Health and Quality of Life Outcomes

BioMed Central

Research

Open Access

Practical methods for dealing with 'not applicable' item responses in the AMC Linear Disability Score project

Rebecca Holman*[1], Cees AW Glas[2], Robert Lindeboom[1], Aeilko H Zwinderman[1] and Rob J de Haan[1]

Address: [1]Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands and [2]Department of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

Email: Rebecca Holman* - r.holman@amc.uva.nl; Cees AW Glas - c.a.w.glas@edte.utwente.nl; Robert Lindeboom - r.lindeboom@amc.uva.nl; Aeilko H Zwinderman - a.h.zwinderman@amc.uva.nl; Rob J de Haan - rob.dehaan@amc.uva.nl

* Corresponding author

Published: 16 June 2004

Health and Quality of Life Outcomes 2004, **2**:29

This article is available from: http://www.hqlo.com/content/2/1/29

Received: 23 April 2004
Accepted: 16 June 2004

Abstract

Background: Whenever questionnaires are used to collect data on constructs, such as functional status or health related quality of life, it is unlikely that all respondents will respond to all items. This paper examines ways of dealing with responses in a 'not applicable' category to items included in the AMC Linear Disability Score (ALDS) project item bank.

Methods: The data examined in this paper come from the responses of 392 respondents to 32 items and form part of the calibration sample for the ALDS item bank. The data are analysed using the one-parameter logistic item response theory model. The four practical strategies for dealing with this type of response 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'.

Results: The item and respondent population parameter estimates were very similar for the strategies involving 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 estimates obtained using the cold deck imputation method were substantially different.

Conclusions: The cold deck imputation method was not considered suitable for use in the ALDS item bank. The other three methods described can be usefully implemented in the ALDS item bank, depending on the purpose of the data analysis to be carried out. These three methods may be useful for other data sets examining similar constructs, when item response theory based methods are used.

Background

When questionnaires consisting of a number of related items are used to measure constructs such as health related quality of life [1,2], cognitive ability [3] or functional status [4], 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 [5]. These range from imputation methods [6,7] to algorithms, which permit parameters to be estimated, whilst ignoring missing data points [8] and frameworks, in which it is possible to construct a joint model for the data and the pattern of missing data points [9]. It is always essential to examine why some responses are missing and whether there is a pattern underlying the missing data for questionnaires [10-12], 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 [4,13,14]. 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 [13] 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 [3,15] and the Sickness Impact Profile [16], treat such responses as a 'negative' category and others, such as the SF-36 [1,2], 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 [17].

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 [18] with around 4000 patients [4,19]. 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 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'.

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 [1,2]. 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 [12]. 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 [16], the Mini-mental state examination and the CAMCOG [15], in which items, to which

Table 1: Item content and parameters.

	Estimates of the item parameters ($\hat{\beta}$)				
Item description	Hot deck 1st run	Cold deck	Items never offered	Including tendency to respond	Mean 5 runs hot deck
Running for more than 15 minutes (++) (2)	3.77 (0.242)	3.49 (0.238)	3.71 (0.242)	3.72 (-)	3.76 (0.242)
Going for a walk in the woods (2)	-1.17 (0.125)	-1.02 (0.120)	-1.15 (0.124)	-1.16 (-)	-1.16 (0.125)
Running for less than 5 minutes (3)	1.37 (0.135)	1.26 (0.129)	1.34 (0.135)	1.34 (-)	1.36 (0.135)
Walking up a hill or high bridge (++) (3)	-2.54 (0.163)	-2.27 (0.156)	-2.50 (0.162)	-2.51 (-)	-2.53 (0.163)
Lifting up a toddler (3)	-1.91 (0.140)	-1.69 (0.134)	-1.87 (0.139)	-1.88 (-)	-1.90 (0.140)
Moving a bed or table (4)	-2.49 (0.160)	-2.20 (0.153)	-2.44 (0.160)	-2.44 (-)	-2.47 (0.160)
Playing with a child on the floor (5)	-1.84 (0.137)	-1.62 (0.132)	-1.82 (0.138)	-1.83 (-)	-1.84 (0.138)
Tightening a screw (+) (5)	-3.23 (0.204)	-2.82 (0.188)	-3.18 (0.204)	-3.18 (-)	-3.21 (0.204)
Going shopping for clothes (++) (6)	-3.11 (0.195)	-2.69 (0.179)	-3.05 (0.195)	-3.06 (-)	-3.10 (0.195)
Change a light bulb in a ceiling lamp (7)	-1.57 (0.131)	-1.37 (0.126)	-1.59 (0.132)	-1.59 (-)	-1.59 (0.132)
Mopping the floor (++) (11)	-3.56 (0.231)	-2.89 (0.193)	-3.55 (0.236)	-3.55 (-)	-3.56 (0.233)
Putting the rubbish out (12)	-3.45 (0.222)	-2.82 (0.188)	-3.47 (0.231)	-3.47 (-)	-3.47 (0.224)
Lifting a box weighting 10 kg (13)	-1.37 (0.128)	-1.11 (0.121)	-1.35 (0.129)	-1.36 (-)	-1.36 (0.128)
Shopping for groceries for a week (13)	0.03 (0.120)	0.15 (0.115)	0.03 (0.122)	0.03 (-)	0.01 (0.120)
Painting a ceiling (14)	1.21 (0.132)	1.17 (0.127)	1.18 (0.134)	1.19 (-)	1.19 (0.132)
Cleaning a bathroom (17)	-1.99 (0.142)	-1.57 (0.131)	-1.98 (0.144)	-1.99 (-)	-2.00 (0.142)
Carrying a heavy bag upstairs (17)	-0.53 (0.120)	-0.31 (0.114)	-0.47 (0.122)	-0.48 (-)	-0.49 (0.121)
Painting a wall (18)	-0.29 (0.120)	-0.08 (0.114)	-0.25 (0.122)	-0.25 (-)	-0.26 (0.120)
Cycling for 15 minutes (24)	-1.84 (0.137)	-1.38 (0.126)	-1.85 (0.142)	-1.86 (-)	-1.89 (0.140)
Change sheets and duvet cover on bed (25)	-2.20 (0.149)	-1.58 (0.131)	-2.16 (0.151)	-2.17 (-)	-2.19 (0.150)
Caring for potted plants on a balcony (25)	-1.65 (0.133)	-1.20 (0.122)	-1.62 (0.137)	-1.62 (-)	-1.63 (0.134)
Vacuuming a flight of stairs (26)	-1.40 (0.128)	-1.02 (0.120)	-1.43 (0.133)	-1.43 (-)	-1.44 (0.130)
Washing a window from the outside (27)	-1.30 (0.127)	-0.84 (0.117)	-1.24 (0.129)	-1.25 (-)	-1.27 (0.126)
Cycling with a heavy load of shopping (30)	-0.74 (0.121)	-0.41 (0.114)	-0.77 (0.125)	-0.77 (-)	-0.76 (0.122)
Pumping up a bicycle tyre (33)	-3.00 (0.188)	-2.02 (0.145)	-2.98 (0.199)	-2.99 (-)	-3.03 (0.193)
Travelling by plane (38)	-2.14 (0.147)	-1.38 (0.126)	-2.10 (0.153)	-2.10 (-)	-2.11 (0.149)
Mopping a flight of stairs (39)	-2.16 (0.147)	-1.38 (0.126)	-2.11 (0.154)	-2.12 (-)	-2.13 (0.147)
Vacuuming the inside of a car (48)	-1.97 (0.141)	-1.15 (0.122)	-1.92 (0.151)	-1.92 (-)	-1.95 (0.142)
Swimming for an hour (+) (54)	-1.25 (0.126)	-0.56 (0.115)	-1.19 (0.134)	-1.20 (-)	-1.18 (0.129)
Washing a car (82)	-1.16 (0.125)	-0.37 (0.114)	-1.22 (0.143)	-1.22 (-)	-1.23 (0.131)
Mowing the lawn (102)	-0.68 (0.121)	0.19 (0.115)	-0.67 (0.140)	-0.67 (-)	-0.71 (0.122)
Repairing a puncture in bicycle tyre (133)	-1.25 (0.126)	0.08 (0.114)	-1.22 (0.156)	-1.23 (-)	-1.25 (0.127)
Cronbach's alpha coefficient for scale	0.87	0.84	0.81	0.81	0.87

Item content and parameters. Item content with the number of patients responding in the 'not applicable' category (in parenthesis) and the estimates of the item parameters (β_i) and their standard errors (in parenthesis) for each of the procedures. Standard errors for the parameters in the 'tendency to respond' model are not currently available in the software. This is indicated by the symbol '-'. Items denoted by (++) demonstrated item misfit across more than one method and items denoted by (+) demonstrated item misfit for one method.

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 [11] incorporating theoretical or observed aspects of the data [12]. 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 [11]. 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 [8]. 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 [20,21], it is possible to treat items, to which no response was made, as if they had never been offered to the respondent [22]. This is equivalent to ignoring the missing responses [21] and is essential in the application of computerised adaptive testing [23,24]. 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 [22]. 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 [25]. This type of model is described in more depth elsewhere [26].

Statistical analysis

In this paper, the one-parameter logistic model [27], 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

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

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 [28].

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 [20,29]. 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 [30,31]. To avoid this, plausible values for the functional status of each patient have been drawn from their own posterior distribution of θ [20]. The item parameters and patients' functional status have been estimated in ConQuest [20]. Other calculations were carried out in S-PLUS [32].

Results

The estimates of the item parameters (β_i) and their standard errors are given in Table 1. 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 1. 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 1. All values are greater than 0.8, indicating that the items reflect a single construct.

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 2. 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 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 parameters are similar over runs of the hot deck imputation procedure. Figure 2 is similar to Figure 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

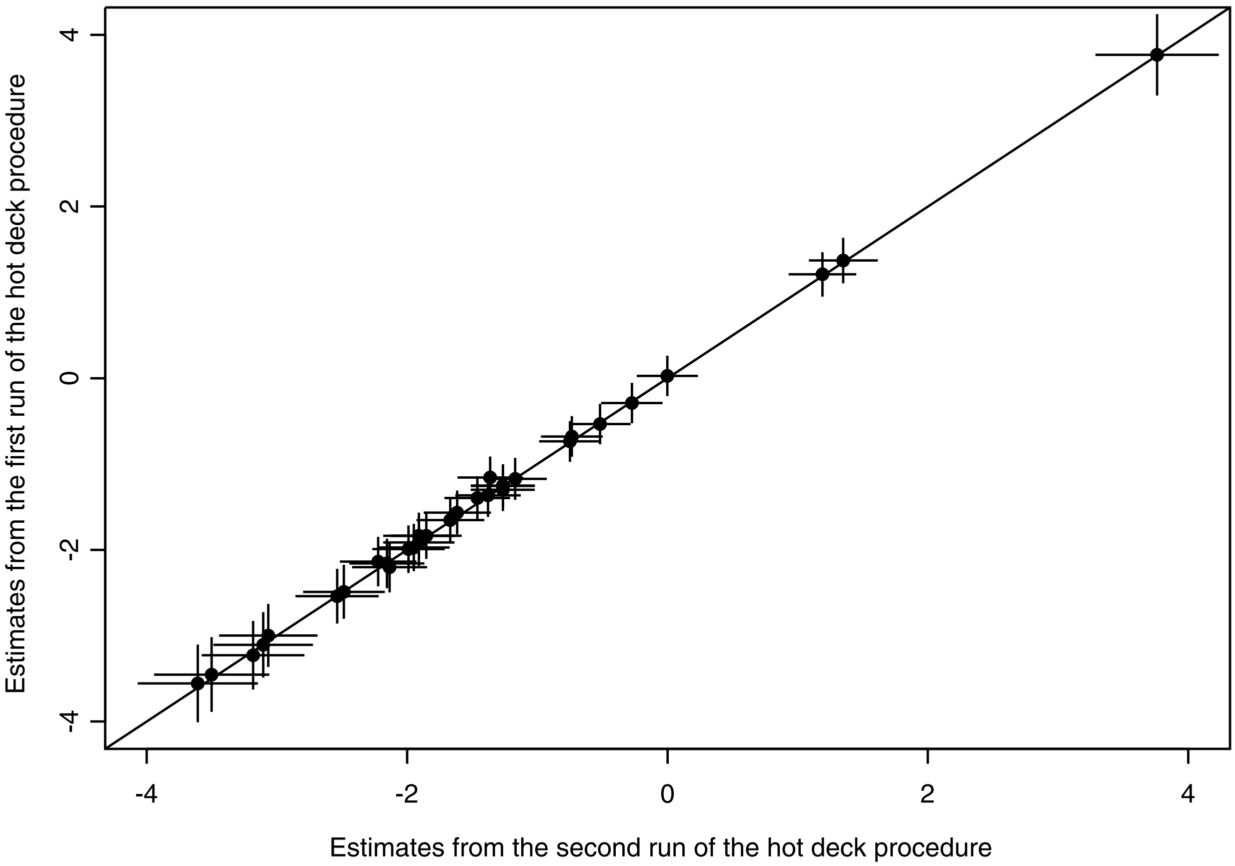

Figure 1
The estimates of the item parameters obtained using the first two runs of the hot deck imputation procedure. The horizontal and vertical lines indicate the 95% confidence intervals for the estimates obtained using the first and second runs, respectively.

obtained from a single run, illustrating the correction made to account for the fact that some data points are imputed.

Re-examining Table 2, 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 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 patients. In contrast to Figures 1 and 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 2 and Figure 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.

The RMSD, in Table 2, 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

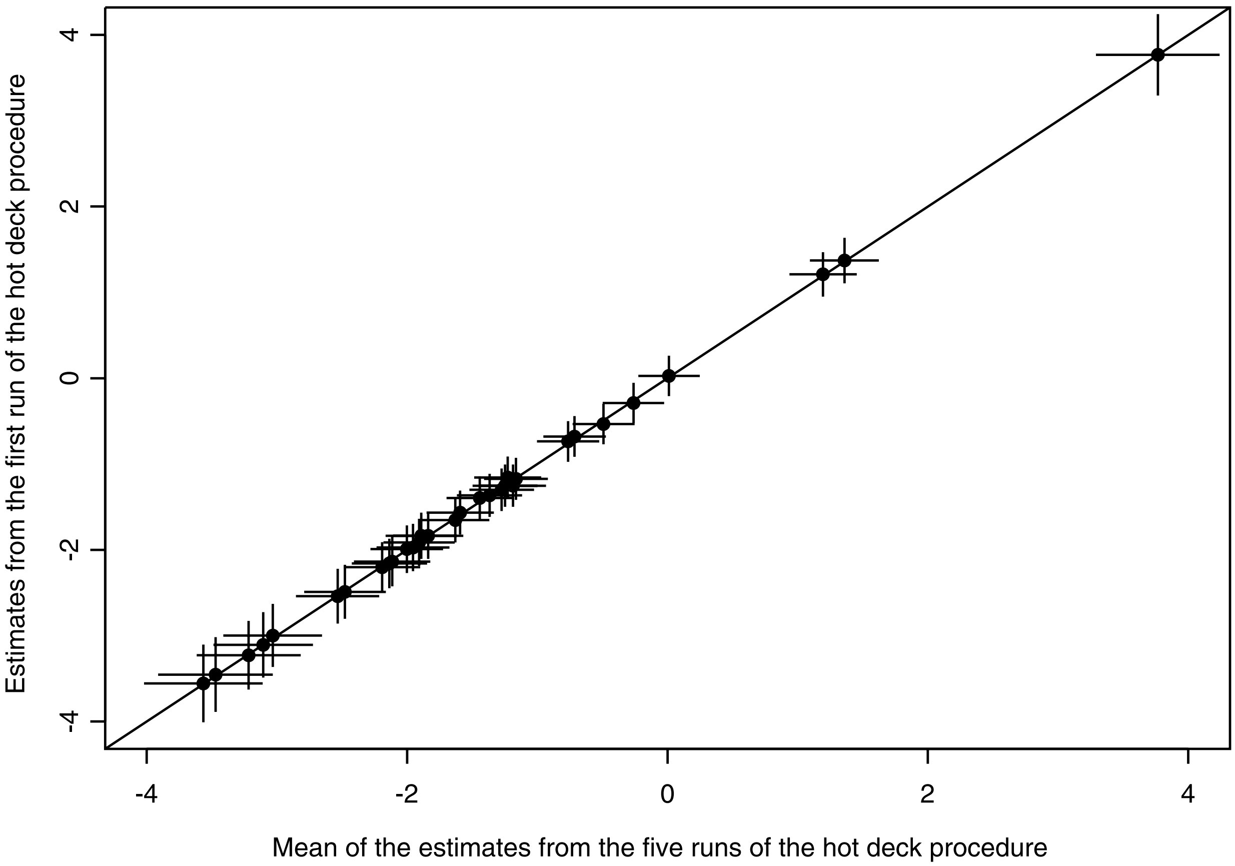

Figure 2
The estimates of the item parameters obtained using the first run and the mean of five runs of the hot deck imputation procedure. The horizontal and vertical lines indicate the 95% confidence intervals for the estimates obtained using the first and second runs, respectively.

Table 2: The root mean squared differences.

	Cold deck	1st run hot deck	2nd run hot deck	Mean 5 runs hot deck	Items never 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

The root mean squared differences. 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'.

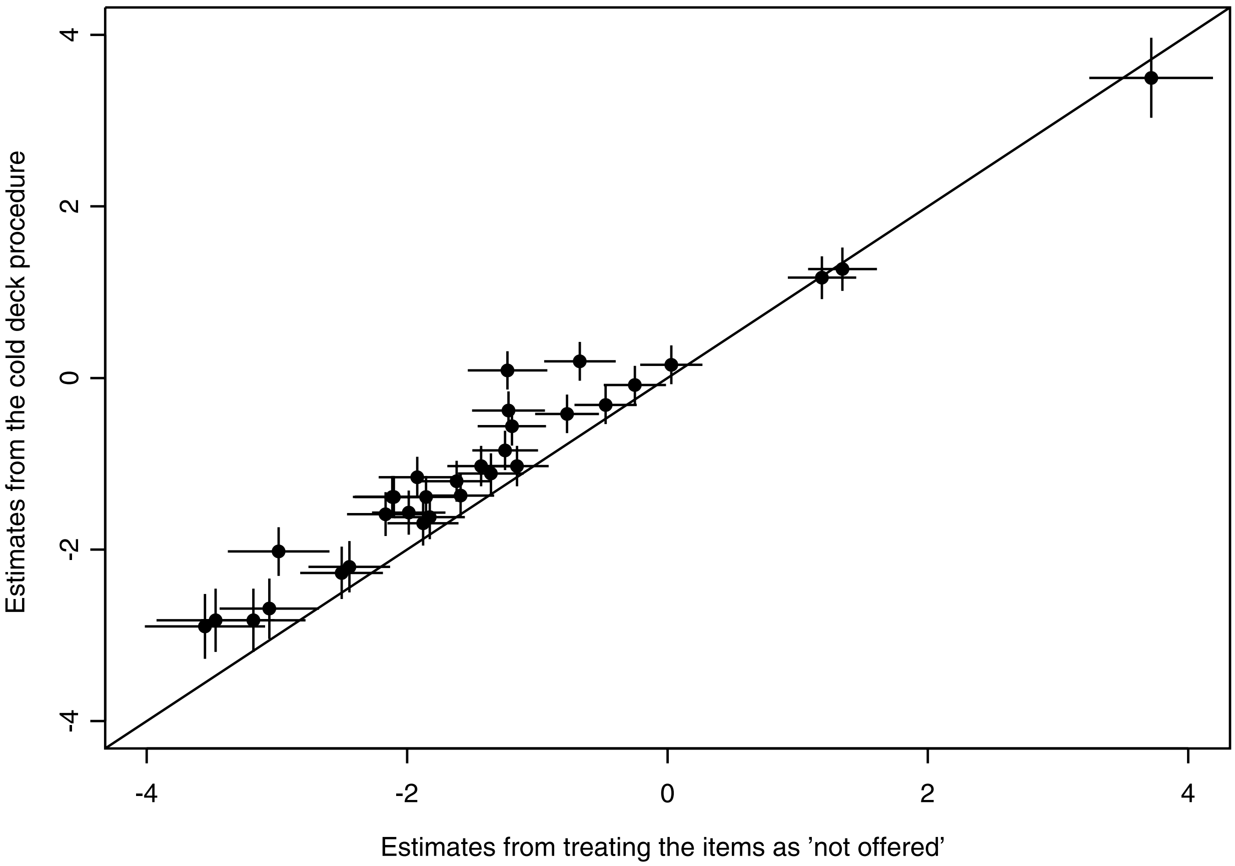

Figure 3
The estimates of the item parameters obtained using the cold deck imputation procedure and by treating the missing item responses as if they had never been offered to the individual patients. The horizontal and vertical lines indicate the 95% confidence intervals for these estimates.

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 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 2 and Figure 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. The mean and standard deviation are lower when cold deck imputation is used

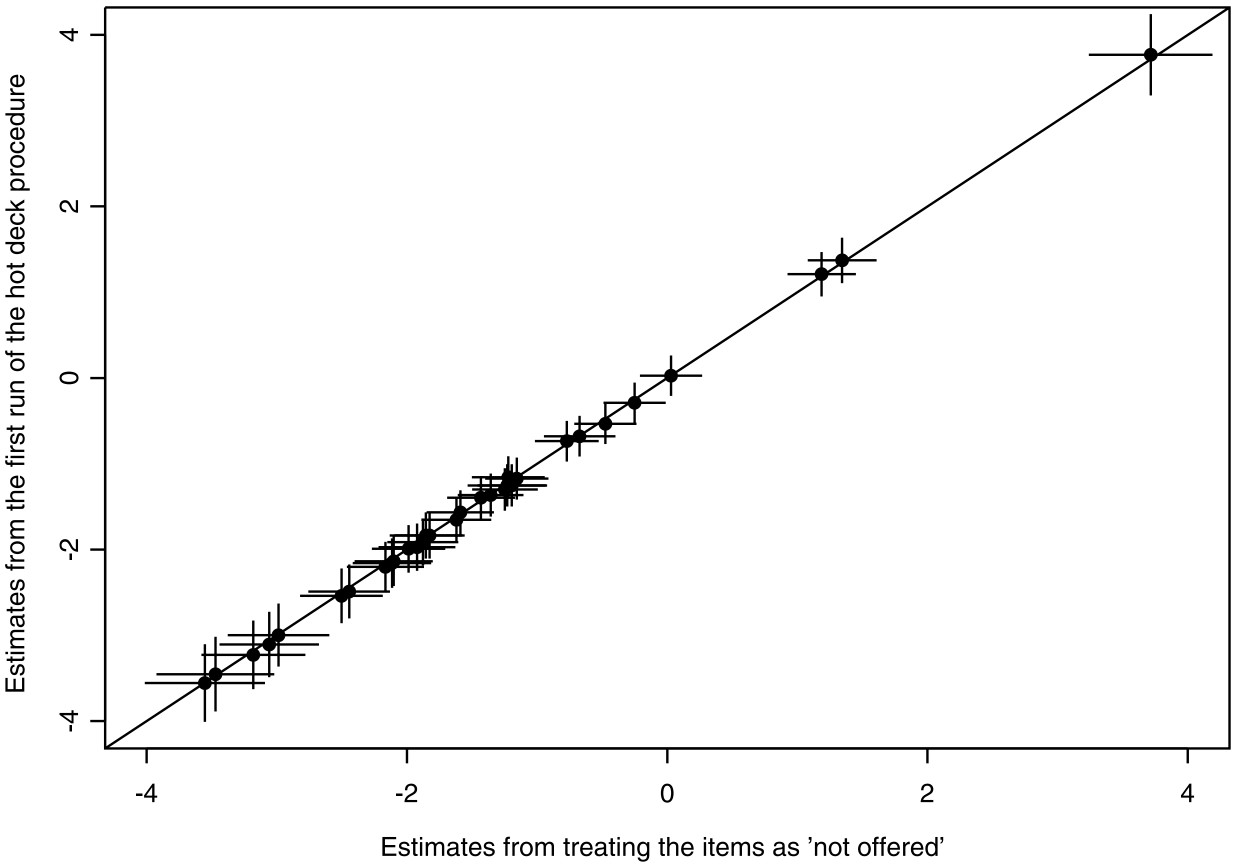

Figure 4
The estimates of the item parameters obtained using the first run of the hot deck imputation procedure and by treating the missing item responses as if they had never been offered to the individual patients. The horizontal and vertical lines indicate the 95% confidence intervals for these estimates.

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 [16].

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

Table 3: The root mean squared differences.

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 root mean squared differences. 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.

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 [19], although there is a large degree of differential item functioning between male and female and between younger and older respondents [14]. 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 [19]. 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,

$$t_k = \frac{m_{1k}}{m_{0k} + m_{1k}} \qquad (2)$$

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

$$r_{ik} = \frac{\exp\left(b_{0i} + b_{1i}t_k\right)}{1 + \exp\left(b_{01} + b_{1i}t_k\right)} \qquad (3)$$

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, $l \in (1, 2,..., K)$, 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

$$\text{s.e.}\left(\bar{\beta}_i\right) = \frac{1}{\sqrt{5}}\sqrt{\sum_{j=1}^{5}\left(\beta_{ij} - \bar{\beta}_i\right)^2 + \text{s.e.}\left(\beta_{ij}\right)^2} \qquad (4)$$

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 [20].

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 [21]. The likelihood, L, of a particular response pattern for the one parameter logistic IRT model can be written

$$L = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(p_{ik}^{I_{ik}}\left(1 - p_{ik}\right)^{1-I_{ik}}\right)^{J_{ik}} \qquad (5)$$

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[21].

The root mean squared difference

The root mean squared difference (RMSD) is defined as

$$\text{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{i,a} - \hat{\beta}_{i,b}\right)^2} \qquad (6)$$

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

Abbreviations

IRT = Item response theory

ALDS = AMC Linear Disability Score

RMSD = Root mean squared difference

Competing interests

None declared.

Funding

RH and RL were supported by a grant from the Anton Meelmeijer fonds, a charity supporting innovative research in the Academic Medical Center, Amsterdam, The Netherlands.

Authors contributions

RH conceived the study, prepared the first draft and carried out the analyses. CAWG, RL, AHZ and RJdH critically reviewed the manuscript. RH prepared the final version.

Acknowledgement

The authors would like to thank Janneke te Marvelde for her help in developing the Figures in this paper.

References

1. 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* 1994, **32:**40-66.
2. Rand Health Sciences Program: *Rand 36-item Health Survey 1.0* Santa Monica, California: Rand Corporation; 1992.
3. 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* 1986, **49:**698-709.
4. 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* 2001, **27:**4-5 [http://www.mapi-research-inst.com/allissue.asp].
5. Fayers PM, Curran D, Machin D: **Incomplete quality of life data in randomized trials: missing items.** *Stat Med* 1998, **15:**679-696.
6. 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* 2001, **20:**305-16.
7. 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* 2002, **55:**184-91.
8. Schafer JL: *Analysis of incomplete multivariate data* New York: Chapman and Hall; 1997.
9. Heckman JJ: **Sample selection bias as a specification error.** *Econometrica* 1979, **47:**153-161.
10. Rubin DB: **Inference and missing data.** *Biometrika* 1976, **63:**581-92.
11. Rubin DB: *Multiple Imputation for Nonresponse in Surveys* New York: Wiley; 1987.
12. Little RJA, Rubin DB: *Statistical analysis with missing data* New York: Wiley; 1987.
13. Lindeboom R, Vermeulen M, Holman R, de Haan RJ: **Activities of daily living instruments in clinical neurology, optimizing scales for neurologic assessments.** *Neurology* 2003, **60:**738-742.
14. 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 [http://www.mapi-research-inst.com/allissue.asp].
15. Fillenbaum GG, George LK, Blazer DG: **Scoring nonresponse on the mini-mental state examination.** *Psychological Medicine* 1988, **18:**1021-5.
16. Bergner M, Bobbitt RA, Carter WB, Gilson BS: **The sickness impact profile: development and final revision of a health status measure.** *Med Care* 1981, **19:**787-805.
17. 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
18. Kolen MJ, Brennan RL: *Test Equating* New York: Springer; 1995.
19. 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* 2003, **4:**19-33.
20. Wu ML, Adams RJ, Wilson MR: *ACER ConQuest: Generalised Item Response Modelling Software* Melbourne: ACER Press; 1998.
21. Thissen D: **Marginal maximum likelihood estimation for the one parameter logistic model.** *Psychometrika* 1982, **47:**175-186.
22. Lord FM: **Maximum likelihood estimation of item response parameters when some responses are omitted.** *Psychometrika* 1983, **48:**477-482.
23. van der Linden WJ, Glas CAW: *Computerized Adaptive Testing. Theory and Practice* Dordrecht, the Netherlands: Kluwer Academic Publishers; 2000.
24. Mislevy RJ, Chang H: **Does addaptive testing violate local independence?** *Psychometrika* 2000, **65:**149-156.
25. Andersen EB: **Estimating latent correlations between repeated testings.** *Psychometrika* 1985, **50:**3-16.
26. Holman R, Glas CAW: **Modelling non-ignorable missing data mechanisms with item response theory models.** *British Journal of Mathematical and Statistical Psychology* in press.
27. Rasch G: **On general laws and the meaning of measurement in psychology.** In *Proceedings of the Fourth Berkely Symposium on Mathematical Statistics and Probability* 1961, **4:**321-34.
28. Cronbach LJ: **Coefficient alpha and the internal structure of tests.** *Psychometrika* 1951, **16:**297-334.
29. Wright BD, Masters GN: *Rating scale analysis: Rasch measurement* Chicago, IL: MESA Press; 1982.
30. May K, Nicewander WA: **Measuring change conventionally and adaptively.** *Educational and Psychological Measurement* 1998, **58:**882-897.
31. Little RJA, Rubin DB: **On jointly estimating parameters and missing data by maximising the complete-data likelihood.** *American Statistician* 1983, **37:**218-220.
32. Pinheiro JC, Bates DM: *Mixed-Effects Models in S and S-PLUS* New York: Springer-Verlag; 2000.