Missing Data in Behavioral Science Research: Investigation of a Collection of Data Sets

Mark Huisman *

Department of Statistics, Measurement Theory, & Information Technology
University of Groningen, The Netherlands.

Abstract

Missing data is a phenomenon which is frequently encountered in empirical research. There are many theoretical reviews about how to handle missing data, but to what extent and in which way does missing data occur in behavioral science research? This paper gives an idea of the extent of item nonresponse in a collection of data sets consisting of answers to test items from empirical research in this field. An important characteristic of the missing data is the (non)randomness of the missing data patterns. However, determining the nature of the patterns based on the information available in the data set, is not easy. There are several methods which can be used to investigate the nature of the missing data patterns. These methods indicate that the data in the collection of data sets are not randomly missing.

Key words: item nonresponse, missing data mechanisms, item response theory.

1 Introduction

In empirical research a researcher is frequently confronted with *missing data*. Even if the research is carefully planned and implemented there are often some individuals who either refuse to participate or do not supply answers to certain questions. This failure to completely measure units can cause serious problems (Roth, 1994, Little & Schenker, 1995). Researchers are often inclined not to pay too much attention to them, or to the methods for handling missing data. Ignoring the missing data by analyzing complete cases only,

*Mark Huisman, Department of Statistics, Measurement Theory, & Information Technology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, telephone: +31 50 3636193, email: M.Huisman@ppsw.rug.nl

however, is misleading because it implicitly subscribes one type of missing data treatment, *i.e.* listwise deletion (Kromrey & Hines, 1994). This particular missing data treatment does not only suffer from loss of information by discarding observed values of incomplete cases, it also ignores the mechanism causing the missingness. This can lead to seriously biased results due to systematic differences between complete and incomplete cases.

Missing data treatments which are commonly found in the literature, are often based on the assumption that the missing data mechanism is ignorable. This roughly means that the analyses can be based on the observed portion of the data, in which all information about the missingness process is contained (a more precise definition will be given in section 3, see also Rubin, 1976). When this assumption holds, the data are said to be *missing at random*, and conditionally on some variables related to the missingness, there are no systematic differences between respondents and nonrespondents. The question now rises whether to expect the data to be missing at random.

In this paper a collection of data sets from the behavioral sciences is investigated to give an idea to what extent and in which way the problem of missing data may occur in this field. Assessing the nature of the missing data, and especially the randomness assumption, is an important step in determining a treatment procedure. Therefore, several methods which use the information available in the data, are applied to the data sets to determine the nature of the missing data patterns. Before investigating the occurrence of missing data in the data sets, the structure of the data will be discussed in section 2. In section 3 the process causing the missingness will be discussed, as well as methods which can be used to assess the nature of the missing data. The results of the investigation of the data sets are presented in the last sections.

2 Item nonresponse in scales

In the behavioral sciences researchers are often interested in person characteristics which are not directly observable. For this purpose scales consisting of items which cover these latent traits are constructed. The responses to all items in the scale are used to make inferences about the unobservable properties (*e.g.* mathematic skills, emotional stability, or attitude towards some topic). Measuring the latent abilities by modeling the item responses is a difficult task. In the presence of item nonresponse, however, this task is even

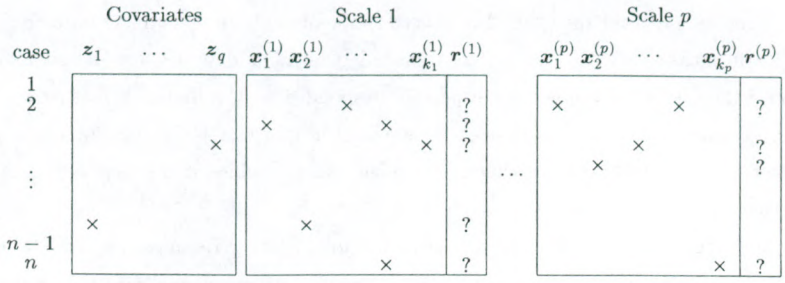

Figure 1: Missing values in a data set consisting of scales. × indicates a missing value, ? indicates the problem of computing a sum score in the presence of missing data.

more difficult.

The data sets investigated in this paper consist of scales measuring latent properties of respondents. The general form of the data is displayed in Figure 1. It consists of covariates $\boldsymbol{z}_h$ $(h = 1, \cdots, q)$ like age or sex, and several (sub)scales, which consist of responses to items $\boldsymbol{x}_i$ $(i = 1, \cdots, k)$ of $v = 1, \cdots, n$ respondents. All items have a fixed number of ordered response categories, and the sum score $\boldsymbol{r}$ of the item responses can be used as an estimate of the latent ability of the respondents.

Missing item responses are not uncommon in empirical research. Guadagnoli & Cleary (1992) and Ying (1989) studied the occurrence of item nonresponse in scales measuring some kind of physical and/or mental health concept. They found percentages of respondents with one or more items missing ranging from 2.2% to 16.3%, and 10.7% to 20.6%, respectively. Ferber (1966), Craig & McCann (1978), and Durand, Guffey, & Planchon (1983) studied the occurrence of item nonresponse in the field of marketing research. They found percentages of respondents with missing values as high as 65.5%, 53.8% to 100%, and 24.3%, respectively.

Measurement models used to estimate the latent properties of respondents can be found in the field of psychometrics. Earlier models used for measurement of the latent abilities were models from *Classical test theory*, more recent developments can be found in *Item response theory* (IRT; see *e.g.* Eggen & Sanders, 1993). An advantage of IRT models is

that they do not only model the relationship between the latent trait and the observed item responses, but that they can also be used in the presence of incomplete data. Because in IRT the measurement of persons is item independent and vice versa, latent abilities can be estimated even though some persons responded to a different set of items than others. In such incomplete test designs the mechanism that causes the incompleteness, *i.e.* the mechanism that determines which respondent is given which items, is specified by the researcher, who thus creates the missing data.

Incomplete test designs are often constructed for efficiency reasons. The estimation of the latent ability of respondents can be made more efficient when the difficulty of the items matches the level of ability of the respondents. These kind of incomplete designs can be found, for example, in targeted testing or adaptive testing (Glas, 1988, Mislevy & Wu, 1996). Another reason to use incomplete designs is that the strain on the respondents will be less because they have to answer a smaller set of items. For an overview of IRT models and incomplete designs see Eggen (1993).

Measurement in the presence of unintentionally missing data, on the other hand, is much more complicated. The problems that arise are twofold. First, results can be biased due to differences between complete and incomplete cases. The mechanism causing the missingness, however, is unknown in contrast with the case of incomplete designs. Second, the incompleteness of the data results in loss of statistical power, and will complicate the analysis with standard statistical techniques designed for complete data sets. Estimation of the latent ability of respondents can be severely biased, and although a number right score can be computed, it is probably not meaningful. Common treatment procedures proposed in the literature (Little & Rubin, 1989, Roth, 1994) like *imputation procedures* and *weighting procedures*, or *direct analysis of the incomplete data*, can be used to handle the missing values of the individual items, after which the latent ability is estimated.

3 The nature of the missing data

There are two important factors influencing the decision how to treat the missing data. These are the amount of missing data and the nature of the missing data mechanism. Roth (1994) gives an overview of suggested missing data techniques according to the amount and pattern of missing data.

3.1 Missing data patterns and mechanisms

An important assumption underlying most of the common missing data techniques, is the assumption that the process resulting in missing data is a random process. However, the simple dichotomy—randomly versus nonrandomly missing data—is not sufficient. It is therefore important to investigate the different patterns of missing data which may emerge in the data. Kim & Curry (1978) give the following categorization:

1. Missing data are randomly generated. This means that missingness of an item does not depend on the (unknown) value of that item nor on the values of other items in the data set.

2. Missingness of an item is dependent on the value of another variable, but not on the (unknown) value of the item itself. This variable on which the missingness depends may be another item in the test or a covariate.

3. Variables which are not observed determine the missingness. For instance nonresponse due to low cognitive skills in an attitudinal survey.

4. Missingness depends on the value of the missing item itself. The nonresponse occurs when the question topic is too sensitive or even threatening for the respondent.

5. Missing data is the product of a particular combination of two or more variables. Kim & Curry (1978) give the example of people with high education who may be unwilling to reveal their low income.

The first missingness pattern contains the assumption that the probability of response is independent of the observed and missing item responses. In this case Rubin (1976) defines the missing data to be missing at random (MAR) and the observed data to be observed at random (OAR) or, more simply, the data are *Missing Completely At Random* (MCAR). This means that the cause of missingness is unrelated to the items, and the observed values form a random subsample of the sampled values (see also Little & Rubin, 1989). When the data are MCAR most common treatments can be used to obtain unbiased results. In this case the problems that arise from the missing data are mainly a matter of statistical power.

A less stringent assumption is made when the missing data are assumed to be *Missing At Random*, but the observed responses are not assumed OAR. This means that the probability of a missing value for an item depends on the observed responses on other items or covariates, but given these, it does not depend on the missing value itself (Rubin, 1976). Within subgroups formed by the observed items (covariates) on which the missingness depends, the data are MCAR. This is pattern 2, described above, and conditional on the observed values of the items (covariates) upon which the missingness depends, analyses will give unbiased results.

A special case of pattern 3 occurs when the missingness depends on the latent trait, which is the topic of the investigation. This kind of missingness can also be seen as a special case of pattern 4, because all items together determine the value of the latent variable and each item contributes to the sum score. However, MAR holds if all information with respect to the missingness process is contained in the observed data. Special cases of these mechanisms are the intentionally created incomplete designs, mentioned in section 2. Here the process causing the missing data is known, and therefore it can be determined whether the data are indeed MAR.

In pattern 4 the missingness is truly nonrandom, *i.e.* the cause of missingness is the unknown value of the item itself. The systematic differences between the respondents and nonrespondents created by these patterns can seriously bias analyses that are based on only the observed cases. Therefore, the missing data mechanism cannot be ignored. Pattern 5, finally, can be seen as a combination of the patterns 2 and 4. This pattern occurs when the data are nonrandomly missing across and within subgroups formed by observed items or covariates, as opposed to MAR, where the data are only nonrandomly missing across subgroups (see also Roth, 1994).

3.2 Ignorable and accessible missing data mechanisms

The performance of any missing data treatment depends heavily on the mechanisms that lead to missing values. Ignoring these mechanisms is the easiest way out, but analyses may be severely biased when the data are not missing at random. Rubin (1976) gives a definition of *ignorability* of the missing data mechanism in terms of the mechanism itself and the method of data analysis. In this definition, characteristics of both the probability

model for the observed data and the missing data are used.

The probability model for missingness can be specified with a missing data indicator matrix $\boldsymbol{M} = [m_{vi}]$. The matrix $\boldsymbol{M}$ describes the patterns of missing data, with $m_{vi} = 1$ if respondent v has a missing value for item i and $m_{vi} = 0$ otherwise. The conditional distribution of $\boldsymbol{M}$ given the data is denoted by $Pr_{\phi}(\boldsymbol{M}|\boldsymbol{Z}, \boldsymbol{X})$, which is the probability model of the missingness with parameter ϕ. Let $(\boldsymbol{Z}, \boldsymbol{X})_{obs}$ denote the observed data and $(\boldsymbol{Z}, \boldsymbol{X})_{mis}$ the missing data. Formally, the data are MAR if

$$Pr_{\phi}(\boldsymbol{M}|\boldsymbol{Z}, \boldsymbol{X}) = Pr_{\phi}(\boldsymbol{M}|(\boldsymbol{Z}, \boldsymbol{X})_{obs}) \text{ for all } (\boldsymbol{Z}, \boldsymbol{X})_{mis} .$$

Now denote the parameters of the observed data model θ. If both MAR holds and the parameters ϕ and θ are not functionally related, Rubin's (1976) ignorability principle states that under direct likelihood inference, the missing data mechanism can be ignored. This means that the analyses can be based on the observed data which contains all information with respect to the missingness. In a similar way ignorability is defined for sampling inference and Bayesian inference.

When the missing data mechanism cannot be ignored it should be included in the data analyses. Graham & Donaldson (1993) define a missing data mechanism to be *accessible* when the cause of missingness has been measured and is available for use in the analysis. They show that when the cause of missingness is included properly in the analysis, accessible, nonrandom mechanisms cause no bias. The term accessible is related to the term ignorable, except that accessible refers only to the missing data mechanism, whereas ignorable refers to a combination of the mechanism and the data analysis.

When the missingness is dependent on an unobserved variable, the mechanism is generally inaccessible and cannot be included in the analysis. In intentionally incomplete test designs, however, missing data mechanisms are accessible although they may be dependent on the latent, and thus unobserved, ability of the respondents. Therefore, the ignorability of the these processes can be determined. Mislevy & Wu (1996) investigate the ignorability of missing data mechanisms in several incomplete designs (*e.g.* adaptive testing and time limit tests) for inferences about the latent ability of respondents. They find that for incomplete test designs in which missingness depends on observed variables or on the

latent trait, the missingness process is ignorable under likelihood inference, assuming item difficulties are known. On the other hand, Glas (1988) and Eggen (1993), find that for the estimation of item difficulties in incomplete test designs, the missingness process cannot always be ignored. For some designs the estimation of item difficulties with particular estimation procedures is even impossible.

Knowledge of the missing data mechanism is the main element in determining a treatment for missing data, and largely determines the performance of this treatment. It is, however, impossible to verify the MCAR assumption and the causes of missingness in practice without additional information. Still, one can investigate the missing data patterns in the data and use the available information to make reasonable guesses about the mechanism.

Information available in the data can be used in two ways. First, the missing data patterns can be investigated by looking at the item responses. Because the items together measure one latent trait there exists a (strong) relationship between them. The information the items give about each other and about the respondents, can be used to make assumptions about the missing data mechanism. Secondly, covariate information can be used to make assumptions about the missing data mechanism. Nonrandomness, in the sense that missing data occurs more frequently in certain subgroups defined by the covariates, can be investigated. The correlates of item nonresponse are determined and can be used in the analyses.

3.3 Information from item responses

As an 'advance organizer', first the random distribution of the blanks in the data matrix is investigated. Although the MAR assumption is not tested, the results will give the researcher an indication of the nature of the missing data. The randomness of the missing data distribution can be studied in several ways. The easiest test is by counting the number of missing data in each row and column of the data matrix and comparing the observed distributions with a suitably chosen hypothetical one.

Kim and Curry (1978) suggest a strategy to investigate the randomness of $k+2$ patterns of missing data. The patterns to consider are $\mathcal{M}_i$—missing only on item i, $\mathcal{MM}$—missing on two or more items, and $\mathcal{NM}$—none missing. For every item i the proportion of cases

with a missing value for this item, q_i can be computed. The expected number of cases with one of the $k+2$ patterns of missing data, under the assumption of independent items and random patterns, are: $\mathrm{E}(\mathcal{NM}) = N \prod_i p_i$, $\mathrm{E}(\mathcal{M}_i) = \frac{q_i}{p_i}\mathrm{E}(\mathcal{NM})$, and $\mathrm{E}(\mathcal{MM}) = N - \mathrm{E}(\mathcal{NM}) - \sum_i \mathrm{E}(\mathcal{M}_i)$, where $p_i = 1 - q_i$. The deviation of the observed frequencies from these expected frequencies can be tested with a χ^2-test with $k+1$ degrees of freedom. In the same way more patterns (like missing values for both item i and item j) can be considered.

The mutual relationships among the indicator variables $\boldsymbol{m}_i$ can be used to detect any significant clustering between the missingness of particular items. This can be done by examining the correlation matrix of the indicator variables (for dichotomous items the correlation equals Cramer's ϕ, for which a χ^2 significance test can be computed, Kim & Curry, 1978). Also factor analysis or some kind of scale analysis can be applied to the indicator variables. It should be noted that it is not fully justified to apply factor analysis to dichotomous variables, but it may give an idea of the degree of clustering and may help to determine the accessibility of the mechanism by identifying the underlying causes of missingness.

After investigating the randomness of the distribution of the blanks, the MCAR assumption should be investigated by testing the association between missingness of one item and the responses on others. For that, the distribution of a particular item based on the complete cases should be compared with the distribution of this item based on the incomplete cases for which it is recorded. In particular, t-tests are used to compare the means of respondents and nonrespondents of an item. For each item i the sample is split into two groups: cases for which item i is observed, and cases for which the item is missing. The differences in the means of the observed values of the other items j in the two groups are then tested. Significant differences between the means indicate that the MCAR assumption is not valid. This procedure yields up to $k(k-1)$ t-tests for test length k. The mean of item j across cases with $\boldsymbol{x}_i$ missing and nonmissing, can also be compared by testing the significance of the regression coefficient β of the regression $\mathrm{E}(\boldsymbol{x}_j) = \alpha + \beta \boldsymbol{m}_i$, where $\boldsymbol{m}_i$ is the missing data indicator variable of item i (Kim & Curry, 1978). In the same way the correlation between $\boldsymbol{m}_i$ and $\boldsymbol{x}_j$ can be used to study the MCAR assumption.

Instead of means, also the correlations between the items can be compared over the

complete and incomplete cases. The difference between the correlation matrix for the complete cases $\boldsymbol{R}_c$ and the incomplete cases $\boldsymbol{R}_m$ can be tested by Bartlett's χ^2-test. For the incomplete cases the correlation matrix is computed by pairwise deletion. The problem which arises when using pairwise deletion, is that the correlations between the items are based on different sample sizes. To overcome this problem, the harmonic mean of the different sample sizes is used to provide for a good compromise sample size.

Another way to use t-tests is in testing whether the mean item score based on observed cases, is related to the number of missing values. The mean item scores are computed for different numbers of missing items, and the association between the number of missing items and the scores can be examined. A clear positive or negative association indicates systematic missing data: missingness is dependent on the latent ability of respondents (see *e.g.* Molenaar, 1997).

3.4 Correlates of Item Nonresponse

When modelling under the MAR assumption the use of covariates, especially completely observed covariates, is important. There are several studies examining correlates of item nonresponse. Most of them examined variables like age, sex, education, and occupation. Higher item nonresponse rates were found for females, older respondents, less educated respondents, and respondents with a lower level job (Craig & McCann, 1978, Durand, Guffey, & Planchon, 1983, Colsher & Wallace, 1989, Ying, 1989). Also item topic (Craig & McCann, 1978, and Colsher & Wallace, 1989) and scale topic (Colsher & Wallace, 1989, and Guadagnoli & Cleary, 1992) were found to be related to item nonresponse. Omura (1983) studied the relation between personality characteristics of respondents and the occurrence of item nonresponse. He found that self-image traits and psychographic traits are related to item nonresponse. In these studies, however, the traits were measured by one item instead of a scale of items, as is usual in IRT.

4 A collection of data sets

The data investigated in this paper come from a number of already performed empirical studies in the behavioral sciences. They are of course not a random sample from all possible

studies, but will be useful to give an impression of the diversity of item nonresponse and missing data patterns that can be encountered in empirical research.

Data set 1: Population screening of hypertension The data in this set have been collected within the framework of a population screening program set up by the Groningen Hypertension Service and the University of Groningen in 1993, to examine the occurrence of hypertension. The questionnaire used contains the *RAND 36-item Health Survey* (RAND-36), which is a self-administered test containing 36 items measuring eight health concepts, and was sent to participants of 60 years and older. Three subscales are used in this investigation, *i.e. Physical Functioning, General Health*, and *Mental Health*. Another test included in the questionnaire is the *Duke Activity Status Index* (DASI), with which the overall functional capacity of the respondents is tested.

Data set 2: VOCL-89 research In this study Dutch students in grade 5 of schools of higher general secondary education (HAVO) and pre-university education (VWO) were asked to fill in a questionnaire containing, among others, questions about courses, homework, and learning. The questions are the same for the HAVO and VWO students except when dealing with the courses the students take (which are different for the two groups). In this paper a scale consisting of 24 four-category items asking about attitudes towards school and learning will be examined for the HAVO students. Two scales are investigated for the VWO students, one asking about the way in which the students study, the other about the effect of strain on performance.

Data set 3: English language test Students of different secondary schools completed a test paper to test their skill in several aspects of the English language. The schools differ in level of education (MAVO, HAVO, VWO) and within each level students of four different schools were tested. The test consists of 40 items covering grammar, textual comprehension, and reading skills. Only the items measuring textual comprehension are multiple choice questions, for which the score is either 0 (incorrect) or 4 (correct). For the other, open ended questions, the obtained score is either 0 when the answer is not correct, or a number between 1 and 4 when the answer is (partially) correct.

Data set 4: Mathematics test The English Language Test discussed above is part of a larger research program. Another part of this program is a test of mathematical skills. The test was developed by the Dutch Testing Service (CITO), as was the English Language Test, and again students of different secondary schools took the test. A total of 884 students provided answers to the 32 items of which 12 are multiple choice items. For every item the score is either 0 (incorrect answer) or a maximum score (correct answer). This maximum score depends on the difficulty of the item and ranges from 1 to 8 points.

Data set 5: Social interactions research The data was collected by the Northern Center for Healthcare Research (NCG) in Groningen, to study the influence of social support on the level of psychological and physical complaints of cancer patients. The questionnaire used contains, among others, the 34 items belonging to the *Social Interactions Scale* (SSLI), and the 48 items of the *Eysenck Personality Questionnaire* (EPQ). Both tests consist of several subscales and are answered by 730 respondents belonging to one of two groups, a patient and control group containing 475 and 255 respondents, respectively.

Data set 6: Ageing and memory Within the framework of a Dutch research program to investigate the consequences of ageing (NESTOR), a questionnaire was sent to 392 participants aged 45 and older in order to study the consequences of ageing on memory. The questionnaire can be divided into three parts: a somatic, memory, and activity part. In the first two parts several topics (like hearing or physical well being, remembering names or faces) were investigated with the help of four items for each topic: comparison with peers and with people of age 25, comparison with yourself at age 25, and expectation for the future. In the activity part of the questionnaire there are three items per topic: frequency of occurrence, comparison with yourself at age 25, and expectation for the future. For all items asking about future expectations a 'don't know' response category was offered.

5 Missing data in the collection of data sets

The occurrence of item nonresponse is investigated in each of the six data sets. The main results are presented in Table 3 (occurrence of item nonresponse) and Table 5 (randomness of the missing data). The results are illustrated with an example.

5.1 Occurrence of missing data

Example 1: Physical Functioning

The *Physical Functioning* scale is a subscale of the *RAND-36*, measuring the functional status of respondents. It consists of 10 items which each have 3 ordered response categories (van der Zee & Sanderman, 1993). The number of respondents who returned the questionnaire is 2773, of whom 78 (2.8%) have missing values for all 10 items. Of the 2695 remaining respondents, 8.6% has one or more items missing, and the mean number of items missing is 0.24. The data matrix consists of 2695×10 values of which 651 are missing (score 9 in Table 1). The overall percentage missing data, Q_0, is therefore 2.4%.

The observed percentages missing for every item can be tested against Q_0. The result of the 10 binomial tests of difference of the percentages missing (testing the homogeneity of the percentages missing) can be found in Table 1. In this table it can be seen that item 1 has the largest amount of missing values and item 10 the smallest ($Q_0^{10} = 1.7\%$ and $Q_0^1 = 4.3\%$). To test the homogeneity of the percentages missing, the 10 tests should not be investigated separately, but treated as a multiple test procedure to prevent capitalizing on chance. This results in one item which differs significantly from Q_0 instead of three, as would follow from the P-values in Table 1 ($\alpha = 0.05$).

The probabilities of missingness from Table 1 are used to compute the expected number

Table 1: Observed marginal frequency distributions of the *Physical Functioning* data.

item	0	1	2	9	proportion missing	two-sided P-value
1	1147	935	497	116	0.043	< 0.0001
2	442	821	1384	48	0.018	0.032
3	546	797	1300	52	0.019	0.100
4	800	965	861	69	0.026	0.625
5	328	634	1671	62	0.023	0.697
6	567	873	1200	55	0.020	0.205
7	825	618	1178	74	0.027	0.264
8	545	607	1482	61	0.023	0.607
9	354	481	1791	69	0.026	0.625
10	88	309	2253	45	0.017	0.012

Table 2: Observed and expected frequency of the number of missing items and the Kim & Curry patterns in the *Physical Functioning* data.

number missing	0	1	2	3	4	5	6	7	8	9	
obs. frequency	2462	113	30	34	8	6	6	18	10	8	
exp. frequency	2110	523	58	4	0	0	0	0	0	0	
K&C patterns	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$	$\mathcal{M}_9$	$\mathcal{M}_{10}$	$\mathcal{MM}$
obs. frequency	54	5	5	10	7	8	13	3	6	2	120
exp. frequency	95	38	42	55	50	44	60	49	55	36	62

of missing items, under the assumption of independence of the probabilities. In Table 2 the observed and expected frequencies of some patterns of missing data can be found.

The large difference between the observed and expected frequencies of the number of missing items contradicts the independence of missingness across items (the χ^2 statistic is rapidly increasing when the expected frequencies become small for large number of missing values). The method proposed by Kim & Curry (1978) checks in more detail the different patterns when there is only one item missing. The difference between the observed and expected frequencies is found to be highly significant ($\chi^2 = 451.2$, $df = 11$). The items for which the largest frequencies missing are found (items 1, 4, and 7) appear to be the 'most difficult' items, *i.e.* the mean item scores are the smallest (0.75, 1.02, and 1.13) which means that many respondents have difficulties performing these tasks (walking more than 1 mile, or climbing several flights of stairs). When looking in more detail at the patterns with three missing values, 29% is caused by items 1, 2, and 3, which are on the same page of the questionnaire. The same is true for the pattern with seven missing values, caused by items 4 to 10 in 94% of the cases. These items also are on the same page of the questionnaire.

Analysis of the correlation matrix of the indicator variables $\boldsymbol{m}_i$ shows that the relationships between missingness of one item and missingness of another item are in some cases very strong. For instance the correlation between $\boldsymbol{m}_2$ and $\boldsymbol{m}_3$ is 0.91. Again, when testing the significance of the $k(k-1)$ correlations, a multiple testing procedures should be used. Because of the large sample size, the power of the individual tests is still as large as 0.95 when the difference between correlations is small (effect size of 0.10, Cohen, 1977). The same holds for a medium effect size of 0.30 with $N = 350$. When a factor analysis is

performed to examine the clustering of the indicator variables, two factors are extracted. It turns out that the items loading highest on each factor are on two different pages of the questionnaire.

The example of the *Physical Functioning* data showed the extent of item nonresponse (INR) encountered in this data set. The results are summarized in Table 3 in the following way:

> There are $N = 2695$ respondents of the *Physical Functioning* scale, which consists of $k = 10$ items with 3 ordered categories, of whom 8.6% has at least one item missing (*units INR*). The average number of items missing for the 2695 respondents is 0.24 (*mean INR*), and the overall percentage missing in the scale is 2.4% (Q_0). The observed percentages missing for every item have a minimum of 1.7% and a maximum of 4.3%. There is one item for which the proportion missing differs significantly from Q_0 ($dQ_0 = 1$), and there are no completely observed items ($nMD = 0$).

The occurrence of item nonresponse is also investigated in some subscales. The *RAND-36* consists of eight subscales, but only the scales with five or more items are taken into consideration. The subscales of the *English language test* and the *EPQ* are all included.

As can be seen in Table 3, the incidence of item nonresponse varies considerably across the tests. The percentage of respondents with at least one item missing ranges from 2.5% to 72.7%. Especially the *English* and *Maths* tests and the *Ageing & Memory* study have a large number of respondents with missing data, although the smallest amount of incomplete cases was found in a subtest of the English test paper. An explanation can be the different levels of education for the different groups of students for the two test papers. In the *Ageing & Memory* study the items asking about future expectations are responsible for the high percentage of item nonresponse and when these items are not taken into consideration, the number of nonrespondents drops significantly.

The overall percentage missing data, Q_0, ranges from 0.3% to 26.6%, with again *Ageing & Memory* having the largest percentage of missing data. Although on occasion the maximum percentage missing on an individual item may be quite high, the overall (or average) percentages for the other tests are all smaller than 10%. Also the homogeneity across items with respect to the percentage missing, differs from test to test. Especially for the *DASI*, *Mathematics*, and *Ageing & Memory* tests the items are very heterogeneous.

Table 3: Item nonresponse in a collection of empirical studies. N is the number of respondents, *units INR* is the percentage of respondents having at least 1 item missing, *mean INR* is the mean of the number of items missing for every respondent, k is the number of items, *cat* is the number of categories, Q_0 is the overall percentage missing, *min* is the minimum percentage missing, *max* is the maximum percentage missing, dQ_0 is the number of items for which the proportion missing differs strongly from Q_0, and *nMD* is the number of items having no missing values.

				units	mean	item nonresponse				
	N	k	cat.	INR	INR	Q_0	min.	max.	dQ_0	nMD
RAND-36	2773	36	2-6	25.4	2.10	5.8	1.1	9.0	10	0
Physical funct.	2695	10	3	8.6	0.24	2.4	1.7	4.3	1	0
General health	2715	5	5-6	6.4	0.18	3.6	0.6	4.8	3	0
Mental health	2645	5	6	4.6	0.09	1.8	1.6	2.2	0	0
DASI	2742	12	2	30.1	0.51	4.3	0.3	21.8	10	0
HAVO learning	837	24	7	5.0	0.10	0.4	0.0	1.8	1	3
VWO learning	1197	19	5	3.6	0.05	0.3	0.0	0.8	1	1
strain[a]	1157	8	8	19.5	0.27	3.4	1.3	10.8	4	0
English test[b]	847	39		36.1	0.92	2.4	0.0	12.4	24	2
grammar	847	10		17.2	0.28	2.8	0.7	9.4	5	0
text.	847	10	3-4[c]	2.5	0.03	0.3	0.0	1.4	1	2
reading	847	19		27.7	0.61	3.2	0.1	12.4	13	0
Mathematics test	884	32		70.2	2.38	7.4	0.0	36.5	30	1
multiple choice	884	12	4-5[c]	19.6	0.29	2.4	0.0	10.5	9	1
SSLI	730	34	4	12.2	0.27	0.8	0.1	1.6	0	0
EPQ	730	48	2	18.8	0.40	0.8	0.0	3.6	1	2
neuroticism	730	12	2	6.0	0.08	0.6	0.2	1.5	0	0
psychoticism	730	12	2	5.4	0.06	0.5	0.0	1.6	2	2
extraversion	730	12	2	10.3	0.15	1.2	0.6	3.6	1	0
social desirability	730	12	2	7.1	0.11	0.8	0.3	1.9	1	0
Ageing&Memory	392	65	3-4	72.7	5.51	8.5	0.0	34.4	65	1
future items[a]	392	19	3-4	72.2	5.05	26.6	11.7	34.4	4	0
other items	392	46	3-4	12.2	0.46	1.0	0.0	4.3	2	1

[a] 'Don't know' option was offered. [b] One item was removed due to layout problems.
[c] The responses have been dichotomized based on correct/incorrect responses.

5.2 Nature of the missing data patterns

Example 1 continued

The nature of the missingness in the *Physical Functioning* data is examined with the methods mentioned earlier. The MCAR assumption is tested with $10 \times 9 = 90$ pairwise t-tests. A stem-and-leaf plot of the t-statistics is shown in Figure 2; for the sample size studied they can be viewed as normal deviates.

When treated as a multiple testing procedure with 90 null-hypotheses of equal means, only two hypotheses are rejected. This does not indicate a very serious violation of the MCAR assumption. However, care should be taken with this simultaneous inference because the t-statistics are correlated. Besides, the power of the individual tests is often very small because of small sample sizes. With a sample size of 25 or lower, which occurs in 31 cases, the power is 0.51 or lower for detecting differences of 0.4 standard deviation units (medium effect size of 0.4, Cohen, 1977). For detecting a small effect size (0.2) the power is even as low as 0.17.

The correlations between the missing data indicator variables $\boldsymbol{m}_i$ and the item values also indicate violations of the randomness assumption in the *Physical Functioning* scale, although the Bartlett χ^2-test shows no significant difference in correlations between the items for the respondents and the nonrespondents ($\chi^2 = 47.4$, $df = 45$).

The last effect that was examined is the effect of missingness on the frequency distribu-

2	-4	.	02
1	-3	.	4
3	-2	.	169
11	-1	.	01122257889
9	-0	.	002222355
24	0	.	001334444445566777788999
19	1	.	0001123444455668889
17	2	.	00122233444566688
4	3	.	1145
0	4	.	

Figure 2: Distribution of 90 pairwise t-statistics for the *Physical functioning* data.

Table 4: Average valid score on the observed items as a function of the number of missing items for the *Physical Functioning* data.

number missing	frequency	score	stand. deviation
0	2462	1.300	0.607
1	113	1.461	0.519
2	30	1.212	0.689
3,4,5	48	1.168	0.672
6,7,8,9	42	1.054	0.732

tion of the mean item scores of the observed items. In Table 4 the average valid scores on the observed items can be found as a function of the number of missing values. It shows a negative association between missing data on some items and mean scores on other items; more missing, lower valid score, although the pattern is not as clear as in Molenaar (1997, p. 43).

The covariates sex and age are used to investigate the occurrence of missing data in the subgroups formed by these variables. Although a small difference in the average number of missing values was found between men and women (women tend to have more missing data), this was not significant ($\alpha = 0.05$). For age a significant difference was found; older persons tend to have more missing observations.

As was argued in section 3, the methods to investigate the nature of the missingness can be divided into two groups; the first group testing the random distribution of the blanks in the data matrix, the second testing the association between missingness on one item and the responses on others. The dependence on the value of the item itself, however, can never be tested because it is unknown. In Table 5 the results of the tests for the collection of data sets are presented; the first group in columns 1 to 4 , the second in columns 5 to 8.

It is hard to draw general conclusions from Table 5 because not all tests tell the same thing about the missingness. Each test investigates only some specific part of the randomness assumption, as it should, because missing data may emerge in different ways. The tests of MCAR in columns 5 to 8 of Table 5, however, are the ones of which the results vary the most and sometimes even contradict.

Table 5 shows that in almost all data sets the missing values are not randomly dis-

Table 5: Results of investigating the nature of the missing data. *MD item* is the test of homogeneity of the fraction missing for very item in the set, *MD case* is the test of the frequency distribution of the number of missing items, the *K&C-test* is the Kim & Curry test for patterns of missing data, *corr* $\boldsymbol{M}$*-*$\boldsymbol{M}$ is the examination of the mutual correlations between the missing data indicator variables, *corr* $\boldsymbol{M}$*-*$\boldsymbol{X}$ is the examination of the correlations between the missing data indicator variables and the items, B is Bartlett's χ^2-test, *valid score* is the examination of the average valid item scores, r indicates randomness, n nonrandomness.

	MD item	MD case	K&C test[a]	corr $\boldsymbol{M}$-$\boldsymbol{M}$[a]	corr $\boldsymbol{M}$-$\boldsymbol{X}$[a]	t-test[a]	B[a]	valid score[d]
RAND-36	n	n	n	n	n	–[bc]	–[b]	–
Physical functioning	r/n	n	n	n	n	n	r	n
General health	n	n	n	n	n	n[c]	n	n
Mental health	r	n	n	n	r	r[c]	n	r/n
DASI	n	n	n	n	n	n[c]	n	n
HAVO	r/n	n	n	n	n	–[bc]	r	r/n
VWO learning	r/n	n	n	n	r/n	–[c]	r	r
strain	n	n	n	r/n	r/n	r	r	r
English test	n	n	n	n	n	–[bc]	n[c]	–
grammar	n	n	n	n	n	n[c]	n	n
text.	r/n	r/n	r	r/n	r/n	–[c]	r[c]	r/n
reading	n	n	n	n	n	n[c]	n	r/n
Mathematics test	n	n	n	n	n	–[bc]	n	–
multiple choice itm.	n	n	n	n	n	n[c]	r	n
SSLI	r	n	n	n	r/n	–[bc]	r	r
EPQ	r/n	n	n	n	n	–[bc]	r	–
neuroticism	r	r/n	n	r/n	r	–[c]	n	r
psychoticism	n	r	r/n	r	r/n	–[c]	–[c]	r
extraversion	r/n	r/n	n	n	n	n[c]	r	r/n
social desirability	r/n	r/n	n	n	n	–[c]	n	r/n
Ageing & Memory	n	n	n	n	n	–[bc]	–[b]	–
future items	n	n	n	n	n	r	r	–
other items	n	n	n	n	n	–[bc]	r[c]	–

[a] items without missing values are not used in the analysis.

[b] not performed due to large number of items.

[c] few data left due to small number of cases with missing data on particular items.

[d] only (sub)scales measuring one latent trait are considered.

tributed over the data matrix (columns 1 to 4). Only in the textual skills part of the *English language* test and in some less degree the subscales of the *EPQ*, the missing data appear randomly distributed. The columns 5 to 8 show that in the scales *Mental Health*, *VWO learning* and *strain*, *SSLI*, and in a less degree *English language text.*, some subscales of the *EPQ*, and the future items of *Ageing & Memory*, the missingness appears to be independent of the values of other items. It should be noted, however, that when the number of incomplete cases is small, difficulties arise using these last four tests. The power of the tests is in some case very small, and sometimes a test even cannot be performed. Therefore the results of the test should be viewed with care.

Another problem which affects the t-tests and also Bartlett's test is caused by the missing data itself. For both tests the sample has to be split into two groups: one containing respondents which answered an item and one containing respondents which did not answer that item. The means and correlations of other items are then tested, but these other items also have missing values. The procedure of pairwise deletion is used to handle this problem. Although this procedure assumes MCAR data, we expect this 'second order missing data effect' not to have a large effect on the results.

Example 2: An artificial data set

A data set consisting of responses of $n = 1000$ respondents to $k = 10$ dichotomous items was generated with a specific IRT model, *viz.* the Rasch model. In these data missing values were created ($Q_0 = 12\%$) in two ways resulting in two incomplete data sets: one in which the data are MCAR, the other in which the mechanism is nonrandom. The nonrandom missingness was modeled as logistic function of $\boldsymbol{X}$ and $\boldsymbol{Z}$, *i.e.* missingness depends on the item responses and the covariates sex and age (selection modeling approach, see Little & Schenker, 1995). In Table 6 the extent of the item nonresponse in the two data sets can

Table 6: Item nonresponse in artificial data sets. MCAR indicates the data with randomly missing data, NMAR the set with nonrandomly missing data.

				units	mean	item nonresponse				
	N	k	cat.	INR	INR	Q_0	min.	max.	dQ_0	nMD
MCAR	1000	10	2	70.7	1.20	12.0	10.1	13.4	0	0
NMAR	1000	10	2	51.8	1.20	12.0	5.0	20.4	10	0

be found.

From Table 6 it follows that the difference in missingness mechanisms causes the distribution of the blanks over the data sets to be different. When the procedures of Table 5 are used to test the mechanisms all tests indicate that the missing values are MCAR in the first data set, except Bartlett's test of difference in correlations. This test shows no clear indication of either randomness or nonrandomness. In the second data set, all tests indicate that the data are nonrandomly missing, except the test on the average valid item scores.

5.3 Correlates of missing data

Analysis of variance is used to determine whether the missing data varied by sex and age (the only covariates available in all data sets). In the *VOCL-89*, *English*, and *Maths* test data the covariate age is not observed because in these studies only students with age 15-20 were tested. In Table 7 the similarity of the average number of missing items (mean INR, Table 3) for respondents belonging to a particular sex or age group is tested. In this table the main effects are reported for the data in which they are significant. Interactions between sex and age were examined, but very few were found significant.

Table 7: Correlates of item nonresponse. Deviations from mean INR: significant main effects only.

	RAND-36	*Phys. funct.*	*DASI*	*SSLI*	*EPQ*	*Ageing memory*
Mean	1.872	0.242	0.511	0.273	0.400	5.510
Sex	$P = 0.310$	$P = 0.376$	$P = 0.005$	$P = 0.166$	$P < 0.005$	$P < 0.001$
male			-0.07		-0.21	-1.42
female			0.06		0.09	0.99
Age	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P = 0.146$	$P = 0.019$
< 29				-0.27		
30-39				-0.06		
40-49				-0.13		-1.76
50-59				-0.17		-0.47
60-69	-0.81	-0.12	-0.09	-0.03		1.01
70-79	0.70	0.14	0.11	0.07		0.73
> 80	2.32	0.26	0.17	1.08		-0.01

A consistent pattern of missing data is evident in Table 7 with respect to sex and age. Older respondents tend to have more missing values, and so do women. The two subscales of the *RAND-36*, *General* and *Mental Health*, have significant age effects in the same order as *Physical Functioning* and the *DASI*. Only the subscales *psychoticism* and *extraversion* showed a significant effect of sex. For this scale, as for the *SSLI*, there are no significant patient/control group effects. In the *English* and *Maths* tests the covariate 'education level' (1 is lowest, 3 is highest) has a significant effect ($P < 0.005$) and the adjusted deviations from the means are for the two tests respectively 0.90, -0.31, -0.54 and 1.48, -0.24, -1.11, indicating that students with a lower education level are more inclined to skip items.

6 Discussion

The data sets examined in this paper show a wide range of missingness. From sets with many respondents having a low overall percentage missing, to sets with few respondents having a considerable amount of missing data. The occurrence of missing data is a problem which every researcher will encounter in some way, even if one tries to avoid it by careful planning and data collection. However, the treatment of these missing data heavily depends on the nature of the missing data pattern.

Assuming randomness

An important assumption about the nature of the missing data which is often made, is that it is missing (completely) at random. However, determining whether the nonresponse can be ignored is a difficult task. Rubin (1987, p. 155) actually states: "*An important feature of the assumption of ignorable nonresponse is that generally there will be no direct evidence in the data to contradict it.* [...] *Since no X values are observed for nonrespondents, without external information there will be no way to judge whether the nonrespondents' missing values are systematically different from the respondents' observed values.*"

On the other hand, assuming that the data are randomly missing can be very dangerous. Table 7, for instance, shows a considerable age effect for the *RAND-36*: older respondents have more missing values than younger respondents. Given that for some subscales of the *RAND-36* like the *Physical Functioning* scale, the item responses (and therefore the scale score) are dependent upon age (the older, the less able), assuming randomness of

the missing data will result in an overestimation of ability. Moreover, as Table 5 and the example of the *Physical Functioning* data show, missingness may be related to the position of the respondent on the latent trait (respondents with more missing values have a lower average responses), which means that the bias in estimating respondents abilities is even larger when the missing data are treated is if they were a random subsample of the data. Actually there exists an interaction effect of age, average valid score, and sex on missingness, making inference under the assumption of randomness unreliable.

Modeling missingness

Although the MCAR assumption cannot be completely verified based on the information available in the data only, effects of covariates and other items on the missingness of one item can be detected. Also the occurrence and distribution of the missing values in the data set can indicate whether to expect the data to be randomly or systematically missing. The investigation of the missing data patterns in the collection of data sets shows that in many cases some kind of nonrandomness occurs.

In the *Maths testpaper* data, for example, missingness is strongly related to the ability of the respondents (respondents with missing data have low average valid scores). A possible reason for this can be that when respondents do not know the answer, they skip the question. When a scale score is computed based only on the observed items (assuming randomly missing values), the score, and therefore the respondent's ability, would be overestimated. Therefore, the nonrandomness of the missing data should be taken into consideration, for instance, by treating the missing values as incorrect answers (imputation) and computing a score based on the observed and the imputed items.

Nonignorable missing data mechanisms should be modeled and included in the analysis. In the *SSLI* data for instance, missingness is related to age and in the *Ageing & Memory* data missingness is strongly related to age and sex. Conditioning on these covariates (modeling under the less strict MAR assumption) will improve the analyses. Rubin, Stern, & Vehovar (1995) recommend to use the MAR assumption as a starting point for analysis in large well-conducted surveys. When the missing data mechanism is clearly nonignorable, a more sophisticated model for the missingness should be included in the analysis and the results compared with the results obtained with the MAR assumption.

Assessing the nature of missing data and the accessibility of the missing data mechanism to model the missing data , however, is a difficult task. The methods demonstrated in this paper are useful tools, but to investigate the difference between observed and missing values extra information is necessary. This can be additional data from respondents who originally have missing values, or information from other sources like theory, logic, or prior data (Graham & Donaldson, 1993). With the extra information about the missing data mechanism, the missing data can be modeled more accurately and the analysis will be less biased.

Acknowledgement

The author is grateful to Ivo Molenaar and two reviewers for their comments on an earlier version of this paper, and to Wilfred Heesen, Hans Kuyper, Eric van Sonderen, and Birgit de Cnodder for supplying the data sets analyzed in this paper. This research is supported by the Netherlands Research Council (NWO), Grant 575-67-048.

References

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

Colsher, P.L. & Wallace, R.B. (1989). 'Data quality and age: health and psychobehavioral correlates of item nonresponse and inconsistent responses'. *Journal of Gerontology, 44*, 45-52.

Craig, C.S. & McCann, J.M. (1978). 'Item nonresponse in mail surveys: extent and correlates'. *Journal of Marketing Research, 15*, 285-289.

Durand, R.M., Guffey, H.J., & Planchon, J.M. (1983). 'An examination of the random versus nonrandom nature of item omissions'. *Journal of Marketing Research, 20*, 305-113.

Eggen, T.J.H.M. & Sanders, P.F. (Eds.). (1993). *Psychometrie in de praktijk* [Psychometrics in practice]. Arnhem: CITO.

Eggen, T.J.H.M. (1993). 'Itemresponsetheorie en onvolledige gegevens' [Item Response Theory and incomplete data]. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* (pp. 239-284). Arnhem: CITO.

Ferber, R. (1966). 'Item nonresponse in a consumer survey'. *Public Opinion Quarterly, 30*, 399-415.

Glas, C.A.W. (1988). 'The Rasch model and multistage testing'. *Journal of Educational Statistics, 13*, 45-52.

Graham, J.W. & Donaldson, S.I. (1993). 'Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data'. *Journal of Applied Psychology, 78*, 119-128.

Guadagnoli, E. & Cleary, P.D. (1992). 'Age-related item nonresponse in surveys of recently discharged patients'. *Journal of Gerontology, 47*, 206-212.

Kim, J.O. & Curry, J. (1978). 'The treatment of missing data in multivariate analysis'. In D.F. Alwin (Ed.), *Survey Design and Analysis* (pp. 91-116). London: Sage Publications.

Kromrey, J.D. & Hines, C.V. (1994). 'Nonrandomly missing data in multiple regression: an empirical comparison of common missing-data treatments'. *Educational and Psychological Measurement, 54*, 573-593.

Little, R.J.A. & Rubin, D.B. (1989). 'The analysis of social science data with missing values'. *Sociological Methods and Research, 18*, 292-326.

Little, R.A.J. & Schenker, N. (1995). 'Missing data'. In G. Arminger, C.C. Clogg, & M.E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 39-75). New York: Plenum Press.

Mislevy, R.J. & Wu, P-K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing* (Research report RR-96-30-ONR). Princeton, NJ.: Educational Testing Service.

Molenaar, I.W. (1997). 'Lenient or strict application of IRT with an eye on practical consequences'. In J. Rost & R. Langeheine (Eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences* (pp. 38-49). Münster: Waxmann.

Omura, G.S. (1983). 'Correlates of item nonresponse'. *Journal of the Marketing Research, 25*, 321-330.

Roth, P.L. (1994). 'Missing data: a conceptual review for applied psychologists'. *Personnel Psychology, 47*, 537-560.

Rubin, D.B. (1976). 'Inference and missing data'. *Biometrika, 63*, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D.B., Stern, H.S., & Vehovar, V. (1995). 'Handling "don't know" survey responses: the case of the Slovenian plebiscite'. *Journal of the American Statistical Association, 90*, 822-828.

van der Zee, K.I. & Sanderman, R. (1993). *Het meten van de algemene gezondheidstoestand met de RAND-36: een handleiding* [Measuring the general state of health with the RAND-36: a manual]. Groningen: Northern Center for Healthcare Research (NCG).

Ying, Y-W. (1989). 'Nonresponse on the center for epidemiological studies-depression scale in Chinese Americans'. *The International Journal of Social Psychiatry, 35*, 156-163.

Ontvangen: 07-02-1997
Geaccepteerd: 14-01-1998