Chapter 4

ARE INCONSISTENT RESPONDENTS CONSISTENTLY INCONSISTENT? A STUDY OF SEVERAL NONPARAMETRIC PERSON FIT INDICES

E.D. de Leeuw and J.J. Hox[1]

4.1. Introduction

Four well-known sources of measurement error in surveys are: the questionnaire (e.g. question wording), the data collection mode (e.g. face to face or telephone communication), the interviewer, and the respondent (Groves 1989). To minimize the overall measurement error the errors resulting from each source should be reduced. Survey methodologists have developed and evaluated methods to improve and test questionnaires (e.g. cognitive laboratory methods, Forsyth and Lessler (1991)), optimize data collection modes (e.g. the Total Design Method for mail and telephone surveys, Dillman (1978)), estimate interviewer error (Groves 1989, chap. 8), and reduce interviewer related error through training and supervision (Fowler 1991).

Unlike the first three sources of measurement error (questionnaire, mode, interviewer) the fourth source (respondent) is difficult to minimize. Respondents can be instructed in what is expected from them (e.g. thinking carefully, use the answer categories provided) and they may be motivated to do one's best. But there are not many ways in which a survey researcher can manipulate a respondent and so reduce the respondent error. Therefore research has concentrated on attempts to identify unique properties of those respondents who produce errors.

A major problem in this type of research is how to measure respondent error. Groves (1989, p. 445-446) summarizes this as follows: "Measurement errors are generally viewed as specific to a particular measure, a question posed to the respondent. Only by identifying response tendencies of respondents over many questions can inference about respondent influences on measurement error be made. Then only by comparing different respondents on the same task can characteristics of the respondents which produce measurement error be identified." One promising approach is the

use of structural modeling on multitrait-multimethod data for different respondent groups (cf. Andrews and Herzog 1986). Another is the application of person fit indices to detect inconsistent respondents.

Person fit indices have been developed in the field of psychological and educational testing. In person fit research persons with unexpected or aberrant response patterns with respect to a test model or with respect to other response patterns in the sample are identified and further examined. For example, Levine and Rubin (1979) discussed the use of person fit indices for the detection of cheating on aptitude tests. Harnish and Linn (1981) used person fit indices to differentiate schools with special curriculum on math and reading. Tatsuoka and Tatsuoka (1982, 1983) could identify students who used a wrong algorithm in problem solving tasks. Finally, Van der Flier (1980) used person fit indices in intercultural research, and Van Tilburg and De Leeuw (1991) used person fit indices in a comparison of different data collection methods. To compute person fit indices one needs data on a reliable multi-item scale. For an overview of person fit indices see Meijer (1994) and also Meijer and Sijtsma in this book.

In this chapter we investigate whether person fit indices are a useful tool in the study of respondent error. To be useful for the study of respondent error in survey research, 'aberrancy' first of all should be a stable characteristic at least for the duration of the interview (cf. Groves op. cit.). This can be translated into the following research question: Is aberrancy or inconsistency in response pattern on a set of items as detected by a person fit index a stable respondent characteristic, i.e. is it independent of the particular set of items used? Secondly, respondent characteristics that might influence aberrancy should be systematically investigated (cf. Groves above). This leads to the second research question: Are certain types of respondents (e.g. the elderly, the lower educated) more prone to give aberrant response patterns?

First, we will give a short description of the person fit indices and data sets used in this study. Next, we will present the major results in three subsections. We end with a conclusion and discussion of the results.

4.2. Method

4.2.1. Indices for the detection of aberrant response patterns

Person fit analysis investigates whether a person exhibits answering behavior that deviates from the behavior predicted by a measurement model or from the answering behavior of the majority in the population to which that person belongs. For persons detected as aberrant the total scale score does

not adequately reflect the attribute that is being measured and further research is needed before any firm conclusion can be drawn about the test performance. For example, if a student answers eight out of 10 questions correctly, one expects that s/he will have missed the two most difficult questions. If the two easiest questions are the ones that are answered incorrectly, her/his response pattern is completely unexpected.

In the literature, two groups of person fit indices can be distinguished. The first group consists of indices that are based on the assumptions of parametric IRT-models, such as the Rasch model. The second group consists of indices that evaluate a response pattern given the assumptions of a nonparametric Item Response Theory (IRT) model, or by means of statistics based on the group to which a person belongs. Scales that fit the strict assumptions of IRT-models such as the Rasch-model are scarce in survey research, so we concentrated on nonparametric person fit indices. We selected four promising nonparametric person fit indices from the literature (for definitions and formulae see Meijer 1994; Meijer and De Leeuw 1992). Each index evaluates the individual response pattern on a multi-item scale.

1. The H_i coefficient (scalability index; range $-\infty$,1). Sijtsma (1988) proposed a person coefficient H_i in the context of Mokken's nonparametric item response theory. $H_i<0$ if the mean covariance between the itemscores of person i and the other persons is negative. $H_i=0$ means that the mean covariance between the item scores of person i and the itemscores of the other persons equals 0. $H_i=1$ if the covariance between the item scores of person i and the other persons equals the maximum covariance given the marginal frequency between the item scores of person i and the item scores of the other persons. Thus negative values indicate that a response pattern clearly deviates from the usual response behavior in the group.

2. The modified caution index C^*_i (deviance index; range 0,1) proposed by Harnisch and Linn (1981). $C^*_i=0$ if the response pattern equals the perfect Guttman pattern and $C^*_i=1$ if the response pattern equals the reversed Guttman pattern. Thus, relatively high values for C^*_i indicate that a pattern deviates from the usual response behavior in the group.

3. The $Q(x_i)$-index (probability response pattern; range: 0,1) and 4. the U^3_i-index (deviance index; range: 0,1), both proposed by van der Flier (1980). $Q(x_i)$ is the one-tailed probability of a response pattern, given the p-values of the questions within the conditional distribution of pattern probabilities. In other words $Q(x_i)$ is the sum of the probabilities of a certain response pattern and all more deviant patterns, given the test score of the respondent; it can be interpreted as the probability (p-value) used in ordinary statistical testing. That is, a small value of $Q(x_i)$ indicates that the

probability to find this pattern is small and therefore this pattern is unexpected or deviant considering the total group of respondents. When a test or multi-item scale consists of a large number of questions the computation of $Q(x_i)$ can take a lot of time and often exceeds the capacity of standard computers. In those cases U^3_i can be computed; U^3_i is monotonely related to $Q(x_i)$. $U^3_i = 0$ if the pattern of itemscores equals the Guttman pattern and $U^3_i = 1$ if the pattern of itemscores equals the reversed Guttman pattern; therefore U^3_i is a deviance index. Van der Flier (1980) showed that U^3_i is approximately normally distributed.

It should be noted that both H_i and $Q(x_i)$ are scalability indices. A large value means that a respondent has a response pattern on a set of questions as could be expected. while a low value indicates that a response pattern of a respondent on a set of questions is unexpected. Conversely, C^*_i and U^3_i are both deviance indices. A high value indicates that an individual response pattern deviates from the ideal Guttman-pattern. Thus, the correlation between H_i and $Q(x_i)$, and between C^*_i and U^3_i should be positive; but correlations between a scalability index and a deviance index (e.g. between H_i and C^*_i) should be negative. One should keep this in mind when inspecting the results.

4.2.2. Data sets used

Two data sets were used to investigate the effectiveness of person fit indices for survey research.

The first data set consisted of 243 (paper and pencil) face to face interviews of adult Dutch (age 18-92).[2] The subject of the questionnaire was psychological and economic well-being. The questionnaire contained four multi-item scales: The 11 item De Jong-Gierveld loneliness scale, a condensed eight-item form of Brinkman's self-evaluation scale, and a balanced extension of Bradburn's affect balance scale consisting of a nine-item positive affect scale and a nine-item negative affect scale. Standard socio-economic background information (e.g. age, sex) of the respondents and several ratings of the respondent by the interviewer were also available.

An analysis of person fit indices is only appropriate if the scale investigated has at least moderate scalability (e.g. Loevinger's $H > .3$; cf. Mokken 1971). To investigate this we conducted a Mokken analysis for each multi-item scale (cf. Meijer, Sijtsma and Smid 1990); as an overall indicator of Mokken scalability Loevinger's H was used. As an indicator of the precision of measurement coefficient alpha (cf. Cronbach 1951) was also computed. In a recent study Meijer and Sijtsma (1993) point out that person fit analysis is only appropriate on reliable items. As an indicator of the item-

reliability of the items in this study we computed the mean item correlation (Spearman-Brown corrected alpha for one item) for each scale. The results are summarized in Table 1.

Table 1. **Psychometric properties of 4 multi-item scales: Loneliness (LO) Self-Evaluation (SE), Positive Affect (PA) and Negative Affect (NA). Data De Leeuw (1992): N=243.**

Scale	*# questions*	*H*	*alpha*	*mean item r*
LO	11	.40	.83	.31
SE	8	.45	.76	.28
PA	9	.27	.65	.17
NA	9	.34	.71	.21

When we inspect Table 1 we see that the scalability and reliability of the loneliness scale and the self-evaluation scale is adequate. The two affect scales perform less well, but seem strong enough for exploratory analysis.

The second data set consists of 4494 (computer assisted) face to face interviews with Dutch elderly (age 55-89).[3] The subject of the questionnaire was living arrangements and social networks. The questionnaire contained two multi-item scales that were also used in the study of De Leeuw, which provides the first data set: The loneliness scale and the self-evaluation scale. Besides standard background information interviewer's ratings of the respondent were available.

Again preliminary psychometric analyses were performed on the two multi-item scales; the results are summarized in Table 2.

Table 2. **Psychometric properties of 2 multi-item scales: Loneliness (LO) Self-Evaluation (SE). Data NESTOR-LSN N=4499.**

Scale	*# questions*	*H*	*alpha*	*mean item r*
LO	11	.33	.82	.29
SE	8	.34	.72	.24

When we compare Table 1 and Table 2 we see that the reliability and scalability of both the loneliness scale and the self-evaluation scale is somewhat lower in the second data set, but still adequate for further analysis.

The person fit indices described above were computed for each of the multi-item scales.[4] All four person fit indices were computed on the first data set. The second dataset was so large that the H_i-coefficient could not be computed.[5]

4.3. Results

4.3.1. Consistency of aberrant response patterns over scales

The first research question was: Is aberrancy or inconsistency in response pattern on a set of items as detected by a person fit index a respondent trait, i.e. is it independent of the particular set of items used? This question can be further divided into three subquestions:

- Is the intercorrelation between the four person fit indices high for each scale? The indices should all measure the same concept (aberrancy) and therefore they should intercorrelate.
- Are the scores on the person fit indices independent of the total score on the multi-item scale used? If the indices should correlate high with the total scale score they would measure the attribute of the scale (e.g. loneliness) instead of aberrancy of response pattern on the scale.
- Is the intercorrelation of person fit indices *across* different multi-item scales high? If not than aberrancy is scale specific and not generalizable over different sets of questions.

To answer the first and third question we computed the correlations between the person fit indices for all multi-itemscales. This was done for both data sets. It should be noted that the second dataset was so large that the H_i-coefficient could not be computed. The results are summarized in Table 3 and Table 4. A specific person fit index computed on a specific scale is indicated by the first letter of the index and two letters indicating the scale. For example, HLO indicates the H_i-coefficient computed on the responses on the Loneliness-scale, CSE indicates the C^*_i-index computed on the Self-Esteem scale.

Table 3. **Pearson correlation coefficients: four person fit indices (H_i, C^*_i, U^3_i, $Q(x_i)$) and four scales (Loneliness, Self-Esteem, Positive Affect, Negative Affect). Data De Leeuw.**

	HLO	*CLO*	*ULO*	*QLO*	*HSE*	*CSE*	*USE*	*QSE*	*HPA*	*CPA*	*UPA*	*QPA*	*HNA*	*CNA*	*UNA*	*QNA*
HLO	1.00	-.93	-.96	.86	-.08	.10	.05	-.04	-.04	.06	.03	-.00	-.04	.03	.03	-.08
CLO	-.93	1.00	.99	-.87	.09	-.14	-.10	.13	.04	-.07	-.05	.03	.03	-.02	-.02	.10
ULO	-.96	.99	1.00	-.88	.10	-.13	-.08	.11	.04	-.07	-.04	.02	.04	-.03	-.03	.10
QLO	.86	-.87	-.88	1.00	-.11	.12	.08	-.07	-.02	.04	.00	-.00	.00	-.01	-.01	-.05
HSE	-.08	.09	.10	-.11	1.00	-.92	-.84	.66	.05	-.04	-.05	.05	.11	-.11	-.12	.16
CSE	.10	-.14	-.13	.12	-.92	1.00	.95	-.76	-.05	.05	.07	-.06	-.05	.06	.07	-.12
USE	.05	-.10	-.08	.08	-.84	.95	1.00	-.78	-.04	.07	.11	-.09	.00	.00	.01	-.09
QSE	-.04	.13	.11	-.07	.66	-.76	-.78	1.00	.01	-.05	-.0	8.3	-.06	.03	.02	.03
HPA	-.04	.04	.04	-.02	.05	-.05	-.04	.01	1.00	-.94	-.91	.82	.19	-.17	-.16	.19
CPA	.06	-.07	-.07	.04	-.04	.05	.07	-.05	-.94	1.00	.97	-.87	-.16	.15	.13	-.17
UPA	.03	-.05	-.04	.00	-.05	.07	.11	-.08	-.91	.97	1.00	-.85	-.15	.14	.13	-.17
QPA	-.00	.03	.02	-.00	.05	-.06	-.09	.13	.82	-.87	-.85	1.00	.10	-.10	-.09	.11
HNA	-.04	.03	.04	.00	.11	-.05	.00	-.06	.19	-.16	-.15	.10	1.00	-.98	-.98	.81
CNA	.03	-.02	-.03	-.01	-.11	.06	.00	.03	-.17	.15	.14	-.10	-.98	1.00	.99	-.80
UNA	.03	-.02	-.03	-.01	-.12	.07	.01	.02	-.16	.13	.13	-.09	-.98	.99	1.00	-.82
QNA	-.08	.10	.10	-.05	.16	-.12	-.09	.03	.19	-.17	-.17	.11	.81	-.80	-.82	1.00

For *each* scale the absolute values of the correlations between the four person fit indices are high; the lowest correlation was between HSE and QSE (.66), the highest was between ULO and CLO (.99). However the correlations for each person fit index *between* scales is disturbingly low; for instance the highest correlation is between HPA and HPA (.19). When we look at Table 4 we see the same pattern. Although the different person fit indices all measure something in common, this is highly scale-specific.

Table 4. **Pearson correlation coefficients: three person fit indices (C^*_i, U^3_i, $Q(x_i)$) and two scales (Loneliness, Self-Esteem). Data NESTOR-LSN.**

	CLO	*ULO*	*QLO*	*CES*	*UES*	*QES*
CLO	1.00	.98**	-.73**	.04	.04	-.06*
ULO	.98**	1.00	-.71**	.04	.05	-.06*
QLO	-.73**	-.71**	1.00	-.03	-.03	.05
CES	.04	.04	-.03	1.00	.99**	-.91**
UES	.04	.05	-.03	.99**	1.00	-.92**
QES	-.06*	-.06*	.05	-.91**	-.92**	1.00

To answer the second subquestion, correlations were computed between the total score on a scale and the person fit indices computed for each scale.

The results are summarized in Table 5.

Table 5. **Pearson correlations total scale score with person fit indices matching that scale. Data De Leeuw: columns 1-4; Data NESTOR-LSN: columns 5-6.**

	De Leeuw				*NESTOR-LSN*	
	LO	SE	PA	NA	LO	SE
H	-.22*	-.38**	-.27**	-.22**	--	--
C	-.05	.05	.06	.22**	.07**	.05
U	.05	-.13	-.09	.15	.19**	.04
Q	-.09	.16	-.00	.03	-.04	.06*

Note. For very large data sets like the NESTOR-LSN data H can not be computed.

The H_i-coefficient correlates rather high with the total scale score in all cases. The other coefficients do not show a consistent pattern of high correlations. Especially the $Q(x_i)$-index performs well and has low correlations with the total scale score.

In sum:
- All person fit indices do intercorrelate within one scale.
- All person fit indices are scale-specific and have low intercorrelations over scales.
- Only H_i correlates disturbingly high with the total scale score.

The conclusion is that aberrant response patterns as detected by person fit indices are NOT consistent over scales. Therefore, the first research question: 'Is aberrancy or inconsistency in response pattern on a set of items as detected by a person fit index a respondent trait, i.e. is it independent of the particular set of items used?' can be answered with a tentative 'no'.

4.3.2. Correlates of inconsistent or aberrant response patterns

Person fit indices do not correlate highly between multi-item scales, and a respondent who is scored as deviant according to the person fit index on one scale is not necessary deviant according to the score on the same person fit index calculated on the response patterns of a second scale. Persons indicated as aberrant based on their response patterns on one scale are not necessary the *same* persons who are marked as aberrant on their responses

to a second scale. Still, they could be the same *type* of persons, that is: they could share the same characteristics. Or as stated as the second research question in the introduction: Are certain types of respondents (e.g. the elderly, the lower educated) more prone to give aberrant response patterns?

To investigate this question we started by classifying respondents into the categories aberrant and not aberrant according to well-known criteria from the literature. These criteria for aberrancy are: $H_i \leq 0$ (cf. Sijtsma 1988), $C^*_i > .5$ (cf. Harnish and Linn 1981; Harnish 1983), standardized $U^3_i > 1.29$ and $Q(x_i) \leq .05$ (Van der Flier 1980). We also constructed an index called 'superab' that indicates whether a respondent showed an aberrant response pattern according to any person fit index on any scale. This is summarized in Table 6 for the De Leeuw data set.

Table 6. **Percentage respondents classified as aberrant on each of four person fit indices (H_i, C^*_i, U^3_i, $Q(x_i)$) and four scales (Loneliness, Self-Esteem, Positive Affect, Negative Affect). Data De Leeuw.**

	LO	*SE*	*PA*	*NA*
H	25.4	6.7	13.6	9.8
C	33.9	6.2	15.2	10.7
U	16.4	10.7	9.6	7.3
Q	12.4	3.9	5.6	8.3

The percentage of respondents classified as aberrant on at least one criterion (SUPERAB > 0) is 47.7%.

When we inspect Table 6 we see that standard criteria for the classification of 'aberrancy' result in markedly different percentages. As a control we computed the intercorrelations between the classification on the person fit indices for all multi-itemscales; also the correlations with total scale score were computed. The same pattern as discussed in 3.1 is seen: high correlations within one scale, low correlations between scales. In general, all correlations were slightly lower, which is to be expected because of the classification (restriction of range).

The correlation of person fit indices across scales is low. Aberrant respondents according to their responses on one specific scale are not necessarily the same as aberrant respondents according to their response patterns on another scale. But they still could share characteristics, they could be the same type of persons instead of the same persons. What type of respondents produces unexpected or aberrant response patterns?

Respondents classified as aberrant were investigated and the correlations between these classifications and background information was computed. The following variables were used: gender, age, education (educat), income, marital status (three dummy variables: married, divorced, widowed), and evaluation of interview by respondent as pleasant (r.pleasant) was computed. Also correlations with the following interviewer ratings were computed: the interview was pleasant (i.pleasant), some questions were too difficult for respondent (q.difficult), the respondent did understand questions (understood), the interview was emotionally difficult (emot.dif), the interview was too long for the respondent (too long), the respondent was sometimes dishonest (dishonest), the respondent was cooperative (cooperative).

Remember, that all person fit indices and the classifications based on these indices correlated to some degree with the total score on the multi-item scales. Thus, a substantial correlation between aberrancy and another variable, for instance age, could be confounded. We therefore constructed a new variable named 'superresid' based on the variable 'superab'; in superresid all correlations with the substantial multi-item scales were partialed out. We then correlated superresid with the background variables.

The results are rather disappointing: a clear pattern could not be distinguished. The correlations between classification as aberrant according to several criteria and background variables did not suggest a certain type of 'aberrant-prone' respondent. The only really strong correlations (i.e. >.30) were found for aberrancy according to the loneliness scale and marital status and age. The combination of aberrancy indicators into more complex variables as 'superab' and 'superresid' did not result in new findings. By way of illustration Table 7 reports the correlations of the classification based on the $Q(x_i)$-index with the background variables.[6]

In sum:

- Standard criteria for the classification of 'aberrancy' result in markedly different percentages of respondents with inconsistent or aberrant response patterns.
- No profile of respondents who give unexpected or aberrant response pattern could be distinguished.
- The only really strong correlations (i.e. >.30) were found for aberrancy according to the loneliness scale and marital status and age. Aberrant respondents were more often older (.45) and widowed (.30). Weaker correlations indicate that aberrant respondents also have a lower education (.15) and have more difficulties in understanding the questions (.16).

Table 7. **Correlations of respondent background variables and interviewer ratings with classification as aberrant according to the criterion Q ≤ .05 on four scales and with SuperAb, SupResid. Data De Leeuw. (loqc, seqc, paqc, naqc indicate classifications on Q based on response pattern on loneliness scale, etc.)**

	loqc	*seqc*	*paqc*	*naqc*	*SuperAb*	*SupResid*
(Respondent variable)						
gender	.12	-.19*	-.04	-.08	-.02	-.02
age	.30**	.09	.05	.05	.22**	.22**
educat	-.15	-.08	-.13	-.04	-.14*	-.13
income	-.11	-.04	-.13	.01	-.15	-.12
married	-.14	.03	-.04	-.03	-.10	-.13
divorced	-.07	.05	.04	.06	.05	.00
widowed	.45**	-.05	.13	.01	.23**	.26**
r.pleasant	.14	.06	.08	.01	.12	-.12
(Interviewer rating)						
i.pleasant	.01	.09	.05	-.00	.05	.03
q.difficult	.02	.06	.00	.05	.04	-.00
understood	-.16*	-.08	-.03	-.03	-.14*	-.04
emot.dif.	.09	.00	.05	.02	.17*	.14
too long	-.00	.08	-.04	-.02	.02	.00
dishonest	.06	-.02	.04	-.02	.00	.04
cooperative	-.07	.06	-.07	-.05	-.07	-.09

Again we repeated the analyses using the second larger data set. This data set contained only two multi-item scales (loneliness and self esteem). First, we classified respondents into aberrant and not aberrant based on their scores on the person fit indices C^*_i, U^3_i, $Q(x_i)$ using the criteria described above. Then the indices 'superab' and 'superresid' were computed. The results are summarized in Table 8.

Table 8. **Percentage respondents classified as aberrant on each of three person fit indices (C^*_i, U^3_i, $Q(x_i)$) and two scales (Loneliness, Self-Esteem). Data NESTOR-LSN.**

	LO	*SE*
Q	5.1	5.3
U	10.8	10.6
C	27.5	21.1

The percentage of respondents classified as aberrant on at least one criterion (SUPERAB > 0) is 44.1%.

When we compare the results of Table 6 and Table 8 we see that again the percentages aberrant respondents do differ, but not as widely as in the first data set. In the second data set the differences are mainly caused by the person fit index and its cutting criterion, not also by the multi-item scale used. Also, in this case we checked the intercorrelations between the classification on the person fit indices for all multi-itemscales and the correlations with total scale. The same now familiar pattern emerged: high correlations within one scale, low correlations between scales.

Next, we once more investigated whether a profile of an aberrant respondent could be detected. The same respondent variables (e.g. age, gender) were available as in the first data set; one additional respondent variable was available in which the respondent indicated how tiring the interview was (r.tiring). A different set of interviewer ratings was used, specifically designed to evaluate the performance of elderly respondents. This set consisted of two subsets. The first set directly asked about the behavior during the interview, for instance, were there any problems during the interview (problems), how much help did the respondent need with the questions (helpneed), did the respondent understand the questions (understo), did the respondent stray during the interview (stray), was the respondent worried about her performance (worries), did the respondent forgot the point of the interview (forgot), how well did the respondent express the answers (express), how honest did the respondent answer (honest), and was the interview toolong for the respondent (too long). The second group of ratings was about the general condition of the respondent (i.e. mobility, memory, concentration, health). All correlations were low and no clear pattern could be discerned (see Table 9).

In sum:

- Standard criteria for the classification of 'aberrancy' again result in different percentages of respondents with inconsistent or aberrant response patterns.
- Again, no profile of those respondents who give unexpected or aberrant response pattern could be distinguished.
- Strong correlations (i.e. >.30) were not discovered in the second data set. A limited number of statistically significant (N ≈ 1500 !) but weak correlations (i.e. <.10) were found. The strong relationships with the variables age and widowed were not replicated. In the first data set aberrant respondents were more often older (.30) and widowed (.45); in the second data set the highest correlation of aberrancy with age was .10 and with widowhood .07.
- No strong correlates of aberrancy were detected; there are weak

indications of a possible influence of age. Therefore, also the second research question: 'Are certain types of respondents more prone to give aberrant response patterns?' can be answered with a tentative 'no'.

Table 9. Correlations of respondent background variables and interviewer ratings with classification as aberrant according to various criteria (Q ≤ .05, U*>1.29, C>.5) on the loneliness and self-esteem scale and with SuperAb, SupResid. Data NESTOR-LSN. (loqc indicates classification on Q based on response pattern on loneliness scale, etc.).

	loqc	*louc*	*locc*	*seqc*	*seuc*	*secc*	*SuperAb*	*SupResid*
(respondent variables)								
gender	-.03	.00	.01	-.02	-.01	.01	-.00	-.00
age	.04*	.02	.0**	-.04	-.02	-.10**	.00	.02
educat	-.03	-.05*	-.05*	-.05*	-.03	-.10**	-.08**	-.07*
income	-.00	-.05*	-.05*	-.04	-.05*	-.10**	-.11**	-.09**
married	.00	-.03	.02	-.03	-.04*	-.08**	-.04	-.01
divorced	.04	.03	.05*	.03	.02	.00	.07*	.06*
widowed	-.04*	.01	-.07**	-.00	.01	.06**	-.02	-.05
r.pleasant	.01	-.00	-.01	.01	.02	.00	-.01	.00
r.tiring	.00	.06**	.01	.02	-.00	.03	.09**	.07*
(interviewer ratings task performance)								
problems	-.00	.03	.03	.03	.02	.09**	.04	.03
helpneed	.01	.03	.00	.03	.04	.11**	.06*	.05
understo	-.02	-.04*	-.02	-.06*	-.05*	-.11**	-.08**	-.06*
stray	.02	.03	-.03	.03	.03	.08**	.05	.04
worries	.02	.00	-.01	-.01	-.03	.00	.00	.00
forgot	.00	.01	-.01	.05*	.00	.07**	.04	.03
express	-.02	-.04	-.03	-.06*	-.05*	-.11**	-.09**	-.08**
honest	-.02	-.03	-.01	-.00	-.01	-.00	-.03	-.02
toolong	-.03	-.00	-.01	.03	.02	.03	.01	.00
(interviewer ratings general condition)								
mobility	.02	-.05*	-.00	-.03	-.01	-.05*	-.06*	-.03
memory	-.01	-.04*	-.01	-.02	-.01	-.08**	-.05*	-.03
concentr	.00	-.03	-.00	-.05*	-.02	-.07**	-.03	-.01
health	.00	-.06**	-.02	-.03	-.00	-.03	-.06*	-.03

4.3.3. Aberrancy and responses to a MTMM Matrix

Until now results have been disappointing for a survey researcher interested in the investigation of respondent error: aberrancy seems to be dependent on the set of questions used and no clear profile of aberrancy-prone

respondents could be distinguished. In what way do aberrant respondents then differ from non-aberrant respondents?.

A different approach to estimate measurement error is the analysis of a multi-trait multi-method matrix (MTMM-matrix). In this approach trait-variance is seen as 'true' variance. Both systematic error-variance (method-variance) and random error-variance (noise) are discerned (cf. Widaman, 1985). The second data set did contain a multi-trait multi-method matrix. Three different traits (i.e. satisfaction with life in general, satisfaction with health, satisfaction with social contacts) was measured using three methods (i.e. Likert scale, ladder, social comparison). See also Andrews and Withey (1976).

The MTMM-model fitted well on the total group (aberrant and nonaberrant) respondents ($p=.001$, GFI = 1.00, AGFI = .99). We then tested this MTMM-model in a two-group LISREL-analysis. The two groups used were: a group of 'non-aberrant' respondents (NAB) which was formed by 582 respondents with value 0 on the variable SUPERAB and a group of 'aberrant' respondents (AB) consisting of 444 respondents (SUPERAB > 0). This also resulted in a satisfactory fit for a large data set (the invariance model resulted in $p=.00$, GFI = .97). In the final step we estimated the trait-variance, the method-variance, and the error-variance associated with the two groups (aberrant and nonaberrant) respondents. The results are summarized in Table 10.

In general, the aberrant and non-aberrant respondents do not differ greatly on trait variance. There are some differences on method variance (i.e. systematic error variance). The main difference is in the proportion error variance. The pattern is also consistent for all variables: if there are any differences than we find that aberrant respondents generate a smaller proportion of systematic (trait or method) variance, and a larger proportion of error variance, It appears that aberrant respondents are not much more sensitive to a certain method or type of question (e.g. likert scale) than non-aberrant respondents, but aberrant respondents do produce more random error.

4.4. Conclusion and discussion

When four well-known person fit indices are computed on the individual response patterns on several multi-item scales it is found that:

1. For each multi-item scale all person fit indices intercorrelate high, i.e. for one set of questions all person fit indices grab the same concept.

2. The correlations between person fit indices estimated on different set of items is low. Person fit or aberrancy is specific for a set of questions or

a specific test and cannot be generalized to a respondent trait.

3. When the relationship between aberrancy and a large number of background variables was investigated, no strong correlates of aberrancy were detected. There are weak indications of a possible influence of age, but no standard type of respondents that give unexpected or aberrant response pattern could be distinguished.

4. The main difference between respondents who were classified as aberrant and those classified as nonaberrant is that the aberrant respondents produce more random error as indicated by a MTMM-model.

Table 10. **Trait-, method- and error-variance for ABerrant and NonABerrant respondents. NESTOR-LSN data. 3*3 MTMM-matrix. Traits: SAtisfaction General, SAtisfaction Health, SAtisfaction Social contacts. Methods: LIkert, LAdder, Social Comparison.**

		Proportion Variance					
Var.	*Type*	*Trait*		*Meth.*		*Error*	
		NAB	*AB*	*NAB*	*AB*	*NAB*	*AB*
1	LI/SAG	.08	.06	.41	.31	.50	.62
2	LI/SAH	.53	.52	.24	.24	.22	.24
3	LI/SAS	.26	.23	.23	.21	.51	.56
4	LA/SAG	.56	.56	.44	.44	-	-
5	LA/SAH	.56	.55	.28	.26	.15	.19
6	LA/SAS	.42	.40	.40	.37	.18	.24
7	SC/SAG	.00	.00	.58	.48	.42	.52
8	SC/SAH	.36	.35	.19	.18	.45	.47
9	SC/SAS	.24	.20	.26	.22	.51	.58

Person fit indices do not seem to be the expected useful tool for measuring pure respondent error in survey research, because aberrancy as measured by such indices does not appear to be a generalizable respondent trait. Several different types of problems associated with the typical properties of survey research data may be at the bottom of this.

First of all, simulation studies have shown that relatively large tests with reliable items are needed in person fit research. The rate of detection of aberrant respondents increases with the reliability of the questions and the number of questions (cf. Meijer 1994; Meijer and Sijtsma, this book). Although the multi-item scales used in this study are not long from a psychological testing point of view (between 8 and 11 questions per scale),

they are rather long from a survey point of view. The aberrancy indices computed on such short scales could simply be too imprecise to be useful. However, the scales used have good psychometric properties and the questions are fairly reliable (see 2.2). Meijer (1994, p.69) points out that the use of reliable questions can compensate for a relative short test length; in those cases relatively high percentages of aberrants could be detected in a simulation study. Furthermore, Van der Flier (1980, p. 162) who used large tests (40 items) in an empirical setting also found small correlations of person fit indices across tests. To test whether the lack of consistently aberrant respondents is the effect of using scales that are to short we did a small simulation experiment. We simulated data on a 10 item-test for 1000 persons, with 100 respondents guessing and 100 respondents cheating on the last two items, in a procedure similar to the one used by Meijer and Sijtsma (this book). Next, we split the scale on an even/odd basis and computed the indices C^*_i, U^3_i, and $Q(x_i)$ for each 5-item part separately. For the simulated data, the mean correlation between these indices within the same part was $r=.88$; between the two parts it was $r=.68$. We conclude that if there were consistently aberrant respondents in the two data sets used here, the procedure we used would have found them.

Another problem is posed by the structural missing data in a person fit study. When a person has either the minimum or the maximum possible total score on a test the resulting response pattern is totally predictable (either all zero's or one's) and thus noninformative. In those cases person fit is not defined, which results in a missing value. In educational and cognitive testing where person fit indices were developed and applied, this does not pose a large problem; after all not many students will fail all items or pass all items. However, in social psychological and survey research this can pose a serious problem. For instance, the distribution of scores on loneliness-scales is extremely skewed: many people are not lonely and therefore have a minimum score on the loneliness scale. A related problem is an occasional missed question in a scale. In cognitive tests this is usually scored as failed, the individual response pattern is complete and person fit indices can be computed. In psychological and survey research an item missing value is assigned and no person fit indices can be computed on this incomplete individual response pattern.

In the De Leeuw dataset ±66 (27%) of the respondents were assigned a missing value on the person fit indices; in the NESTOR-LSN data ± 1900 or 42%. This could lead to a serious restriction of range, which could cause artificially low correlations. But the study of Van der Flier (cf. van der Flier 1980, p. 171) makes it unlikely that restriction of range could be the main cause of the low intercorrelations. In his study aberrancy on cognitive tests

did also show low correlations with background variables (range .05-.19).

From our data we conclude that *THE* aberrancy prone respondent probably does not exist, at least not in the sense that there is an identifiable group of respondents that have the stable psychological trait to produce aberrant answers to a variety of measuring instruments. Aberrancy is probably be more a characteristic of the question posed, or of the interaction between respondent and question characteristics. Of course, specific respondents may be more prone to a react 'aberrant' to certain questions as has been shown in educational research (cf. section 1), but such aberrancy does depend on the set of questions used and is not a generalizable respondent characteristic. This is akin to the situation regarding social desirability bias and bias caused by response sets like yes-saying or acquiescence. In the early 1960's these were considered as personality characteristics and certain types of respondents were thought to be more prone to produce response bias than others. At present the general accepted view is that social desirability and response sets are heavily influenced by characteristics of the question posed, and is more a characteristic of the question than of the respondent (Groves 1989, chap 9.6). As a result, we do not recommend the application of person fit indices to detect problematic respondents. This does not disqualify the use of person fit indices altogether. The study of aberrant individual response patterns can be very useful when constructing and evaluating questions. Unusual response patterns can indicate questions that pose problems for certain (types of) respondents.

NOTES

1. The authors thank Wim Jansen, Klaas Sijtsma and Theo van Tilburg for their comments.

2. The data collection was funded by the Netherlands Organization for Scientific Research (NWO) under grant number 500278008. For details on the data collection procedure, the questionnaire, and the 4 multi-item scales used see De Leeuw (1992).

3. The NESTOR-LSN data were collected in the context of the research program 'Living arrangements and social networks of older adults.' This research program is conducted at the departments of Sociology & Social Gerontology and Social Research Methodology of the Vrije Universiteit in Amsterdam, and the Netherlands Interdisciplinary Demographic Institute in the Hague. The research is supported by a program grant from the Netherlands Program for Research on aging (NESTOR), funded by the

Ministry of Education and Sciences and the Ministry of Welfare, Health and Cultural Affairs. For details on the questionnaire and the data collection see Van Tilburg, Dykstra & Liefbroer (1993) and Knipscheer et al. (Forthcoming).

4. The computations were done with an special pc-adaptation of a program developed by Rob Meijer. This adaptation was necessary to accommodate large datasets. For more information about this version contact the second author.

5. To compute H_i all data must be stored in memory, which limits the size of the data set that can be analyzed.

6. In addition exploratory analyses were conducted which included many substantive variables, such as happiness-score and well-being. These analyses all confirmed the conclusion that no clear profile of an aberrant respondent could be distinguished.

REFERENCES

Andrews, F.A. and S.B. Withey (1976),
Social indicators of well-being. American perceptions of life quality, New York (Plenum).

Andrews, F.A. and A.R. Herzog (1986),
'The quality of survey data as related to age of respondent', *Journal of the American Statistical Association*, 81, p. 403-410.

Cronbach, L.J. (1951),
'Coefficient alpha and the internal structure of tests', *Psychometrika*, 16, p. 297-334.

Dillman, D.A. (1978),
Mail and telephone surveys; The total design method, New York (John Wiley and Sons).

Forsyth, B.H. and J.T. Lessler (1991),
'Cognitive laboratory methods: A taxonomy', in: Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, *Measurement errors in surveys*, New York (Wiley).

Fowler, F.J. jr. (1991),
'Reducing interviewer related error through interviewer training, supervision and other means', in: Biemer, P.P., R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman, *Measurement errors in surveys*, New York (Wiley).

Groves, R.M. (1989),
Survey errors and survey costs, New York (Wiley).

Harnish, D.L. (1983),
'Item response patterns: Applications for educational practice', *Journal of Educational Measurement*, 20, p. 191-205.

Harnish, D.L. and R.L. Linn (1981),
'Analysis of item response patterns: Questionable test data and dissimilar curriculum practices', *Journal of Educational Measurement*, 18, p. 133-146.

Leeuw, E.D. de (1992),
Data quality in mail, telephone and face to face surveys, Amsterdam (TT-Publikaties).

Levine, M.V. and D.B. Rubin (1979),
'Measuring the appropriateness of multiple-choice test score', *Journal of Educational Statistics*, 4, p. 269-290.

Knipscheer, C.P.M., J. De Jong-Gierveld, T.G. van Tilburg, P.A. Dykstra,
'Living arrangements and social networks of the elderly in The Netherlands: First results' [forthcoming].

Meijer, R.R. (1994),
Nonparametric person fit analysis, Ph.D.-thesis, Vrije Universiteit, Amsterdam.

Meijer, R.R. and E.D. de Leeuw (1993),
'Person fit in survey research: The detection of respondents with unexpected response patterns', in: Oud, J.H.L. and R.A.W. van Blokland-Vogelesang (eds), *Advances in longitudinal and multivariate analysis in the behavioral sciences*, Chapter 17, Nijmegen ITS, p. 236-245.

Meijer, R.R. and K. Sijtsma (1993),
'Reliability of item scores and its use in person fit research', in: Steijer, R., K.F. Wender and K.F. Widamann (eds), *Psychometric methodology*, Proceedings of the 7th European meeting of the Psychometric Society, Stuttgart (Gustav Fisher Verlag).

Meijer, R.R., K. Sijtsma and N.G. Smid (1990),
'Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT', *Applied Psychological Measurement*, 14, p. 283-298.

Mokken, R.J. (1971),
A theory and procedure of scale analysis, The Hague (Mouton).

Sijtsma, K. (1988),
Contributions to Mokken's nonparametric item response theory, Amsterdam (Free University Press).

Tatsuoka, K.K. and M.M. Tatsuoka (1982),
'Detection of aberrant response patterns', *Journal of Educational Statistics*, 7, p. 215-231.

Tatsuoka, K.K. and M.M. Tatsuoka (1983),
'Spotting erroneous rules of operation by the individual consistency index', *Journal of Educational Measurement*, 20, p. 221-230.

Flier, H. van der (1980),
Vergelijkbaarheid van individuele testprestaties (Comparability of individual test performance), Lisse (Swets and Zeitlinger).

Tilburg, T.G. van, and E.D. de Leeuw (1991),
'Stability of scale quality under various data collection procedures: A mode comparison on the "De Jong-Gierveld loneliness scale"', *International Journal of Public Opinion Research*, 3, p. 69-85.

Tilburg, T. van, P. Dykstra and A. Liefbroer (1993),
Questionnaire and documentation data; Nestor-program living arrangements and social networks of older adults main study 1992-network study 1992-1993, Internal report Departments of Sociology and social gerontology and Social research methodology, Vrije Universiteit, Amsterdam and Netherlands Interdisciplinary Demographic Institute, The Hague.

Widaman, K.F. (1985),
'Hierarchically tested covariance structure models for multitrait-multimethod data', *Applied Psychological Measurement*, 9, p. 1-26.