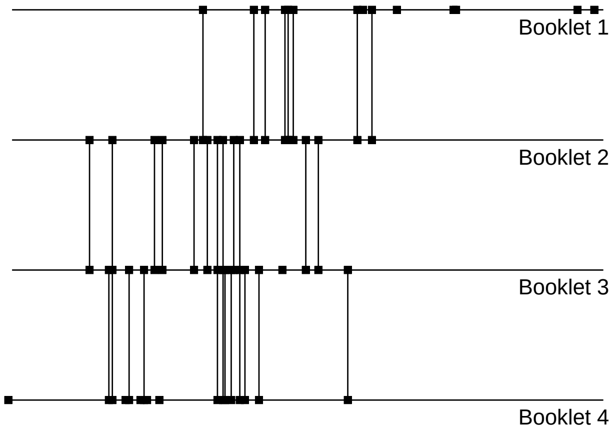

Figure 2.5: the anchors following the data analysis

The extended one-parameter model was fitted to the 48 items remaining in the calibration. Seven items were removed from the model, due to large values of the S_i statistic, meaning that 41 items proceeded in the analysis. Finally, the two-parameter logistic model was fitted to the remaining items using the original, pre-imputation data set. The anchors following the data analysis and expressed in terms of the values of β_i are given in Figure 2.5. There are still at least eight items in each anchor, meaning that the anchors between the booklets have not been substantially eroded by the removal of items during the calibration process. It is also apparent that the averages of the 'difficulty' of the items in each of booklets 2, 3 and 4, is remarkably similar, while booklet 1 remains more difficult than the others.

The results of the calibration process, for a selection of items, are illustrated in Figure 2.6. In the lower half of this Figure, the probability, modelled using the two-parameter logistic model, that a patient is able to perform a given activity, given their functional ability, is plotted. The content of the items is identified by arrows

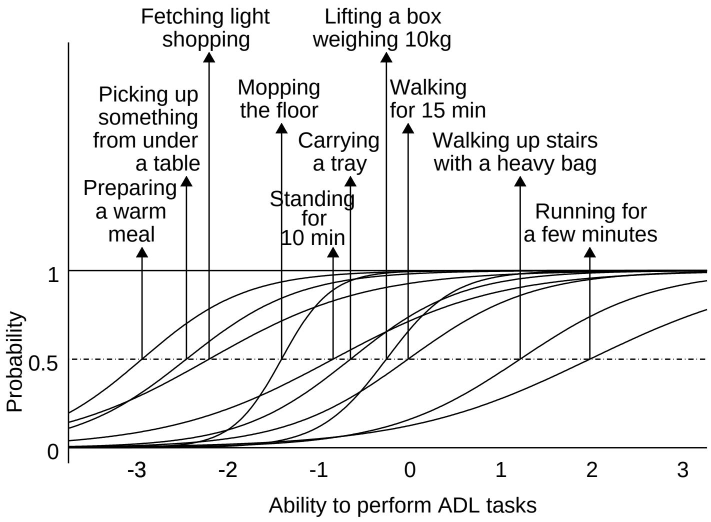

Figure 2.6: the results of the calibration process

pointing to labels in the upper part of the Figure. The value of the β_i parameter for a particular item is at the point where the curve for that item crosses the horizontal broken line. The relative difficulty of the items is usually expressed in terms of the ordering on this line. For instance, the item ‘picking up something from under a table’ is easier than ‘standing for 10 minutes’, which in turn, is easier than ‘walking for 15 minutes’. The differences in the α parameters can be seen the the variation of the steepness of the item curves. For example, the item ‘lifting a box weighing 10kg’ has a larger value of α than ‘standing for 10 minutes’.

Discussion

This article has developed a methodology for calibrating an item bank to measure functional health status using item response theory (IRT) methods. The majority of publications on the use of IRT in health status assessment have used data collected in the framework of a study primarily carried out for another purpose. In contrast, this article has presented the methodology and techniques required, when the primary aim of a study is to develop an item bank. Data from the AMC Linear Disability Score project[7] were used as an illustration.

Once an item bank has been constructed and the calibration process completed, the item bank can be used in a number of ways to assess the health status of patients. An important characteristic of a calibrated item bank is that the health status of two patients can be compared, even if they are assessed using disjoint sets of items, facilitating the use of computerised adaptive testing. When computerised adaptive testing algorithms are implemented, the items administered to patients depends on the responses they gave to previous items[6]. For example, more 'difficult' items will be administered to healthier patients, whilst 'easier' items are administered to sicker patients. Properly administered, these algorithms can lead to accurate estimates of health status whilst keeping the burden of testing on the patient as low as possible. The ways in which these procedures can be implemented are only limited by the imagination of those administering item banks and the willingness of clinicians to accept new ways of measuring latent constructs.

The authors expect that, once the ALDS item bank has been calibrated, it will be used as a basis for using a variety of testing procedures, including computerised adaptive testing, to assess the functional status in individuals, groups and populations in a wide variety of clinical studies. Selections of the items are currently being used in studies to investigate the effectiveness of a range of medical interventions. The data collected in these and future studies will be stored and used to update the estimates of the item parameters and to examine whether the ALDS item bank performs in the same way in actual clinical studies as in the calibration process.

Chapter 3

Practical methods for dealing with responses in the category 'not applicable'

This chapter is adapted from the following article:

> Holman R, Glas CAW, Zwinderman AH, de Haan RJ. Practical methods for dealing with 'not applicable' item responses in the AMC Linear Disability Score project. *Health Qual Life Outcomes.* 2004; **2**: 29.

and is available from `http://www.hqlo.com/content/2/1/29`

Background

When questionnaires consisting of a number of related items are used to measure constructs such as health related quality of life[25, 62], cognitive ability[63] or functional status[7], it is likely that some patients will omit responses to a subset of items. A variety of ways of dealing with missing item responses in this type of questionnaires have been proposed[58]. These range from imputation methods[64, 65] to algorithms, which permit parameters to be estimated, whilst ignoring missing data points[66] and frameworks, in which it is possible to construct a joint model for the data and the pattern of missing data points[67]. It is always essential to examine why some responses are missing and whether there is a pattern underlying the missing data for questionnaires[68, 69, 70], but particularly when an item bank is being calibrated. A calibrated item bank is a large collection of questions, for which the measurement properties, in the framework of item response theory, of the individual items are known and should form a solid foundation for measuring the construct of interest. This foundation could be weakened if the treatment of missing item responses had not been properly examined.

The AMC Linear Disability Score (ALDS) item bank aims to measure functional status, as defined by the ability to perform activities of daily life[7, 10, 71]. Items for inclusion in the ALDS item bank were obtained from a systematic review of generic and disease specific instruments for measuring the ability to perform activities of daily life[10] and supplemented by diaries of activities performed by healthy adults. The ALDS items were administered by specially trained nurses. Two response categories were used: ‘I could carry out the activity’ and ‘I could not carry out the activity’. If patients had never had the opportunity to experience an activity a not applicable response was recorded. In the context of the ALDS item bank, it is not immediately clear how responses in the category ‘not applicable’ should be analysed. Some instruments, such as the CAMCOG neuropsychological test battery[63, 72] and the Sickness Impact Profile[2], treat such responses as a ‘negative’ category and others, such as the SF-36[25, 62], impute a response based on those given to the other items. In this paper, responses to the ‘not applicable’ category in the ALDS project have been examined in the wider context of missing data[14].

In this paper, four practical, missing data based strategies for dealing with responses in the category ‘not applicable’ are examined in the context of item response theory. The four strategies are: cold deck imputation; hot deck imputation; treating the missing responses as if these items had never been offered to those individual patients; and using a model which takes account of the ‘tendency to respond to items’. The results will be used to make recommendations about the choice of procedure in the ALDS project and other measures of functional status, which are analysed with item response theory.

Methods

Data

The whole ALDS item bank, consisting of approximately 200 items, is currently being calibrated using an incomplete design[55] with around 4000 patients[7, 8]. Since this paper concentrates on the utility of four missing data techniques, rather than on fitting an item response theory model, the data described come from a single subset 32 items and the responses from 392 patients. In Table 3.1, a short description of the content in each of the 32 items used in this analysis is given, along with the number of the 392 patients responding in the category ‘not applicable’. The number of responses per item in this category varies from 2 (1%) to 133 (34%). Fourteen of the 32 items have more than 20 (5%) responses in the category ‘not applicable’. Of the 392 patients, 108 had no responses in the category ‘not applicable’ and 284 patients responded to between 1 and 12 of the 32 items in this category. Of the 284 patients with ‘not applicable’ responses, 94 had four or more ($> 10\%$) and 20 seven or more ($> 20\%$) responses in this category. Overall, 841 of the 12544 (7%) responses are ‘not applicable’. Thus, a substantial proportion of the data points in this subset of the data used to calibrate the ALDS item bank can be classified as ‘omitted’.

Table 3.1: Item content with the number of patients responding in the 'not applicable' category in parenthesis. Items denoted by (++) demonstrated item misfit across more than one method and items denoted by (+) demonstrated item misfit for one method.

Item Number	Item description	Responses in 'not applicable' category	Item misfit indicator
1	Running for more than 15 minutes	2	++
2	Going for a walk in the woods	2	
3	Running for less than 5 minutes	3	
4	Walking up a hill or high bridge	3	++
5	Lifting up a toddler	3	
6	Moving a bed or table	4	
7	Playing with a child on the floor	5	
8	Tightening a screw	5	+
9	Going shopping for clothes	6	++
10	Change a light bulb in a ceiling lamp	7	
11	Mopping the floor	11	++
12	Putting the rubbish out	12	
13	Lifting a box weighting 10 kg	13	
14	Shopping for groceries for a week	13	
15	Painting a ceiling	14	
16	Cleaning a bathroom	17	
17	Carrying a heavy bag upstairs	17	
18	Painting a wall	18	
19	Cycling for 15 minutes	24	
20	Change sheets and duvet cover on bed	25	
21	Caring for potted plants on a balcony	25	
22	Vacuuming a flight of stairs	26	
23	Washing a window from the outside	27	
24	Cycling with a heavy load of shopping	30	
25	Pumping up a bicycle tyre	33	
26	Travelling by plane	38	
27	Mopping a flight of stairs	39	
28	Vacuuming the inside of a car	48	
29	Swimming for an hour	54	+
30	Washing a car	82	
31	Mowing the lawn	102	
32	Repairing a puncture in bicycle tyre	133	

Dealing with ‘not applicable’ item responses

This section describes the four strategies for dealing with these responses: cold deck imputation; hot deck imputation; treating the missing responses as if these items had never been offered to those individual patients; and using a model which takes account of the ‘tendency to respond to items’. These strategies were chosen because they are implemented in instruments measuring similar constructs and the authors regarded them as representing clinically plausible mechanisms. The strategies will be compared by examining the root mean squared difference, as defined in the Appendix, between estimates of the item parameters and by comparing estimates of the mean functional status in the group.

Cold deck imputation replaces each missing data point with a pre-determined constant. This may be the same for each data point or vary with factors internal or external to the data. For example, it has been recommended that missing item responses in the SF-36 be replaced by the mean of the responses to other items in the same sub-scale[25, 62]. Imputing the same value for all missing data points can be attractive because of its apparent simplicity or because researchers feel that they have a strong justification for the choice of constant in the context of the data. However, this method artificially reduces the amount of variability in the data, possibly leading to substantial bias in parameter estimates. In addition, statistical theory provides little support for this method[70]. The cold deck imputation procedure used in this paper replaces all responses made in the category ‘not applicable’ with ‘cannot’. This is consistent with some other questionnaires for measuring aspects of functional status, such as the Sickness Impact Profile[2], the Mini-mental state examination and the CAMCOG[72], in which items, to which patients make no response, are coded in a ‘negative’ category.

Hot deck imputation replaces each missing value with a value drawn from a plausible distribution[69] incorporating theoretical or observed aspects of the data[70]. Clinicians may feel that hot deck imputation procedures introduce an unnecessary random element into their data, and hence be wary of these methods. However, if the hot deck procedure is run a number of times and each data set is analysed in the same way, differences in the results can be used to make inferences about the effect of the imputation procedure[69]. In this paper, the hot deck imputation procedure has been run five times, resulting in five complete data sets, and is based on logistic regression and closely mirrors the one-parameter logistic IRT model described above. The procedure is constructed, so that patients with a higher level of functional status have a higher probability of having responses in the category 'can carry out the activity' imputed than patients with a lower level of functional status. Similarly, responses imputed for more difficult items are more likely to be in the category 'cannot carry out the activity' than those for easier items. Technical details of the hot deck imputation procedure are given in the Appendix.

In some circumstances, it may be desirable to act as if the researchers had no intention of collecting the missing data points[66]. This avoids any potential bias or reduction of variability introduced by an imputation procedure. Care should be taken that only the data points that are actually missing are 'ignored', rather than that the whole case, or unit, is removed from the analysis, as occurs in many standard procedures. When using IRT and marginal maximum likelihood estimation procedures[73, 46], it is possible to treat items, to which no response was made, as if they had never been offered to the respondent[74]. This is equivalent to ignoring the missing responses[46] and is essential in the application of computerised adaptive testing[6, 75]. This procedure is explained in more depth in the Appendix.

A number of models have been proposed, which directly incorporate the pattern of 'missing' item responses into the model used to examine the data. These models rest on the assumption that two, perhaps related, processes are at work when an item is presented to a patient. The first process can be described as the tendency to judge items to be applicable to one's own situation or the tendency to respond to items[74]. The second process reflects the patients' functional status. These two processes can be modelled jointly by using the one-parameter logistic IRT model for each process individually and assuming that the health status of a patient and the tendency to judge items to be applicable is correlated[76]. This type of model is described in more depth elsewhere[15].

Statistical analysis

In this paper, the one-parameter logistic model[41], sometimes known as the Rasch model, is used as a tool to analyse the response patterns given by patients to a set of items. This model examines the probability P_{ik} that patient k, with functional status equal to θ_k, responds to item i in the category 'can carry out', where

$$P_{ik} = \frac{\exp(\theta_k - \beta_i)}{1 + \exp(\theta_k - \beta_i)} \tag{3.1}$$

and β_i describes the 'difficulty' of item i in relation to the construct functional status. It is unlikely that this model would fit functional status data satisfactorily enough to be used as a final model for an instrument, but since the aim of this study is to compare the performance of a number of methods for dealing with missing data, this simpler model is acceptable. The extent to which all items represented a single construct was examined using Cronbach's alpha coefficient[60].

In this paper, a two stage procedure was used to estimate the parameters in the one-parameter logistic model. Firstly, the item parameters (β_i) were estimated. In this process it was assumed that the values of the functional status (θ_k) formed a Normal distribution, resulting in marginal maximum likelihood estimates. Secondly, estimates of the patients' functional status (θ_k) were obtained.

The fit of the model to the data was assessed using weighted residual based indices transformed to approximately standard Normal deviates[77, 73]. Values above 2.54 (1% level) were regarded as indicative of item misfit. Estimates of the item difficulty parameters (β_i) obtained using the different procedures for dealing with missing data were compared using the root mean squared difference, as described in the Appendix.

The best estimates of functional status for individual patients are usually obtained using maximum likelihood methods. However, clinical studies are often more concerned with inferences based on groups of patients. It has been shown that using maximum likelihood estimates of the functional status (θ_k) in standard statistical techniques can lead to substantial biases[78, 79]. To avoid this, plausible values for the functional status of each patient have been drawn from their own posterior distribution of θ[73]. The item parameters and patients' functional status have been estimated in ConQuest[73]. Other calculations were carried out in S-PLUS[80].

Results

The estimates of the item parameters (β_i) and their standard errors are given in Table 3.2. Standard errors for the parameters in the 'tendency to respond' model are not currently available in the software. This is indicated by the symbol '-' in Table 3.2. Items denoted by (++) demonstrated item misfit across more than one method and items denoted by (+) demonstrated item misfit for one method. The values of Cronbach's alpha coefficient for each procedure are given in the bottom row of Table 3.2. All values are greater than 0.8, indicating that the items reflect a single construct.

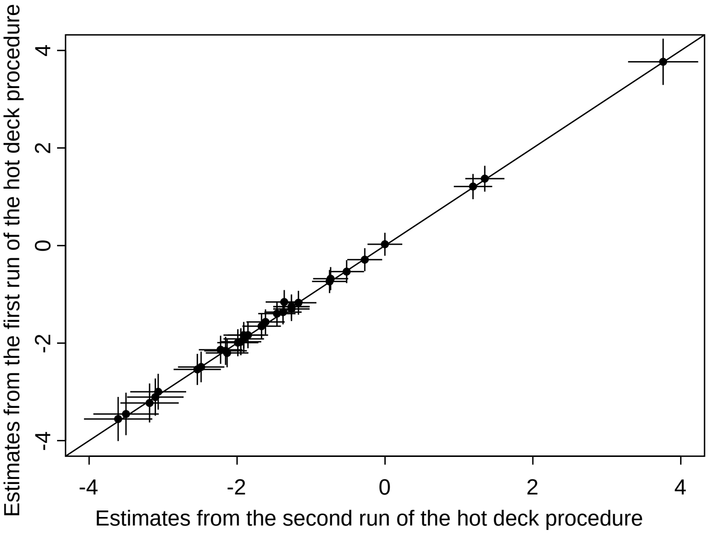

Figure 3.1: The estimates of the item parameters obtained using the first two runs of the hot deck imputation procedure. The horizontal and vertical lines indicate the 95% confidence intervals for the estimates obtained using the first and second runs, respectively.

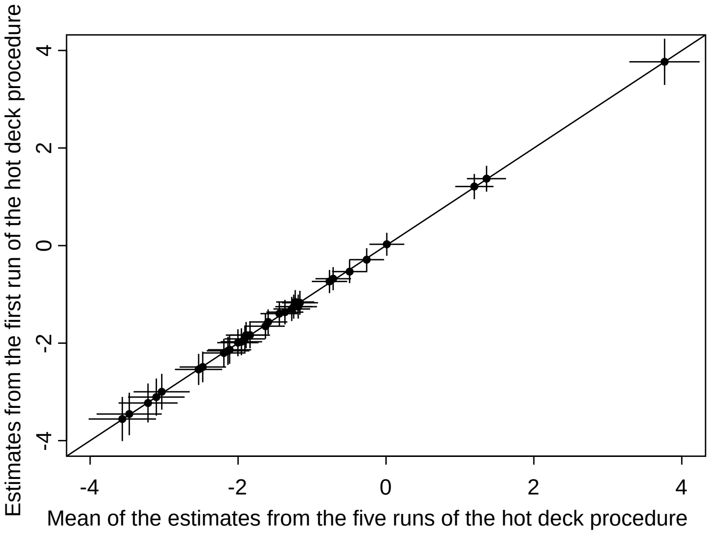

Figure 3.2: The estimates of the item parameters obtained using the first run and the mean of five runs of the hot deck imputation procedure. The horizontal and vertical lines indicate the 95% confidence intervals for the estimates obtained using the first and second runs, respectively.

The root mean squared differences (RMSD) between the estimates of the item parameters obtained using the cold deck imputation procedure, the first and second runs of the hot deck imputation procedure, treating the missing responses as if these items had never been offered to those individual patients and using a model which takes account of the 'tendency to respond to items' are given in Table 3.3. The values of the RMSD between the estimates obtained from the first and second runs of the hot deck imputation procedure are lower. This indicates that the different runs of the hot deck imputation procedure result in very similar point estimates of the item difficulty parameters. The 95% confidence intervals of these point estimates are plotted in Figure 3.1. The diagonal line indicates where the confidence intervals would cross if the estimates from the two runs were identical. Both 95% confidence intervals for all items cross this line and the lengths of the confidence intervals for both runs are similar, indicating that interval estimates of the item difficulty

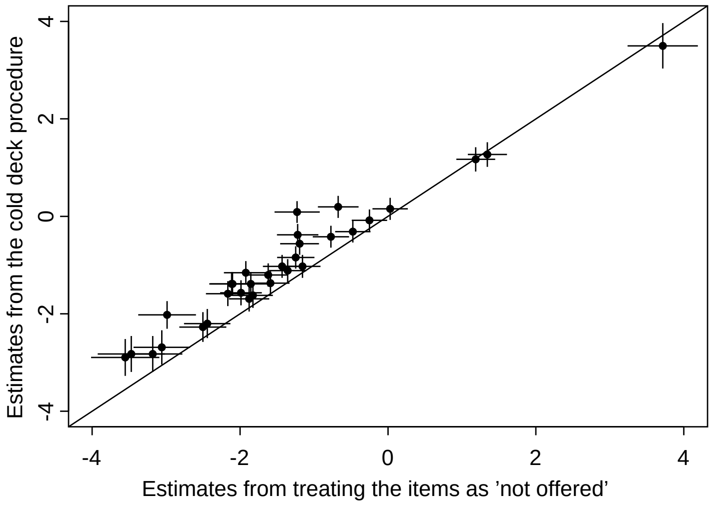

Figure 3.3: The estimates of the item parameters obtained using the cold deck imputation procedure and by treating the missing item responses as if they had never been offered to the individual patients. The horizontal and vertical lines indicate the 95% confidence intervals for these estimates.

parameters are similar over runs of the hot deck imputation procedure. Figure 3.2 is similar to Figure 3.1, but compares the interval estimates obtained in the first run of the hot deck imputation procedure with those obtained by combining the five estimates obtained in the five runs of the hot deck imputation procedure. The interval estimates for the mean of the five runs are slightly wider than those obtained from a single run, illustrating the correction made to account for the fact that some data points are imputed.

Re-examining Table 3.3, it can be seen that the RMSD, which result from comparing the cold deck imputation procedure with the other procedures are over ten times the size of the RMSD, which result from comparing the estimates obtained from other combinations of procedures. Figure 3.3 is a plot of the estimates using the cold deck imputation procedure against the estimates obtained when the missing responses were treated as if these items had never been offered to those individual

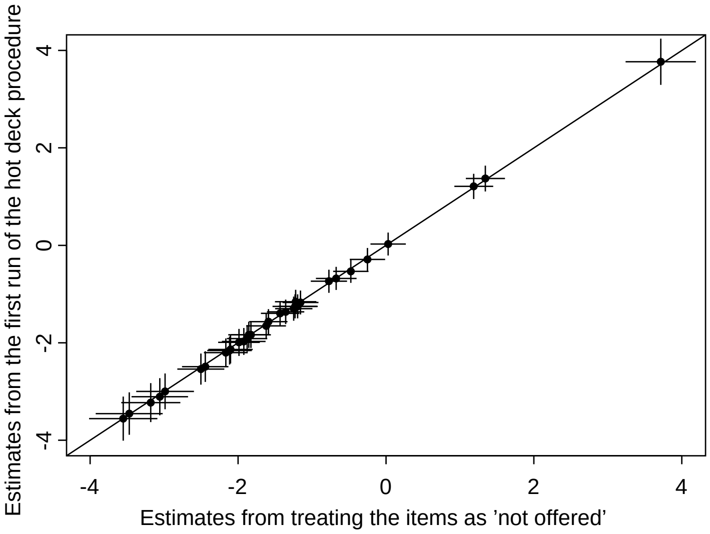

Figure 3.4: The estimates of the item parameters obtained using the first run of the hot deck imputation procedure and by treating the missing item responses as if they had never been offered to the individual patients. The horizontal and vertical lines indicate the 95% confidence intervals for these estimates.

patients. In contrast to Figures 3.1 and 3.2, the 95% confidence intervals of the two estimates intersect above the diagonal line for the majority of items. In addition, for 18 items, both confidence intervals do not cross the diagonal line. The results in Table 3.3 and Figure 3.3 indicate that both point and interval estimates obtained using the cold deck imputation procedure are very different and systematically biased from the estimates obtained using the other procedures. Plots of the estimates obtained using the cold deck imputation procedure against those obtained from the remaining procedures have a similar appearance to Figure 3.3.

Table 3.2: The estimates of the item parameters (β_i) and their standard errors (in parenthesis) for each of the procedures. In addition, Cronbach's alpha coefficient (CAC) is given for each procedure. Standard errors for the parameters in the 'tendency to respond' model are not currently available in the software. This is indicated by the symbol '-'.

	Estimates of the item parameters ($\hat{\beta}$)				
	Hot			Including	Mean
	deck		Items	tendency	5 runs
Item	1st	Cold	never	to	hot
Number	run	deck	offered	respond	deck
1	3.77 (0.242)	3.49 (0.238)	3.71 (0.242)	3.72 (–)	3.76 (0.242)
2	-1.17 (0.125)	-1.02 (0.120)	-1.15 (0.124)	-1.16 (–)	-1.16 (0.125)
3	1.37 (0.135)	1.26 (0.129)	1.34 (0.135)	1.34 (–)	1.36 (0.135)
4	-2.54 (0.163)	-2.27 (0.156)	-2.50 (0.162)	-2.51 (–)	-2.53 (0.163)
5	-1.91 (0.140)	-1.69 (0.134)	-1.87 (0.139)	-1.88 (–)	-1.90 (0.140)
6	-2.49 (0.160)	-2.20 (0.153)	-2.44 (0.160)	-2.44 (–)	-2.47 (0.160)
7	-1.84 (0.137)	-1.62 (0.132)	-1.82 (0.138)	-1.83 (–)	-1.84 (0.138)
8	-3.23 (0.204)	-2.82 (0.188)	-3.18 (0.204)	-3.18 (–)	-3.21 (0.204)
9	-3.11 (0.195)	-2.69 (0.179)	-3.05 (0.195)	-3.06 (–)	-3.10 (0.195)
10	-1.57 (0.131)	-1.37 (0.126)	-1.59 (0.132)	-1.59 (–)	-1.59 (0.132)
11	-3.56 (0.231)	-2.89 (0.193)	-3.55 (0.236)	-3.55 (–)	-3.56 (0.233)
12	-3.45 (0.222)	-2.82 (0.188)	-3.47 (0.231)	-3.47 (–)	-3.47 (0.224)
13	-1.37 (0.128)	-1.11 (0.121)	-1.35 (0.129)	-1.36 (–)	-1.36 (0.128)
14	0.03 (0.120)	0.15 (0.115)	0.03 (0.122)	0.03 (–)	0.01 (0.120)
15	1.21 (0.132)	1.17 (0.127)	1.18 (0.134)	1.19 (–)	1.19 (0.132)
16	-1.99 (0.142)	-1.57 (0.131)	-1.98 (0.144)	-1.99 (–)	-2.00 (0.142)
17	-0.53 (0.120)	-0.31 (0.114)	-0.47 (0.122)	-0.48 (–)	-0.49 (0.121)
18	-0.29 (0.120)	-0.08 (0.114)	-0.25 (0.122)	-0.25 (–)	-0.26 (0.120)
19	-1.84 (0.137)	-1.38 (0.126)	-1.85 (0.142)	-1.86 (–)	-1.89 (0.140)
20	-2.20 (0.149)	-1.58 (0.131)	-2.16 (0.151)	-2.17 (–)	-2.19 (0.150)
21	-1.65 (0.133)	-1.20 (0.122)	-1.62 (0.137)	-1.62 (–)	-1.63 (0.134)
22	-1.40 (0.128)	-1.02 (0.120)	-1.43 (0.133)	-1.43 (–)	-1.44 (0.130)
23	-1.30 (0.127)	-0.84 (0.117)	-1.24 (0.129)	-1.25 (–)	-1.27 (0.126)
24	-0.74 (0.121)	-0.41 (0.114)	-0.77 (0.125)	-0.77 (–)	-0.76 (0.122)
25	-3.00 (0.188)	-2.02 (0.145)	-2.98 (0.199)	-2.99 (–)	-3.03 (0.193)
26	-2.14 (0.147)	-1.38 (0.126)	-2.10 (0.153)	-2.10 (–)	-2.11 (0.149)
27	-2.16 (0.147)	-1.38 (0.126)	-2.11 (0.154)	-2.12 (–)	-2.13 (0.147)
28	-1.97 (0.141)	-1.15 (0.122)	-1.92 (0.151)	-1.92 (–)	-1.95 (0.142)
29	-1.25 (0.126)	-0.56 (0.115)	-1.19 (0.134)	-1.20 (–)	-1.18 (0.129)
30	-1.16 (0.125)	-0.37 (0.114)	-1.22 (0.143)	-1.22 (–)	-1.23 (0.131)
31	-0.68 (0.121)	0.19 (0.115)	-0.67 (0.140)	-0.67 (–)	-0.71 (0.122)
32	-1.25 (0.126)	0.08 (0.114)	-1.22 (0.156)	-1.23 (–)	-1.25 (0.127)
CAC	0.87	0.84	0.81	0.81	0.87

Table 3.3: Using the root mean squared difference to compare the estimates of item parameters obtained in the different procedures. ‘Cold deck’ denotes cold deck imputation, ‘1st hot deck’ and ‘2nd hot deck’ the first and second runs of the hot deck imputation procedure, respectively, ‘Mean hot deck’ the mean of all 5 runs of the hot deck imputation procedure, ‘Never offered’ the procedure treating ‘not applicable’ responses as if the item had never been offered to the patient and ‘Tendency’ the model taking account of the tendency to respond to items’.

				Mean	Items
	Cold	1st run	2nd run	5 runs	never
	deck	hot deck	hot deck	hot deck	offered
1st run hot deck	0.5462				
2nd run hot deck	0.5712	0.0518			
Mean 5 runs hot deck	0.5493	0.0280	0.0396		
Items never offered	0.5317	0.0358	0.0496	0.0249	
Tendency to respond	0.5316	0.0351	0.0494	0.0242	0.0020

Table 3.4: Estimates of the mean and standard deviation of the functional status obtained using the a variety of procedures to estimate the functional status for the individual patients and the measurement characteristics of the items.

Procedure used to deal with NA responses	Mean	Standard deviation	95% Confidence interval for mean
Cold deck imputation	1.17	1.21	(1.05, 1.29)
Hot deck imputation	1.67	1.57	(1.52, 1.83)
Treating ‘NA’ as if the items had never been presented	1.65	1.52	(1.50, 1.80)

The RMSD, in Table 3.3, which result from comparing the first run and mean estimates over the five runs of the hot deck imputation procedure, treating the missing responses as if these items had never been offered to those individual patients and using a model which takes account of the ‘tendency to respond to items’, are even lower than the value of the RMSD used to compare the first and second runs of the hot deck imputation procedure. Figure 43.4 is a plot of the estimates using the first run of the hot deck imputation procedure against the estimates obtained when treating the missing responses as if these items had never been offered to those individual patients. The 95% confidence intervals of the two estimates intersect very close to and cross the diagonal line for all items. The results in Table 3.3 and Figure 3.4 indicate that the point and interval parameter estimates obtained using the two procedures are very similar. Other plots of the estimates obtained using the first run of the hot deck imputation procedure, treating the missing responses as if these items had never been offered to those individual patients and using a model which takes account of the ‘tendency to respond to items’ had a similar appearance. The correlation between estimates of the functional status of a patient and of the ‘tendency to respond to items’ was 0.136. This shows that patients with a higher functional status are marginally more likely to omit items than patients with a lower functional status.

Estimates of the mean and the standard deviation of the level of functional status, obtained using different procedures for dealing with responses in the category ‘not applicable’, are given in Table 3.4. The mean and standard deviation are lower when cold deck imputation is used than for the other methods, which result in broadly similar estimates.

Discussion

In the ALDS project, ‘not applicable’ item responses occur when patients have never had the opportunity to attempt to perform the activity described. This means that it is not possible to assess whether a respondent would be able to perform an activity if they had had an opportunity to do so. Hence, there is no theoretical evidence to support the use of the cold deck imputation procedure described in this article, even though comparable methods are used in some, broadly similar, questionnaires such as the Sickness Impact Profile[2].

The procedures for dealing with missing item responses, which use hot deck imputation or treat the missing responses as if these items had never been offered to those individual patients and are described in this article, could both be useful in the calibration phase of an item bank based on item response theory. The latter method can be implemented if marginal maximum likelihood or some Bayesian estimation methods are applied to avoid any bias caused by the imputation method. The hot deck imputation procedure may be valuable in situations where a complete data matrix is required. However, it should be noted that there are three reasons that the hot deck imputation procedure performs so well for the data in this paper. Firstly, the hot deck imputation procedure closely resembles the IRT model used. Secondly, the model fits the data fairly well. Finally, 32 items have been used. It is highly likely that a poor outcome for the hot deck imputation procedure would have resulted if these conditions had not pertained.

However, it should be noted that it may be impractical to repeat exploratory analyses a number of times, reducing the attractiveness of true multiple hot deck imputation, although results obtained using a single run of a hot deck imputation procedure should be treated with care. Finally, if the aim of a study is to make inferences on the functional status of patients, the procedure, which takes account of the 'tendency to respond to items' may be a valuable tool. However, in a calibration study to estimate item difficulty parameters this model does not provide any more useful information than when hot deck imputation is implemented or the missing responses were treated as if these items had never been offered to those individual patients.

There were almost no true missing item responses in the data described in this paper. The nurse interviewers were instructed to ensure that they had a response on each item and the response forms were machine readable. These procedures illuminated two important causes of missing data. The 'not applicable' option was only selected after the nurse-interviewer had made extensive inquiries into the experiences of the respondent. Hence, it seems reasonable to assume that the 'not applicable' category was used for the reason described. However, qualitative research on the reasons why respondents used this category would be needed to be sure about this. Given the relatively low level of responses in the category 'not applicable', the authors feel unable to make recommendations about the use of these procedures in data sets with much higher proportions of missing data. All four methods are relatively practical and can be implemented fairly easily. However, the hot and cold deck imputation methods are more suitable if analysis using software requiring a complete data matrix is to be carried out.

The ALDS item bank is currently under development. This means that the dimensionality and measurement properties of the item bank are still being investigated, although preliminary results suggest that a selection of items reflect a single latent trait[8], although there is a large degree of differential item functioning between male and female and between younger and older respondents[71]. It has been decided that items for which more than 10% of responses are in the category 'not applicable' are not suitable for inclusion in the item bank[8]. Hot deck imputation and the procedure treating the items as if they had never been presented to the respondents have been implemented in different types of analysis of the ALDS data.

Conclusions

This article has examined four strategies to deal with responses in a ‘not applicable’ category in the context of missing data when item response theory is used to analyse the data resulting from multi-item questionnaires. These were cold and hot deck imputation, treating the missing responses as if these items had never been offered to those individual patients and using a model which takes account of the ‘tendency to respond to items’. The four procedures were implemented on data from the AMC Linear Disability Score project. This project aims to develop an item bank to measure the functional status of chronically ill patients. In the first part of this study, estimates of the item parameters were obtained and compared using a numerical and a graphical method. The results show that the point and interval estimates obtained are very similar when the procedures based on hot deck imputation, treating the missing responses as if these items had never been offered to those individual patients and using a model which takes account of the ‘tendency to respond to items’ are used. The estimates obtained following the cold deck imputation procedure were substantially different to the estimates obtained using the other strategies.

In the second part of the study, the effects of the type of procedure on estimates of the functional status of patients was examined. It appears that cold deck imputation leads to significantly different estimates of the mean functional status in a group of patients than either hot deck imputation or treating the missing responses as if these items had never been offered. Differences between estimates obtained using the latter two methods were not significant. These results confirm that, in clinical studies, it is necessary to consider the method for dealing with responses in a ‘not applicable’ category in the context of the data.

Appendix

Hot deck imputation

In the hot deck imputation procedure implemented in this paper, the functional status of patient k is estimated by t_k,

$$t_k = \frac{m_{1k}}{m_{0k} + m_{1k}} \tag{3.2}$$

where m_{1k} and m_{0k} are the number of questions patient k responded to in the categories 'can' and 'cannot', respectively. Using the data from patients that had responded to item i, the probability, r_{ik} that patient k responded in category 'can' was modelled using

$$r_{ik} = \frac{\exp(b_{0i} + b_{1i}t_k)}{1 + \exp(b_{0i} + b_{1i}t_k)} \tag{3.3}$$

where the parameters b_{0i} and b_{1i} describe the relationship between the functional status, estimated by t_k, and the probability of responding in category 'can' of item i. In turn, if patient l, $l \in (1, 2, \ldots, K)$, did not respond to item i, the values of $\hat{b}_{0i}$, $\hat{b}_{1i}$ and t_l were used in r_{ik} to obtain an estimate of r_{il}. This probability is used to obtain an observation on a Binomial distribution, $B(1, \hat{r}_{il})$, which is imputed to replace the missing observation on item i for patient l.

In this paper, this procedure was implemented five times, resulting in five 'complete' data sets. The mean of the five estimates of β_i was taken to obtain $\bar{\beta}_i$. The standard error of $\bar{\beta}_i$ is defined as

$$\text{s.e.}\left(\bar{\beta}_i\right) = \frac{1}{\sqrt{5}} \sqrt{\sum_{j=1}^{5} \left(\beta_{ij} - \bar{\beta}_i\right)^2 + \text{s.e.}\left(\beta_{ij}\right)^2} \tag{3.4}$$

where j denotes the run of the hot deck imputation procedure, β_{ij} the estimate of β obtained for item i in run j of the imputation procedure and s.e.(β_{ij}) the standard error of β_{ij} obtained directly from the likelihood in the estimation process[73].

Treating the missing responses as if those items were never offered to the individual patients

In order to examine the effect of treating responses to individual items in the category 'not applicable' as if those items were never offered to the individual patients, the item parameters, β_i, will be estimated using a marginal maximum likelihood estimation procedure[46]. The likelihood, L, of a particular response pattern for the one parameter logistic IRT model can be written

$$L = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(p_{ik}^{I_{ik}} \left(1 - p_{ik}\right)^{1-I_{ik}} \right)^{J_{ik}} \tag{3.5}$$

where p_{ik} is as defined in the section on statistical analysis. In addition, I_{ik} is an indicator variable taking the value 1 if patient k responds to item i in the category 'can carry out', the value 0 if patient k responds to item i in the category 'cannot carry out' and the value c if if patient k responds to item i in the category 'not applicable'. Furthermore, J_{ik} is an indicator variable taking the value 0 if patient k responds to item i in the category 'not applicable' and the value 1 otherwise. In order to estimate β_i and θ_k a number of assumptions have to be made. Firstly, the item parameters have to be identified in relation to the latent trait. In this article, the mean of the distribution of θ, μ_θ, will be assumed to be 0. An increase in the number of subjects from k to $k+1$ results in a corresponding increase in the number of parameters to be estimated, meaning that parameter estimates may not be consistent. It is common to assume that the values θ_k are observations on a particular, often Normal, distribution. This results in marginal maximum likelihood estimates of β_i[46].

The root mean squared difference

The root mean squared difference (RMSD) is defined as

$$\text{RMSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_{i,a} - \hat{\beta}_{i,b})^2} \tag{3.6}$$

where $\hat{\beta}_{i,a}$ and $\hat{\beta}_{i,b}$ are estimates of β_i, $i = 1, 2, \ldots, n$, obtained under two different procedures for dealing with item responses in the category 'not applicable'.

Chapter 4

Modelling non-ignorable missing data mechanisms with item response theory models

This chapter is adapted from the following article:

> Holman R, Glas CAW. Modelling non-ignorable missing data mechanisms with item response theory models. Accepted for publication in *British Journal of Mathematical and Statistical Psychology.*

The *British Journal of Mathematical and Statistical Psychology* is available from the British Psychological Society's website: `http://www.bps.org.uk`.

Background

Whenever data are collected, however carefully, the possibility, origin and treatment of missing responses should be considered. Even if care is taken to ensure that all appropriate respondents are contacted and provide some data, responses on individual variables may be missing, uncodeable or in a category such as 'don't know' or 'not applicable'. If a data set contains missing observations, then the mechanism causing the incompleteness can be characterized according to its degree of randomness[68].

Let $\tilde{d}$ be a realization of some missing data indicator, and let $x_{(0)}$ and $x_{(1)}$ be the unobserved and observed data, respectively[68]. If the probability, $g_{\xi}(\tilde{d}|x_{(0)}, x_{(1)})$, of the missing data pattern $\tilde{d}$ depends on neither $x_{(0)}$ nor $x_{(1)}$, that is if $g_{\xi}(\tilde{d}|x_{(0)}, x_{(1)}) = 1$, then the data are both 'missing at random' (MAR) and 'observed at random' (OAR). Such data can also be described as 'missing completely at random' (MCAR). If $g_{\xi}(\tilde{d}|x_{(0)}, x_{(1)})$ does not depend on the unobserved data $x_{(0)}$ and the parameter of the missing data process, ξ, is distinct from the parameter θ of the distribution of $x_{(0)}$ and $x_{(1)}$, then the data are MAR. If the data are MCAR or MAR, the missing data mechanism is ignorable for likelihood based inferences. This means that it is not necessary to incorporate the incompleteness mechanism into models for the observed data process. In addition, if the data are MCAR, the missing data mechanism is ignorable for sample based inferences. Data entry errors, lost pages of responses and respondents following instructions incorrectly, on which items they should respond to, are examples of mechanisms leading to MCAR data. On the other hand, in data sets where the probability of the missing data pattern $\tilde{d}$ depends on θ, the missing data mechanism is not ignorable[70]. In such cases, the missing data mechanism must be modelled alongside the relationships of direct interest. These mechanisms can also be described in Bayesian terms. If the posterior distribution of x_0 does not include a specification of the response mechanism, then the mechanism is ignorable[69].

Missing data in data sets potentially suitable for analyzing using IRT techniques can be split into four types. The first type consists of the missing observations resulting from a priori fixed incomplete test and calibration designs. Since the design is a priori fixed, it is inherently independent of $x_{(0)}$ or $x_{(1)}$, and the data are MCAR. The second type consists of a class of response-contingent designs, such as two stage and multistage testing[32] and computerized adaptive testing. Here the choice of the items administered is completely governed by the responses actually observed, and independent of the unobserved responses. As a consequence, the data collected in these designs are MAR. These designs have been discussed extensively[75, 81].

The third and fourth types of missing data result from unscalable responses such as 'don't know' or 'not applicable' [82]. The third type concerns situations where the scalability of the response does not depend on the latent variable to be measured. Thus the data are MAR and may also be OAR. Procedures for analyzing data subject to this kind of missing mechanism were proposed[82], who examined the imputation of partially correct item scores, and [45], who proposed treating omitted responses as another response category. However, it has been shown that when marginal maximum likelihood estimation methods are used on data of this type, omitted responses can be ignored in the analysis [83]. The fourth type of missing data are similar to the third type but result from a non-ignorable missing mechanism. This type of data may be produced when low-ability respondents fail to produce a response, as a result of discomfort or embarrassment, or simply because they have skipped items. Another example are missing responses due to time constraints. [84] show that ignoring this kind of missing data process leads to bias in parameter estimates. Therefore, the mechanism causing the incompleteness has to be included in the analysis of this type of data. [74] suggested that whether a student gave a scalable response to a particular item depended on both the ability of the examinee and a latent trait representing 'temperament'. He went on to consider ways of incorporating this information into a model. In this article, these suggestions will be elaborated and their usefulness tested in a number of simulation studies.

This article will present four general IRT models for taking non-ignorable missing data mechanisms into account. These models are reformulations of the models proposed by [85, 86, 87, 88, 89, 90, 91]. In the formulation presented here, the models support a simple framework to assess, explicitly, the extent to which the missing data are non-ignorable. In addition, the relationship between the present models and the model by [85] will be outlined in more detail in an appendix on the identification of the model.

The approach presented here will be applied to the estimation of parameters in IRT models. Simulation studies will be carried out to compare the mean squared error of the estimates obtained by ignoring the missing data process and explicitly modelling the missing data process. Finally, the feasibility of the method will be demonstrated using data from the Amsterdam Linear Disability Score project[8].

A general IRT model for missing data processes

Consider a two-dimensional persons by items data matrix X with entries x_{ik}, $i = 1, \dots, N$, and $k = 1, \dots, K$. If a combination of i and k has been observed, x_{ik} is equal to the observation, otherwise it is equal to some arbitrary constant. At this point, no assumptions are made about the range of the values of x_{ik}. We define a design matrix D of the same dimensions as X with elements

$$d_{ik} = \begin{cases} 0 & \text{if } x_{ik} \text{ was not observed,} \\ 1 & \text{if } x_{ik} \text{ was observed.} \end{cases} \tag{4.1}$$

Our objective is to make inferences about parameters partitioned into two sets: structural parameters $\beta_h, h = 1, \dots, H$ and incidental parameters θ_i, $i = 1, \dots, N$. The latter set of parameters is called incidental because their number increases in proportion to the number of observations, which, in general, leads to inconsistency [92]. It is assumed that H is a constant that does not depend on N. It has been shown that, under fairly reasonable regularity conditions, consistent estimators of structural parameters can be obtained by assuming a common distribution for the incidental parameters and integrating them out of the likelihood[93]. Inferences about the incidental parameters are then made given the estimated values of the structural parameters.

A general model for the missing data process is given by a multidimensional

IRT model [94, 95, 96, 97]. The probability of an observation is given by

$$p(d_{ik} = 1|\xi_i, \delta_k) = \frac{\exp(\sum_q \delta_{kq}\xi_{iq} - \delta_{k0})}{1 + \exp(\sum_q \delta_{kq}\xi_{iq} - \delta_{k0})}. \tag{4.2}$$

This model has the Rasch model [98] and the two-parameter logistic model (2PLM) [99] as special cases. In many instances, the amount of missing data will be small, which suggests using a model with few parameters such as the Rasch model.

If the probability of a particular observation, say $p(x_{ik}|d_{ik} = 1, \theta_i, \beta)$, does not depend on ξ, and θ and ξ are independent, then the missing data are ignorable. In this situation and assuming local independence, a straightforward model for the data and the missingness processes is

$$G_1 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \beta)p(d_{ik}|\xi_i, \delta_k)g_1(\xi_i)g_2(\theta_i), \tag{4.3}$$

where $p(x_{ik}|d_{ik}, \theta_i, \beta)$ and $p(d_{ik}|\xi_i, \delta_k)$ are the density of the outcome variable and the design variable, respectively, and $g_1(\xi_i)$ and $g_2(\theta_i)$ are the density of ξ_i and θ_i, respectively. To model non-ignorable missing data, it will be assumed that θ and ξ have a common distribution $g(\xi_i, \theta_i|\phi)$ that is indexed by a parameter ϕ, that is,

$$G_2 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \beta)p(d_{ik}|\xi_i, \delta_k)g(\xi_i, \theta_i|\phi). \tag{4.4}$$

Note that $p(x_{ik}|d_{ik}, \theta_i, \beta)$ does not depend on ξ_i. The obvious alternative is that the observations may depend on ξ, which leads to

$$G_3 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \beta)p(d_{ik}|\xi_i, \theta_i, \delta_k)g(\xi_i, \theta_i|\phi). \tag{4.5}$$

or that both the observed data and the missing data indicators depend on both θ and ξ, that is,

$$G_4 = \prod_{i,k} p(x_{ik}|d_{ik}, \theta_i, \xi_i, \beta)p(d_{ik}|\xi_i, \delta_k)g(\xi_i, \theta_i|\phi). \tag{4.6}$$

The models given by (4.4), (4.5) and (4.6) are not necessarily different, there may be transformations of θ and ξ that transform one model into the other. However, the model given by (4.4) is conceptually the simplest: the distributions of the observed data x_{ik} and the missing data indicator d_{ik} are parameterized by distinct sets of parameters, which have a common distribution with parameters ϕ. In the sequel, it will be shown that the parameters ϕ can be used to index the extent to which ignorability holds. The relations between the formulations (4.4), (4.5) and (4.6) will be outlined further below.

Using these models to analyze both the data and missingness processes together can provide extra information on the mechanisms underlying a particular data set, even if the missing data process is ignorable. They may, for instance, give indications about the quality of the definitions of the variables collected. The models could be used in conjunction with a wide range of statistical models. However, this article will concentrate on the use of these models in conjunction with item response theory, developing the idea that the probability of a missing response depends on a separate personality trait as well as ability, proposed by [74].

The combination with IRT models for observed data

In the previous section, the missing data process was modelled with an IRT model, but the model for the observations was left unspecified. In this section, it will be set out in detail how the model can be combined with an IRT model for the observed data. In that case, both θ_i and ξ_i are latent variables. The elements of matrix D will be necessarily dichotomous, whereas those in the matrix X may be either dichotomous or polytomous. Depending on the model chosen for the combined process of θ and ξ, the elements of X and D may reflect either or both latent traits.

In the examples given below, both items with dichotomous and polytomous responses will be considered. These responses will be analyzed with the generalized partial credit model (GPCM) [100]. In the GPCM the probability, $p(x_{ik} = j|\theta_i, \alpha_k, \beta_k)$, that respondent i responds to item k in category j, $(j = 1, \ldots, m_k)$, is denoted by

$$p(x_{ik} = j|\theta_i, \alpha_k, \beta_k) = \frac{\exp(j\alpha_k\theta_i - \beta_{kj})}{\sum_{j=0}^{m_k} \exp(j\alpha_k\theta_i - \beta_{kj})}, \tag{4.7}$$

where β_k is a vector of item parameters $(\beta_{k0}, \beta_{k1}, \ldots, \beta_{kj}, \ldots, \beta_{km_k})$, with $\beta_{k0} = 0$ to ensure that the estimates of β_k are unique. The two-parameter logistic model [39] is the special case for $m_k = 1$. further, the model given by (4.7) specializes to the partial credit model (PCM) [101, 102] upon setting $\alpha_k = 1$, and specializes further to the Rasch model for dichotomous items by setting $m_k = 1$. Extending (4.7) to include a second the latent trait ξ gives

$$p(x_{ik} = j|\theta_i, \xi_i, \beta_k) = \frac{\exp(j(\alpha_{k1}\theta_i + \alpha_{k2}\xi_i) - \beta_{kj})}{\sum_{j=0}^{m_k} \exp(j(\alpha_{k1}\theta_i + \alpha_{k2}\xi_i) - \beta_{kj})}. \qquad (4.8)$$

The examples given below examine the bias in the estimates of the item parameters when ignoring the missing data mechanism and the possibility of avoiding this bias using the models given in equations (4.3) to (4.6).

It will be assumed that θ and ξ have a multivariate normal distribution with density $g(\xi, \theta|\mu, \mathbf{\Sigma})$. The mean μ will be set equal to zero to identify the model and the covariance matrix $\mathbf{\Sigma}$ is an estimand. To obtain consistent estimates, a likelihood that is marginalized with respect to θ and ξ is maximized, that is, we maximize

$$\log L(\beta, \delta, \mathbf{\Sigma}) = \sum_{i,k} \log \int, \ldots, \int p(x_{ik}|d_{ik}, \theta, \beta_k) p(d_{ik}|\xi, \delta_k) g(\xi, \theta|\mu, \mathbf{\Sigma}) d\theta d\xi, \quad (4.9)$$

with respect to the item parameters β and δ and the covariance matrix $\mathbf{\Sigma}$. In the framework of IRT, this method is known as maximum marginal likelihood [83]. The adaptation of the marginal likelihood to (4.3), (4.5) and (4.6) is straightforward. Procedures for maximizing this likelihood were developed by [76] [103] and [104]. For the multidimensional Rasch and partial credit model, parameter estimates can be computed using the computer program ConQuest [73]; more complicated models, such as models involving (4.2) can be estimated using Testfact [105].

The model considered in (4.9) is a special case of the model by [85]. These authors restrict the covariance matrix to an identity matrix, but in Appendix A it is shown that this does not imply a restriction: the model can always be reparameterized such that the covariance matrix becomes a free estimand. On the other hand, [85] allow the observed data and the missing data indicators to depend on θ and ξ concurrently, that is, they pursue model (4.6). As a result of the identity covariance matrix and the lack of separation between the parameters of the model for the observations and the model for the missing data process, the extent to which ignorability holds becomes hard to assess. In the present formulation, the correlation between θ and ξ explicitly indexes ignorability, and this will be used in the presented simulation study to assess the bias imposed when ignorability is used unjustified.

A simulation study

A simulation study was carried out to assess the effect of a missing data process as described in equation (4.4) on estimates of item parameters. In this study, data were simulated using the OAR missing process given by model G_2 and analyzed using the MCAR process described in model G_1 and also using model G_2. For sample sizes of $n = 500, 1000, 2000$, latent trait values (θ_i, ξ_i) were drawn from a bivariate Normal distribution $g(\xi_i, \theta_i|\mathbf{\Sigma})$ with means 0, variances 1 and correlation ρ, where $\rho = 0.0, 0.1, \ldots, 0.9$. These sample sizes were chosen because they reflect the numbers of respondents, which could be reasonably included in a medical study similar to the AMC Linear Disability Score project used as an example in this paper. Tests consisted of $K = 10$, $K = 20$ and $K = 30$ dichotomously scored items. The values d_{ik} were drawn from $p(d_{ik}|\xi_i, \delta_k)$ and values x_{ik} were drawn from $p(x_{ik}|d_{ik}, \theta_i, \beta_k)$. The data were used to compute $\hat{\beta}_k$, from model G_1, and $\hat{\beta}'_k$ and $\hat{\delta}_k$ from model G_2.

The values of $\hat{\beta}_i$, $\hat{\beta}'_i$ and $\hat{\delta}_i$ were compared with the values of the parameters used to generate the data using the mean squared error (MSE) and mean absolute error (MAE). The MSE for δ_k is defined as

$$\mathrm{MSE}(\delta_k) = \sum_{r=1}^{R} \sum_{i=1}^{n} \left(\hat{\delta}_i - \delta_k\right)^2 \tag{4.10}$$

where $r = 1, 2 \ldots, R$ denote the replications of the simulation process and $\hat{\delta}_i$ the estimate of δ_k. The MAE is defined as

$$\mathrm{MAE}(\delta_k) = \sum_{r=1}^{R} \sum_{i=1}^{n} \left|\hat{\delta}_i - \delta_k\right| \tag{4.11}$$

Ten replications were made for each combination of $K = 10, 20, 30$, $n = 500, n = 1000$ and $n = 2000$ and $\rho = 0.1, 0.2, 0.3, \ldots, 0.9$. In the first set of simulations, the item parameters were $\delta_k = \beta_k = 0$ for all k. As a result, about 50% of the data were missing. The values of the MAE and the MSE for δ (solid lines), β' (dashed lines) and β (dotted lines) are given in Figures 1 and 2, respectively.

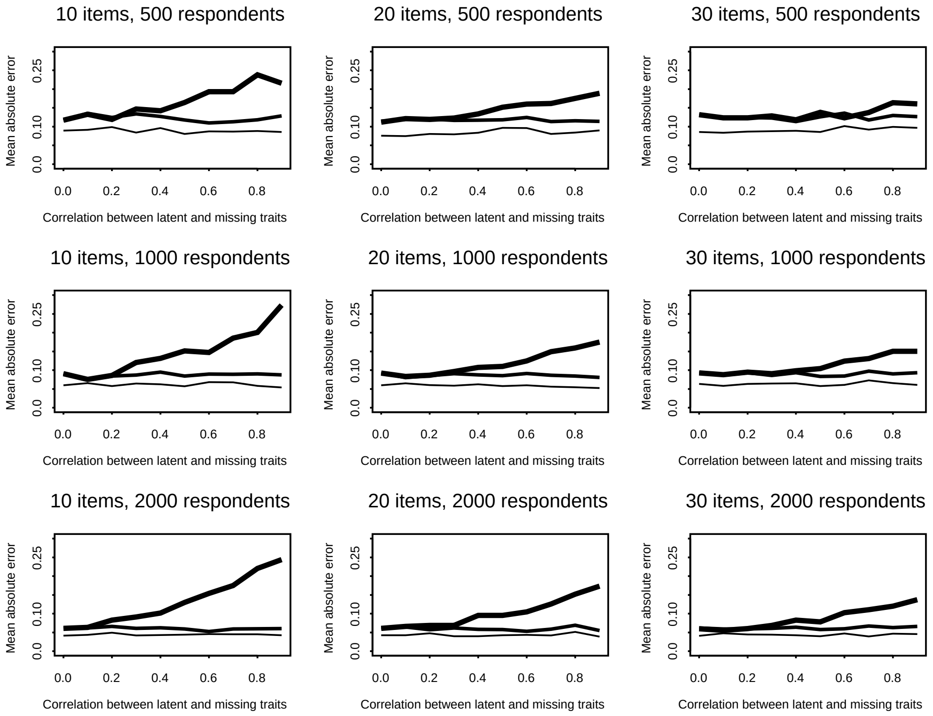

Figure 4.1: Comparing the values of $\hat{\delta}$ (narrowest lines), $\hat{\beta}'$ (medium lines) and $\hat{\beta}$ (widest lines) with the values used to generate the data using the mean absolute error (MAE) for $\beta_k = \delta_k = 0$

It is apparent that, for $\rho = 0.0$, the MAE and MSE of the estimates of β_k and β'_k are very close. This indicates that the estimates of β_k and β'_k are equal within random fluctuations and confirms that the data mechanism is ignorable, and hence MCAR for $\rho = 0.0$. This is in spite of ML estimates of IRT item parameters being slightly biased [106], confirming that the bias is negligible with respect to the standard errors. It can be seen that the MAE and MSE of the estimates of β_k, incorrectly assuming that the data are MCAR, increase with ρ, whilst the estimates of β'_k, remain stable apart from random fluctuations. This effect was most noticeable for tests with 10 items and least apparent for tests with 30 items. Partitioning the MSE into squared bias and estimation variance showed that the inflation of MSE

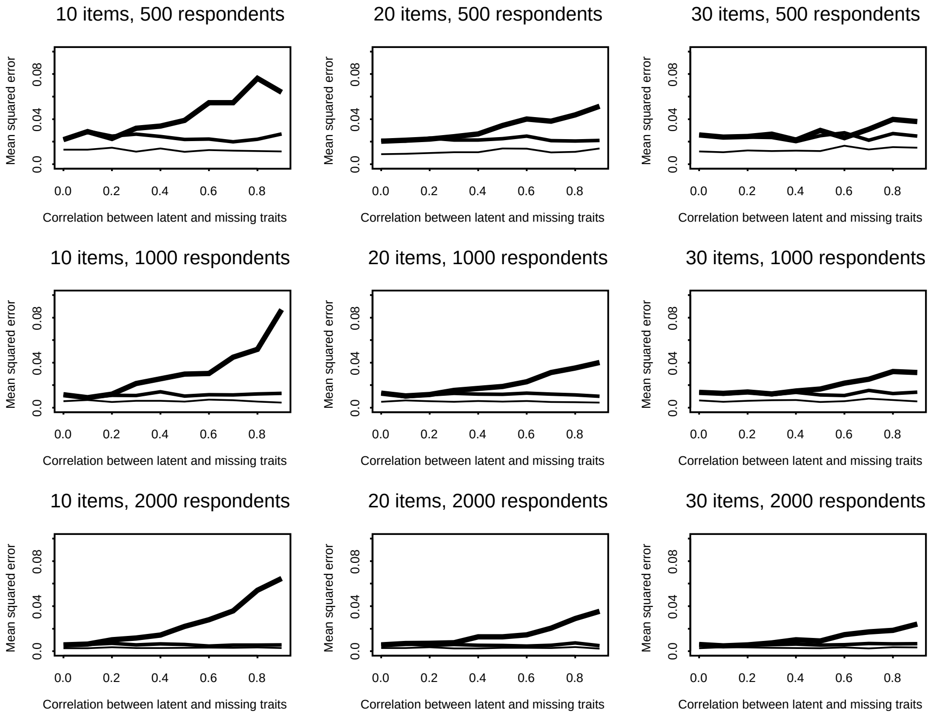

Figure 4.2: Comparing the values of $\hat{\delta}$ (narrowest lines), $\hat{\beta}'$ (medium lines) and $\hat{\beta}$ (widest lines) with the values used to generate the data using the mean squared error (MSE) for $\beta_k = \delta_k = 0$

was completely due to an inflation of bias.

In the second set of simulations the MSE and MAE were calculated over 10 replications each with $n = 1000$ respondents, $K = 30$ items, and $\delta_k = 0$ for all k for $\rho = 0.0, 0.1, \ldots, 0.9$. Three sets of parameter values were chosen: for the first set $\beta_k = 1$ for all k, for the second $\beta_k = 2$ for all k and for the third $\beta_1 = \ldots = \beta_6 = 2$, $\beta_7 = \ldots = \beta_{12} = 1$, $\beta_{13} = \ldots = \beta_{18} = 0$, $\beta_{19} = \ldots = \beta_{24} = -1$ and $\beta_{25} = \ldots = \beta_{30} = -2$. The results of these simulations are given in Figure 3. The relationship between ρ and the MAE and MSE of β_k are similar to those displayed in Figures 1 and 2, indicating that the bias in the estimation of β_k caused by the non-ignorable missing mechanism is similar across different sets of item parameters.

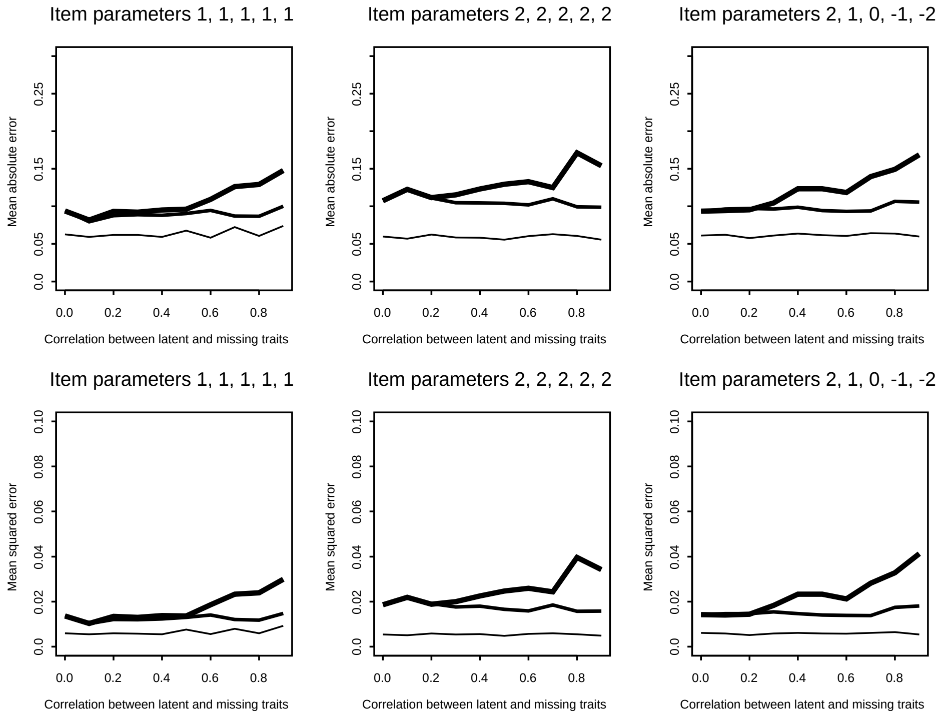

Figure 4.3: Comparing the values of $\hat{\delta}$ (narrowest lines), $\hat{\beta}'$ (medium lines) and $\hat{\beta}$ (widest lines) with the values used to generate the data using the mean absolute error (MAE) and mean squared error (MSE) for the values of β_k given

The AMC Linear Disability Score project

The AMC Linear Disability Score (ALDS) project aims to develop an item bank to measure disability to perform activities of daily life [7, 8]. The ALDS item bank consists of about 200 items, each describing an activity that a healthy adult might perform in the course of daily life. They range from very easy (sitting up in bed) to difficult (jogging for 15 minutes). When patients are presented with the items they are asked to respond in one of three ordered response categories: '*I cannot perform the activity*'; '*I can perform the activity, but find it difficult*' or '*I can perform the activity*'. If patients have never performed an activity, then they are instructed to respond in a further category: '*not applicable*'. For instance patients who have never held a driving licence are instructed to respond to the item 'driving a car' in this way. Responses in this last category can be seen as missing, since they are not directly scalable. The data were collected by specially trained nurses.

The data used in this article form two distinct parts of the sample being used to calibrate the item bank. The parts result from offering Test 1 and Test 2, each consisting of 32 items, to samples of 171 and 179 patients, respectively. The tests had no items in common. In Test 1, 27 items had missing responses, the number of missing responses per item ranged from 7 to 56, with a mean of 16.1. In Test 2, 25 items had missing responses, the number ranging from 1 to 68 with a of mean 10.4.

In order to obtain an impression of the missing data pattern, the relation between the respondents' score levels and the amount of missing responses was examined, by computing the correlation between $\text{logit}(\sum_k d_{ik}x_{ik}/\sum_k d_{ik}m_k)$ and $\text{logit}(\sum_k d_{ik}/K)$. Note that $m_k = 2$ for all k and $K = 32$. The correlations were -0.04 for Test 1 and -0.12 for Test 2. This indicates that the amount of missing responses went up with the proficiency level. A possible explanation is that patients with a higher proficiency level tended to boost their rating by failing to respond, while the patients of low proficiency were less inclined or motivated to impress the nurses.

The data were modelled using G_1, G_2, G_3 and G_4, described above, and a model where the missing data process and the observed data loaded on the same latent trait, that is, a model where the correlation between ξ_i and θ_i equals one. This model will be labelled G_0. These models were elaborated further in two versions. In the first version, the Rasch model was used to model d_{ik} and the PCM to model the observed responses x_{ik}. In this case the parameter α_i in (4.7) and (4.8) was set equal to one. In the second version, α_i was estimated, meaning that the 2PLM and the GPCM were used to model d_{ik} and x_{ik}, respectively.

Table 4.1: Results of fitting models G_0, G_1, G_2, G_3 and G_4 to the ALDS data. The number of parameters is denoted by 'NP'.

Data	Model		Type of missing process	Deviance	NP	ρ	p-value of χ^2 test against G_1
Test 1	PCM	G_0		11665.7	92	1.000	0.000
		G_1	MCAR	11509.9	93	0.000	-
		G_2	OAR	11508.7	94	-0.124	0.273
		G_3	OAR	11508.7	94	-0.896	0.273
		G_4	non-MAR, non-OAR	11495.6	94	-0.547	0.000
Test 1	GPCM	G_0		11032.1	150	1.000	-
		G_1	MCAR	10950.3	150	0.000	-
		G_2	OAR	10923.5	151	-0.104	0.000
		G_3	OAR	10808.0	177	-0.678	0.000
		G_4	non-MAR, non-OAR	10801.8	182	-0.543	0.000
Test 2	PCM	G_0		11515.1	90	1.000	0.000
		G_1	MCAR	11308.3	91	0.000	-
		G_2	OAR	11293.5	92	-0.424	0.000
		G_3	OAR	11293.5	92	-0.758	0.000
		G_4	non-MAR, non-OAR	11292.7	92	-0.909	0.000
Test 2	GPCM	G_0		11295.2	146	1.000	-
		G_1	MCAR	11279.5	146	0.000	-
		G_2	OAR	11223.5	147	-0.445	0.000
		G_3	OAR	11168.7	171	-0.778	0.000
		G_4	non-MAR, non-OAR	11148.5	178	-0.834	0.000

An overview of the results of the analyses is given in Table 4.1. The column labelled 'Deviance' gives the log-likelihood of the model multiplied by minus two. The last column gives the estimates of the correlation between the dimensions for the models G_2, G_3, and G_4, and the fixed values for the other two models. Models can be compared by subtracting the deviance of a special model from the deviance of a more general model. The resulting test statistic has an asymptotic χ^2 distribution with degrees of freedom equal to the difference of the numbers of parameters estimated. So in the case of the PCM, G_1 is a significant improvement over G_0, both for Test 1 and 2. In the case of the GPCM, the hypothesis cannot be formally tested, because the numbers of parameters in G_0 and G_1 are equal. However, the deviance of G_1 is also substantially lower. In all cases G_2, G_3 and G_4 are a significant improvement on G_1, except for the combination of Test 1 with the PCM, where only G_4 is significantly better. Finally, it can be seen that the GPCM has a better overall fit than the PCM, but the estimates of ρ in both models are comparable. Observed correlations are usually attenuated by unreliability, which in turn, is related to the number of items in the test. Comparing the estimates of the correlations in G_2 with the correlation between $\text{logit}(\sum_k d_{ik}x_{ik}/\sum_k d_{ik}m_k)$ and $\text{logit}(\sum_k d_{ik}/K)$ reported above, it can be seen that this is also true in the present case. In Test 1 the observed correlation was -0.04, while the latent correlation was -0.124 for the PCM and -0.104 for the GPCM. In Test 2 these figures were -0.12, -0.424 and -0.445, respectively. Hence, the latent correlations make the association between the proficiency level and the missing data process more manifest.

The final remark concerns the fit of the model. Methods for the analysis of model fit of multidimensional IRT models are not readily available. Therefore, two components making up the complete model, say $p(x|d,\theta,\beta)$ and $p(d|\xi,\delta)$, were evaluated separately. The analyses were made using the computer program OPLM [53]. The program computes conditional maximum likelihood estimates of the item parameters and computes a test statistic R_{1c} [107, **?**]. The test evaluates whether the item response probabilities implied by the IRT model used, properly describe the observed response proportions. For this test, the score range is partitioned into a number of categories and the observed and expected response frequencies of item responses are combined into an asymptotically χ^2 distributed test statistic. If the test rejects the IRT model (say the Rasch model or the PCM), a more general model (say the 2PLM or the GPCM) is needed. The results are shown in Table 4.2, following the labels '1PLM' and 'PCM'.

Table 4.2: Goodness of fit of the PCM and GPCM to the ALDS data for the latent and missingness traits in model G_1

Data	Model		Trait	R_{1c}	*df*	*p*-value
Test 1	PCM	$p(x\|d,\theta,\beta)$	latent	317.8	189	0.0000
	1PLM	$p(d\|\xi,\delta)$	missingness	112.2	78	0.0063
	GPCM	$p(x\|d,\theta,\beta)$	latent	248.9	189	0.0021
	2PLM	$p(d\|\xi,\delta)$	missingness	60.6	69	0.7694
Test 2	PCM	$p(x\|d,\theta,\beta)$	latent	399.4	189	0.0000
	1PLM	$p(d\|\xi,\delta)$	missingness	63.4	48	0.0657
	GPCM	$p(x\|d,\theta,\beta)$	latent	392.6	189	0.0000
	2PLM	$p(d\|\xi,\delta)$	missingness	20.4	22	0.5782

Overall, model fit is far from perfect. The 2PLM and GPCM were then used as alternatives. Using the OPLM program the item difficulties β_{kj} were re-estimated using conditional maximum likelihood and the R_{1c} test statistic was computed as above[53, 107]. The estimation and testing procedure used in OPLM entails rounding the item discrimination parameters α_k to the nearest integer [108]. Since the rounding produces an approximation to the 2PLM and the GPCM, the procedure results in a conservative test. The results are shown in Table 4.2 after the label 'GPCM'. It appears that the model fit for the missing data process is now quite acceptable.

These results suggest that the missing process in the ALDS data cannot, in general, be satisfactorily modelled using models G_0 or G_1, although models G_2 and G_3 seem to be reasonable in most cases. This suggests that the data are OAR but not MAR, meaning that the missing process is ignorable for inferences on β but not for inferences on θ. In addition, the results in Table 4.2, suggest that the fit of the 2PLM to the missingness process is satisfactory, whereas the 1PL is not. In addition, neither the PCM or the GPCM seem to fit the data sufficiently. This indicates that items need to be removed from these booklets, before inferences can be made on the values of θ.

Discussion and Conclusion

A variety of methods for dealing with ignorable and non-ignorable missing data in practical situations have been proposed[66]. These range from imputation methods to algorithms which permit parameters to be estimated, whilst ignoring missing observations. The development of models, in which the primary data and missing processes are considered jointly[67], is particularly interesting. These models can be useful in situations where it is thought that the mechanism causing the missing data is not ignorable. A model based procedure for handling non-ignorable missing data using IRT models is presented that is formulated is such a way that the extent to which ignorability is violated can be easily assessed. Four general IRT models for missing data mechanisms are proposed. As an example, these models are worked out in detail in conjunction with item response data modelled by the partial credit and generalized partial credit models. In a number of simulation studies it was shown that ignoring the missing data process results in considerable bias in the estimates of the item parameters. This bias increases as a function of the correlation between the proficiency to be measured and the latent variable governing the missing data process. Further, it was shown that this bias can be reduced using the models presented above. The feasibility of the procedure was demonstrated using data from a calibration study of a medical disability scale. The correlation between the proficiency and the latent variable of the missing data processes was significant and using the missing data models significantly increased model fit.

This approach can be generalized by the inclusion of covariates in the missing data model. IRT models with manifest covariates were proposed by [104, 109, 110]. Finally, test statistics are needed to evaluate the appropriateness of the models presented above. Therefore, this provides another incentive for the development of evaluation methods for fit to multidimensional IRT models.

Appendix: Identification and relations between various formulations of the model

In this appendix, a multidimensional IRT model will be considered for observations y_{ik}, where $(i = 1, \ldots, N,, \ k = 1, \ldots, K)$, y_{ik} assumes integer values $(0, 1, \ldots, M)$ and

$$p(y_{ik} = j) = \frac{\exp(j \sum_{q=1}^{Q} \alpha_{kq}\theta_{iq} - \beta_{kj})}{1 + \sum_{h=1}^{M} \exp(h \sum_{q=1}^{Q} \alpha_{kq}\theta_{iq} - \beta_{kh})}. \tag{4.12}$$

In addition, the latent variables θ_i, $\theta_i^{'} = (\theta i1, \ldots, \theta_{iq}, \ldots, \theta_{iQ})$, have a multivariate normal distribution with mean μ and covariance matrix Σ. [111, 112, 113] point out that this model can be viewed as a factor analysis model with Q factors, where $\theta_{i1}, \ldots, \theta_{iq}, \ldots, \theta_{iQ}$ are factor scores and $\alpha_{11}, \ldots, \alpha_{kq}, \ldots, \alpha_{KQ}$ are factor loadings.

The model can be identified in two ways:

1. using the restrictions $\mu = \mathbf{0}$, and $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$, in which case Σ is a free estimand, or,

2. using the restrictions $\mu= \mathbf{0}$, $\Sigma= \mathbf{I}$, and $\alpha_{jq}^{o} = 0$, for $j = 1, \ldots, Q - 1$ and $q = j + 1, \ldots, Q$.

Let $\mathbf{A}$ and $\mathbf{A}^o$ be the matrices of discrimination parameters for the former and latter parameterization, respectively. That is, $\mathbf{A}$ is defined as a $K \times Q$ matrix with elements α_{iq}, and $\mathbf{A}^o$ is defined analogously. For example and with $K = 5$ and $Q = 3$, the first parametrization results in a matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} \\ \alpha_{51} & \alpha_{52} & \alpha_{53} \end{bmatrix}, \tag{4.13}$$

and a free matrix $\mathbf{\Sigma}$, while the second parameterization results in a matrix A given by

$$\mathbf{A}^o = \begin{bmatrix} \alpha_{11}^o & 0 & 0 \\ \alpha_{21}^o & \alpha_{22}^o & 0 \\ \alpha_{31}^o & \alpha_{32}^o & \alpha_{33}^o \\ \alpha_{41}^o & \alpha_{42}^o & \alpha_{43}^o \\ \alpha_{51}^o & \alpha_{52}^o & \alpha_{53}^o \end{bmatrix}, \tag{4.14}$$

and a covariance matrix $\mathbf{\Sigma}$ that is equal to the identity matrix. In both cases, the number of restrictions is equal to Q^2. In the first example $\mathbf{A}$ has 9 restrictions, while in the second, $\mathbf{A}^o$ has 3 restrictions and $\mathbf{\Sigma}$ has 3 restrictions on diagonal and 3 on off-diagonal elements.

The parameters θ_i can be transformed to θ_i^o by $\theta_i^o = \mathbf{L}^{-1}\theta_i$, where $\mathbf{L}$ is the Cholesky decomposition of Σ. Because $\mathbf{L}$ is lower triangular and $\mathbf{A}\theta_i = \mathbf{AL}\theta_i^o = \mathbf{A}^o\theta_i^o$, the restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$, are transformed into restrictions $\alpha_{jq}^o = 0$, for $j = 1, \ldots, Q-1$ and $q = j+1, \ldots, Q$. On the other hand, defining the lower triangular matrix $\mathbf{F}$ as the first Q rows of $\mathbf{A}^o$ and applying $\theta_i = \mathbf{F}\theta_i^o$, results in $\Sigma = \mathbf{FF}^t$ and $\mathbf{A} = \mathbf{A}^o\mathbf{F}^{-1}$, which in turn produces restrictions $\alpha_{jq} = 1$, if $j = q$, and $\alpha_{jq} = 0$, if $j \neq q$, for $j = 1, \ldots, Q$ and $q = 1, \ldots, Q$. Hence, the two parameterizations of the model are easily interchanged.

In the application discussed above, the variables y_{ik} can either be the observations x_{ik} or missing data indicators d_{ik}. For instance, suppose that the Rasch model holds for both the observations x_{ik} and missing data indicators d_{ik}. Then, for a test with 3 items, the matrix $\mathbf{A}$ can be defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{4.15}$$

where the rows 1, 3 and 5 relate to responses and the rows 2, 4, and 6 to missing data indicators. The responses load on the first factor only, and the missing data indicators load on the second factor only. The correlation between the factors indexes the dependence between the observations and the missing data indicators, that is, the extent to which the missing data process is ignorable.

In the formulation given by [85], the covariance matrix is an identity matrix. In that case, the model in this example can be identified using

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & 0 \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \\ \alpha_{41} & \alpha_{42} \\ \alpha_{51} & \alpha_{52} \\ \alpha_{51} & \alpha_{62} \end{bmatrix}. \tag{4.16}$$

Note that both the responses and the missing data indicators load on both dimensions. The model can be transformed to a version with a free covariance matrix, but this would, in most cases, not lead to separate dimensions for the responses and missing data indicators. Hence, models with this property are a special case of the general model proposed [85].

Chapter 5

Differential item functioning with regard to gender and age

This chapter is adapted from the following article:

> Holman R, Lindeboom R, de Haan RJ. Gender and age based differential item functioning in the AMC Linear Disability Score project. *Quality of Life Newsletter* 2004; **32**; 1–4.

and is available from `http://www.mapi-research-inst.com`

Background

An important assumption in item banking is that the items included in a calibrated item bank have the same measurement characteristics for all subgroups in the population. Differential item functioning occurs when patients in two groups have different probabilities to respond to an item in a given way, even though they have the same level of health status or quality of life. If differential item functioning is ignored, inaccurate measurements are obtained and true differences between patient groups may be obscured and non-existent differences 'created'.

Differential item functioning has been identified in a wide range of instruments, designed to quantify aspects of health related quality of life. Items from the Mini-Mental State Examination have been shown to function differently for education and gender based groups in a community based sample[114]. Items from the RAND-36, based on the SF-36, had different measurement characteristics for patients with multiple sclerosis, rheumatic diseases and COPD[115], while items from the SF-12 function differently in age and education level based groups in respondents to the 2000 Medical Expenditure Panel Survey[116]. The Rivermead Mobility Index items functioned differently for age based but not for gender based groups in a sample of lower limb amputees[117]. The Bath Ankylosing Spondylitis Functional Index and Revised Leeds Disability Questionnaire both demonstrated differential item functioning with regard to age and perceived duration of illness[118]. The items from the National Health Interview Survey Disability Supplement in a sample of civilian, non-institutionalised persons requiring help or supervision with at least one activity[119] and the items form the PDQ-2 39 and Nottingham Health Profile in a sample of patients with Parkinson's disease[120] function differently with respect to age and gender based groups. The conclusions drawn in the studies discussed, indicate that it is essential to examine items, in instruments to quantify health related quality of life, to see whether they have the same measurement properties for subgroups of respondents.

The AMC Linear Disability Score project aims to construct and calibrate an item bank to measure functional status, as expressed by the ability to perform ‘activities of daily life’[7]. Each item in the item bank describes an activity of daily life essential for living independently or in an appropriate care setting. Necessarily, a large proportion of items describe domestic activities, which are traditionally, and currently, primarily performed by either men or women. This paper uses non-parametric item response theory methods[121, 122] to compare the measurement properties of the items from this item bank across gender and age based groups in the responses given by residents of nursing and care homes in and around Amsterdam, The Netherlands. The results will extend knowledge of differential item functioning in the items in this and similar item banks[16].

Respondents and items

The data described in this article come from the responses of long-term residents of sheltered housing schemes, care and nursing homes in Greater Amsterdam, The Netherlands, to sets of items presented in an interview carried out by specially trained nurses. The interviews were conducted between December 2002 and July 2003. Of the 551 respondents, 440 (80%) were female and 277 (50%) were aged 84 years or under. Furthermore, 525 (95%) of the respondents were aged 70 or older.

Each of the 160 items described an activity of daily life. Participants responded to each item in one of the categories ‘I can carry out the activity’, ‘I cannot carry out the activity’ or ‘This item is not applicable to me’. The final category was only used if the respondent had never experienced the activity. For example, respondents, who had never had a full driving licence, responded in this way to the item ‘driving a car’. Four sets of 80 items were used. The sets of items were compiled in such a way that they were equally difficult and were allocated to patients randomly. Each set was linked to two other sets using 39 or 40 common items[7]. In order to obtain the maximum amount of information on the measurement properties of the items described in this paper, four data arrays, based on the common items between pairs of sets of items, were constructed. Each item appears in a single data array, but about half of each patient’s responses appear in one data array and the other half in another data array. Full details of the methodology employed has been given elsewhere[7, 8].

Statistical analysis

Although most instruments measuring health related quality of life are analysed using sum score based methods, interest in item response theory (IRT) based methods, including item banking, is currently increasing[123]. The majority of research using IRT, concentrates on the well-known parametric models, such as the Rasch and the two-parameter logistic models. However, non-parametric IRT models have also been developed. Non-parametric models have the advantage that it is easier to differentiate between differential item functioning and item misfit. In this article, the probability of persons with a particular level of functional status responding to a given item in the category 'I can carry out this activity' is plotted against their functional status. These probabilities are then smoothed using a kernel method. Differential item functioning is quantified as the mean of the distance between the smoothed probability for the reference group, say women, and the smoothed probability for the focus group, say men, weighted by the distribution of the latent variable in the reference group[121]. This type of model can be fitted using the TestGraf software[122], although the standard errors for the estimates of differential item functioning are too small by a factor of 0.4 [personal communication - J.O. Ramsay, 2003].

Before the IRT analysis commenced, items with more than 90% or fewer than 10% of responses in the category 'cannot' were excluded from further analysis. Responses by patients in the category 'not applicable' were treated as if the item had not been presented to the patient[14] as described in Chapter 3 of this thesis. Items were regarded as demonstrating differential item functioning if the P-value of the appropriate statistic was less than 0.01.

Results

A total of 160 items were included in this study. Two items (1%) were excluded because they were only in a single data set, meaning that there were not enough responses to evaluate them. In addition, 35 items (22%) were excluded because more than 90% or less than 10% of the responses were in the category ’cannot’. This means that there was little information on the behaviour of these items and that they were unsuitable for quantifying the functional status of residents of care and nursing homes. A total of 15 items (9%) were easier for women and 18 items (11%) were easier for men. A further 4 items (3%) were easier for respondents aged 84 or under and 2 items (1%) were easier for men and easier for respondents aged 84 or under. The remaining 84 items (54%) did not display statistically significant amounts of differential item functioning.

The types of item, for which the majority have the same measurement characteristics for men and women and for respondents aged 84 or younger and respondents aged 85 or older are given in Table 5.1. Of the 27 items reflecting mobility and agility, 25 have the same characteristics for all groups and 2 are easier for men than women. All 4 items on caring for plants have the same characteristics for all groups. Of the 19 items reflecting dressing and self-care, 16 have the same characteristics for all groups and three are easier for men than for women. Two of these three items are particularly interesting: ‘showering and washing ones hair’ and ‘putting on lace-up shoes’. When considering these items, it be useful to note the large proportion of women in this age group, who have a highly coiffured hair style requiring ‘setting’ by a hairdresser, and the unpopularity of lace-up shoes in women of this age group. Furthermore, of the 16 items reflecting activities outside the home, 14 have the same characteristics for all groups and two are easier for respondents aged 84 or younger. One of these, ‘going to the pharmacy’, may be explained by the fact that pharmacies in the Amsterdam region provide a free delivery service.

Table 5.1: types of item, for which the majority have the same measurement characteristics for men and women and for respondents aged 84 or younger and respondents aged 85 or older. The letter ‘M’ denotes that the item is easier for men than for women, the letter ‘F’ that the item is easier for women that for men and the letter ‘Y’ that the item is easier for respondents aged 84 or younger.

Mobility and agility – 26 items
Walking up a flight of stairs
Standing for 10 minutes
Walking up two flights of stairs
Going for a long walk
Going down a flight of stairs
Getting out of bed into a chair
Walking for less than 15 minutes
Sitting up in bed
Sitting on the edge of the bed
Moving between two dining chairs
Moving between two easy chairs
Reaching into a high cupboard
Putting a dining chair up to table
Locking a door
Using a lift
Opening and closing a door
Wiping feet on doormat
Getting a book off a shelf
Answering the door bell
Reaching into a low cupboard
Picking something up off the floor
Opening and closing a window
Opening a high window
Opening and closing curtains
Going for a walk in the woods
Getting into a car (M)
Walking up a high bridge (M)

Caring for plants – 4 items
Re-potting a houseplant
Putting flowers in a vase
Watering a houseplant
Caring for plants on a balcony

Dressing and self-care – 19 items
Putting on a coat
Putting on a blouse or shirt
Putting on a T-shirt or vest
Putting on a scarf and gloves
Putting trousers on
Washing upper body at sink (taken)
Washing lower body at sink (taken)
Washing upper body at sink
Washing lower body at sink
Rubbing in body lotion
Shaving or putting make-up on
Using a public toilet
Going to the toilet in your own home
Cutting finger nails
Taking oral medication
Pulling a blanket around oneself
Showering and washing hair (M,Y)
Cutting toe nails (M)
Putting on lace-up shoes (M)

Activities outside the home – 16 items
Eating in a restaurant
Visiting an outpatients’ clinic
Visiting a dentist
Voting in an election
Visiting a general practitioner
Travelling by bus
Going to the bottle bank
Travelling (not driving) in a car
Going to a concert
Shopping for clothes
Going to a birthday party
Visiting a hairdresser
Fetching groceries
Visiting the neighbours
Going to the pharmacy (Y)
Crossing the road (Y)
Having a drink in a cafe (F)

Table 5.2: types of item, for which a large proportion do not have the same measurement characteristics for men and women or for respondents aged 84 or younger and respondents aged 85 or older. The letter ‘M’ denotes that the item is easier for men than for women, the letter ‘F’ that the item is easier for women that for men and the letter ‘Y’ that the item is easier for respondents aged 84 or younger.

Housework, cleaning and laundry – 24 items

Cleaning a bathroom
Mopping the floor
Sweeping the floor
Using a dustpan and brush
Cleaning a fridge
Clearing the table after a meal
Vacuum cleaning
Washing up
Shaking a table cloth out
Polishing shoes
Hanging clothes in a cupboard
Putting crockery away
Cleaning a toilet (F)
Changing the sheets on a bed (F)
Changing a duvet cover (F)
Making a bed (F)
Cleaning kitchen worktops (F)
Cleaning a bathroom sink (F)
Dusting (F)
Using a washing machine (F)
Washing underwear by hand (F)
Hanging the washing out (F)
Ironing clothes (F)
Folding clean laundry up (F)

Household maintenance – 5 items

Changing the bulb in a side lamp
Plugging in an electrical appliance
Tightening a screw (M)
Hanging something on a nail (M)
Unblocking a sink (M)

Eating and food preparation – 13 items

Eating a meal
Making a fried egg sandwich
Making coffee or tea
Preparing a bowl of cereal
Peeling an apple
Warming tinned soup up
Making a cheese sandwich
Preparing porridge
Preparing a light meal (F)
Preparing a warm meal (F)
Opening a bottle of fizzy drink (M)
Pulling a cork out of a wine bottle (M)

Lifting and carrying – 7 items

Lifting a box weighing 10kg (M)
Lifting a toddler up (M)
Doing weekly grocery shopping (M)
Putting a bag of rubbish out (M)
Carrying a tray (M)
Moving a bed or table (M)
Picking up something under table (M)

Household administration – 10 items

Setting an alarm clock
Reading a newspaper
Writing a letter
Fetching and opening letters
Wrapping up a present
Filling in an official form (M)
Going to the post office (M)
Using an ATM (M,Y)
Replacing a passport (Y)
Posting a letter (Y)

The types of item, for which a large proportion do not have the same measurement characteristics for men and women or for respondents aged 84 or younger and respondents aged 85 or older, are given in Table 5.2. Of the 24 items reflecting housework, cleaning and laundry, 12 are easier for women. It is interesting to note that all five items reflecting washing clothes are easier for women than men. Of the 5 items reflecting household maintenance tasks, 3 are easier for men than women. Furthermore, of the 12 items reflecting eating and food preparation, 2 are easier for women and 2 are easier for men. Women appear to find preparing a whole meal easier and men find preparing drinks easier. All 7 items reflecting activities with a large element of lifting or carrying are easier for men than women. Finally, of the 10 items reflecting household administration, 2 are easier for men, another 2 items are easier for respondents aged 84 or younger and one item is easier for both men and respondents aged 84 or younger.

Discussion

Due to the relatively small sample size and the low significance level used in this study, only items with a substantial amount of differential item functioning have been identified. It is plausible that even more items would have demonstrated different measurement characteristics for subgroups if a larger sample had been available or a higher significance level used.

A lot of items in instruments to assess functional health status reflect activities traditionally performed primarily by either men or women. These items are likely to have different measurement characteristics for men and women. Examples include: housework, particularly items relating to laundry; preparing meals; lifting and carrying heavy objects; and household maintenance. A number of items will describe activities, which have only become common place in recent years. These items are likely to have different measurement characteristics for younger and older respondents. Examples are ‘withdrawing money using an automatic teller machine’ and ‘using a mobile telephone’. Notice should be taken of local customs when interpreting results. These results were obtained from a sample of respondents from the Netherlands. Activities of daily life seen as the preserve of men or of women differ, even between countries with very similar cultures, meaning that a sample from another country or culture may result in very different results. The items, which behave in different ways for different subgroups, will either be removed from the item bank or included with different measurement properties for different subgroups. The latter approach may seem complicated, but is straightforward in the framework of a computerised item bank.

Chapter 6

The AMC Linear Disability Score project in a population requiring residential care: psychometric properties

This chapter has been published as:

> Holman R, Lindeboom R, Vermeulen M, de Haan RJ. The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. *Health and Quality of Life Outcomes* 2004, 2:42.

and is available from `http://www.hqlo.com/content/2/1/42`

Background

It is now widely accepted that examining quality of life is an important aspect in the treatment and evaluation of many conditions. Functional status is seen as an important determinant of quality of life. A wide variety of instruments have been developed to quantify functional status[10]. These instruments tend to have a fixed length and all items are administered to the whole group of patients under scrutiny. However, currently interest is moving towards the more flexible framework offered by item banks. An item bank is a collection of items, for which the measurement properties of each item are known[124, 125]. When using an item bank, it is not essential for all respondents to be examined using all items. This enables the burden of testing to be considerably reduced for both patients and researchers. It is even possible to select the 'best' items for individual patients using computerised adaptive testing algorithms[6]. Furthermore, results from studies using different selections of items from an item bank can be directly compared. Item banks, measuring concepts such as quality of life[124, 126], the impact of headaches[127] or functional status[7, 128], have been developed.

The AMC Linear Disability Score (ALDS) project item bank was developed to quantify functional status[7, 8]. The ALDS item bank covers a large number of activities, which are suitable for assessing respondents with a very wide range of functional status and many types of chronic condition. The item bank is particularly suitable for use in the Netherlands. The ALDS items were obtained from a systematic review of generic and disease specific functional health instruments[10]. Five psychometric aspects of the ALDS item bank need to be considered before it can be implemented. These are: (a) there needs to be enough variation in the response categories used for each item[8]; (b) estimates of the item response theory model parameters should not depend on patient characteristics such as age or gender[16, 71]; (c) estimates of the item response theory model parameters, which are stable across different subsets of items from the instrument and based on a sufficiently large sample[40] of respondents, should be available[8]; (d) an examination of the extent to which the ALDS items represent a single construct;

and (e) testing whether a simpler item response theory model is suitable for the set of items.

This paper examines these five aspects of the ALDS item bank using the responses given by residents of supported housing schemes, residential care and nursing homes in and around Amsterdam, the Netherlands. This, mainly elderly, population has been chosen because they generally experience some level of functional restriction and consume a large amount of health care services.

Methods

Data collection

This paper considers 160 items, which were considered to be applicable in a residential care setting. Each item has two response categories: 'I could carry out the activity' and 'I could not carry out the activity'. If a respondent had never had the opportunity to experience an activity 'not applicable' was recorded. In the analysis, responses in the category 'not applicable' were treated as if the individual items had not been presented to the individual respondents[129]. It was felt that presenting all 160 items to each respondent would place an unnecessary and unacceptable burden on those responding to the items. Therefore, the data described in this paper were collected using an incomplete, anchored calibration design[7, 8, 55, 27] with four sets of 80 items. Item sets A and B have half their items in common, as do item sets B and C, item sets C and D and item sets A and D. The items in common between two sets of items are known as 'anchors' and allow all items and patients to be calibrated on the same scale. The patterns of missing data in this type of design are, in statistical terms, ignorable[75]. The item sets were administered randomly to 150 respondents (item set A), 143 respondents (item set B), 138 respondents (item set C) and 124 respondents (item set D).

Respondents

A total of 555 residents of supported housing, residential care and nursing homes were interviewed. The median age was 84 years (range 37 to 101 years), while 444 (80%) were female. Since the respondents were interviewed ‘at home’, accurate data on medical conditions were not available. All respondents gave informed consent. The study was approved by the medical ethics committee in our hospital.

The item response theory models

In this paper the data were analysed using the two-parameter logistic item response theory model[7, 8, 39, 43]. In this model, the probability, P_{ik}, that patient k responds to item i in the category ‘can’ is modelled using

$$P_{ik} = \frac{\exp(\alpha_i(\theta_k - \beta_i))}{1 + \exp(\alpha_i(\theta_k - \beta_i))} \tag{6.1}$$

where θ_k denotes the ability of patient k to perform activities of daily life. The discrimination parameter (α_i) and difficulty parameter (β_i) describe the measurement characteristics of item i. The larger the value of β_i, the more difficult item i is. In addition, the larger the value of α_i, the better an item is a distinguishing between abilities above and below β_i. If the values of α_i are constrained to be equal for all items, the model in equation 6.1 becomes the one-parameter logistic item response theory model[98]. The model in equation 6.1 can be extended to test whether the values of β_i for, say males and females, are significantly different. If the values of β_i for different groups of respondents are significantly different, then there is evidence of differential item functioning. Full-information factor analysis also uses an extension of the model in equation 6.1. These approaches are described in mathematical terms in the Appendix.

In this paper, estimates of α_i and β_i were obtained using a marginal maximum likelihood based procedure[83]. This method assumes that the ability parameters (θ_k) follow a Normal distribution and can account for incomplete designs, as described in the Appendix. Expected a posteriori methods were used to estimate θ_k[130].

Statistical analysis

To achieve the objectives of this study, there were five steps in the statistical analysis. In step (a), the amount of variation in the response categories used for each item[8] was considered and items demonstrating too little variation were removed. Items were excluded from further analysis if fewer than 10% or more than 90% of the patients responded in the category ‘cannot’. In step (b), the items were examined to investigate whether the value of the item difficulty parameter (β_i) was similar for male and female patients and for patients younger than 85 years and those aged 85 or older. The model is described in depth in the Appendix. Items were excluded from further analysis if the value of the item difficulty parameter was significantly different (1% level) between gender or aged based groups. In this step, the fit of the model to the data from each item was not assessed. In step (c), estimates of the item parameters (α_i and β_i) were obtained. The fit of the model to the data from each item was assessed using G^2 statistics[131]. Items, for which the fit statistic had a p-value of less than 0.01, were excluded from the item bank. In addition, the stability of the estimates of the item parameters over different sets of items was examined using the model from step (b). Items were excluded from further analysis if the value of the item difficulty parameter was significantly different (1% level) between item sets A and B, B and C, C and D or A and D. Furthermore, a Kolmogornov-Smirnov test was carried out to examine whether the ability parameters (θ_k) were Normally distributed. In step (d), the dimensionality of the item bank was examined using item response theory based full information factor analysis[131, 103, 43]. The number of latent roots greater than 1 is regarded as an indicator of the number of factors in the data set. This method is described in more depth in the Appendix. Four exploratory factor analyses were carried out, one on each of the anchors between item sets A and B (293 respondents), B and C (281 respondents), C and D (262 respondents) or A and D (274 respondents). A fifth, confirmatory, factor analysis was carried out on the whole data set (555 respondents). In addition, Cronbach’s coefficient alpha was calculated for each anchor and the whole data set[60, 132]. In step (e) the one-parameter logistic item response theory model was fitted to the

Table 6.1: The number of and examples of items removed at each stage of the psychometric analysis.

Stage of analysis	Number of items	Reason	Examples
	1	Concerns about the way the item was presented	
(a)	28	$< 10\%$ or $> 90\%$ of responses in 'cannot'	Reaching for a cup and taking a sip of water Combing hair at a sink Cycling on a heavily laden bicycle
(b)	26	Significant difference between M and F and/or under and over 85	Washing up (easier for over 85) Crossing the street (easier for under 85) Preparing a warm meal (easier for females)
(c)	26	Item fit p-value < 0.01 or estimates of β_i not stable	Taking oral medication Cycling Getting money out of the bank using an ATM
In item bank	79		See Table 2
Total	160		

remaining items. The differences between the -2log likelihoods of this model and the two-parameter model fitted in step (c) was tested using a χ^2 test. The analysis in steps (a), (b), (c) and (e) was carried out in Bilog, version 3.0[131]. The analysis in step (d) was carried out using TESTFACT, version 4.0[131].

Results

Of the 160 items included in the item bank, one was removed because it was worded differently in two different item sets. Of the 159 remaining items, 77 were removed from the item bank. This process is described in Table 6.1. In step (a), 28 items were excluded from further analysis because fewer than 10% or more than 90% of responses were in the category ‘cannot’. In step (b), 26 items were removed because they had significantly different estimates of the item difficulty parameter (β_i) for for males and females and/or for younger and older respondents. Of these 26 items, 19 had different measurement characteristics for females and for males, 5 items had different measurement characteristics for those aged under 85 and for those aged 85 or over, and 2 items had different measurement characteristics for both males and females and for older and younger respondents. In step (c), 23 items had an item fit statistic p-value of less than 0.01. In addition, 3 items were excluded from further analysis because the value of the item difficulty parameter (β_i) was significantly different between two item sets of items. Hence, 79 psychometrically sound items remained in the item bank. A short description of the content of the 79 items in the final version of the calibrated item bank, together with estimates of the dispersion (α) and difficulty (β) parameters and their standard errors, are given in Tables 6.2 and 6.3. Following step (c) of the analysis, the anchors between the sets of items contained between 13 and 23 items. In addition, there was no evidence to suggest that estimates of θ do not follow a Normal distribution (Kolmogorov-Smirnov test, p-value $= 0.637$).

In step (d), the full information factor analysis indicated that, for three of the four anchors between the item sets, there was only one latent root of the correlation matrix larger than 1. In the fourth item set, a second latent root was marginally above 1. The percentage of the variance explained by the first factor varied between 67% and 72%. The values of Cronbach’s alpha coefficient for the four anchors were between 0.86 and 0.93. The confirmatory factor analysis carried out on the whole data set indicated that 70% of the variance was explained by the first factor. Cronbach’s alpha coefficient for the whole data set equalled 0.98.

In step (e), the one-parameter logistic item response theory model was fitted to the 79 items remaining after step (c). This model fitted the data significantly less well than the two-parameter model (p-value < 0.0001). For 3 items, the item fit statistic had p-value < 0.01. After removal of these items, the two-parameter model was still significantly better than the one-parameter model (p-value < 0.0001).

Table 6.2: The 79 items remaining in the calibrated item bank. The number of respondents, to whom the item was presented (PT), the number responding in the categories 'not applicable' (NA), 'can' (can) and 'cannot' (CN) are given. The discrimination (α) and difficulty (β) parameters are given along with their standard errors in parentheses.

		Response			Location	Discrim.
		category			parameter	parameter
Description of item content	PT	NA	can	CN	(β)	(α)
Walking upstairs with bag	262	0	19	243	-3.607 (2.404)	1.122 (0.892)
Mopping a flight of stairs	262	5	16	241	-2.830 (1.708)	0.447 (0.411)
Cleaning top of cupboard	281	2	27	252	-2.816 (1.946)	0.480 (0.393)
Cleaning a bathroom	293	1	37	255	-2.621 (2.061)	0.323 (0.338)
Vacuuming	274	0	33	241	-2.408 (1.844)	0.287 (0.280)
Walking the woods	281	0	31	250	-2.343 (1.636)	0.362 (0.340)
Fetching groceries	293	0	36	257	-2.262 (1.623)	0.353 (0.343)
Mopping the floor	281	2	41	238	-2.225 (1.902)	0.339 (0.374)
Caring for balcony plants	262	1	32	229	-2.108 (1.616)	0.314 (0.325)
Travelling by bus or tram	281	0	44	237	-2.093 (1.835)	0.308 (0.370)
Walking up 2 flights of st.	274	0	39	235	-1.921 (1.532)	0.277 (0.298)
Cleaning a fridge	293	2	64	227	-1.406 (1.464)	0.171 (0.236)
Going to a restaurant	293	3	60	230	-1.335 (1.238)	0.159 (0.188)
Carrying a tray	281	1	75	205	-1.304 (1.774)	0.222 (0.326)
Going for a long walk	281	0	72	209	-1.290 (1.629)	0.178 (0.277)
Going to the dentist	293	18	75	200	-1.283 (1.881)	0.205 (0.299)
Sweeping the floor	262	1	70	191	-1.239 (1.902)	0.196 (0.313)
Cutting toe nails	262	0	45	217	-1.175 (0.786)	0.136 (0.144)
Walking up a hill	281	3	73	205	-1.133 (1.425)	0.137 (0.198)
Walking up 1 flight of st.	274	2	67	205	-1.127 (1.382)	0.155 (0.222)
Going to a concert	262	0	57	205	-0.996 (0.860)	0.113 (0.128)
Going to the pharmacist	262	2	75	185	-0.976 (1.572)	0.131 (0.201)
Hanging washing out	293	10	84	199	-0.960 (1.514)	0.144 (0.240)
Going to the post office	274	0	88	186	-0.948 (1.881)	0.151 (0.248)
Going to a party	281	1	69	211	-0.924 (0.878)	0.109 (0.129)
Filling an official form in	281	1	68	212	-0.896 (0.790)	0.106 (0.119)
Using a washing machine	281	6	95	180	-0.851 (1.743)	0.153 (0.244)
Visiting outpatients' clinic	293	0	95	198	-0.815 (1.317)	0.112 (0.161)
Going to the bottle bank	281	5	108	168	-0.675 (1.840)	0.153 (0.312)
Short walk	274	0	95	179	-0.645 (1.358)	0.108 (0.166)
Putting the rubbish out	293	5	108	180	-0.573 (1.338)	0.116 (0.198)
Reaching (high cupboard)	274	0	95	179	-0.569 (1.097)	0.099 (0.172)
Using a dustpan + brush	262	2	103	157	-0.537 (1.865)	0.135 (0.335)
Opening a high window	281	0	140	141	-0.078 (1.290)	0.096 (0.166)
Fetching groceries (1 day)	262	0	128	134	-0.043 (1.373)	0.099 (0.184)
Using a public toilet	293	5	159	129	0.139 (1.835)	0.110 (0.241)
Putting flowers in a vase	293	2	162	129	0.169 (1.787)	0.107 (0.240)
Frying an egg	281	3	154	124	0.178 (1.982)	0.115 (0.261)

Table 6.3: (*Table 6.2 continued*)

		Response			Location	Discrim.
		category			parameter	parameter
Description of item content	PT	NA	can	CN	(β)	(α)
Warming up soup	293	1	164	128	0.203 (1.919)	0.113 (0.232)
Cleaning a toilet	262	0	149	113	0.308 (1.528)	0.105 (0.203)
Putting lace up shoes on	281	1	167	113	0.314 (1.165)	0.093 (0.157)
Changing a light bulb	281	1	177	103	0.533 (1.541)	0.116 (0.174)
Cleaning a bathroom sink	281	5	170	106	0.564 (2.089)	0.126 (0.302)
Cutting finger nails	262	0	168	94	0.605 (1.337)	0.114 (0.173)
Rubbing lotion into body	262	4	164	94	0.627 (1.469)	0.115 (0.184)
Reaching (low cupboard)	274	0	184	90	0.672 (1.131)	0.106 (0.152)
Picking something up	262	0	172	90	0.712 (1.466)	0.129 (0.198)
Making porridge	293	2	191	100	0.714 (1.704)	0.119 (0.216)
Getting into a car	281	3	185	93	0.738 (1.656)	0.132 (0.209)
Shaking a tablecloth out	274	2	190	82	0.906 (1.438)	0.125 (0.204)
Making a bed	281	0	193	88	1.003 (2.028)	0.152 (0.292)
Preparing lunch	262	1	186	75	1.117 (1.729)	0.169 (0.279)
Using a lift	262	2	199	61	1.208 (1.299)	0.158 (0.186)
Setting an alarm clock	281	4	216	61	1.319 (1.431)	0.165 (0.199)
Pulling a blanket up	293	0	253	40	1.485 (0.898)	0.167 (0.149)
Visiting neighbours	293	1	231	61	1.548 (1.685)	0.221 (0.252)
Travelling in a car	274	3	230	41	1.592 (1.126)	0.222 (0.192)
Shaving face	274	1	233	40	1.593 (1.075)	0.180 (0.164)
Watering a house plant	262	3	204	55	1.600 (1.681)	0.226 (0.259)
Opening low a window	262	0	201	61	1.735 (2.137)	0.246 (0.343)
Putting trousers on	293	2	224	67	1.821 (2.372)	0.295 (0.406)
Making coffee or tea	293	0	235	58	1.832 (1.936)	0.237 (0.290)
Peeling an apple	281	1	233	47	1.859 (1.631)	0.226 (0.219)
Making cereal	281	1	225	55	1.860 (1.921)	0.222 (0.256)
Eating a meal	293	0	255	38	2.081 (1.509)	0.253 (0.225)
Hanging clothes up	262	0	203	59	2.105 (2.595)	0.344 (0.481)
Opening curtains	262	0	216	46	2.129 (1.958)	0.366 (0.357)
Moving between chairs	281	0	237	44	2.214 (1.905)	0.389 (0.375)
Putting gloves on	293	1	259	33	2.364 (1.617)	0.306 (0.246)
Making a sandwich	281	1	243	37	2.416 (1.856)	0.392 (0.333)
Sitting down on a bed	262	1	231	30	2.457 (1.658)	0.309 (0.244)
Putting a coat on	274	0	227	47	2.463 (2.323)	0.425 (0.395)
Putting a shirt on	262	0	228	34	2.495 (1.842)	0.360 (0.287)
Washing upper body	274	0	243	31	2.705 (1.875)	0.420 (0.327)
Answering the door	274	1	233	40	2.792 (2.373)	0.481 (0.449)
Getting out of bed	262	0	232	30	3.019 (2.132)	0.581 (0.448)
Washing lower body	293	1	241	51	3.037 (3.098)	0.722 (0.761)
Putting a T-shirt on	293	2	257	34	3.440 (2.664)	0.718 (0.630)
Locking a door	262	0	230	32	3.366 (2.512)	0.970 (0.749)

Discussion

In this study, the psychometric properties of the item bank have been examined using a sample of 555 respondents and an incomplete calibration design. Each item was presented to between 262 and 293 respondents. These figures are above the minimum, of 200 respondents, regarded as necessary to implement the models used in this paper[40]. It could be argued that it would have been desirable for all respondents to be presented with all items, but this would have placed an unacceptable burden on the, often frail, population in this study. Incomplete calibration designs are regularly implemented in the development and maintenance of item banks used in educational testing[6, 55] and have gained some recognition in health related applications[27]. Developments in psychometric theory mean that it is now possible to perform the same types of analysis on data resulting from incomplete designs, as is performed on data from complete calibration designs[131, 103, 132]. The number of items in the anchors following the analysis, indicate that the design was still amply linked[8].

One of the major assumptions underlying the use of the item response theory models described in this paper is that the items reflect a single latent trait (θ). This has been examined using item response theory based full-information factor analysis. Part of the full-information factor analysis was performed on sub-sets of the data, as exploratory analyses on incomplete designs may lead to instable results. However, the confirmatory factor analysis was performed on all data. The results, together with the high level of internal consistency, as measured by Cronbach's alpha, and the acceptable fit of the two-parameter logistic item response theory model to the data indicate that the items presented in this paper probably represent a unidimensional construct in a population of respondents requiring residential care.

Another important assumption when using item response theory models in conjunction with marginal maximum likelihood estimation procedures is that the values of the latent trait (θ) follow a pre-specified, usually Normal, distribution. In this study, there was no evidence that these values did not follow a Normal distribution. This is in contrast to many previously published studies into health and quality of life outcomes, where a strongly skewed distribution was found. The authors feel that there are two reasons for this contrast. Firstly, in this study, the respondents all had some level of restriction in their ability to perform activities of daily life. Secondly, the item bank includes items well above and well below the level of functional status enjoyed by the respondents. This means that the item bank did not have a ceiling or floor effect with respect this this population.

In this study, 81 (51%) of 160 items were removed from the item bank because they did not conform to the psychometric standards required of the item bank. This is a much higher level than would be expected in the calibration of an item bank for use in educational measurement. However, when the results are examined more carefully, 28 items were removed because they were too difficult or too easy for the population in this study. In addition, 26 items were removed because they had different item parameters for different groups of respondents. These problems would have been identified much earlier in an educational item bank. Hence, only 26 (25%) of 106 items were removed due to item misfit. The number of items retained in the item bank may have been higher if a more flexible model, based on, for example, non-parametric smoothing techniques had been used[121]. However, this type of model is less suitable as a base for implementing modern testing algorithms, such as computerised adaptive testing. In addition, it is possible that more items could be made available if the items demonstrating differential item functioning were included in the item bank with different item location parameters (β_i) for males and females or for younger and older respondents. This may seem complicated, but is straightforward in the framework of a computerised item bank.

This paper has concentrated on the two-parameter logistic item response theory model. However, the one-parameter logistic item response theory model was also fitted to the 79 items remaining in the item bank. This model fitted the data significantly less well than the two-parameter model, even after 3 items demonstrating misfit at the item level were removed. This confirms the choice of the two-parameter model. This model was chosen because it allows the probability of responding in the category ‘can’ to be modelled more flexibly than when the one-parameter logistic model is used. This enables a more realistic model for the data to be built than when the more restrictive approach associated with the one-parameter model is chosen[43].

This paper has examined the calibration of the ALDS item bank in a population requiring residential care. It has been shown that the item bank has sound psychometric properties and could form a stable base for a wide range of applications. However, it is possible that the items will have different measurement characteristics for patients requiring treatment for specific chronic conditions or in other countries. Hence, it is important that the ALDS item bank is tested carefully before it is used to assess the functional status in other groups of respondents or in other countries.

Conclusions

Now that the measurement properties of the ALDS item bank have been examined carefully, the item bank can be used as a foundation for quantifying functional status. If modern algorithms, such as computerised adaptive testing, are implemented, it will be possible to obtain accurate measurements, whilst keeping the burden of testing on respondents and interviewers to a minimum. Items can be selected for use in further research, for allocation individuals to appropriate care settings and for calculating institutional funding based on the actual care load. It is hoped that the ALDS item bank will play an important part in the implementation of computerised adaptive testing of functional status.

Appendix

Differential item functioning

The model in equation 6.1 can be extended to test whether different groups of respondents to have different values of β_i. This is known as differential item functioning. For instance, if interest is in possible differences in β_i between males and females, then the probability, P_{ik}, that patient k responds to item i in the category 'can' is written as

$$P_{ik} = \frac{\exp(\alpha_i(\theta_k - \beta_{iM} - I_k\beta_{iF-M}))}{1 + \exp(\alpha_i(\theta_k - \beta_{iM} - I_k\beta_{iF-M}))} \tag{6.2}$$

where β_{iM} is the item difficulty for male respondents, β_{iF-M} is the difference between the item difficulty for males and for females and I_k is an indicator variable taking the value 0 if respondent k is male and the value 1 if respondent k is female. The hypothesis $\mathrm{H}_0 : \beta_{iF-M} = 0$ can be tested to examine whether item i has the same measurement characteristics for males and for females.

Item parameter estimation in incomplete designs

In this study, the item parameters (α_i and β_i) were estimated using marginal maximum likelihood methods. The likelihood, L, over n items and K ($K = 555$) respondents can be written as

$$L = \prod_{k=1}^{K}\prod_{i=1}^{n} I_{ik} P_{ik}^{J_{ik}} \left(1 - P_{ik}\right)^{1-J_{ik}} \tag{6.3}$$

where I_{ik} is an indicator variable taking the value 1 if respondent k was offered item i and the value 0 otherwise and where J_{ik} is an indicator variable taking the value 1 if respondent k responded to item i in the category 'can' and the value 0 otherwise. Furthermore, the probability, P_{ik}, that respondent k responded to item i in the category 'can' is as in equation 6.1, or, where appropriate, as in equation 6.2 or 6.4. In the estimation process, the values of θ_k or θ_{km} were assumed to follow a Normal distribution with mean equal to 0 and unknown variance, σ^2, and were integrated out of the likelihood to obtain the marginal likelihood. The marginal likelihood was maximised using an EM algorithm[83].

Full information factor analysis

Full information factor analysis is a technique based on multidimensional item response theory models where the ability is represented by M variables, denoted θ_{km} where $m = 1, 2, \ldots, M$[131, 103]. The model, in equation 6.1, for the probability,

P_{ik}, that person k responds to item i in the category 'can' can be extended to

$$P_{ik} = \frac{\exp\left(\left(\sum_{m=1}^{M} \alpha_{im}\theta_{km}\right) - \delta_i\right)}{1 + \exp\left(\left(\sum_{m=1}^{M} \alpha_{im}\theta_{km}\right) - \delta_i\right)} \tag{6.4}$$

where θ_{km} denotes the value of the latent variable θ_m associated with person k and α_{im} denotes the discrimination parameter for item i with respect to the latent variable θ_m. Furthermore, δ_i is a difficulty type parameter. The loading, a_{im} of item i on factor m can be calculated using

$$a_{im} = \frac{\alpha_{im}}{\sqrt{\sum_{m=1}^{M} \alpha_{im}^2}}. \tag{6.5}$$

The value of the standard difficulty parameter, (β_i), can be calculated using

$$\beta_i = \frac{\delta_i}{\sqrt{\sum_{m=1}^{M} \alpha_{im}^2}}. \tag{6.6}$$

Generally, the parameters α_{im} and δ_i are estimated using marginal maximum likelihood methods.

Chapter 7

How does varying the number of patients or number of items affect the power to detect a treatment effect?

This chapter has been previously published. Reprinted with permission from Elsevier from

Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. *Controlled Clinical Trials*. 2003; **24**: 390–410.

and is available from `http://dx.doi.org/doi:10.1016/S0197-2456(03)00061-8`

Introduction

In recent years, there has been an enormous increase in the use of patient relevant outcomes, such as functional status and quality of life, as endpoints in medical research, including randomised controlled clinical trials (RCT). Many patient relevant outcomes are measured using questionnaires designed to quantify a theoretical construct, often modelled as a latent variable. When a questionnaire is administered to a patient, responses to individual items are recorded. Often the scores on each item are summed to obtain a single score for each patient. The reliability and validity of sum scores are usually examined in the framework of classical test theory. This framework is widely accepted and applied in many areas of medical assessment[19]. However, following dissatisfaction with these methods, interest in the use of an alternative paradigm, known as item response theory (IRT), has grown[5]. IRT was developed as an alternative to the use of classical test theory when analysing data resulting from school examinations. An overview of IRT methods is given in this paper, while in depth descriptions in general[36, 133] and in medical applications[23] are given elsewhere.

Advantages of using IRT to analyse a RCT include proper modelling of ceiling and floor effects, solutions to the problem of missing data and straightforward ways of dealing with heteroscedasticity between groups. However, the main advantage is that it is not essential to assess all patients with exactly the same items. For instance, if sufficient information on the measurement characteristics of the SF36[134] health survey in a particular patient population were available, it would be possible to obtain completely comparable estimates of health status using only the items most appropriate to each individual patient. An extension of this is adaptive testing, in which each patient potentially receives a different, computer administered, questionnaire in which the questions offered to each patient depend on the responses given to previous questions [6, 33]. A pre-requisite of this type of testing is access to a large item bank, which has been calibrated using responses from comparable patients[7, 10].

Ethical considerations require that as few patients as possible are exposed to the ‘risk’ of a novel treatment during a RCT, but it is important to ensure that enough patients are included to have a reasonable power of detecting the effect of interest. For this reason, calculation of the minimal sample size required to demonstrate a clinically relevant effect has become integral to the RCT literature[135]. However, since IRT was developed as a tool for analysing data resulting from examinations, most technical work has concentrated on the statistical challenges found in this field. For example, when assessing the effects of an educational intervention, in some ways similar to a RCT, thousands of pupils, even whole cohorts, are often included. This means that minimal sample size and power calculations in relation to questionnaires analysed with IRT have received very little attention. In addition, IRT offers a framework, in which the number of items used to assess patients can be easily varied. Thus, sample size calculations need to consider not only the number of patients, but also the number of items used. Some work has touched on these issues[136] or considered it as a sideline of another aspect[78], but no guidance on sample size calculations for RCT in the context of IRT has been published.

In this paper, the relationship between the number of patients in each treatment arm, the number of items used to assess the patients and the power to detect given effect sizes will be examined using a simulation study. The results will be used to develop guidelines for the number of patients to be used in each arm of a RCT, when a questionnaire analysed using IRT, is used as the primary outcome. The methods are illustrated in a population of patients with end stage renal disease 12 months after starting dialysis[137, 138] using the SF36[134], the SF12[139] and the SF8[140] health status surveys. In addition, to provide a framework, in which data from RCTs can be analysed using IRT, the behaviour of asymptotic methods developed to compare the mean level of the latent trait in two groups will be considered in the relatively small samples encountered in RCTs.

Item response theory in a randomised clinical trial

Item response theory is used to model the probability that a patient will respond to a number of items related to a latent trait in a certain way. The two parameter logistic model[39] is for data resulting from items with two response categories, '0' and '1' and is one of many IRT models developed[37]. In this model, the probability, $p_{ik}(\theta)$ that patient k, with latent trait equal to θ_k, will respond to item i in category '1' is given by

$$p_{ik}(\theta) = \frac{\exp\left(\alpha_i(\theta_k - \beta_i)\right)}{1 + \exp\left(\alpha_i(\theta_k - \beta_i)\right)} \tag{7.1}$$

where α_i and β_i are known as item parameters. The more widely known Rasch model [41], is similar to the two-parameter logistic model, but with the parameters α_i assumed equal to 1. An important assumption of IRT models is that of local independence, meaning that the probability of a patient scoring '1' on a given item is independent of them scoring '1' on another item, given their value of θ. This means that the correlations between items and over patients are fully explained by θ. Models have also been developed for data resulting from items with more than two response categories[133, 56]. The generalised partial credit model[100] is a fairly well known example, in which the probability $p_{ijk}(\theta)$ that a patient with latent trait equal to θ_k, will respond to an item i, with $(J_i + 1)$ response categories, in category j, $j > 0$, is given by

$$p_{ijk}(\theta) = \frac{\exp\left(\sum_{v=0}^{j} \alpha_i(\theta_k - \beta_{iv})\right)}{1 + \sum_{j=0}^{J_i} \exp\left(\sum_{v=0}^{j} \alpha_i(\theta_k - \beta_{iv})\right)} \tag{7.2}$$

where α_i is the discrimination parameter of item i and β_{iv} indicates the point at which the probability of choosing category j or category $j-1$ is equal.

In IRT, the item and patient parameters are usually estimated in a two stage procedure. Firstly, the item parameters are estimated, often by assuming that the θ_k follow a Normal distribution and integrating them out of the likelihood. Secondly, maximum likelihood estimates of θ_k are obtained using the previously estimated item parameters. In this study, it will be assumed that the items under consideration form part of a calibrated item bank[7], meaning that the item parameters have been previously estimated[83] from responses given by comparable patients to the items and are assumed known for all items. It is theoretically possible to estimate the item parameters from the responses given to the items by patients included in a RCT, but accurate estimates can only be obtained from large samples of patients, say over

500. Since this figure is rarely attained in RCTs, it is will often not be practical to estimate the item parameters in an RCT.

In a straightforward RCT, the patient sample is randomly divided into two groups, say A and B. Each group receives a different treatment regime and primary and secondary outcomes are assessed once, at the end of the study. The main interest is in whether the null hypothesis, H_0, that the mean level of the primary outcome, say θ, is equal in both groups. This can be written as

$$\mathrm{H}_0 : \mu_A - \mu_B = 0 \tag{7.3}$$

where μ_A and μ_B denote the mean of the distribution of θ in groups A and B, respectively. Clinically, the two groups are said to differ if the ratio of the difference $\mu_A - \mu_B$ and the standard deviation of θ is larger than a given 'effect size'. A lot of work has been carried out examining the clinical relevance of particular effect sizes in given situations. However, in practice, interest is often in examining the, arbitrarily defined, minimal, moderate and substantial effect sizes of 0.2, 0.5 and 0.8 on continuous variables [135]. The number of patients required to detect a given effect size with a particular power depends on the values of the effect size and the standard errors of μ_A and μ_B.

Now consider a RCT, in which the primary outcome is θ, measured, at the end of the study, using a questionnaire with n items, each with two response categories, analysed using item response theory. Let us assume that patients $1, 2, \ldots, K$ are in group A and patients $K+1, K+2, \ldots, 2K$ in group B, meaning that the total sample size is $2K$. For group A, we can rewrite equation (7.1) as

$$p_{ik}(\theta) = \frac{\exp\left(\alpha_i(\mu_A + \epsilon_k - \beta_i)\right)}{1 + \exp\left(\alpha_i(\mu_A + \epsilon_k - \beta_i)\right)} \tag{7.4}$$

where $\theta_k = \mu_A + \epsilon_k$ and μ_A is the mean of θ_k in group A. Hence, ϵ_k have mean 0 and standard deviation σ_A. For group B, equation (7.1) can be rewritten in a similar way, with $\theta_k = \mu_B + \epsilon_k$ and the standard deviation of ϵ_k equal to σ_B. For RCTs carried out using a questionnaire consisting of items with more than two response categories, equation (7.2) can be rewritten in a similar way. The main interest is in examining the null hypothesis in equation (7.3) by obtaining $\hat{\mu_A}$ and $\hat{\mu}_B$ and testing whether they are significantly different from each other. When using IRT based techniques, it is inadvisable to estimate the values of θ_k for all patients and then perform standard analysis, such as t-tests, on these estimates[78], since this ignores the measurement error inherent to θ_k. This means that using standard methods for calculating sample size may lead to inaccurate conclusions. The estimation equations for μ_A and μ_B are complex and the values of s.e.(μ_A) and s.e.(μ_B) depend not only on the number of patients in each arm of the RCT, but also on the number of items used to assess the patients and the relationship between α_i and β_i and the distribution of θ. This means that it is not possible to write down a straightforward equation for either $\hat{\mu}_A$ and $\hat{\mu}_B$ or the number of patients required to detect a given sample size with a particular power. Consistent estimates of μ_A and μ_B are obtained by maximising a likelihood function, which has been marginalised with respect to θ [107, 141]. These estimation methods are described in more detail in the Appendix.

The marginal maximum likelihood estimates of μ_A and μ_B can be combined, with their standard errors to obtain test statistic

$$Z_L = \frac{\hat{\mu}_A - \hat{\mu}_B}{\text{s.e.}(\hat{\mu}_A - \hat{\mu}_B)} \tag{7.5}$$

where $\hat{\mu}_A$ and $\hat{\mu}_B$ denote the estimates described above. It has been proven that, under the hypothesis $\mu_A - \mu_B = 0$ and for large sample sizes, say $2K > 500$, Z_L follows an asymptotic standard Normal distribution[107]. However, since these methods were developed in the field of educational measurement, where interventions are assessed using very large samples, the behaviour of Z_L for the smaller samples common in RCTs is as yet unknown.

A simulation study

In this study, we simulated data from RCTs to examine the behaviour of Z_L and the power to detect a number of effect sizes with a given number of patients and items. We were particularly interested in RCTs where there were 30, 40, 50, 100, 200, 300, 500 or 1000, denoted K, patients in each arm and where these patients were assessed using 5, 10, 15, 20, 30, 50, 70 or 100, denoted n, items each with two response categories. These values were chosen to reflect the range of sample sizes often encountered in clinical research and the number of items, with which it is acceptable to assess patients in a variety of situations. In total, there were 64 ($= 8 \times 8$) different combinations of sample size and number of items in the study.

The study was carried out by simulating 1000 'RCTs' at each of the 64 different combinations. In each RCT, a group of K values of θ were sampled from each of $N(0,1)$ and $N(\mu_B, 1)$ distributions, to represent the latent trait levels of patients in groups A and B, respectively. Since the standard deviation of θ is equal to one in both groups, the effect size in a RCT is equal to $\mu_A - \mu_B$. In addition, for each RCT n values of β and of α were generated from $N(0,1)$ and $\log N(0.2, 1)$ distributions, respectively, to represent the items. These values were chosen as they give a reasonably high level of statistical information on the values of θ used in this study and thus represent a carefully chosen questionnaire in terms of item characteristics. Each RCT was 'conducted' by calculating the probability, p_{ik}, that patient k would respond to item i in category '1', given their value of the latent trait, θ_k, and the item parameters α_i and β_i using the formula in equation (7.1). The response, x_{ik}, 'made' by patient k on item i was obtained by taking an observation on a $Bi(1, p_{ik})$ distribution. This was repeated for all $2K$ patients in a RCT and resulted in a data matrix with $2K$ rows and n columns. The responses, x_{ik}, were used, together with the two-parameter logistic model, the 'known' values of α and β and marginal maximum likelihood estimation methods to obtain estimates of μ_A, μ_B and the associated standard errors. These estimates were combined to obtain a value of Z_L for each RCT. The simulations were carried out in a program adapted by the authors from OPLM, a commercially available program for estimating parameters in IRT models[53].

The distribution of Z_L under the null hypothesis

In order to provide a framework for examining the small sample behaviour of Z_L, under the null hypothesis H_0: $\mu_A - \mu_B = 0$, the values of μ_A and μ_B were set equal to 0 and 1000 RCTs carried out at each of the 64 combinations of K and n. The distribution of the values of Z_L obtained for each combination of K and n

was examined by calculating the mean, $\bar{Z}_L$, and standard deviation, s.d.(Z_L), of the values of Z_L obtained from the 1000 RCTs conducted with each combination of K and n. In addition, the values of Z_L were tested to see whether there was evidence that they did not form a sample from a Normal distribution, with mean $\bar{Z}_L$ and standard deviation, s.d.(Z_L), using the Kolmogornov-Smirnov statistic.

Sample size, number of items and power

The main objective of the simulation study was to examine the power of a RCT using a questionnaire with n items as a primary endpoint and K patients in each arm of the RCT to detect minimal (0.2), moderate (0.5) or substantial (0.8) effect sizes. Hence, the whole simulation process, with 64 combinations of k and n was repeated three times. The value of μ_A was set to 0 and $\sigma_A = \sigma_B = 1$. Hence, the values of μ_B were set at 0.2, 0.5 and 0.8 for the first, second and third repetitions, respectively.

The 1000 values of Z_L obtained for each of the 192 combinations of sample size, number of items and effect size and were compared to the critical values of the appropriate Normal distribution to determine how many of the test statistics were significant at the (2-sided) 90%, 95% and 99% levels. For 'RCTs' with more than 100 patients in each arm, a standard Normal distribution was used, meaning that the critical values were ±1.64, ±1.96 and ±2.58 for the 90%, 95% and 99% levels, respectively. The critical values for 'RCTs' with up to than 100 patients in each arm were obtained in the first part of this simulation study and are given in the section 'Results of the simulation study'.

Results of the simulation study

The distribution of Z_L under the null hypothesis

The mean, $\bar{Z}_L$, s.d.(Z_L) and the P-value of the Kolmogornov-Smirnov test for Normality, P(KS), of the 1000 values of Z_L produced at each combination of n and

Table 7.1: the mean, standard deviation and Kolmogornov-Smirnov P-value, P(KS), for Z_L when $\mu_A = \mu_B = 0$. The number of items is denoted by n, the number of patients in each of groups A and B by K and the the standard deviation by sd.

n	$\bar{Z}_L$	sd(Z_L)	P(KS)	n	$\bar{Z}_L$	sd(Z_L)	P(KS)
$K = 30$				$K = 40$			
5	0.0963	1.0330	0.027	5	0.0006	1.0432	0.465
10	-0.0031	1.3092	0.012	10	0.0252	1.1260	0.230
15	-0.0058	1.2144	0.318	15	0.0202	1.1482	0.327
20	0.1366	1.2138	0.429	20	0.0246	1.2059	0.182
30	0.0764	1.1905	0.872	30	-0.0063	1.0984	0.296
50	0.0794	1.2374	0.289	50	0.0037	1.1648	0.339
70	-0.1169	1.2262	0.292	70	-0.0576	1.1728	0.294
100	0.0151	1.2146	0.374	100	-0.0005	1.1425	0.879
$K = 50$				$K = 100$			
5	0.0269	1.0886	0.570	5	0.0158	1.0267	0.263
10	0.0152	1.0405	0.639	10	0.0024	1.0015	0.826
15	-0.0050	1.1383	0.807	15	-0.0228	1.0515	0.119
20	-0.0230	1.1729	0.407	20	0.0557	1.0325	0.826
30	-0.0115	1.1075	0.320	30	-0.0255	1.0175	0.853
50	-0.0222	1.1356	0.571	50	0.0336	1.0484	0.336
70	-0.0252	1.0612	0.194	70	0.0178	0.9801	0.776
100	-0.0070	1.1250	0.728	100	0.0097	1.0267	0.615
$K = 200$				$K = 300$			
5	-0.0197	1.0406	0.970	5	-0.0068	0.9882	0.322
10	0.0342	1.0428	0.860	10	-0.0634	0.9778	0.610
15	-0.0561	1.0167	0.860	15	-0.0142	1.0064	0.652
20	-0.0051	0.9980	0.472	20	-0.0319	0.9914	0.515
30	0.0112	1.0263	0.567	30	0.0156	1.0141	0.966
50	-0.0114	1.0102	0.980	50	0.0144	0.9661	0.471
70	0.0142	0.9972	0.687	70	-0.0364	0.9680	0.289
100	0.0126	0.9889	0.541	100	0.0382	1.0377	0.561
$K = 500$				$K = 1000$			
5	0.0086	1.0181	0.793	5	-0.0420	1.0183	0.959
10	-0.0287	1.0516	0.994	10	-0.0713	0.9775	0.302
15	-0.0223	0.9574	0.782	15	0.0069	1.0059	0.219
20	-0.0146	0.9898	0.446	20	0.0896	0.9679	0.942
30	0.0114	0.9612	0.782	30	-0.0235	0.9465	0.317
50	-0.0201	0.9817	0.362	50	-0.0330	1.0285	0.894
70	-0.0227	0.9950	0.153	70	-0.0286	1.0209	0.279
100	-0.0114	0.9898	0.697	100	0.0094	0.9488	0.661

K for $\mu_A = \mu_B = 0$ are given in Table 7.1. The mean value of Z_L over all 64000 replications is 0.0004 and only two of the 64 values of P(KS) are less than 0.10. This indicates that there is no reason to suspect that Z_L does not attain it asymptotic Normal distribution, with mean 0, for samples consisting of two groups, each of between 30 and 1000 patients.

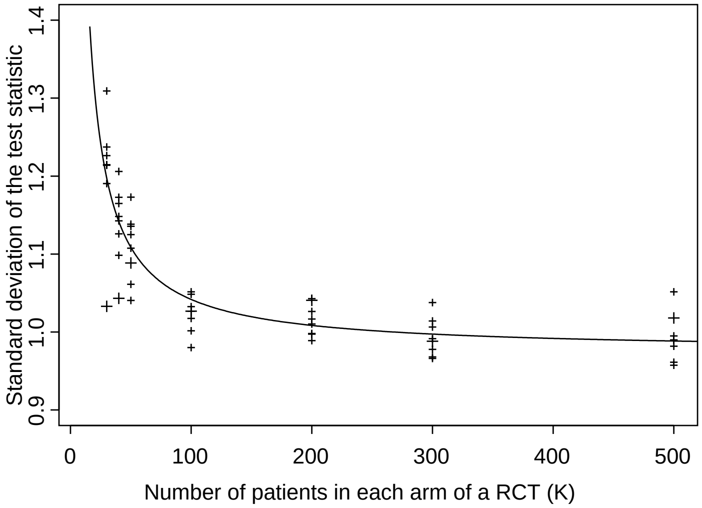

Figure 7.1: Standard deviation of Z_L against the number of patients in each arm of a RCT

Table 7.2: Modelled distribution for Z_L and critical values for RCTs with small sample sizes

Number of patients per arm (K)	Modelled distribution of Z_L	Critical values 90%	95%	99%	Type I error rate 90%	95%	99%
30	$N(0, 1.19^2)$	±1.96	±2.33	±3.07	0.1016	0.0565	0.0139
40	$N(0, 1.14^2)$	±1.88	±2.23	±2.94	0.1005	0.0545	0.0139
50	$N(0, 1.10^2)$	±1.81	±2.16	±2.83	0.1025	0.0558	0.0158
100	$N(0, 1.04^2)$	±1.71	±2.04	±2.68	0.0933	0.0470	0.0086
≥200	$N(0, 1)$	±1.64	±1.96	±2.58			

The variation of the standard deviation of the 1000 values of Z_L produced at

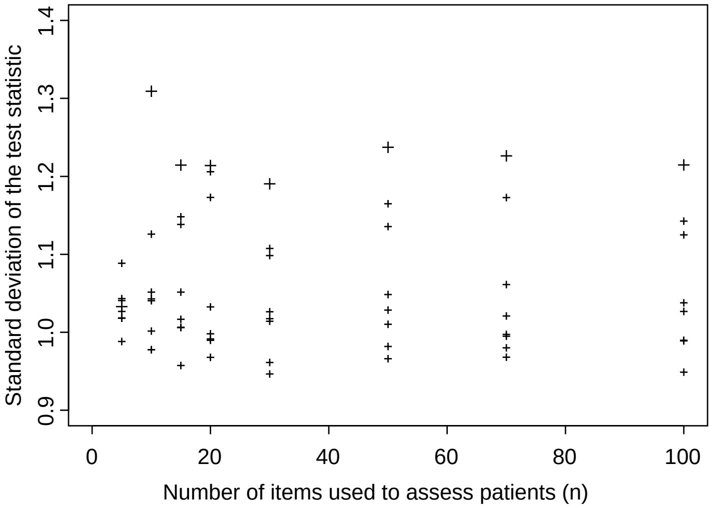

Figure 7.2: Standard deviation of Z_L against the number of items used to assess the patients

each combination of conditions is illustrated with respect to K and n in Figures 7.1 and 7.2, respectively. It does not appear that n has a substantial effect on the standard deviation of Z_L. However, s.d.(Z_L) increases as the sample size decreases. We modelled the relationship between s.d.(Z_L) and K, under the null hypothesis H_0: $\mu_A = \mu_B$, using

$$\text{s.d.}(Z_L) = 0.97 + 6.67\frac{1}{K} \tag{7.6}$$

Equation 7.6 was obtained using regression analysis, since the variance of statistic is often related to the reciprocal of the number of patients used to calculate the statistic. The usual assumptions were tested and the data did not violate them. The asymptote was not forced down to 1, since Equation 7.6 gave the best fit to data obtained from 30, 40, 50 and 100 patients, and primary interest was in studies with these numbers of patients, rather than obtaining an accurate representation of smaller deviations from a $N(0, 1)$ distribution. The correlation between s.d.(Z_L) and $\frac{1}{K}$ was 0.8843, meaning that $R^2 = 0.7820$. The modelled distribution for Z_L along with critical values and type I error rates based on the 8000 'trials' with an effect size of 0.0 carried out using the eight test lengths for the 90%, 95% and 99% levels for RCTs using 30, 40, 50 and 100 patients per arm are given in Table 7.2.

Sample size, number of items and power

Tables 7.3, 7.4 and 7.5 contain the number of the 1000 RCTs carried out under each of the 192 combinations of K, n and effect size, which resulted in a value of the Z_L statistic beyond the appropriate critical values given in Table 7.2. In order to facilitate comparisons with the situation where θ were known, similar to obtaining an estimate using an infinite number of items, the theoretical number of 1000 RCTs, which would result in a value of the T statistic beyond the appropriate critical values have been calculated using standard procedures[142]. These are presented in rows labelled ∞. If the numbers in Tables 7.3, 7.4 and 7.5 are divided by 1000, then an estimate of the power under the given combination of factors is obtained. It can be seen that the power to detect the given effect size increases with the effect size and the number of items used to asses the patients. It is also apparent that while increasing the number of items used to assess the patients from five to ten and from ten to twenty results in substantial increases in the power, increasing the number of items beyond thirty increases this figure only minimally. In addition, the power obtained using 100 items does not even approach the theoretical maximum, using an infinite number of items, for $K \leq 40$, indicating that the correction introduced in Table 7.2, may not be sufficient for very small RCTs.

Using the results in Tables 7.3, 7.4 and 7.5 and linear interpolation, the values of K required to detect the effect size at the two-sided 5% level and with a power of 80% using a given n was calculated and displayed in Table 7.6. The significance level and power were chosen as these values are regularly used when analysing RCTs. It can be seen that, as long as at least 20 items are used, the number of items barely effects the number of patients required in each arm of a RCT to detect effect sizes of 0.5 and 0.8 with a power of 80%. However, the number of items used to assess

Table 7.3: number of 1000 'RCTs' in which Z_L was greater than the appropriate critical value for the two-sided significance level given for effect size 0.2. The number of items used is denoted by n and the number of patients in each of group A and group B K.

	Significance level								
n	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
	$K = 30$			$K = 40$			$K = 50$		
5	126	64	24	136	76	18	149	77	26
10	145	74	15	139	72	31	188	108	28
15	141	87	24	133	81	30	208	117	27
20	155	91	28	163	99	29	180	126	37
30	158	86	24	186	109	36	190	130	33
50	170	109	25	158	85	28	210	126	41
70	157	78	33	181	117	44	215	120	36
100	139	96	37	165	100	33	211	138	48
∞	190	110	30	220	140	40	260	160	50
	$K = 100$			$K = 200$			$K = 300$		
5	216	139	42	372	268	112	515	375	164
10	276	174	63	492	350	170	625	508	266
15	305	223	74	536	395	189	675	556	317
20	339	215	66	539	424	209	688	565	334
30	344	236	87	561	414	211	736	619	394
50	336	227	87	598	462	267	773	657	409
70	345	247	98	585	473	246	750	659	401
100	365	258	100	612	489	262	765	659	414
∞	400	290	120	630	510	270	780	680	440
	$K = 500$			$K = 1000$					
5	649	523	328	900	822	610			
10	810	712	474	970	949	825			
15	849	763	534	982	963	896			
20	880	802	564	992	980	923			
30	901	827	625	995	990	929			
50	904	853	673	994	989	943			
70	926	871	718	998	995	962			
100	920	879	697	994	991	966			
∞	930	880	710	1000	1000	970			

patients has more effect on the number of patients required to detect an effect size of 0.2 with a power of 80%. For instance, if only five randomly selected items are used, it is necessary to include 950 patients in each arm, but if 50 items are used, 450 are required in each arm.

Table 7.4: number of 1000 'RCTs' in which Z_L was greater than the appropriate critical value for the two-sided significance level given for effect size 0.5. The number of items used is denoted by n and the number of patients in each of group A and group B K.

	Significance level								
n	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
	$K = 30$			$K = 40$			$K = 50$		
5	174	104	30	344	247	90	433	314	126
10	267	181	69	438	297	132	575	445	208
15	280	170	91	492	364	145	579	463	225
20	303	185	70	467	349	140	606	488	267
30	294	193	88	550	416	180	650	510	283
50	291	208	89	559	426	208	622	492	274
70	305	208	67	549	428	213	663	536	287
100	317	208	88	557	415	197	674	550	296
∞	600	470	240	710	590	340	790	690	450
	$K = 100$			$K = 200$			$K = 300$		
5	724	610	361	960	924	802	985	976	931
10	837	742	542	994	985	923	1000	998	986
15	876	800	611	998	992	957	1000	1000	997
20	905	854	702	995	990	965	1000	1000	999
30	927	862	704	997	991	967	1000	1000	1000
50	939	885	706	995	994	982	1000	1000	1000
70	947	900	735	1000	997	991	1000	1000	1000
100	935	889	727	996	995	986	1000	1000	1000
∞	960	940	820	1000	1000	1000	1000	1000	1000

Table 7.5: number of 1000 'RCTs' in which Z_L was greater than the appropriate critical value for the two-sided significance level given for effect size 0.8. The number of items used is denoted by n and the number of patients in each of group A and group B K.

	Significance level								
n	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
	$K = 30$			$K = 40$			$K = 50$		
5	389	262	98	655	538	297	805	689	422
10	497	354	163	774	669	425	900	852	639
15	489	384	179	846	747	482	911	817	649
20	475	371	194	872	774	497	931	892	690
30	538	416	205	856	789	581	950	897	750
50	573	461	217	875	804	574	933	903	771
70	535	428	229	862	773	549	952	919	799
100	571	427	214	891	822	618	940	899	775
∞	920	860	660	970	940	820	1000	970	910
	$K = 100$			$K = 200$					
5	973	949	815	1000	999	995			
10	995	991	956	1000	1000	1000			
15	1000	997	984	1000	1000	1000			
20	997	994	985	1000	1000	1000			
30	999	999	993	1000	1000	1000			
50	1000	1000	992	1000	1000	1000			
70	998	998	994	1000	1000	1000			
100	1000	1000	995	1000	1000	1000			
∞	1000	1000	1000	1000	1000	1000			

Table 7.6: Approximate number of patients required in each arm (K) of a RCT to demonstrate a given effect size at the 5% level with 80% power

	Number of items used					
Effect size	5	10	20	50	100	∞
0.2	950	680	500	450	440	394
0.5	160	125	95	90	87	64
0.8	70	50	45	40	39	26

An illustration using the Short Form instruments from the Medical Outcomes Study

In order to illustrate the methods described in this paper, data were used from the Netherlands Co-operative Study on Dialysis (NECOSAD). The Medical Outcomes Study Short-Form Health Survey with 36 items (SF36) was developed, from the original Medical Outcomes Study survey[143], to measure health status in a variety of research situations[134]. The SF36 is a reliable measure of health status in a wide range of patient groups[25] and has been used with classical[144] and IRT based[30] psychometric models. In addition, the SF12, with 12 items [139], and a preliminary version of the SF8, with 9 items[140], have been developed. In this article, the Dutch language version of the SF36[145] was used. The SF36 uses between two and six response categories per item and, on the majority of items, a higher score denotes a better health state. The number of scoring categories is one more than the number of β parameters per item. Items 1, 2, 6, 7, 8, 9a, 9d, 9e, 9h, 11b and 11d were re-scored so that a higher score denotes a better health state on all items. A brief description of each item is given in Table 7.7, while full details are available elsewhere[19]. The items in the SF12 are 1, 3b, 3d, 4b, 4c, 5b, 5c, 6, 8, 9d, 9e and 9f. The items in the SF8 are 1, 2, 3d, 4d, 5b, 6, 7, 9e and 9f. The measurement properties of the SF36 in a sample of patients with chronic kidney failure 12 months after the start of dialysis[137, 138] have been examined using the generalised partial credit model described in Equation 7.2 and are summarised in Table 7.7. Questionnaires were sent to 1046 patients, of whom 978 completed at least one question on the SF36. Of the 978 patients, 583 were male and 395 were female, while 615 were on hemodialysis and 363 on peritonital dialysis. The patients were aged between 18 and 90 years, with a median of 61 years. When IRT techniques were used to examine the quality of life, the mean, μ, was 0 and the standard deviation was $\sigma = 0.692$. It should be emphasised that the model parameters given in this article are for illustration purposes only and should not be regarded as a definitive calibration of these items as the fit of the model has not been extensively tested, particularly with respect to differential item characteristics between subgroups of patients.

Table 7.7: The characteristics of the items in the SF36 questionnaire. The item number is denoted by 'IN'. Estimates of the item parameters α, β_1, β_2, β_3, β_4 and β_5 are given.

IN	Item content	α	β_1	β_2	β_3	β_4	β_5
1	General health status	2.64	-3.89	-3.81	0.09	4.83	
2	Compared with year ago	0.62	-1.76	-2.86	-2.01	-0.93	
	Does your health limit						
3a	vigorous activities	1.73	1.30	4.50			
3b	moderate activities	2.77	-1.03	1.02			
3c	lifting or carrying groceries	2.25	-1.37	-0.03			
3d	climbing 2+ flights of stairs	2.35	-0.89	0.61			
3e	climbing one flight of stairs	2.28	-2.00	-2.13			
3f	kneeling or stooping	1.92	-1.27	-0.38			
3g	walking more than a mile	2.16	-0.20	0.81			
3h	walking several blocks	2.25	-1.21	-1.30			
3i	walking one block	2.08	-1.87	-2.52			
3j	bathing or dressing yourself	2.41	-3.78	-5.50			
	Problems with work as a result of physical health						
4a	worked less	2.77	0.44				
4b	did less than hoped	3.09	1.08				
4c	limited in kind of work	3.50	1.29				
4d	difficultly in working	3.00	1.39				
	Problems with work as a result of emotional health						
5a	worked less	2.59	-0.44				
5b	did less than hoped	2.73	0.05				
5c	worked less carefully	2.30	-0.68				
6	Reduced social activities	1.83	-2.62	-4.43	-5.87	-5.59	
7	Amount of bodily pain	0.84	-1.32	-3.20	-3.19	-3.13	-3.24
8	Pain interfere with work	1.66	-2.29	-4.06	-4.93	-4.64	
	Emotions felt						
9a	full of pep	0.63	-1.48	-2.64	-1.70	-2.60	-1.24
9b	very nervous	0.62	-1.06	-2.26	-3.76	-3.78	-3.53
9c	down in the dumps	1.22	-2.45	-4.76	-6.61	-7.13	-7.29
9d	calm and peaceful	0.99	-1.93	-3.27	-2.92	-3.58	-2.03
9e	lots of energy	1.35	-1.74	-2.17	-1.10	-0.58	1.88
9f	downhearted and blue	1.09	-1.63	-3.58	-5.40	-5.36	-4.95
9g	worn out	1.45	-2.01	-3.83	-4.64	-3.52	-1.93
9h	a happy person	0.59	-1.40	-2.82	-2.47	-3.44	-2.61
9i	tired	1.51	-1.23	-2.11	-2.24	0.02	2.46
10	Reduced social activities	1.70	-2.33	-3.96	-3.62	-2.91	
	Rate relevance of statements						
11a	Sicker than other people	0.67	-0.77	-1.28	-0.66	-0.35	
11b	As happy as others	0.76	-0.56	-0.01	0.11	1.95	
11c	Expect health to worsen	0.64	-1.21	-1.89	-0.12	0.52	
11d	Health is excellent	1.08	-0.53	0.70	0.00	2.44	

In order to examine the power, with which studies using the Short Form instruments could detect treatment effects in a population of patients with chronic kidney failure 12 months after starting dialysis, a simulation study was carried out. The simulations were carried out in a similar way to those described in Section 7 of this article. We were interested in detecting effect sizes of 0.2, 0.5 and 0.8, denoted ξ, in RCTs with 30, 40, 50, 100, 200, 300, 500 or 1000 patients in each arm (K) using the SF36, the SF12 or the SF8 as the primary endpoint, meaning that there were 24 different combinations of ξ and K. In each run, K values of θ were sampled from each of $N(\mu, \sigma^2)$ and $N(\mu + \xi\sigma, \sigma^2)$, where ξ represents the effect size expected in the study and σ the standard deviation of quality of life, to denote the control and treatment groups, respectively. These values were combined with the item parameters in Table 7.7 and Equation 7.2 to generate 'responses' by 'patients' to the SF36. These data were used to obtain the statistic Z_L using the items in the SF36, in the SF12 and in the SF8. One thousand runs, or RCTs, were carried out at each of the 24 combinations of K and ξ. It was assumed that the item parameters were known and equal to those in Table 7.7.

The number of runs at each combination of K, ξ and Short Form instrument resulting in a value of Z_L more extreme than the appropriate critical value, in Table 7.2, for known item parameters is recorded in Table 7.8. In general, the SF36 detected a given treatment effect more often than the SF12, which in turn detected a treatment effect more often than the SF8. The number of patients needed in each arm of a RCT using the SF36, SF12 and SF8 to detect standard treatment effects in the underlying latent trait with a power of 80% were obtained using similar graphical methods as in Section 7 and are given in Table 7.9. It appears that the number of patients required when using the SF36 to detect an effect of 0.2 is less than if θ were known. This is not true, but is a result of the inherent error involved in a simulation study with only 1000 trials at each combination of K and ξ. It can also be seen that the differences between the numbers of patients required in each arm are less marked than in Table 7.6. This is because the items on the SF36 have up to 6 response categories leading to a total of 149, 47 and 40 different possible sum scores

Table 7.8: The number of 1000 RCTs using the SF36, SF12 and SF8, in which Z_L was greater that the appropriate critical value for a test at the two-sided significance level given. The number of patients in each of groups A and B is denoted by K.

Effect					Significance level					
size		0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
		$K = 30$			$K = 40$			$K = 50$		
0.2	SF36	155	92	29	210	137	38	252	168	57
0.2	SF12	146	80	25	197	110	34	240	160	70
0.2	SF8	152	84	23	184	114	32	221	138	49
		$K = 100$			$K = 200$			$K = 300$		
0.2	SF36	430	312	156	666	561	353	841	782	549
0.2	SF12	415	295	141	636	515	300	808	739	525
0.2	SF8	385	265	116	637	516	288	772	687	489
		$K = 500$			$K = 1000$					
0.2	SF36	952	917	786	998	995	982			
0.2	SF12	928	891	751	997	997	972			
0.2	SF8	936	891	742	995	990	959			
		$K = 30$			$K = 40$			$K = 50$		
0.5	SF36	493	352	147	690	571	332	769	695	452
0.5	SF12	495	363	166	618	513	284	765	659	392
0.5	SF8	469	314	142	611	490	260	750	627	391
		$K = 100$			$K = 200$					
0.5	SF36	976	949	838	1000	998	988			
0.5	SF12	960	933	838	998	997	995			
0.5	SF8	944	895	726	999	999	993			
		$K = 30$			$K = 40$			$K = 50$		
0.8	SF36	858	755	538	961	914	785	980	955	897
0.8	SF12	834	724	500	945	893	716	986	964	856
0.8	SF8	840	714	451	925	860	663	977	959	858
		$K = 100$								
0.8	SF36	1000	999	997						
0.8	SF12	999	999	997						
0.8	SF8	999	998	989						

on the SF36, SF12 and SF8 respectively. The items used in the first simulation study described in this paper had only two response categories, meaning that 149, 47 and 40 items would have been necessary to obtain the same range of possible sum scores.

Table 7.9: Approximate number of patients required in each arm (K) of a RCT to demonstrate a given effect size at the 5% level with 80% power

	Instrument used			
Effect size	SF8	SF12	SF36	∞
0.2	410	380	325	394
0.5	82	75	71	64
0.8	36	35	33	26

Discussion

This paper has described simulation based studies into the the use of IRT to analyse RCTs using a questionnaire consisting of items with two response categories and designed to quantify a theoretical variable as the primary outcome. The study into the small sample behaviour of asymptotic results, provides a framework, in which data from RCTs can be analysed using IRT. It had been proven that the Z_L statistic closely approximated a $N(0,1)$ distribution under the null hypothesis for studies in which each arm contained at least 500 patients[107]. The results in this paper show that Z_L follows a $N(0,\sigma)$ distribution for small RCTs, but that $\sigma \neq 1$. The variance appears to be a function of the reciprocal of the number of patients in each arm of a RCT. This means that, as in many situations, in which asymptotic results are used, the procedure for testing whether an effect is significant needs to be adjusted for the relatively small sample sizes often found in RCTs when IRT is used.

The main simulation study examined the relationship between the number of patients in each arm of a RCT, the number of items used to assess the patients and the power to detect given effect sizes when using IRT. As ever, the smaller the effect size, the larger the number of patients needed in each arm to detect the effect with a given power. In addition, increasing the number of items used to assess the patients can mean that given effects can be detected using fewer patients. The number of patients required to detect a minimal effect using five items is more than twice the number required when 50 items are used. The results suggest that reductions in the number of patients required are minimal for more than 20 carefully chosen items, indicating that a maximum of 20 items, with good measurement qualities, is sufficient to assess patients. However, if the items have poor measurement properties, such as very low values of α_i or values of β_i more extreme than $\mu_\theta \pm 2\sigma_\theta$, indicating a lack of discrimination and floor or ceiling effects in a questionnaire, respectively, then many more items may be required. These results also indicate that if a RCT is to be designed to detect small effects, it is inadvisable to use very short instruments consisting of items with two response categories analysed using IRT. Again, it should be noted that these results are based on the assumption that the items used do not have poor measurement properties, as discussed above.

It is common wisdom that, to detect effect sizes of 0.2, 0.5 or 0.8 using a t-test with a significance level of 0.05 and a power of 80%, it is necessary to include 394, 64 or 26 patients, respectively, in each arm of a RCT. The results presented in this paper, showed that 450, 90 and 40 patients are required in each arm to detect the same effects using 50 items and IRT. The main reason for the differences between these figures is that when a t-test is used, it is assumed that the variable of interest is measured without error, where as IRT takes into account that a latent variable cannot be measured without error. In addition, the sample sizes obtained using IRT are conservative as linear interpolation has been used between observations, whereas it is reasonable to assume that the true function would curve towards the top left hand corner, indicating that the true numbers of patients required are marginally lower than these results suggest. It should also be emphasised that, if researchers require very accurate estimates of the number of patients required in each arm of a RCT using a particular set of items, they should carry out their own simulation study. There are a number of advantages of the use of IRT in the analysis of RCTs. Firstly, the Z_L test described in this article can, in contrast to the t-test, always be used in the same format, regardless of whether the variance of the latent trait is the same in both arms of the RCT or not. Secondly, it is not essential that all patients are assessed using exactly the same questionnaire. The questionnaire can be shortened for particular groups of patients, whilst the estimates of the latent trait remain comparable over all patients.

The value of the standard errors of μ_A and μ_B depend on a complex relationship between the values of the latent trait of the patients included in a RCT and the parameters of the items used to assess them. This means that the results obtained in this study are, theoretically, local to the combination of patients and items used. However, we re-selected all parameters for each of the 1000 individual 'RCTs' carried out at each combination of effect size, number of patients per arm and number of items, in order to give a more general picture of the number of patients required. In addition, item parameters for the SF36, SF12 and SF8 quality of life instruments were estimated from a data set collected from patients undergoing dialysis for chronic kidney failure and used to illustrate the methods described.

This article has examined the sample sizes required in a RCT, when the primary outcome is a latent trait measured by a questionnaire analysed with IRT. It has been proven that it is possible, following a minor transformation, to use the statistics developed for analysing large scale educational interventions in the much smaller sample sizes encountered in RCTs in clinical medicine. In addition, the relationship between the number of items used and the number of patients required to detect particular effects was examined. It is hoped that this article will contribute to the understanding of IRT, particularly in relation to RCTs.

Appendix: obtaining $\hat{\mu}_A$, $\hat{\mu}_B$ and their standard errors

The values of μ_A and μ_B and their standard errors can be estimated using marginal maximum likelihood methods. The likelihood L can be written as

$$L = \prod_{g=A,B} \prod_{k=1}^{K_g} \int_{-\infty}^{\infty} \prod_{i,j} p_{ijk}(\theta)^{x_{ijk}} g(\theta|\mu_g, \sigma_g) d\theta \qquad (7.7)$$

where p_{ijk} is as defined in Equations (7.1) and (7.2), $g(\theta|\mu_g, \sigma_g)$ is a Normal density function with mean and standard deviation equal to μ_g and σ_g^2 for $g = A, B$, respectively. Furthermore, x_{ijk} is an indicator variable taking the value 1 if patient k responds in category j of item i and the value 0 otherwise. In the study described in this paper, we have assumed that the item parameters α_i and β_i are known. In practice, this would mean that the items came from a calibrated item bank. It has been shown that μ_A can be estimated using

$$\hat{\mu}_A = \frac{1}{K}\sum_{k=1}^{K} E(\theta_k|X_k) \tag{7.8}$$

where X_k is a vector with n elements containing the responses, x_{ik}, of patient k to the items and $E(\theta_k|X_k)$ is the posterior expected value of θ_k for patient k given their pattern of responses, X_k[83, 107, 141]. It has also been shown that, if the item parameters are considered known, as in this study, the asymptotic standard error s.e.$(\hat{\mu}_A - \hat{\mu}_B)$ is equal to s.e.$(\hat{\mu}_A) +$ s.e.$(\hat{\mu}_B)$, where

$$\text{s.e.}(\hat{\mu}_A) = \frac{1}{\hat{\sigma}_A^{-4}\sum_{k=1}^{K}(E(\theta_k|X_k) - \hat{\mu}_A)^2} \tag{7.9}$$

and $\hat{\sigma}_A$ is the estimated standard deviation of θ in group A. The definition of s.e.$(\hat{\mu}_B)$ is analogous to that of s.e.$(\hat{\mu}_A)$. The expectation $E(\theta_k|X_k)$ is as defined in equation (7.8) and given by

$$E(\theta_k|X_k) = \int_{-\infty}^{\infty} \theta f(\theta_k|X_k)\partial\theta \tag{7.10}$$

where the posterior distribution of θ is

$$f(\theta_k|X_k) = \frac{P(X_k|\theta_k)g(\theta_k|\mu_A, \sigma_A)}{\int_{-\infty}^{\infty} P(X_k|\theta_k)g(\theta_k|\mu_A, \sigma_A)\partial\theta_k} \tag{7.11}$$

where $P(X_k|\theta_k)$ is the probability of the response pattern made by patient k given θ_k.

Chapter 8

How does item selection procedure affect power and sample size when using an item bank to measure health status?

This chapter is adapted from the following article:

> Holman R. How does item selection procedure affect power and sample size when using an item bank to measure health status? *Quality of Life Newsletter* 2004; **Special issue**; 9–11.

and is available from `http://www.mapi-research-inst.com`

Introduction

In the last two decades, there has been an enormous increase in interest in examining how treatment regimes affect the patient as a whole, rather than as a collection of physiological parameters. The effect on the patient as a whole is often expressed on a latent trait, such as ‘quality of life’, ‘cognitive ability’ or ‘functional health status’, and measured using a questionnaire[19]. In order to obtain an interpretable estimate of a latent trait, it is usually necessary to administer a questionnaire in its entirety to all patients in a study. Shorter versions of widely used instruments, such as the Health Outcomes Study Short Form Instruments[25, 139, 140] and Health Assessment Questionnaire[146], are available. These clearly reduce the burden of testing on patients and researchers, but it can be difficult to compare studies using different versions of the same instrument[146, 10]. Item banks form an attractive alternative to fixed length tests. An item bank is a collection of items, all measuring the same construct, for which the individual measurement properties are known[124, 125, 7]. Once an item bank has been constructed, it is possible to assess the health status of patients using selections of items. For example, an ‘easy’ set of items can be chosen for patients with a severe manifestation of a condition and a ‘difficult’ set of items for patients with a mild manifestation. It is even possible to select the ‘best’ items for individual patients using computerised adaptive testing algorithms[6, 147]. Furthermore, results from studies using different selections of items from an item bank can be directly compared. A number of item banks are being developed for clinical applications. These measure concepts such as quality of life[124, 126], the impact of headaches[127] or functional health status[7].

The majority of item banks are developed using item response theory (IRT). IRT techniques have been accepted in many areas of medical research, including cognitive screening[23], physical functioning[25] and mobility[26]. Item response theory and other techniques associated with the construction, maintenance and implementation of item banks have been developed in educational theory. This means that the majority of work has concentrated on the challenges found in that field. In particular, work has concentrated on assessing the ability of individuals rather than comparing groups, as occurs in the majority of clinical studies.

This chapter uses a simulation study to examine the power, with which standard treatment effects can be detected, when using an item bank to measure functional health status[7, 8] in four ways. Firstly, all items in an item bank are administered to the patients. This is equivalent to treating the item bank as a fixed length instrument. Secondly, two selections of items spanning the whole range of the item bank are administered. This is similar to the way many shortened versions of well known instruments are constructed. Thirdly, a selection of items chosen to be appropriate for the patient population will be made. Finally, items clearly inappropriate for the patient population will be used. This is close to the situation, which arises, when an instrument is used, which was developed for a very different patient populations. The selections of items are analysed using IRT based techniques. The results are presented in terms of the power to detect a treatment effect when a range of numbers of patients in each arm of the RCT are used. In addition, guidelines are given on the approximate numbers of patients required in each arm of an RCT when using different item selection procedures.

Methods

A straightforward randomised clinical trial

In a straightforward RCT, the patient sample is randomly divided into two equally sized groups. The members of each group receive a different treatment regime and all outcomes are assessed once, at the end of the study. Clinically, the two groups are said to differ if the ratio of the difference between the mean health status in the two groups and the standard deviation of health status in the population under consideration is larger than a given treatment effect size. Interest is often in examining a continuous variable and the, arbitrarily defined, minimal, moderate and substantial treatment effect sizes of 0.2, 0.5 and 0.8, respectively. The number of patients required to detect a given effect size with a particular power depends on the values of the effect size, the standard errors of the estimates of mean health status in the two groups and the significance level used[135].

Item response theory

Item response theory models can be used to describe the probability that a patient will respond to a set of items in a particular way[133, 23]. The two-parameter logistic IRT model[39] was developed for data resulting from items with two response categories, say ‘cannot’ and ‘can’. In this model, the probability, P_{ik}, that patient k, with health status equal to θ_k, will respond to item i in the category ‘can’ is given by

$$P_{ik} = \frac{\exp\left(\alpha_i(\theta_k - \beta_i)\right)}{1 + \exp\left(\alpha_i(\theta_k - \beta_i)\right)} \tag{8.1}$$

where α_i and β_i are item parameters, which describe the measurement characteristics of item i in relation to the construct ‘health status’. The larger the value of β_i the more difficult an item is and the larger the value of α_i the better the item can discriminate between levels of health status better or worse than β_i.

In IRT, the item and patient parameters are usually estimated in a two stage procedure[83]. Firstly, the item parameters are estimated, often by assuming that the health status of patients follows some known distribution and integrating them out of the likelihood. Secondly, maximum likelihood estimates of the health status of individual patients are obtained using the previously estimated item parameters. Concurrent estimation of the item parameters and the level of health status in a population using the two-parameter logistic IRT model less than 200 patients can lead to substantial inflation of the standard errors of the mean health status[40], meaning that RCTs with fewer than 100 patients per treatment arm can only be analysed using this model if the item parameters have been estimated from a previous data set and are assumed known[5]. Since a smaller number of patients are included in many RCTs, this study concentrates on the use of IRT when a calibrated item bank is available. When using IRT based techniques, it is inadvisable to estimate health status, θ, for each individual patient then perform standard analysis, such as t-tests, on these estimates, since this ignores the inherent estimation error in the latent variable[78]. Instead, consistent estimates of the mean health status in each treatment group should be obtained directly from the data using marginal maximum likelihood methods[107, 141]. The difference between these estimates can be divided by its standard errors to obtain test statistic Z_L. It has been shown that this statistic follows an asymptotic standard Normal distribution under the null hypothesis that the mean health status is equal in both groups[107].

The AMC Linear Disability Score project

The AMC Linear Disability Score project, currently being carried out in our hospital, was set up to construct an item bank to measure the functional health status, defined as the ability to perform 'activities of daily life'[7, 8]. In the project, two item response categories are used: 'I can carry out the activity' and 'I cannot carry out the activity'. The parameters associated with the 60 items used in this article, given in Tables 8.1 and 8.2, were obtained from the responses of 2513 patients spanning all levels of disability using the two-parameter logistic IRT model in two commercially available programs[40, 53]. Since each patient was assessed using only a selection of the items, the number of responses to each item varied between 133 and 866. The parameters for these items are given to illustrate the type of parameters obtained for this type of item and are not to be regarded as the definitive parameters for these items. This study concentrates on patients living independently in their own homes without professional care. The mean and standard deviation of the distribution of functional status in the sub-group of the 1179 such patients, who participated in the calibration study, were 1.447 and 0.705, respectively.

Table 8.1: A selection of items from the AMC Linear Disability Score project. The parameters α and β denote the discrimination and difficulty parameters, respectively. 'WR' denotes the selection of items covering the whole range of the item bank, 'S' the selection chosen to be suitable for the patient population and 'US' the selection chosen to be unsuitable for the patient population.

				WR		S		US	
	Item content	α	β	10	20	10	20	10	20
1	Answering the phone	0.49	-6.90						+
2	Drinking from a beaker	1.02	-5.52		+			+	+
3	Opening a drawer	0.48	-5.16						
4	Walking between rooms	1.19	-4.47						+
5	Sitting up in bed	0.62	-4.26	+	+			+	+
6	Opening a door	2.16	-3.72						
7	Sitting on the edge of a bed	0.80	-3.71						+
8	Moving between two chairs	1.71	-3.17		+			+	+
9	Peeling an apple	1.59	-2.99						
10	Putting on a shirt	1.61	-2.80						+
11	Cleaning a bathroom sink	1.33	-2.67	+	+			+	+
12	Reaching (low cupboard)	1.33	-2.61						
13	Using a lift	2.03	-2.59						+
14	Opening a window	1.48	-2.58		+			+	+
15	Fetching and opening letters	1.95	-2.46						
16	Making a pot of coffee	1.43	-2.30						+
17	Opening a soft drink bottle	1.21	-1.93	+	+			+	+
18	Picking something up	1.09	-1.91						
19	Making porridge	1.90	-1.89						+
20	Cutting your finger nails	1.33	-1.83		+			+	+
21	Cleaning a toilet pot	1.00	-1.57						
22	Warming up tinned soup	3.06	-1.41						+
23	Clearing the table	1.65	-1.40	+	+			+	+
24	Putting lace-up shoes on	0.69	-1.35						
25	Frying an egg	2.62	-1.29						+
26	Making a bed	1.38	-1.25		+			+	+
27	Getting into and out of a car	0.99	-1.18						
28	Putting flowers in a vase	1.95	-1.10						+
29	Shopping for a loaf of bread	0.87	-0.97	+	+			+	+
30	Preparing a warm meal	0.50	-0.75						
31	Eating in a restaurant	0.71	-0.53						
32	Reaching under table	1.32	-0.53		+				
33	Moving between easy chairs	0.99	-0.53						
34	Reaching (high cupboard)	1.21	-0.45						
35	Using a dustpan and brush	1.35	0.03	+	+				
36	Going up a flight of stairs	1.22	0.20						
37	going to the bottle bank	0.75	0.29						
38	Posting a letter	0.67	0.43		+				

Table 8.2: *Table 8.1 continued*

				WR		S		US	
	Item content	α	β	10	20	10	20	10	20
39	Hanging out the washing	0.93	0.52						
40	Opening a high window	0.96	0.68				+		
41	Standing for 10 minutes	1.00	0.99	+	+	+	+		
42	Cleaning a fridge	1.35	1.41				+		
43	Going to a birthday party	0.64	1.42			+	+		
44	Travelling by bus or tram	1.00	1.51		+		+		
45	Carrying a tray with cups	1.11	1.56			+	+		
46	Mopping the floor	0.79	1.57				+		
47	Travelling by train	1.32	1.85	+	+	+	+		
48	Pushing a bed or table	1.08	2.26				+		
49	Vacuuming the floor	1.14	2.36			+	+		
50	Cleaning a bathroom	0.97	2.81		+		+		
51	Cleaning a kitchen cupboard	1.14	2.97			+	+		
52	Going for a woodland walk	0.67	3.28				+		
53	Lifting a box (10 kg)	1.03	3.33	+	+	+	+		
54	Replacing a ceiling light	0.97	3.35				+		
55	Vacuuming a flight of stairs	1.12	3.51			+	+		
56	Carrying a bag up stairs	1.02	4.39		+		+		
57	Painting the wall of a room	1.09	4.66			+	+		
58	Fetching heavy shopping	1.06	4.85				+		
59	Running to catch a bus	0.75	6.05	+	+	+	+		
60	Jogging for 15 minutes	0.26	10.86						

Choosing items from an item bank

One way of using an item bank is to assess the health status of all patients in a RCT using all items in an item bank. This will lead to the maximum power to detect treatment effects, which it is possible to achieve using the item bank. If it is not desirable to assess the health status of the patients in an RCT using all items in an item bank, a single selection of items covering the whole range of the item bank can be made and presented to all patients in a RCT. This procedure is simple to implement and does not require specific knowledge of the level of functional status in the patient population. The selections of 10 items and of 20 items spanning the item bank used in this article are indicated in Tables 8.1 and 8.2 in the column 'WR'.

Even if only limited information is available on the patient population, it should be possible to judge roughly, which items are more suitable for assessing health status in the population under examination. For instance, if a RCT is to be carried out on mildly disabled patients living independently, then the more difficult items in Table 8.2 are likely to provide more information on their health status, than the easier ones in Table Tables 8.1. The selections of 10 and 20 items judged to be suitable for the patient population under consideration in this article are given in in Table 8.2 in the column 'S'. Likewise, it is also possible to identify items, which will clearly provide less information on the health status of patients in the population. In this article, a selection of items suitable for patients in nursing or care homes were presented to patients living independently. The selections of 10 and of 20 'unsuitable' items, used in this article are indicated in Table 8.1 in the column 'US'.

A simulation study

In this simulation study, a large number of 'RCTs', based on a population of chronically ill patients but living independently without professional care, were simulated. In each RCT, the functional status, at the end of the study, of the 'patients' in the control group was simulated using the mean and standard deviation observed in the original data. The functional status of the 'patients' in the treatment group was simulated using the same standard deviation, but with the mean equal to the original mean plus a treatment effect determined by the effect size under investigation multiplied by the standard deviation of the population. Each RCT was conducted by presenting all 'patients' with the selection of items in the testing procedure being examined. One thousand RCTs were 'performed' at informative combinations of 30, 40, 50, 100, 200, 500 and 1000 'patients' in each treatment arm and effect sizes of 0.2, 0.5 and 0.8. The difference in the mean functional health status in the two treatment arms of the RCTs was assessed using IRT techniques, assuming that the item parameters were equal to those in Tables 8.1 and 8.2, and the Z_L statistic. To enable the simulation study to be repeated, a more technical description is given in the Appendix.

Results

Figure 8.1 presents the relationship between the number of patients in each arm of a RCT and the number of the 1000 'trials', which demonstrated a statistically significant difference between the two treatment groups. The power of a particular type of RCT can be obtained by dividing the number of the trials, which demonstrated a statistically significant difference, by 1000. For purposes of comparison, the Figure also displays the results obtained, when it was assumed that functional health status could have been measured without any estimation error. This is equivalent to assuming that the patients had been presented with all possible theoretical items to measure functional health status. It can be seen that, as with any type of statistical technique, the power to detect a given effect increases with the size of the effect and the number of patients in each arm of the RCT. The power of a RCT is highest if functional health status has been measured using all 60 items, followed by RCTs using a selection of 20 items suitable for the population are used. In general, a selection of 20 items spanning the whole item bank provides the next best power, followed by 10 items suitable for the population and a selection of 10 items spanning the whole item bank. Using selections of 20 or 10 items unsuitable for the population, results in a substantially lower power to detect treatment effects.

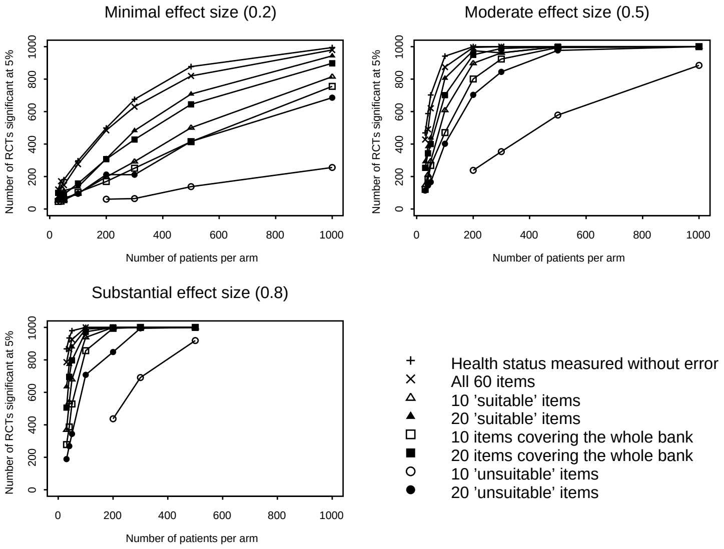

Figure 8.1: The relationship between the number of patients in each arm of a RCT and the number of the 1000 'RCTs', which demonstrated a statistically significant difference at the two-sided 5% level between the two treatment groups for the three treatment effect sizes, using various selections of items and IRT based analysis.

Interestingly, for a minimal effect size, the power remains noticeably below the level, which could have been attained if it had been possible to measure functional health status without error, even when all 60 items are used.

In order to gain a more practical insight into these results, estimates of the number of patients required in each arm of a RCT to detect treatment effects of 0.2, 0.5 and 0.8 in underlying functional health status with 80% power at the 5% level have been obtained using linear interpolation. This combination of power and level of statistical test have been chosen because they are commonly used in practice in RCTs. The results for the various selections of items and for the two methods of analysis are given in Table 8.3 and can be compared with the standard numbers of patients required in each arm of a RCT if it were possible to measure functional status directly and without error. The smallest number of patients is required to

Table 8.3: approximate numbers of patients required in each arm of a RCT using the item selections described to detect standard treatment effect sizes at the 5% significance level in functional health status with 80% power

	Item response theory		
	Effect size		
Item selection	0.8	0.5	0.2
All 60 items in the item bank	32	88	480
The 'suitable' section of 20 items	43	100	690
The 'suitable' section of 10 items	73	165	975
The selection of 20 items spanning the item bank	51	139	805
The selection of 10 items spanning the item bank	91	200	> 1000
The 'unsuitable' section of 20 items	165	270	> 1000
The 'unsuitable' section of 10 items	395	860	> 1000
If functional status could be measured directly and without error	26	64	394

detect an effect if all 60 items are used. In addition, it can be seen that the numbers of patients required if either 10 or 20 items are used depends on the type of items used. If a selection of items suitable for the population is used, fewer patients are required than if a selection spanning the whole item bank is used.

Discussion

This article has investigated the power, with which standard treatment effects can be detected, when functional health status is the primary outcome in a RCT and is assessed using either all items in an item bank, items spanning the whole item bank, items chosen to be appropriate for the patient population or items clearly inappropriate for the patient population analysed using IRT techniques. The power, with which standard treatment effects could be detected, varied greatly between RCTs using the three different types of selection of items. For a given number of items, the selection of items chosen to be 'suitable' for the population of patients under consideration resulted in more power to detect treatment effects than the selection spanning the whole range of the item bank, which in turn provided substantially more power than the 'unsuitable' selection of items. This reenforces the intuitive feeling that many researchers have, that choosing an instrument displaying a substantial floor or ceiling effect to measure an endpoint in a RCT, results in a loss of power to detect treatment effects.

Appendix: technical details of the simulation study

In the simulation study, 'RCTs' were carried out by drawing K values of θ from a $N(\mu, \sigma^2)$ distribution and K values of θ from a $N(\mu+\sigma\xi, \sigma^2)$ distribution to represent the functional status of 'patients' in the control and treatment arms, respectively. Here $\mu = 1.447$ and $\sigma = 0.705$ to correspond with the values found when analysing the original data and the effect size, denoted by ξ, was equal to 0.2, 0.5 or 0.8 as appropriate. Each RCT was conducted by presenting each of the $2K$ 'patients' with the items in the testing procedure being examined. The probability, p_{ik}, that patient k, $k = 1, 2, \ldots, 2K$, would respond in the category 'can' to item i, $i = 1, 2, \ldots, n$, was calculated using θ_k, the item parameters in Tables 8.1 and 8.2 and Equation 8.1. The response that patient k made to item i was a single observation on a $Bi(1, p_{ik})$ distribution. One thousand RCTs using each combination of K, $K = 30, 40, 50, 100, 200, 500, 1000$, and ξ were carried out for each of the selection of items.

Chapter 9

Epilogue

This thesis has examined some of the statistical and methodological challenges encountered when calibrating and using an item bank. These issues have not been previously considered in conjunction with health status assessment in enough depth to provide a stable base for the AMC Linear Disability Score (ALDS) project item bank. The majority of the papers that have contributed to this thesis have used the ALDS item bank, which aims to quantify functional status as expressed by the ability to perform activities of daily life, as an example, although the Short Form instruments from the Medical Outcomes Study[143, 139, 140] were also briefly considered in Chapter 7 [18].

The first part of this thesis examines problems encountered during the calibration phase of the AMC Linear Disability Score project item bank. It is essential to have a clear plan of the way the research is to be carried out and the statistical analysis performed. This is provided in Chapter 2[8]. Chapter 3 considers four practical ways of dealing with responses in the 'not applicable' category[14]. One of these methods is examined in more mathematical detail in Chapter 4[15]. If an item bank is to be used for various groups of patients, it is important to compare the measurement properties of the items in different groups of patients. In Chapter 5 the measurement properties for men and women and for patients aged 84 or under and patients aged over 84 are compared[16]. Finally, in Chapter 6, the measurement properties of the ALDS item bank are examined in a group of respondents requiring residential care. This Chapter concludes that the AMC Linear Disability Score project item bank has acceptable psychometric properties for this population.

The second part of this thesis considers how the number and type of items affects the the number of patients required to demonstrate the effectiveness of a novel treatment when item response theory based methods are used. Chapter 7 examines how varying the number of items used to assess the functional status of patients affects the number of patients required in a study[18]. Chapter 8 examines the effect of methods of selecting items from an item bank on the power to detect differences between treatment groups[17].

Simulation studies

The results in Chapters 4, 7, and 8 were obtained using simulation studies. This method was chosen because when item response theory models are used, the information matrix on the person parameters (θ) has a very complex form. This matrix, and hence the standard errors of estimated individual and population parameters, are highly dependent on the values of the item parameters (α and β) and have a complicated mathematical structure. This makes it difficult to obtain results using analytic techniques for the power to detect a treatment effect or the number of patients required in a particular study when using an item bank to quantify an endpoint.

However, results obtained in simulation studies have to be treated with some caution. The accuracy of the results depend on a number of factors. These include the skills of the person writing and operating software used to perform the simulations, the number of replications examined and the choice of parameter values. In addition, when generalising the results to other situations, it has to be remembered that the results obtained are only truly valid for the combinations of parameter values used in the simulation studies. Theoretically, different parameter values could lead to very different results. Only an analytical solution would prove that these results can be extended to other situations. However, it is thought that the results described in this thesis are broadly indicative of the types of studies examined and will provide useful guidelines to researchers carrying out similar studies in the future.

Implications

The results presented in this thesis may have implications for future research in the development and implementation of item banks in clinical outcome measurement. Chapters 2, 3, 5 and 6 examine a number of issues in the development of an item bank in a clinical context. The results in these chapters indicate that it is possible to use item response theory in this context, but that a number of issues, such as item non-response, differential item functioning and model fitting, need to be tackled in a slightly different way than in the field, where these methods were originally used.

The results in Chapters 7 and 8 may be important, when clinical studies using item banks are being set up. These chapters indicate that, when using item response theory in combination with an item bank, it is important to consider both the number of patients in a study and the number and type of items carefully.

Future research

There has been an enormous increase in interest in the application of item response theory in clinical measurement in recent years. A substantial number of papers have appeared using and extending these methods. However, there are two areas, in which a lot of work needs to be carried out, before item response theory can be widely implemented in clinical outcome measurement.

Firstly, proper statistical methods need to be developed to combine item response theory methods and designs commonly used in clinical studies, such as randomised clinical trials and crossover and cohort studies. Previous work has highlighted that simply estimating patient parameters (θ) and using these estimates in standard statistical techniques, such as t-tests, may lead to serious bias in results. Once appropriate statistical methods have been developed, it is important that they are implemented in software that is easy to access for researchers with a variety of levels of technical expertise. It would also be appropriate to examine some of the power and sample size issues considered in simulation studies in this thesis in a more analytical framework.

Secondly, clinicians and more technically orientated researchers need to combine their individual strengths to develop clinically useful item banks, which have been developed to the highest possible psychometric standards. Only when such item banks have been developed and made available, will it be possible to truly exploit the potentials of item response theory in clinical outcome measurement.

Bibliography

[1] Lind J. *A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject.* Printed by Sands, Murray and Cochran for A. Kincaid and A. Donaldson. Avaliable from www.jameslindlibrary.org, Edinburgh, United Kingdom, 1753.

[2] Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. *Med Care*, 19:787–805, 1981.

[3] Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol*, 9:789–93, 1982.

[4] Collin C, Wade DT, Davies S, Horne V. The Barthel ADL index: a reliability study. *Int Disabil Stud*, 10:61–3, 1988.

[5] Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. *Med Care*, 38(9 Suppl):II66–72, 2000.

[6] van der Linden WJ, Glas CAW. *Computerized Adaptive Testing. Theory and Practice*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2000.

[7] Holman R, Lindeboom R, Vermeulen M, Glas CAW, de Haan RJ. The Amsterdam Linear Disability Score (ALDS) project. The calibration of an

item bank to measure functional status using item response theory. *Quality of Life Newsletter*, 27:4–5, 2001.

[8] Holman R, Lindeboom R, Glas CAW, Vermeulen M, de Haan RJ. Constructing an item bank using item response theory: the AMC linear disability score project. *Health Services and Outcomes Research Methodology*, 4:19–33, 2003.

[9] Verbrugge LM, Jette AM. The disablement process. *Soc Sci Med*, 38:1–14, 1994.

[10] Lindeboom R, Vermeulen M, Holman R, de Haan RJ. Activities of daily living instruments in clinical neurology. optimizing scales for neurologic assessments. *Neurology*, 60:738–742, 2003.

[11] Glass TA. Conjugating the "tenses" of functioning: Disconcordance among hypothetical, experimental, and enacted function in older adults. *Gerontologist*, 38:101–112, 1998.

[12] Karagiozis H, Gray S, Sacco J, Shapiro M, Kawas C. The direct assessment of functional abilities (DAFA): a comparison to an indirect measure of instrumental activities of daily living. *Gerontologist*, 38(1):113–21, 1998.

[13] Sager MA, Dunham NC, Schwantes A, Mecum L, Halverson D, Harlowe K. Measurement of activities of daily living in hospitalized elderly: a comparison of self-report and performance-based methods. *J Am Geriatr Soc*, 40(5):457–62, 1992.

[14] Holman R, Glas CAW, Zwinderman AH, de Haan RJ. The treatment of not applicable responses in an item bank to measure functional status using item response theory. Poster presented at the 23rd meeting of the International Society for Biostatistics. Held in Dijon, France. 11-13 September 2002.

[15] Holman R, Glas CAW. Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, In press.

[16] Holman R, Lindeboom R, Vermeulen M, Glas CAW, de Haan RJ. The Amsterdam Linear Disability Score (ALDS) project. Differential item functioning with regard to gender. *Quality of Life Newsletter*, 29:13–14, 2002.

[17] Holman R, Lindeboom R, Vermeulen M, de Haan RJ. The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. *Health Qual Life Outcomes*, 2:42, 2004.

[18] Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. *Control Clin Trials*, 24:390–410, 2003.

[19] McDowell I, Newall C. *Measuring Health: a guide to rating scales and questionnaires.* Oxford University Press, Oxford, 1996.

[20] van Straten A, de Haan RJ, Limburg M, et al. Clinical meaning of the stroke-adapted sickness impact profile-30 and the sickness impact profile-136. *Stroke*, 31:2610–5, 2000.

[21] Walters SJ, Campbell MJ, Paisley S. Methods for determining sample sizes for studies involving health-related quality of life measures: a tutorial. *Health Services and Outcomes Research Methodology*, 2:83–99, 2001.

[22] Teresi JA, Golden RR, Cross P, Gurland B, Kleinman M, Wilder D. Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *J Clin Epidemiol*, 48:473–83, 1995.

[23] Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Stat Med*, 19:1651–1683, 2000.

[24] Gibbons RD, Clark DC, von Ammon Cavanaugh S, Davis JM. Application of modern psychometric theory in psychiatric research. *J Psychiatr Res*, 19:43–55, 1985.

[25] McHorney CA, Ware JE, Lu JF, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care*, 32:40–66, 1994.

[26] MacKnight C, Rockwood K. Rasch analysis of the hierarchical assessment of balance and mobility (HABAM). *J Clin Epidemiol*, 53(12):1242–7, 2000.

[27] van Buuren S, Hopman-Rock M. Revision of the ICIDH severity of disabilities scale by data linking and item response theory. *Stat Med*, 20:1061–76, 2001.

[28] Breithaupt K, McDowell I. Considerations for measuring functioning of the elderly: IRM dimensionality and scaling analysis. *Health Services and Outcomes Research Methodology*, 2:37–50, 2001.

[29] Badia X, Prieto L, Roset M, Diez-Perez A, Herdman M. Development of a short osteoporosis quality of life questionnaire by equating items from two existing instruments. *J Clin Epidemiol*, 55:32–40, 2002.

[30] Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley NK, Aaronson SM, Apolone G, Bech P, Brazier JE, Bullinger M, Sullivan M. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQoLA project. international quality of life assessment. *J Clin Epidemiol*, 51:1203–14, 1998.

[31] Cook KF, Rabeneck L, Campbell NP, Wray CJ. Evaluation of a multidimensional measure of dyspepsia-related health for use in a randomized clinical trial. *J Clin Epidemiol*, 52(5):381–92, 1999.

[32] Lord FM. *Applications of item response theory to practical testing problems.* Erlbaum, Hillsdale, NJ, 1980.

[33] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care*, 38:II28–II42, 2000.

[34] Kosinski M, Bjorner JB, Ware JE, Batenhorst A, Cady RK. The responsiveness of headache impact scales scored using 'classical' and 'modern'

psychometric methods: a re-analysis of three clinical trials. *Qual Life Res*, 12:903–12, Dec 2003.

[35] McHorney CA, Haley SM, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10): II. comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol*, 50:451–61, Apr 1997.

[36] Fischer GH, Molenaar IW. *Rasch models: Foundations, Recent Developments and Applications.* Springer, New York, 1995.

[37] Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika*, 51:567–577, 1986.

[38] Hambleton RK. Emergence of item response modelling in instrument development and data analysis. *Medical Care*, 38:II60–II65, 2000.

[39] Birnbaum A. *Statistical theories of mental test scores.* Some Latent trait models and their use in inferring an examinee's ability. Addison-Wesley, Reading, MA, 1968.

[40] Zimowski MF, Mukari E, Mislevy RJ, Bock RD. *BILOG-MG. Multiple group IRT analysis and test maintenance for binary items.* Scientific Software International, Inc. `www.ssicentral.com/irt.htm`. Chicago, IL, 1996.

[41] Rasch G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkely Symposium on Mathematical Statistics and Probability*, volume 4, pages 321–34, 1961.

[42] Lord FM. *New horizons in testing*, Small N ustifies Rasch model. Academic Press, New York, NJ, 1983.

[43] Thissen D, Wainer H. *Test Scoring.* LEA, Mahwah, NJ, 2001.

[44] Verhelst ND, Glas CAW. *Rasch models: their foundations, recent developments and applications.* The One-parameter logistic model. Springer, New York, 1995.

[45] Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51, 1972.

[46] Thissen D. Marginal maximum likelihood estimation for the one parameter logistic model. *Psychometrika*, 47:175–186, 1982.

[47] Molenaar IW. *Rasch models: their foundations, recent developments and applications*, Estimation of item parameters. Springer, New York, 1995.

[48] Hoijtink H, Boomsma A. *Rasch models: their foundations, recent developments and applications.* On person parameter estimation in the dichotomous Rasch model. Springer, New York, 1995.

[49] Thissen D. *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory.* Scientific Software., Chicago, 1991.

[50] Glas CAW. Detection of differential item functioning using lagrange multiplier tests. *Statistica Sinica*, 8:647–667, 1998.

[51] Yen W. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5:245–262, 1981.

[52] McKinley R, Mills C. A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9:49–57, 1985.

[53] Verhelst ND, Glas CAW, Verstralen HHFM. *OPLM: computer program and manual.* Cito, Arnhem, The Netherlands, 1995.

[54] Streiner DL, Norman GR. *Health Measurement Scales: a practical guide to their development and use.* Oxford University Press, Oxford, 1995.

[55] Kolen MJ, Brennan RL. *Test Equating.* Springer, New York, 1995.

[56] Holman R, Berger MPF. Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioural Statitics*, 26:361–380, 2001.

[57] van den Wollenberg AL. Two new tests for the Rasch model. *Psychometrika*, 47:123–140, 1982.

[58] Fayers PM, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. *Stat Med*, 15:679–696, 1998.

[59] Ebel RL, Frisbie DA. *Essentials of educational measurement.* Prentice-Hall, Engelwood Cliffs, NJ, 1986.

[60] Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[61] Orlando M, Thissen D. Likelihood-based item-fit indicies for dichotommous item respons theory models. *Applied Psychological Measurement*, 24:50–64, 2000.

[62] Rand Health Sciences Program. *Rand 36-item Health Survey 1.0.* Santa Monica, California: Rand Corporation. Avaliable at `www.qualitymetric.com/sf36`, 1992.

[63] Roth M, Tym E, Mountjoy CO, Huppert FA, Hendrie H, Verma S, Goddard R. CAMDEX. A standardised instrument for the diagnosis of mental disorder in the elderly. *British Journal of Psychiatry*, 49:698–709, 1986.

[64] Hunsberger S, Murray D, Davis CE, Fabsitz RR. Imputation strategies for missing data in a school-based multi-centre study: the pathways study. *Stat Med*, 20:305–16, 2001.

[65] Faris PD, Ghali WA, Brant R, Norris CM, Galbraith PD, Knudtson ML. Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *J Clin Epidemiol*, 55:184–91, 2002.

[66] Schafer JL. *Analysis of incomplete multivariate data.* New York: Chapman and Hall. 1997.

[67] Heckman JJ. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

[68] Rubin DB. Inference and missing data. *Biometrika*, 63:581–92, 1976.

[69] Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York, 1987.

[70] Little RJA, Rubin DB. *Statistical analysis with missing data.* New York: Wiley., 1987.

[71] Holman R, Lindeboom R, de Haan RJ. Gender and age based differential item functioning in the AMC Linear Disability Score project. *Quality of Life Newsletter*, 32:1–4, 2004.

[72] Fillenbaum GG, George LK, Blazer DG. Scoring nonresponse on the mini-mental state examination. *Psychological Medicine*, 18:1 021–5, 1988.

[73] Wu ML, Adams RJ, Wilson MR. *ACER ConQuest: Generalised Item Response Modelling Software.* ACER Press, Melbourne, 1998.

[74] Lord FM. Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48:477–482, 1983.

[75] Mislevy RJ, Chang H. Does addaptive testing violate local independence? *Psychometrika*, 65:149–156, 2000.

[76] Andersen EB. Estimating latent correlations between repeated testings. *Psychometrika*, 50:3–16, 1985.

[77] Wright BD, Masters GN. *Rating scale analysis: Rasch measurement.* MESA Press, Chicago, IL, 1982.

[78] May K, Nicewander WA. Measuring change conventionally and adaptively. *Educational and Psychological Measurement*, 58:882–897, 1998.

[79] Little RJA, Rubin DB. On jointly estimating parameters and missing data by maximising the complete-data likelihood. *American Statistician*, 37:218–220, 1983.

[80] Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York, 2000.

[81] Mislevy RJ, Wu PK. *Missing responses and irt ability estimation: omits, choice, time limits, and adaptive testing. ets research report.* Technical Report RR-96-30-ONR, Educational Testing Service, Princeton, NJ, 1996.

[82] Lord FM. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 39:247–264, 1974.

[83] Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46:443–459, 1981.

[84] Bradlow ET, Thomas N. Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, 23:236–243, 1998.

[85] O'Muircheartaigh C, Moustaki I. Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society, Series A*, 164:177–194, 1999.

[86] Moustaki I, O'Muircheartaigh C. A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *Statistica*, :259–276, 2000.

[87] Moustaki I, Knott M. Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A*, 163:445–459, 2000.

[88] Bernaards CA, Sijtsma K. Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, 34:277–313, 1999.

[89] Bernaards CA, Sijtsma K. Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, 35:321–364, 2000.

[90] Conaway MR. The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association*, 87:817–24, 1992.

[91] Park T, Brown MB. Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, 89:44–52, 1994.

[92] Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32, 1948.

[93] Kiefer J, Wolfowitz J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27:887–903, 1956.

[94] Reckase MD. The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, pages 401–412, 1985.

[95] Reckase MD. *Handbook of Modern Item Response Theory*, A linear logistic multidimensional model for dichotomous item response data. Springer, New York, 1997.

[96] Ackerman TA. Developments in multidimensional item response theory. *Applied Psychological Measurement*, 20:309–310, 1996.

[97] Ackerman TA. Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20:311–329, 1996.

[98] Rasch G. *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research, Copenhagen, Denmark, 1960.

[99] Lord FM, Novick MR. *Statistical theories of mental test scores.* Addison-Wesley, Reading, 1968.

[100] Muraki E. A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16:159–176, 1992.

[101] Masters GN. A Rasch model for partial credit scoring. *Psychometrika*, 47:149–174, 1982.

[102] Masters GN, Wright BD. *Handbook of Modern Item Response Theory.* The partial credit model. Springer, New York, 1997.

[103] Bock RD, Gibbons RD, Muraki E. Full-information factor analysis. *Applied Psychological Measurement*, 12:261–280, 1988.

[104] Adams RJ, Wilson MR, Wang WC. The random coefficients multinomial logit model. *Applied Psychological Measurement*, 21:1–25, 1997.

[105] Wilson DT, Wood R, Gibbons R. *TESTFACT: Test scoring, Item statistics, and Item Factor Analysis.* Scientific Software International, Inc, Chicago, IL, 1991.

[106] Verhelst ND, Glas CAW, van der Sluis A. Estimation problems in the Rasch model: the basic symmetric functions. *Computational Statistics Quarterly*, 1:245–262, 1984.

[107] Glas CAW, Verhelst ND. Extensions of the partial credit model. *Psychometrika*, 54:635–659, 1989.

[108] Verhelst ND, Glas CAW. *Rasch models: their foundations, recent developments and applications*, The generalized one parameter model: OPLM. Springer, New York, 1995.

[109] Zwinderman AH. A generalized Rasch model for manifest predictors. *Psychometrika*, 56:589–600, 1991.

[110] Zwinderman AH. *Handbook of Modern Item Response Theory.* Response models with manifest predictors. Springer, New York, 1997.

[111] McDonald RP. Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, 6:379–396, 1982.

[112] Mellenbergh GJ. Generalized linear item response theory. *Psychological Bulletin*, 115:300–307, 1994.

[113] Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52:393–408, 1987.

[114] Jones RN, Gallo JJ. Education and sex differences in the mini-mental state examination: effects of differential item functioning. *J Gerontol B Psychol Sci Soc Sci*, 57(6):548–58, 2002.

[115] Moorer P, Suurmeijer TP, Foets M, Molenaar IW. Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Qual Life Res*, 10:637–45, 2001.

[116] Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: true differences or differential item functioning? *Med Care*, 41(7 Suppl):III75–III86, 2003.

[117] Ryall NH, Eyres VC, Neumann SB, Bhakta BB, Tennant A. Is the Rivermead mobility index appropriate to measure mobility in lower limb amputees? *Disabil Rehabil*, 25:143–53, 2003.

[118] Eyres S, Tennant A, Kay L, Waxman R, Helliwell PS. Measuring disability in ankylosing spondylitis: comparison of bath ankylosing spondylitis functional index with revised Leeds disability questionnaire. *J Rheumatol*, 29:979–86, 2002.

[119] Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. *J Gerontol B Psychol Sci Soc Sci*, 57(5):S275–84, 2002.

[120] Hagell P, Whalley D, McKenna SP, Lindvall O. Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham health profile. *Mov Disord*, 18:773–83, 2003.

[121] Ramsay JO. Kernal smoothing approaches to non-parametric item characteristic curve estimation. *Psychometrika*, 56:611–630, 1991.

[122] Ramsay JO. *TestGraf. A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data.* Avaliable from `ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/` accessed 28 November 2003.

[123] Ware JE, Bayliss MS. The future of item “banking”: who should make the deposits and how should they be used? *Quality of Life Newsletter*, 31:15–18, 2003.

[124] Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. *Arch Phys Med Rehabil*, 84): S52–60, 2003.

[125] McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Ann Intern Med*, 139: 403–9, 2003.

[126] `http://www.amihealthy.com/static/dynamicsf36info.asp`. Website. Accessed 29th October 2003.

[127] `http://www.headachetest.com/`. Website. Accessed 29th October 2003.

[128] Webster K, Cella D, Yost K. The functional assessment of chronic illness therapy (FACIT) measurement system: properties, applications, and interpretation. *Health Qual Life Outcomes*, 1:79, 2003.

[129] Holman R, Glas CAW, Zwinderman AH, de Haan RJ. Practical methods for dealing with ‘not applicable’ item responses in the amc linear disability score project. *Health Qual Life Outcomes*, 2:29, 2004.

[130] Bock RD, Mislevy RJ. Adaptive eap estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6:431–444, 1982.

[131] du Toit M. (editor). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact.* Scientific Software International, Inc, Lincolnwood, IL, 2003.

[132] Harvey WR. Estimation of variance and covariance components in the mixed model. *Biometrics*, 26:485–504, 1970.

[133] van der Linden WJ, Hambleton RK. *Handbook of Modern Item Response Theory.* Springer, New York, 1996.

[134] Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. conceptual framework and item selection. *Med Care*, 30:473–83, 1992.

[135] Cohen J. *Statistical power analysis for the behavoural sciences.* Lawernce Erlbaum Associates., Hillsdale, NJ, 1988.

[136] Formann AK. Measuring change in latent subgroups using dichotomous data-unconditional, conditional and semiparametric maximum-likelihood-estimation. *J Am Stat Assoc*, 89:1027–1034, 1994.

[137] Korevaar JC, Merkus MP, Jansen MA, Dekker FW, Boeschoten EW, Krediet RT. Validation of the KDQOL-SF: a dialysis-targeted health measure. *Qual Life Res*, 11:437–47, 2002.

[138] van Manen JG, Korevaar JC, Dekker FW, Boeschoten EW, Bossuyt PM, Krediet RT. How to adjust for comorbidity in survival studies in esrd patients: a comparison of different indices. *Am J Kidney Dis*, 40:82–9, 2002.

[139] Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Med Care*, 34:220–33, 1996.

[140] `http://www.sf36.com/tools/sf8.shtml`. Website. Accessed 2nd January 2003.

[141] Glas CAW. Modification indices for the 2-PL and the nominal response model. *Psychometrika*, 64:273–294, 1999.

[142] Cochran WG. *Sampling Techniques.* Wiley, New York, 1977.

[143] McHorney CA, Ware JE, Rogers W, Raczek AE, Lu JF. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth coop charts. Results from the medical outcomes study. *Med Care*, 30(5 Suppl):MS253–65, 1992.

[144] McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care*, 31:247–63, 1993.

[145] Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, Sprangers MAG, te Velde A, Verrips E. Translation, validation, and norming of the Dutch language version of the SF-36 health survey in community and chronic disease populations. *J Clin Epidemiol*, 51:1055–68, 1998.

[146] Wolfe F. Which HAQ is best? a comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. *J Rheumatol*, 28:982–9, 2001.

[147] McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med*, 127(8 Pt 2):743–50, 1997.

Chapter 10

The interesting bits

Summary in English

Chronic illness are unlikely to be fatal in the short term, but also cannot be cured completely. Conditions such as asthma, arthritis, heart failure and stroke can lead to a dramatic reduction in a person's health status and thus their enjoyment of and participation in everyday life. The effectiveness of a number of treatments for a chronic condition is often compared by examining the resulting changes in a person's health status. Health status is a broad construct, but can be narrowed down into a number of clearly defined concepts, such as cognitive status or functional ability. This thesis concentrates on health status as defined by the ability to perform activities of daily life. Activities of daily life are defined as the tasks one needs to carry out in order to live independently in the community or an appropriate care based setting. Examples are washing and dressing oneself, mobility, food preparation and housework.

The ability to perform activities of daily life can be quantified using specially designed questions. Traditionally, these questions have been assembled in fixed length questionnaires and a numerical score allocated to each response category of each question. Each patient's responses were summarised by adding together the item scores on each item. Recently, interest in an alternative measurement

paradigm, item response theory, has increased. Item response theory was developed as a tool for analysing pupils' responses to examination questions and is often used in conjunction with item banks. An item bank is a collection of items, reflecting a single concept, for which the measurement characteristics are known. This thesis examines the development and use of the AMC Linear Disability Score (ALDS) project item bank to measure the ability to perform activities of daily life. The first part of this thesis, consisting of **Chapters 2 to 6**, examines the development of the AMC Linear Disability Score project item bank. The second part of this thesis, consisting of **Chapters 7 and 8**, considers how many patients and what type of items should be used when comparing the effectiveness of two treatments in the framework of item response theory.

Chapter 2 describes and illustrates the methods used in the construction of the AMC Linear Disability Score project item bank. The issues surrounding the definition of health status used, the data collection and statistical analysis are considered in depth and presented in terms of a protocol for the development of an item bank to measure clinical outcomes.

Patients' responses to the items in the AMC Linear Disability Score project item bank were recorded in three categories: 'I can carry out the activity'; 'I cannot carry out the activity'; and 'I have never experienced this activity, it is not applicable to me'. **Chapter 3** examined four practical methods for dealing with responses in the category 'not applicable' in the context of missing data. The first method used cold deck imputation to replace responses in the 'not applicable' category with 'I cannot carry out the activity'. The second method used hot deck imputation to replace responses in the 'not applicable' category with either 'I can carry out the activity' or 'I cannot carry out the activity' with probabilities related to the ability of the individual patient and the difficulty of the item. The third method treated items to which a patient responded in the 'not applicable' category as if they had not been offered to that patient. The fourth method modelled the probability of responding to an item and the ability to perform activities of daily life jointly. When estimating the difficulty of the items or the ability of the patients, the first method produced results, which were substantially different from those given by the second, third and fourth methods. It was concluded that, for clinical reasons, the cold deck method was unsuitable for the data and that the second and third methods were suitable and useful in the analysis of the data. The fourth method would be useful for more in depth analyses of the processes leading to patients using the ‘not applicable’ category.

The fourth method described in **Chapter 3** was examined in greater depth in **Chapter 4**. This method, and a number of related models, can be used to avoid bias in the estimates of the difficulty of items.

When developing an item bank, it is desirable that items should have the same measurement properties for all subgroups in the population. The introduction of new technologies, such as mobile telephones and automatic teller machines has changed how activities of daily life are carried out and a large proportion of activities of daily life have been traditionally carried out by either men or women. Hence, the differences between the probabilities for younger and older respondents and the differences in probabilities between men and women were examined in **Chapter 5**. The results indicated that the majority of items reflecting mobility, agility, dressing, self-care and activities outside the home had the same measurement characteristics for all groups. A large proportion of items concerning housework were easier for women than men, while a substantial number of the items reflecting household maintenance or lifting and carrying were easier for men than women. Some tasks reflecting household administration were more difficult for older respondents.

Chapter 6 examines the psychometric properties of the AMC Linear Disability Score project item bank in a population requiring residential care. The results indicate that the item bank has sound properties for this group of patients.

Once the AMC Linear Disability Score project item bank has been fully calibrated, it can be used as an instrument to assess the patients’ health status in a range of clinical studies. Whenever a clinical study is instigated, it is important to examine how many patients are required to have a reasonable chance of being able to detect the effect expected.

Chapter 7 considers how the number of patients included in each group in a randomised clinical trial and the number of items used to assess their functional status affect the power to detect a treatment effect. This chapter uses the SF-36, SF-12 and SF-8 outcomes as examples.

The ways in which items can be selected from an item bank are examined in **Chapter 8**. Three types of item selection from the AMC Linear Disability Score project item bank were considered. These were: items drawn to cover the whole range of the item bank; items, which in terms of difficulty, were well-suited to the patient population under consideration; and items, which in terms of difficulty, were ill-suited to the patient population. Using items suited to the population resulted in the highest statistical power to detect treatment effects, while using items unsuited to the population resulted in the lowest power. These effects were particularly stark in studies with fewer than 200 patients.

Chapter 9 discusses the results presented in this thesis in a wider context and examines their application in clinical outcome measurement. In addition, a number of directions that could be taken in future research are considered.

Samenvatting in het Nederlands

Chronische aandoeningen zoals astma, reuma, hartfalen en beroerte kunnen tot een dramatische vermindering van de patiënt zijn kwaliteit van leven leiden en zijn maatschappelijke participatie belemmeren. De effectiviteit van een medische behandeling wordt vaak vastgesteld door de functionele gezondheidstoestand van patiënten te meten en groepsgewijs te vergelijken. Functionele gezondheid is een breed begrip dat verwijst naar lichamelijke beperkingen, cognitieve stoornissen en psychosociaal welzijn. In dit proefschrift wordt de functionele toestand van de patiënt gedefinieerd in termen van Activiteiten van het Dagelijks Leven (ADL). Deze activiteiten zijn noodzakelijke randvoorwaarden om zelfstandig in de maatschappij te functioneren of in een geschikte zorgomgeving te kunnen wonen. Voorbeelden zijn wassen en aankleden, eten, mobiliteit in en buiten het huis en huishoudelijke taken.

Het kunnen uitvoeren van ADL wordt traditioneel gekwantificeerd door middel van vragenlijsten waarin een van tevoren gedefinieerd aantal vragen is opgenomen. De totaalscore van een patiënt wordt verkregen door de scores op de afzonderlijke vragen op te tellen. Tegenwoordig hebben onderzoekers een groeiende belangstelling voor een andere kwantitatieve methode, de item respons theorie. Item respons theorie (IRT) is indertijd ontwikkeld als statistische methode voor het analyseren van de antwoorden van scholieren op toetsvragen. Item respons theorie wordt vaak toegepast in combinatie met een zogenaamde itembank. Een itembank is een verzameling van vragen waarvan de meeteigenschappen gedetailleerd bekend zijn.

Het onderhavige proefschrift beschrijft de ontwikkeling en toepassing van de AMC Linear Disability Score (ALDS) itembank. Het eerste deel van het proefschrift (**hoofdstukken 2 tot en met 6**) beschrijft de ontwikkeling van de itembank. Het tweede deel (**hoofdstukken 7 en 8**) richt zich, binnen de context van IRT, op de vraag hoeveel patiënten in een klinische trial moeten worden bestudeerd en hoeveel vragen en wat voor type vragen bij hen moet worden afgenomen om met voldoende statistische zekerheid een uitspraak te kunnen doen over de effectiviteit van een behandeling.

Hoofdstuk 2 beschrijft de methoden die bij de ontwikkeling van de ALDS itembank zijn gebruikt. De gehanteerde definitie van de functionele gezondheidstoestand, de methode van dataverzameling en de geëigende statistische analyses worden in dit hoofdstuk uitgebreid beschreven. Daarnaast wordt een protocol voorgesteld om sturing te geven aan het ontwikkelen van een itembank om klinische uitkomsten te kunnen meten.

In de ALDS itembank zijn drie antwoordcategorieën mogelijk: ik kan de activiteit uitvoeren; ik kan de activiteit niet uitvoeren; en ik heb deze activiteit nooit gedaan, het is niet op mij van toepassing. **Hoofdstuk 3** bestudeert vier statistische technieken om de antwoordcategorie 'niet van toepassing' te analyseren. Een van deze methoden wordt in **hoofdstuk 4** nader uitgewerkt.

Het is belangrijk dat de vragen in een itembank dezelfde meeteigenschappen hebben voor diverse subgroepen van patiënten en bijvoorbeeld niet verschillen tussen mannen en vrouwen of tussen jongeren en ouderen. In **hoofdstuk 5** worden de meeteigenschappen van de ALDS itembank in de genoemde subgroepen met elkaar vergeleken. De resultaten laten zien dat de meeste vragen over mobiliteit, aankleden en zelfzorg dezelfde meeteigenschappen in de subgroepen hebben. Wel vinden vrouwen huishoudelijk activiteiten relatief makkelijker dan mannen, terwijl het dragen van zware voorwerpen minder moeilijk is voor mannen. Administratieve taken zijn in het algemeen moeilijker voor oudere patiënten.

Hoofdstuk 6 beschrijft de meeteigenschappen van het ALDS itembank voor oudere personen die in een beschermde woonomgeving wonen. De resultaten laten zien dat in deze specifieke populatie de itembank over goede psychometrische eigenschappen beschikt.

Na de totstandkoming van de ALDS itembank is het de bedoeling dat de ALDS als uitkomstmaat wordt gebruikt in het klinisch wetenschappelijk onderzoek. Voordat een klinische trial van start gaat, is het belangrijk te weten hoeveel patiënten moeten worden onderzocht om een verschil in gezondheidstoestand tussen behandelgroepen statistisch te kunnen aantonen. **Hoofdstuk 7** onderzoekt hoe bepalend het aantal patiënten en het aantal afgenomen vragen zijn om een aanwezig verschil tussen de behandelarmen te kunnen detecteren. In dit hoofdstuk zijn de SF-36, SF-12 en SF-8 vragenlijsten als voorbeelden gebruikt.

In **hoofdstuk 8** worden drie verschillende methoden onderzocht om vragen uit een itembank te selecteren. Deze zijn: vragen die de gehele breedte van de itembank bestrijken; vragen die, gelet hun moeilijkheidsgraad, ongeschikt zijn voor de betreffende patiëntengroep; en vragen die, gelet hun moeilijkheidsgraad, juist geschikt zijn. Het laatste soort vragen (moeilijkheidsgraad aangepast aan de kenmerken van de patiënten) geeft de meeste kans om verschillen tussen twee behandelgroepen te detecteren. Dit is vooral te merken in onderzoekingen met minder dan 200 patiënten.

Hoofdstuk 9 bespreekt de onderzoeksresultaten van dit proefschrift in een bredere context en beschrijft hoe die bevindingen in het meten van klinische uitkomsten toegepast kunnen worden. Tot slot, worden een aantal suggesties gegeven voor toekomstig onderzoek op het terrein van IRT in het patiëntgebonden onderzoek.

Acknowledgements

This thesis is the result of five enjoyable years spent in the department of Clinical Epidemiology and Biostatistics at the AMC. Many people have supported me in this period and I would like to thank some of them personally.

Firstly, I would like to thank my ‘promoters’ and ‘co-promotor’. Rob de Haan, I have learnt a hell of a lot from you, especially about how to use statistical techniques to benefit clinical research. I am particularly grateful that you have taught me how to present complex statistical ideas in a clinical context. I am really glad that you were my ’promotor’ and I look forward to working with you again in the future. Cees Glas, I am amazed at your knowledge of item response theory. Whenever I ask a question, you immediately pluck the appropriate article from the cupboard. You have taught me about measurement theories, in general, and item response theory, in particular. Without your help, I would have found it much more difficult to write this thesis and provide statistical support for the ALDS project. Rien Vermeulen, as the clinician on the project, you can always explain why I need to put all this effort into making the ALDS item bank work.

I would like to thank the other members of my ‘promotiecommissie’, Professor M.P.F. Berger, Professor G.J. Bonsel, Professor M.H. Prins, Professor M.A.G. Sprangers and Professor J.G.P. Tijssen, for taking the time to read my thesis.

I would also like to thank Patrick Bossuyt and Koos Zwinderman, the head and deputy head of the department. I am still really impressed by your knowledge of the whole fields of epidemiology and biostatistics.

Gre de Vries and Petra Lampe, you are always there to help me find a new pen, use the binder, fill in a form or just for a chat. Thanks a lot for the moral support! Noor van den Bosch, even though you have now left the KEB, I still remember your friendly smile. You always knew the best time to speak to Rob. Now, you are still much better than me at judging when Rob will really have time to do something.

Robert Lindeboom and Nadine Weisscher, you were my main colleagues on the ALDS project. Robert, I learnt a lot about health related quality of life questionnaires from you. I am glad that you helped me understand how these questionnaires are used in residential care and clinical settings. Nadine, you have lots of good ideas and I am pleased that we will be working together in the future. I would also like to thank Nadine Fleitour, Miranda Postma en Annick Gorssen for collecting most of the data used in this thesis. Thanks a lot for your time and effort.

Rien de Vos, during my first few years in the department, I enjoyed your quite companionship, when we were both sitting at our desks at 7:30 in the morning. You also knew just how to pour oil on troubled waters, when I had too many supervisors going in too many different directions. Kimberly de Boer and Helene Thygesen, thanks for the great conversations in room J1b-207 (previously known as J2-212). Sandra van Abswoude and Janneke ter Marvelde, I really enjoyed our conversations on item response theory and appreciated your impartial views on the ALDS project. I would also like to thank all the other people, who work or worked in the department of Clinical Epidemiology and Biostatistics at the AMC. I could come and ask you for help or advice at almost any point. It is also always good to have a little diversion in the form of cake or just a chat, when you are working hard.

Professor H.L. Hardman and Professor F.A.H. van Harmelen, Frank and Lynda. I am really pleased that you are my paranimfen. Now you get to watch a 'promotie' from a slightly different angle! You have given me a lot of support and advice on my thesis and even a place to sleep when the 200 km between Maastricht and Amsterdam really got to me.

Finally, I would like to thank all my friends and family for their support during the last five years. I am especially grateful to my mother, whose strength is an inspiration to me, my brother and sister, and Jan and Thomas, who is the most fascinating person I have ever met.

Published articles

1. van der Crabben SN, Heymans HS, van Kempen AA, Holman R, Sauerwein HP. [Qualitative malnutrition due to incorrect complementary feeding in Bush Negro children in Suriname] *Ned Tijdschr Geneeskd.* 2004 148:1093–7.

2. Holman R. How does item selection procedure affect power and sample size when using an item bank to measure health status? *Quality of Life Newsletter* 2004; **Special issue**; 9–11.

3. Holman R, Glas CAW, Lindeboom R, Zwinderman AH, De Haan RJ. Practical methods for dealing with 'not applicable' item responses in the AMC Linear Disability Score project. *Health Qual Life Outcomes.* 2004; 2:29.

4. Holman R, Lindeboom R, de Haan RJ. Gender and age based differential item functioning in the AMC Linear Disability Score project. *Quality of Life Newsletter* 2004; **32**; 1–4.

5. Holman R, Lindeboom R, Vermeulen M, de Haan RJ. The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. *Health Qual Life Outcomes.* 2004 Aug 03;2(1):42.

6. Hung LQ, De Vries PJ, Binh TQ, Giao PT, Nam NV, Holman R, Kager PA. Artesunate with mefloquine at various intervals for non-severe plasmodium falciparum malaria. *Am J Trop Med Hyg* 2004. **71**:160–166.

7. Lindeboom R, Holman R, Dijkgraaf MG, Sprangers MA, Buskens E, Diederiks JP, De Haan RJ. Scaling the sickness impact profile using item response theory: an exploration of linearity, adaptive use, and patient driven item weights. *J Clin Epidemiol.* 2004 **57**: 66–74.

8. Lindeboom R, Schmand B, Holman R, de Haan RJ, Vermeulen M. Improved brief assessment of cognition in aging and dementia. *Neurology.* 2004 Aug 10;63(3):543–6.

9. Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. *Control Clin Trials.* 2003; 24: 390-410.

10. Holman R, Lindeboom R, Glas CAW, Vermeulen R, de Haan RJ. Constructing an item bank using item response theory: the AMC linear disability score project. *Health Services and Outcomes Research Methodology* 2003: 4; 19-33.

11. Lindeboom R, Vermeulen M, Holman R, De Haan RJ. Activities of daily living instruments: optimizing scales for neurologic assessments. *Neurology.* 2003; 60: 738-42.

12. de Haan RJ, Vermeulen M, Holman R, Lindeboom R. [Measuring the functional status of patients in clinical trials using modern clinimetric methods] *Ned Tijdschr Geneeskd.* 2002; 146:606-11.

13. Holman R, Lindeboom R, Vermeulen R, Glas CAW, de Haan RJ. The Amsterdam Linear Disability Score (ALDS) project. Differential item functioning with regard to gender. *Quality of Life Newsletter* 2002; 29; 13-14.

14. Kuiken SD, Lei A, Tytgat GN, Holman R, Boeckxstaens GE. Effect of the low-affinity, noncompetitive N-methyl-d-aspartate receptor antagonist dextromethorphan on visceral perception in healthy volunteers. *Aliment Pharmacol Ther.* 2002; 16: 1955-62.

15. Holman R, Berger MPF. Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioural Statistics.* 2001: 26; 361-380.

16. Holman R, Lindeboom R, Vermeulen R, Glas CAW, de Haan RJ. The Amsterdam Linear Disability Score (ALDS) project. The calibration of an item bank to measure functional status using item response theory. *Quality of Life Newsletter* 2001; 27; 4-5.

About the author

Rebecca Holman was born on 1st March 1976 in Bristol, England. She grew up in Penuwch, near Tragaron, Wales, Sibford Gower, Oxfordshire, England and Sheffield, South Yorkshire, England. She attended Abbeydale Grange School, Sheffield from 1987 to 1992 and obtained General Certificate of Secondary Education (GCSE) awards in English language, English literature, mathematics, integrated humanities, dual science, economics, German and art and design. She attended King Edward VII School from 1992 to 1994 and obtained A-levels in general studies, economics, German and mathematics with statistics. Rebecca studied mathematics, statistics and languages at the University of Sheffield from 1994 to 1998. She also studied statistics at the University of Amsterdam from 1996 to 1997. In 1998 she was awarded the degree of Masters of Mathematics (MMath) with first class honours. From 1998 to 1999, Rebecca worked as a junior researcher in the Department of Methodology and Statistics, Faculty of Health Sciences, University of Maastricht. From 1999 to 2005, she worked in the Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam. Following the defence of this thesis, she will work in the Department of Neurology, Academic Medical Center, Amsterdam.

Rebecca has one son, Thomas, born on 25th March 2003.