Journal of Official Statistics, Vol. 21, No. 4, 2005, pp. 657–674

An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview

Herbert Matschinger[1], *Sebastian Bernert*[1], *and Matthias C. Angermeyer*[1]

This article is concerned with the explanation of unexpectedly low prevalence rates in the course of an epidemiological mental health survey. It is assumed that the probability of the endorsement of important screening questions is at least partially responsible for this, since it decreases in the course of the work of the interviewer. First the decrease in the probability of the screening variable is demonstrated by means of a conditional fixed within model, and then the probability of the screening variable is analysed by a mixture logistic regression within latent classes. Two models with 4 and 5 classes are estimated and compared with respect to the difference in interviewer behaviour. It is shown that only a small segment of the set of interviewers is responsible for the effect of the sequence of interviews on the probability of the screening variable and that the experience with the CAPI system is moderately associated with the latent classes. By means of this model, those interviewers could be identified who are responsible for the artefact under study.

Key words: Epidemiological survey; random coefficient model; latent class model; mixture model; logistic regression; interviewer effect.

1. Introduction

Mental health surveys aimed at assessing a wide range of mental disorders often use voluminous questionnaires. Applying the whole questionnaire to all respondents would lead to long interviews without increasing the amount of information gathered. Therefore screening questions are implemented to reduce the average interview time without losing information. For example, in a mental health questionnaire the respondent will be asked about his or her lifetime experiences of different mood states. Only if such an experience is reported will further questions follow. Preceding screening questions have been found to be an efficient questionnaire design characteristic when it comes to minimizing the average interview time.

In an epidemiological survey covering six European countries (Alonso et al. 2002) it was found that, for the German sample, not only were the prevalence rates of certain mental disorders implausibly low, but also the endorsement of screening questions decreased considerably over the fieldwork period. The fraction of positively answered screening questions within each month of the study dropped from about 80% to only 7.3% in the last month (Table 1).

Even though we would assume that the probability of a positively answered screening question should not systematically vary during the fieldwork, the drop over the period of

[1] University of Leipzig, Department of Psychiatry, Johannisallee 20/1, D- 04317 Leipzig, Germany.
Email: math@medizin.uni-leipzig.de
Acknowledgments: We would like to thank the four anonymous reviewers and the editor for their valuable comments.

Table 1. Percentage of screening questions for each month of the survey period 1. line: frequencies 2. line: row percentages

	Screening question	
Month of the survey	No	Yes
11. 2000	18	68
	20.93	79.07
12. 2000	68	153
	30.77	69.23
1. 2001	213	309
	40.80	59.20
2. 2001	398	342
	53.78	46.22
3. 2001	558	333
	62.63	37.37
4. 2001	266	133
	66.67	33.33
5. 2001	361	109
	76.81	23.19
6. 2001	246	56
	81.46	18.54
7. 2001	189	38
	83.26	16.74
8. 2001	206	45
	82.07	17.93
9. 2001	164	13
	92.66	7.34
Total	2,687	1,599
	62.69	37.31

the fieldwork is not necessarily an indicator of defective data. Since the probability of a mental disorder can never be a design criterion, it might be possible that healthier people are investigated later. Therefore an analysis was started to find out what might have caused the very low prevalence rates. After excluding other possible explanations of the decrease (CAPI program error, question wording, etc.), the investigation was focused on the interviewer as a possible source of variance in handling screening questions. We should be aware of that the time-economic benefit of employing screening variables (shortening the interview without losing information) also applies to the interviewer. Shortening an interview by skipping a screening question will have an economic effect for the interviewer if, for instance, he or she is paid for the whole interview and not by the hour. Of course, shortening interviews in that way does not necessarily lead to interviews which are, on average, shorter than other ones. Screening question might be skipped to reduce the interview time to an acceptable level. Another reason to skip a screening question is to avoid an early end of the interview if the section following applies to more delicate topics, which might prompt the respondent to refuse further cooperation. However, the occurrence of this type of interviewer behaviour will have a devastating effect on the data, since it would lead to an underestimation of the prevalence rates of disorders in the population.

Of course, interviewer effects as a specific type of measurement error have been widely acknowledged in survey research and therefore the impact of interviewer behaviour on survey results has always been a matter of interest and concern (Fowler and Mangione 1998; Sudman, Bradburn, and Schwarz 1996). Among the various types of interviewer effects that can affect the survey response are:

1. Characteristics of the interviewer, i.e., race, class, sex, and particular experience or training;
2. Interaction between respondent and interviewer characteristics (same race, same gender, same attitudes vs. different race, different gender, attitudes, etc.);
3. Interaction between the respondent and the interviewer which leads to interviewer-specific bias of the responses;
4. Inaccurate recording of responses;
5. Inappropriate application of the interview (i.e., not reading questions according to their wording; direct probing for responses; learning effects, which lead to a systematic change of interviewing behaviour, etc.).

Research literature on interviewer effects has mainly examined the influence of interviewer characteristics (Finkel, Guterbock, and Borg 1991; Hox and De Leeuw 2002; Schuman and Converse 1971) and the interaction between respondent and interviewer characteristics on survey responses (Hox, De Leeuw, and Kreft 1991; van der Zouwen, Dijkstra, and Smit 1991; van der Zouwen and van Tilburg 2001). Also interactional problems in regular face-to-face interviews have been investigated (Suchman and Jordan 1990), as well as the effect of changing interviewers in panel studies (Campanelli and O'Muircheartaigh 2002; O'Muircheartaigh and Campanelli 1999). To combine both individual and interviewer characteristics the application of multilevel analysis was discussed (Hox 1994; Pickery and Loosveldt 2004). Until now, the inappropriate application of the interview resulting from an individual change in conducting the interviews over time has rarely been investigated (Biemer and Stokes 1989; Harrison and Krauss 2002; Roth 1966; Schnell and Kreuter 2000).

This article is aimed at presenting a strategy to assess and control for the variation in response due to a particular type of interviewer behaviour, addressed in the introduction. If an interviewer has got into the habit of avoiding parts of the interview by skipping screening variables, we should observe a considerable decrease in the probability for these variables between the first and last interview. Therefore a decrease in the probability should not only be observable over the period of the fieldwork, but certainly "within" the interviewer from the first to the last interview. The effect produced by this behaviour – which will result in considerable artefacts – will be called a "**sequence effect**" in the following. The probability of a positive answer depends on the number of interviews carried out by an interviewer before a particular interview. Of course, we should not expect the behaviour on the part of all interviewers, but rather on the part of certain – perhaps very small – subgroups. We refrain from calling it cheating, since from the mere data we never know why a particular interviewer changes his or her behaviour when dealing with a set of screening questions. Presenting a statistical approach in order to identify these subgroups of interviewers is the main topic of this article.

2. Data

The data examined in this study were collected as part of a general population mental health survey in Germany, as mentioned in the introduction. The survey encompassed all persons aged 18 years and older residing in the Federal Republic of Germany in private households. Persons sampled were selected from the registers of residents' registration offices. A total of 4,802 interviews were completed. After exclusion of proxy interviews, a total of 4,286 interviews remained for the analysis presented below. This dataset was not used in the final stage of this project. On account of the results presented below, the survey was repeated in 2002/2003. The type of interview is the *C*omposite *I*nternational *D*iagnostic *I*nterview (Robins et al. 1988), a comprehensive, fully structured diagnostic interview designed to assess mental disorders. This system, called CIDI 2000, includes screening questions (Lifetime Psychiatric Screening Instrument) that are administered to all respondents. These screening questions enquire about experiences of specific disorders, such as mood disorders (i.e., depression and dysthymia) and anxiety disorders (i.e., panic disorder). All respondents who answered any of these questions positively had to complete the CIDI section about the corresponding disorder in the course of the interview. We do not regard the sequence of these questions, but instead want to explain the occurrence of any positive answer out of a set of screening questions.

All interviews were conducted by lay interviewers using Computer Assisted Personal Interview (CAPI). Prior to the fieldwork, all interviewers received a three-day training course to learn how to administer the interview. Interviewers were allowed to conduct interviews only if they completed this training course successfully. Only 7 out of 95 interviewers had prior experience with the CIDI interview. The number of interviews carried out ranged from 1 to 260 for all the 95 interviewers. The experienced interviewers carried out many more interviews than the interviewers without any experience. The medians are 127 and 22, respectively. These experienced interviewers also elicited significantly fewer positive answers on the screening questions: 44.7% of the 3,241 interviews carried out by the 88 interviewers without interviewing experience provided a positive response on the screening questions, while only 14.5% of the 1,045 interviews done by the seven interviewers with previous interviewing experience exhibited a positive response. Therefore this group of interviewers was suspected of being responsible for the low prevalence rate, and CIDI experience was included in the model described below. Table 2 shows the relation between CIDI experience and the number of interviews.

Table 2. Number of interviews and interviewers

CIDI-experienced	interviews	percent	interviewers	percent	median mean
no	3,241	75.62	88	92.63	22 36.9
yes	1,045	24.38	7	7.37	127 149.3
Total	4,286	100.00	95	100.00	

3. Methods

To estimate the effect of the sequence on the probability of a screening question a special structure of the data is required. Table 3 gives a sketch of this structure for three interviewers. The interviews are ordered by date and time within each interviewer. The variable "*screening question*" indicates whether or not a positively answered screening question was observed for a particular interview. To predict the screening question, a variable – called "*sequence pointer*" in the following – was generated, which runs from 0 for the first interview of each interviewer to the maximum number of interviews a single interviewer has carried out, starting with 0 for each new interviewer in the sorted data file. Since the number of interviews carried out by each interviewer varies between one and 260 interviews, the number of interviews for each point in the chronological sequence decreases. We have a total of 95 interviewers, so we got 95 first interviews. Since ten interviewers carried out one interview only, only 85 interviewers are left for a second interview, and the number of second interviews drops down to 85. Five interviewers conducted two interviews, therefore 80 interviews, that is to say interviewers, are available for a third interview. The variable which comprises this information is called "*# of interviews*" (Table 3).

Tabulating the "*sequence pointer*" against the screening variable, it can be seen, for instance, that there are still four interviewers who carried out 194 or more interviews, thus there are 4 interviews, but none of them exhibits a positively answered screening question. As it comes, for instance, to the 256th interview, there are only two interviewers (and interviews) left. Of these two interviews, one shows a positively answered screening question.

Table 3. Data for three interviewers sorted by date and time

Screening question	# of interviews	Sequence pointer	Total number of interviews per interviewer	Time	Date	Interviewer code
0	95	0	259	12:09:00	03 Jan 01	5
1	85	1	259	16:56:10	03 Jan 01	5
0	80	2	259	09:08:01	04 Jan 01	5
1	76	3	259	10:32:29	04 Jan 01	5
0	74	4	259	13:15:46	04 Jan 01	5
0	72	5	259	14:00:21	04 Jan 01	5
1	95	0	38	09:43:35	06 Dec 00	21
1	85	1	38	13:05:19	09 Dec 00	21
1	80	2	38	10:02:59	11 Dec 00	21
1	76	3	38	17:03:05	11 Dec 00	21
1	74	4	38	18:27:03	11 Dec 00	21
1	72	5	38	17:39:00	12 Dec 00	21
0	70	6	38	10:05:32	13 Dec 00	21
1	69	7	38	14:18:58	18 Dec 00	21
1	95	0	3	04:48:31	30 Jan 01	26
0	85	1	3	05:13:54	03 Mar 01	26
1	80	2	3	21:19:46	04 Mar 01	26

Owing to this condition, the probability of a screening variable necessarily decreases and is negatively associated with the "*sequence pointer*". Therefore it is of vital importance to control for the number of interviews ("*# of interviews*") at each point in the sequence in order to obtain a correct partial estimate of the sequence effect. If all interviewers had conducted the same number of interviews, this figure would be a constant and could be omitted from the analysis.

3.1. Fixed model

We first estimated three conditional fixed-effect models with a logit link function and a binomial error structure because we wanted to show the effect of the sequence within each interviewer. This model can be written as:

$$\Pr(y_{it} = 1|x_{it}) = F(\alpha_i + x_{it}\beta) \tag{1}$$

and the cumulative logistic distribution:

$$F(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{2}$$

The subscripts i and t denote the independent units, the interviewers, and the individual interviews for each interviewer, respectively (Collet 1991; Greene 2003; Hamerle and Ronning 1995). The two variables "*sequence pointer*" and "*# of interviews*" at each point in the sequence were adopted as predictors. This analysis will provide a first insight into the effect of the sequence, assuming the homogeneity of the sample with respect to the model parameters. The estimations were carried out using STATA 8.2 (StataCorp 2003).

3.2. Mixture model

Whatever the result of these estimations, it must be assumed that the set of interviewers is heterogeneous with respect to these effects, and it would be possible to estimate the potential variance of all the model parameters by means of a random coefficient logit model (Anderson and Aitkin 1985; Bryk and Raudenbush 1992; Goldstein 1995). Since the goal is to identify groups of interviewers responsible for the presumed artefact, we will assume that these latent factors are discrete, employing a mixture model with latent classes of interviewers. Therefore a logistic regression within latent classes will be estimated (Hagenaars and McCutcheon 2002; Vermunt 1997; Vermunt and Magidson 1999; Wedel and DeSarbo 1994).

The simplest probability structure for a latent class of this kind is (Everitt and Hand 1981; McLachlan and Basford 1988):

$$f(y_1) = \sum_{x_1} \pi(x_1) f(y_1|x_1) \tag{3}$$

This model without any predictor makes it possible to describe the unobserved heterogeneity for the screening question y_1 with respect to an unobserved latent variable x_1. Usually two kinds of predictors are employed: variables to predict the probability of the latent classes x_1 and those to predict the dependent variable y_1. The former will be called z^c and the latter z^p (Kamakura, Wedel, and Agrawal 1994). The "*sequence pointer*" and

the "*number of interviews*" at each point in the sequence (called z_1^p and z_2^p respectively) are adopted as predictors for the screening variable. The variable indicating the CIDI experience (z^c) of the interviewer is used to predict the latent class probabilities. The LC regression model can now be written as:

$$f\left(y_1|z^c z_1^p z_2^p\right) = \sum_{x_1} \pi(x_1|z^c) \prod_t f\left(y_1|x_1 z_1^p z_2^p\right) \tag{4}$$

Since the dependent variable is dichotomous, the probability density for $f(y_1|z^c z_1^p z_2^p)$ is assumed to be binomial. $\pi(x_1|z^c)$ is the probability of belonging to a certain class (having a certain value on the latent variable x_1) given an observed value on z^c (for a detailed description see Vermunt and Magidson 2000, p. 160). It is highly recommended to employ both kinds of predictors simultaneously and not to predict the class membership in a second step, since otherwise the model might be incorrectly specified (Muthén 2001). Of course, this extension of the structural model by predictors of the latent class membership might result in a different structure (a different number of classes or a different allocation of observations to the same number of classes) with respect to an optimal likelihood.

The primary goal of this analysis is less to determine a parsimonious model with as few parameters as possible and to test this model against a specified alternative, but rather to detect aberrant patterns of interviewer behaviour. Therefore a mere inspection of the model fit is not sufficient. On the contrary, it is of specific interest to detect even very small subgroups of interviewers. Furthermore, testing for an optimally fitting model with respect to the number of classes by means of likelihood ratio tests is impossible, since the models are not nested and the assumptions for a likelihood ratio test do not hold (Titterington, Smith, and Makov 1985). For the mixture model, the degrees of freedom are not uniquely determined. Therefore a simple strategy based on one criterion only is not available. Fortunately, information criteria like the BIC (Bayes Information Criterion) and the AIC (Akaikes Information Criterion) allow ranking models with different degrees of freedom (Weakliem 2004). Both information criteria, the BIC and the AIC, were adopted to decide for a particular number of classes, since the BIC will tend to favour the more parsimonious model (Nagin 1999; Raftery 1995; Weakliem 1999). A thorough discussion of both criteria and their use for model decision can be found in Bauer and Curran (2004), Burnham and Anderson (2004), Huberty (1993), Kouha (2004) and Nagin and Tremblay (2001). Instead of selecting only one (best) model in accordance with the Schwarz criterion (Kass and Raftery 1995; Kass and Wasserman 1995; Schwarz 1978), we prefer to report and compare two very similar solutions. All the models will be estimated by LatentGold 3.03 (Vermunt and Magidson 2000; Vermunt and Magidson 2003).

4. Results

We already know from Table 2 that the seven experienced interviewers carried out many more interviews than did 88 interviewers without experience. Since the sum of screening questions is highly related to the sum of interviews ($r = 0.74$), the average sum of positively answered screening questions is 21.5, compared to 16.4 for the 88 inexperienced interviewers. However, these figures do not take intoconsideration the chronological order of the interviews, which will be done in the following.

4.1. The sequence effect evaluated by a fixed model

Table 4 shows the results for three conditional fixed effects models with and without control for the number of interviews at each point of the sequence ("*# of interviews*") and the interaction between the sequence of interviews and the interviewer specific variable "*CIDI – experience*." To estimate these models, 16 interviewers who carried out 27 interviews had to be discarded because they did not show any variance with respect to the screening variable. They either showed no positive responses or only positive responses for all their interviews. If the number of interviews is not considered to be a predictor, the negative effect of the sequence is nothing but the total effect of the sequence on the screening variable (Model 1). Since the regression of "*# of interviews*" on the sequence is negative (-0.33146 $p = 0.000$, Wald $CHI^2_{(1)} = 6{,}259.19$), the change in the screening probability within the sequence seems to be dependent on the number of interviews at each point in the sequence only (Model 2). The estimates of the second model clearly show that for the entire sample no sequence effect can be observed, provided that the number of interviews at each point in the sequence is controlled for. Model 3 further includes the interaction between the "*sequence pointer*" and the dichotomous variable, which indicates the experience with the CIDI system. Again it is shown that the seven experienced interviewers do not differ with respect to their behaviour in treating the screening questions. The interaction term is negative as the decreasing slope for the experienced interviewers is a bit steeper than that for those interviewers who only received the three-day training, but the coefficient is rather small and far from being significant. Thus we must conclude that the heterogeneity of the interviewers could not be sufficiently explained by the experience with the CIDI system.

4.2. Latent heterogeneity of the 95 interviewers

So far, we have investigated the effects by means of a fixed model. To explore the latent heterogeneity, we estimated the logistic regression as described in Section 3.2 within one to seven classes. Looking at the 1-class solution first, no considerable effect for the "*sequence pointer*" can be observed. The R^2 is about 0.10 if both the "*sequence pointer*" and the "*number of interviews*" are employed as predictors, but the coefficient for the former is only 0.0009 ($p = 0.44$). The decline of the predicted probabilities of giving

Table 4. Conditional fixed effects model for cross-sectional time series data. Minimum number of interviews = 2; Average number of interviews = 53.9; Maximum number of interviews = 260

Variable	Model 1	Model 2	Model 3
Sequence pointer	−.0058***	.0018	.0024
# of interviews		.0220***	.0219***
CIDI*Sequence p.			−.0027
N	4,259	4,259	4,259
LL	−2,083.8182	−2,060.323	−2,059.439
CHI^2	41.5302	88.520	90.288

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$
Likelihood-test Model 1 = Model 2 $chi^2_{(1)} = 46.99$
Likelihood-test Model 2 = Model 3 $chi^2_{(1)} = 1.77$

a positive answer depends on the latter variable only (0.294 $p = 0.000$) and is shown in Figure 1. Each data point in Figure 1 represents the predicted probability of a positive answer for a particular point in the sequence from the first interview to the 260th interview.

Table 5 indicates that both the 4-class and the 5-class solution turned out to be sufficient, the BIC, as expected, giving preference to the more parsimonious model. It is necessary to compare these two solutions for different numbers of classes as the identification of a "correct" number of classes always remains arbitrary (Bauer and Curran 2003).

Classification statistics show that the error of allocation to one of the two categories of the screening variable is improved by 58%. Nevertheless, the error with regard to allocating interviewers to one of the classes is still 23%, but it cannot be improved by extending the number of latent classes. In other words, the two models do not differ very much with respect to the criteria mentioned above. The fraction of positively answered screening questions decreases with the class size for both the 4- and the 5-class solution.

The probability of being an experienced CIDI interviewer is shared by Classes 3 and 4, or by Classes 4 and 5, thus indicating that all seven interviewers belong to two classes only, a result which holds both for the 4- and the 5-class solution (Tables 6 and 7). However, there is no latent class that comprises solely the entire group of CIDI-experienced interviewers. Tables 8 and 9 provide the coefficient of the mixture models for 4 and 5 classes, respectively. Table 8 shows that for Classes 1, 3, and 4, the sequence effect is negative even after controlling for the number of interviews. The effect is significant for Class 3 only. The regression parameters for the CIDI variable are positive for Classes 3 and 4, thus the probability of belonging to one of these latent classes is increased for this type of interviewers. This is in accordance with the results presented in Table 6. The Wald statistic for equality of parameters is 15.77 ($p = 0.001$), indicating that the sequence effects of the four classes are really different. The overall R^2 (0.27) is surprisingly high. This shows that the effect of the pure ordering of the interviews within the interviewers is particularly important.

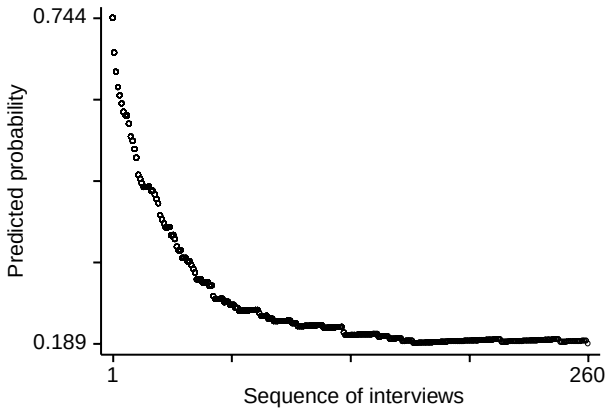

Fig. 1. Predicted probability of the screening variable against the sequence controlling for the # of interviews (1- class model)

Table 5. Criteria of model selection

Number of classes	LL	BIC(LL)	AIC(LL)
1	−2,609.8592	5,233.3801	5,225.7185
2	−2,369.6472	4,775.7255	4,755.2944
3	−2,324.5865	4,708.3734	4,675.1730
4	−2,307.5555	**4,697.0807**	4,651.1109
5	−2,299.9397	4,704.6186	**4,645.8794**
6	−2,296.2615	4,720.0316	4,648.5230
7	−2,292.9245	4,736.1269	4,651.8490

Table 6. Latent class probabilities and proportion of positively answered screening questions (four classes)

	Class 1	Class 2	Class 3	Class 4
Class Size	0.4526	0.2398	0.1945	0.1131
SCREEN				
proportion	0.7047	0.4169	0.2285	0.0918
CIDI-experience				
no	0.4885	0.2588	0.1641	0.0886
yes	0.0000	0.0000	0.5778	0.4221

Table 7. Latent class probabilities and proportion of positively answered screening questions (five classes)

	Class 1	Class 2	Class 3	Class 4	Class 5
Class size	0.4528	0.1786	0.1284	0.1211	0.1191
SCREEN					
proportion	0.7085	0.4000	0.3490	0.2103	0.0949
CIDI-experience					
no	0.4885	0.1927	0.1386	0.0852	0.0950
yes	0.0000	0.0000	0.0000	0.5759	0.4240

For the 5-class solution (Table 9), three classes exhibit a considerable decrease of the probability. The Wald-statistic for equality of parameters (22.06 $p = .000$) indicates again that the effect is not identical for all the five classes. Classes 3, 4, and 5 show a negative effect of the sequence, which is statistically relevant for the third class only. The regression of the class membership on the dichotomous CIDI variable shows that only for Classes 4 and 5, a considerable effect can be reported. Surprisingly enough, the probability for a CIDI-experienced interviewer to be in the 3rd class is zero (cf. Table 7). The segmentation of the 95 interviewers into four or five groups is remarkable also for the 2nd class, which exhibits a positive effect of the "*sequence pointer*" for both solutions, which means that the probability of a positive answer increases with the number of interviews already carried out by a particular interviewer. We will further inspect this in Section 4.3

Table 8. Parameter estimates of the logit model for four latent classes

	Class 1	s.e.	Class 2	s.e.	Class 3	s.e.	Class 4	s.e.	Overall	
									Mean	Std.Dev.
R^2	0.0068		0.0663		0.0553		0.0533		0.2654	
Intercept	0.5948	0.8618	−2.2013	0.3312	−1.4044	0.3613	−2.8262	1.2317	−0.8513	1.3726
Sequence pointer	−0.0012	0.0113	0.0076	0.0019	−0.0051	0.0020	−0.0068	0.0130	−0.0005	0.0050
# of interviews	0.0085	0.0115	0.0354	0.0058	0.0133	0.0060	0.0221	0.0174	0.0174	0.0109
Model for Classes										
Intercept	−0.4125	1.1274	−0.7311	1.1329	0.8030	0.5726	0.3407	0.5884		
CIDI-exp.	−1.265	1.1.27	−0.846	1.133	1.029	0.472	1.182	0.586		

Table 9. Parameter estimates of the logit model for five latent classes

	Class 1	s.e.	Class 2	s.e.	Class 3	s.e.	Class 4	s.e.	Class 5	s.e.	Overall	
											Mean	Std.Dev.
R^2	0.0073		0.0612		0.1572		0.0332		0.0606		0.2685	
Intercept	0.5232	0.8710	−2.3424	0.4030	0.9566	0.9818	−1.2653	0.4156	−2.9393	1.1996	−0.5619	1.4590
Sequence pointer	−0.0004	0.0114	0.0084	0.0022	−0.0329	0.0099	−0.0049	0.0028	−0.0061	0.0127	−0.0042	0.0119
# of interviews	0.0097	0.0116	0.0360	0.0077	−0.0075	0.0150	0.0078	0.0072	0.0245	0.0169	0.0137	0.0131
Model for Classes												
Intercept	−0.1599	1.3489	−0.6260	1.3603	−0.7910	1.3657	0.8387	0.6541	0.7382	0.6417		
CIDI-exp.	−1.2698	1.3489	−0.8024	1.3603	−0.6355	1.3658	1.4552	0.6533	1.2525	0.6405		

by cross-tabulating the modal allocations to the latent classes for the 4- and the 5-class solution.

To further demonstrate the effect of the "*sequence pointer*" on the probability of a person's positively responding to the screening question, a plot of the predicted probabilities against the sequence within each of the four or five classes is provided in Figures 2 and 3. There is still variance within each class, but the different groups can easily be identified with respect to their different behaviour over time. Each latent class is marked with its particular number and the y-axis is labelled between the minimum and maximum overall probability. Figure 3 clearly shows that the most prominent decrease in the probability over time is observed for Class 3. Both ends of the distribution are marked with the number 3.

4.3. Relation between the 4- and the 5-class solution

As already shown in Table 2, 7% of the interviewers carried out more than 24% of the interviews. Therefore any effect generated by these few interviewers will be disproportionately great. Table 10 shows the cross-tabulation between the latent classes for both solutions and the number of interviews performed by the experienced and inexperienced interviewers. As expected, the 7 CIDI-experienced interviewers are all members of 2 classes only: Classes 3 and 4 and Classes 4 and 5 for the 4- and the 5-class solution, respectively. As shown in Tables 8 and 9, these classes are all characterized by a negative effect of the "*sequence pointer*." Therefore we could assume that these interviewers are at least partially responsible for the decreasing probability of the screening variable.

The third column (Class 3) of the cross-tabulation shows that the 4-class solution does not tell us the whole story, as the interviewers of the third class are distributed across all other classes of the 5-class solution, except for the first class. One interviewer (115 interviews) "moved" to the second class, where the sequence effect is more positive. Two interviewers, who carried out 145 and 151 interviews, are now to be found in Class 3, where the sequence effect is extremely negative (cf. Figure 3); neither of them had former experience with the CIDI system. The majority of interviews and interviewers are located in Class 4, and two interviewers are to be found in Class 5. Classes 4 and 5 each comprise 575 interviews conducted by 8 interviewers, three of whom exhibit experience with the CIDI (see the 4th column of Table 10). Also, the second class seems rather heterogeneous, since the results from Table 8 show a positive sequence effect, but two out of 16 interviewers "moved" to Class 3 of the 5-class solution, where the sequence effect is definitely negative. Those four interviewers of Class 3 (3rd row of Table 10) should be inspected carefully.

5. Discussion

The approach outlined above was adopted to identify the reasons for unexpectedly low prevalence rates observed in an epidemiological mental health survey. It has been shown that these low figures result from a decline in the probability of a screening variable within interviewers. We were able to demonstrate that the heterogeneity of the set of interviewers must be investigated carefully, particularly if the fixed effect of the sequence within each

interviewer seems to be negligible, at least in statistical terms. If relevant characteristics of the interviewers are not available, segmentizing by means of a mixture model turns out to be a good strategy (Allenby and Rossi 1999; Wedel and DeSarbo 1995; Wedel and Kamakura 2000). Two models with four and five latent classes, respectively, provide the information to identify those interviewers who are at least partially responsible for the decrease. Since the classes differ considerably with respect to their behaviour over time, any fixed model to explain the probability has to be rejected. The mixture model clearly

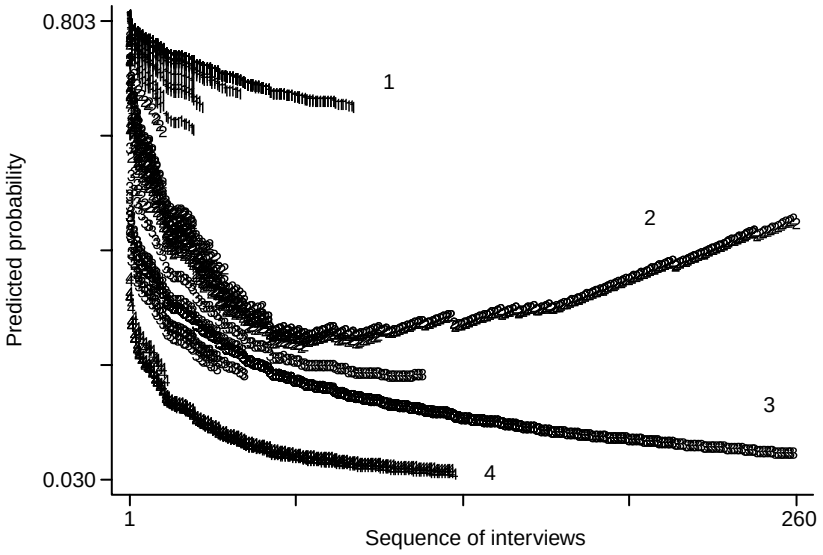

Fig. 2. Predicted probability of the screening variable against the sequence for a four class solution

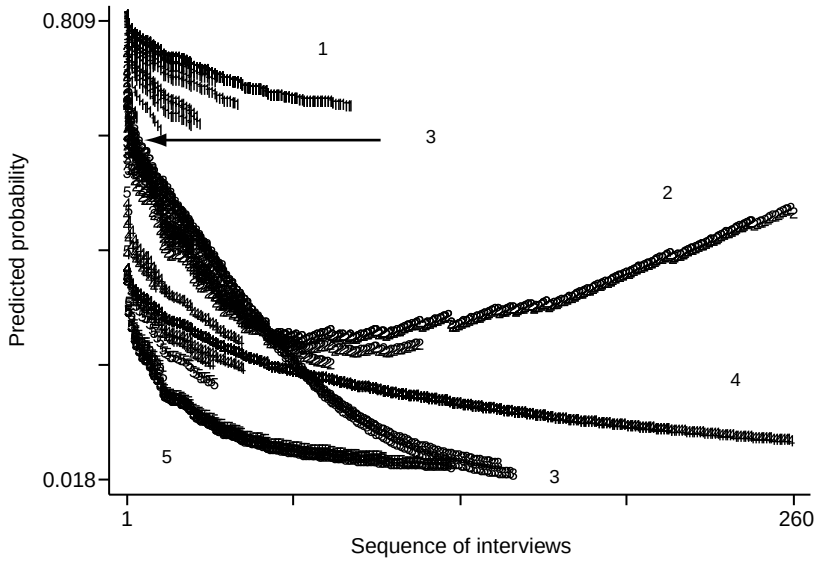

Fig. 3. Predicted probability of the screening variable against the sequence for a five class solution

shows that even if we do not observe any sequence effect "on average" by means of a fixed model, we should not trust in the homogeneity of the set of interviewers, but rather try to explain the heterogeneity by means of a latent class model. It was suspected from a mere inspection of the data on the interviewer level that a particular subgroup of interviewers, characterized by a special training, could be responsible for the low prevalence rates. This turned out to be only partially true. The five-class solution furthermore indicates a subgroup of four interviewers for which the "*sequence effect*" reaches its maximum. The comparison of the 4- and the 5-class solutions also demonstrates the promising opportunities of analysing the latent heterogeneity by a latent class analysis. Modelling the manifest heterogeneity by means of individual characteristics sometimes turns out to be insufficient. Of course, it would be of great interest to further explain these potential artefacts and the manifest heterogeneity of the interviewers by other characteristics of the interviewers, but this was not the core aim of our approach.

Any sequence effect estimated by a within model must be attributed to an artefact resulting from an undesirable behaviour of the interviewer. We would assume that the probability of a positively answered screening question stays constant over time. At least it should not be possible to explain any change by means of individual characteristics or a particular experience of the interviewer. However, the investigation in the course of the survey, in order to control for the artefact as early as possible, is still problematic. Since the estimation of the model parameters depends on the variance of the number of observations (interviews) in each group (interviewer), the number of interviews should not vary too much. Particularly, no singletons or interviewers with only two interviews should be observed, which is highly probable at the beginning of a survey. Ideally, all interviewers should carry out the same number of interviews within the same period of time. However, this will never be possible in practice. If the survey design comes close to it, this sort of a quality control can be carried out at an early stage of the campaign. Furthermore, any learning effect will diminish only if a small number of interviews are performed by each interviewer, as can be seen from the performance of the 56 interviewers

Table 10. Cross tabulation between the 4 and the 5-class solution. 1. row = number of interviews. 2. row = number of interviewers

Class	1	2	3	4	Total
1	1,115	20	0	0	1,135
	56	2	0	0	58
2	0	904	115	0	1,019
	0	12	1	0	13
3	0	151	296	0	447
	0	2	2	0	4
4	0	0	1,073	0	1,073
	0	0	**6/4**[1]	0	10
5	0	0	37	575	612
	0	0	2	**5/3**[1]	10
Total	1,115	1,075	1,521	575	4,286
	56	16	15	8	95

[1] Interviewer with prior experience with the CIDI system

in Class 1, where the maximum number of interviews is 88. Nevertheless, it will always be possible to investigate the effect when the fieldwork is completed.

It is questionable whether or not the identification of subgroups of interviewers can be used to "clean" the available data simply by discarding the interviews generated by the interviewers of a particular "incriminated" class. In our view, this is not possible for at least three reasons. First, there is no criterion available to decide which interview is biased from which point in the sequence since the segmentation has been done with respect to the set of interviewers. Second, and even more important, we have to consider the probabilistic nature of the latent class model. Each interviewer is characterised by a probability for each latent class, which sum up to one. The modal allocation to only one of these classes necessarily has to ignore this. On the other hand, it is not possible to discard interviews because of this probability. Third, if interviews are removed from the original sample, the remaining set cannot be considered a random, representative sample any more, since the selection criterion is a systematic one. As already mentioned the data used for the analysis above had not been used for any other, substantial analysis. The study for Germany was repeated in 2002/2003.

6. References

Allenby, G.M. and Rossi, P.E. (1999). Marketing Models of Consumer Heterogeneity. Journal of Econometrics, 89, 57–78.

Alonso, J., Ferrer, M., Romera, B., Vilagut, G., Angermeyer, M., Bernert, S., Brugha, T.S., Taub, N., McColgen, Z., De Girolamo, G., Polidori, G., Mazzi, F., De Graaf, R., Vollebergh, W.A., Buist-Bowman, M.A., Demyttenaere, K., Gasquet, I., Haro, J.M., Palacin, C., Autonell, J., Katz, S.J., Kessler, R.C., Kovess, V., Lepine, J.P., Arbabzadeh-Bouchez, S., Ormel, J., and Bruffaerts, R. (2002). The European Study of the Epidemiology of Mental Disorders (ESEMeD/MHEDEA 2000) Project: Rationale and Methods. International Journal of Methods in Psychiatric Research, 11, 55–67.

Anderson, D.A. and Aitkin, M. (1985). Variance Component Models with Binary Response: Interviewer Variability. Journal of the Royal Statistical Society, Series B, 47, 203–210.

Bauer, D.J. and Curran, P.J. (2003). Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes. Psychological Methods, 8, 338–363.

Bauer, D.J. and Curran, P.J. (2004). The Integration of Continuous and Discrete Latent Variable Models: Potential Problems and Potential Opportunities. Psychological Methods, 9, 3–29.

Biemer, P.P. and Stokes, S.L. (1989). The Optimal Design of Quality Control Samples to Detect Interviewer Cheating. Journal of Official Statistics, 5, 23–39.

Bryk, A.S. and Raudenbush, S.W. (1992). Hierarchical Linear Models; Applications and Data Analysis Methods. Newbury Park: Sage Publications.

Burnham, P.K. and Anderson, R.D. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods and Research, 33, 261–304.

Campanelli, P. and O'Muircheartaigh, C. (2002). The Importance of Experimental Control in Testing the Impact of Interviewer Continuity on Panel Survey Nonrepsonse. Quality and Quantity, 36, 129–144.

Collet, D. (1991). Modelling Binary Data. London: Chapman and Hall.

Everitt, B.S. and Hand, D.J. (1981). Finite Mixture Distributions. London: Chapman and Hall.

Finkel, S.E., Guterbock, T.M., and Borg, M.J. (1991). Race of Interviewer Effects in a Preelection Poll: Virginia 1989. Public Opinion Quarterly, 55, 313–330.

Fowler, F.J. and Mangione, T.W. (1998). Standardized Survey Interviewing Minimizing Interviewer-related Error. Newbury Park: Sage.

Goldstein, H. (1995). Multilevel Statistical Models. London, Sydney, Auckland: Arnold.

Greene, W.H. (2003). Econometric Analysis; International Edition. (5th ed.) Upper Saddle River, New Jersey: Prentice Hall.

Hagenaars, J.A. and McCutcheon, A. (2002). Advanced Latent Class Analysis. Cambridge: Cambridge University Press.

Hamerle, A. and Ronning, G. (1995). Panel Analysis for Qualitative Variables. In Handbook of Statistical Modelling for the Social and Behavioural Sciences G. Arminger, C.C. Clogg, and M.E. Sobel (eds). New York, London: Plenum Press, 401–451.

Harrison, D.E. and Krauss, S.I. (2002). Interviewer Cheating: Implications for Research on Entrepreneurship in Africa. Journal of Developmental Entrepreneurship, 7, 319–330.

Hox, J.J. (1994). Hierarchical Regression-Models for Interviewer and Respondent Effects. Sociological Methods and Research, 22, 300–318.

Hox, J.J., DeLeeuw, E.D., and Kreft, I.G. (1991). The Effect of Interviewer and Respondent Characteristics on the Quality of Survey Data: A Multilevel Model. In Measurement Errors in Surveys, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley, 439–461.

Hox, J. and DeLeeuw, E.D. (2002). The Influence of Interviewers' Attitude and Behaviour in Household Survey Nonresponse: An International Comparison. In Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J. Little (eds). New York: Wiley, 103–120.

Huberty, C.J. (1993). Historical Origins of Statistical Testing Practices: The Treatment of Fisher Versus Neyman-Pearson Views in Textbooks. Journal of Experimental Education, 61, 317–333.

Kamakura, W.A., Wedel, M., and Agrawal, J. (1994). Concomitant Latent Variable Latent Class Models for the External Analysis of Choice Data. International Journal of Marketing Research, 11, 541–564.

Kass, R.E. and Raftery, A.E. (1995). Bayes Factor. Journal of the American Statistical Association, 90, 773–795.

Kass, R.E. and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. Journal of the American Statistical Association, 90, 928–934.

Kouha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. Sociological Methods and Research, 33, 188–229.

McLachlan, G.J. and Basford, K.E. (1988). Mixture Models. New York: Marcel Dekker.

Muthén, B.O. (2001). Latent Variable Mixture Modelling. In Advanced Structural Equation Modelling: New Developments and Techniques, A. Marcoulides and R.E. Schumacker (eds). Lawrence Erlbaum, 1–33.

Nagin, D.S. (1999). Analysing Developmental Trajectories: A Semi-Parametric, Group-Based Approach. Psychological Methods, 4, 139–157.

Nagin, D.S. and Tremblay, R.E. (2001). Analysing Developmental Trajectories of Distinct but Related Behaviours. A Group Based Method. Psychological Methods, 6, 18–34.

O'Muircheartaigh, C. and Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. Journal of the Royal Statistical Society, Series A, 162, 437–446.

Pickery, J. and Loosveldt, G. (2004). A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviews. Journal of Official Statistics, 20, 77–89.

Raftery, A.E. (1995). Bayesian Model Selection in Social Research. In Sociological Methodology, A.E. Raftery (ed.). Oxford: Blackwell, 111–163.

Robins, L.N., Wing, J., Wittchen, H.U., Helzer, J.E., Babor, T.F., and Burke, J. (1988). The Composite International Diagnostic Interview. An Epidemiologic Instrument for Use in Conjunction with Different Diagnostic Systems and in Different Cultures. Archives of General Psychiatry, 45, 1069–1077.

Roth, J.A. (1966). Hired Hand Research. The American Sociologist, 1, 190–196.

Schnell, R. and Kreuter, F. (2000). Untersuchungen zur Ursache unterschiedlicher Ergebnisse sehr ähnlicher Viktimisierungssurveys. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 52, 96–117. [In German]

Schuman, H. and Converse, J.M. (1971). The Effects of Black and White Interviewers on Black Responses in 1968. Public Opinion Quarterly, 35, 44–68.

Schwarz, G. (1978). Estimating Dimensions of a Model. Annals of Statistics, 6, 461–464.

StataCorp (2003). Stata Statistical Software: Release 8. College Station, TX: Stata Corp LP.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-To-Face Survey Interviews. Journal of the American Statistical Association, 85, 232–253.

Sudman, S., Bradburn, N., and Schwarz, N. (1996). Thinking about Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions. New York: Wiley.

van der Zouwen, J., Dijkstra, W., and Smit, J.H. (1991). Studying Respondent-Interviewer Interaction: The Relationship Between Interviewing Style, Interviewer Behaviour, and Response Behaviour. In Measurement Errors in Surveys, P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: Wiley, 419–437.

van der Zouwen, J. and van Tilburg, T. (2001). Reactivity in Panel Studies and Its Consequences for Testing Causal Hypotheses. Sociological Methods and Research, 30, 35–56.

Vermunt, J.K. (1997). Log-linear Models for Event Histories. Thousand Oaks, London, New Delhi: Sage Publications.

Vermunt, J.K. and Magidson, J. (1999). Exploratory Latent Class Cluster, Factor, and Regression Analysis: The Latent Gold Approach. In Proceedings EMPS'99 Conference (Anonymous). Lüneburg.

Vermunt, J.K. and Magidson, J. (2000). Latent GOLD Users Guide 2.0. Belmont MA: Statistical Innovations Inc.

Vermunt, J.K. and Magidson, J. (2003). Addendum to the Latent GOLD User's Guide: Upgrade Manual for Version 3.0. Belmont, Massachusetts: Statistical Innovations Inc.

Weakliem, D.L. (1999). A Critique of the Bayesian Information Criterion for Model Selection. Sociological Methods and Research, 27, 359–397.

Weakliem, D.L. (2004). Introduction to the Special Issue on Model Selection. Sociological Methods and Research, 33, 167–187.

Wedel, M. and DeSarbo, W.S. (1994). A Reviews of Recent Developments in Latent Class Regression Models. In Advanced Methods of Marketing Research, R.P. Bagozzi (ed.). Cambridge: Blackwell Publishers, 352–388.

Wedel, M. and DeSarbo, W.S. (1995). A Mixture Likelihood Approach for Generalized Linear Models. Journal of Classification, 12, 21–55.

Wedel, M. and Kamakura, W.A. (2000). Market Segmentation; Concepts and Methodological Foundations. 2nd ed. Dordrecht: Kluwer Academic Publishers.

Received November 2003
Revised May 2005