The Effect of Probing "Don't Know" Responses on Measurement Quality and Nonresponse in Surveys

Jouni Kuha* Sarah Butt Myrsini Katsikatsou Chris J. Skinner

April 19, 2017

Abstract

In survey interviews, "Don't know" (DK) responses are commonly treated as missing data. One way to reduce the rate of such responses is to probe initial DK answers with a follow-up question designed to encourage respondents to give substantive, non-DK responses. However, such probing can also reduce data quality by introducing additional or differential measurement error. We propose a latent variable model for analyzing the effects of probing on responses to survey questions. The model makes it possible to separate measurement effects of probing from true differences between respondents who do and do not require probing. We analyze new data from an experiment which compared responses to two multi-item batteries of questions with and without probing. In this study, probing reduced the rate of DK responses by around a half. However, it also had substantial measurement effects, in that probed answers were often weaker measures of constructs of interest than were unprobed answers. These effects were larger for questions on attitudes than for pseudo-knowledge questions on perceptions of external facts. The results provide evidence against the use of probing of "Don't know" responses, at least for the kinds of items and respondents considered in this study.

KEY WORDS: Interviewing; Latent class model; Latent variable model; Measurement error; Missing data; Questionnaire design

*Jouni Kuha is Associate Professor (E-mail: *J.Kuha@lse.ac.uk*), Myrsini Katsikatsou is Postdoctoral Researcher (E-mail: *M.Katsikatsou@lse.ac.uk*), and Chris J. Skinner is Professor (E-Mail: *C.J.Skinner@lse.ac.uk*), Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Sarah Butt is Research Fellow, Department of Sociology, City, University London, Northampton Square, London EC1V 0HB, United Kingdom (E-mail: *sarah.butt.1@city.ac.uk*). This work was supported by a grant from the UK National Centre for Research Methods under its program of Methodological Innovation Projects. Fieldwork was conducted using the ESS Innovation Sample administered by Ipsos MORI as part of *The European Social Survey: Data in a Changing Europe* project (ESS DACE) supported by the European Union under Framework Programme 7 (Research Infrastructures), GA number 262208. We are grateful to the Editor, Associate Editor, and two reviewers for their comments, which substantially improved the article.

1 INTRODUCTION

One of the common problems in survey research is item nonresponse, that is missing data in individual survey questions ('items'). In this paper we consider in particular "Don't know" (DK) responses to survey items. Such responses can be meaningful, as for example in reply to a question about intended vote in an election or as an expression of genuinely absent attitudes (Converse 1964; Krosnick 2002). For many items, however, DK responses are regarded not as valid answers but as nonresponse, leading to the usual problems of missing data such as potential bias and loss of efficiency.

Item nonresponse can have many causes, and we may try to minimise it in different ways, such as through good question design and by limiting interview length to avoid respondent fatigue. Another possible way of reducing DK responses is by discouraging them during the questioning process. Adopting terminology of Luskin and Bullock (2011), three ways of asking a question are: *DK-encouraging* — respondents are offered an explicit DK option and encouraged to use it if they have difficulty answering the question; *DK-neutral* — no DK option is offered but a spontaneously volunteered DK is accepted without discouragement or encouragement; and *DK-discouraging* (probing) — respondents who give a spontaneous DK are 'probed' with a follow-up question, designed to elicit a substantive answer.

Although a DK-discouraging approach may seem the most desirable, on the grounds of minimizing missing data, it also has potential disadvantages. Probing takes up interview time and adds to the overall cost of a survey, and it can increase respondent burden (Bradburn, 1978) from the extra time and by adding effort and stress from being pressed harder for answers. It may also have unwanted effects on the measurement process (Bishop et al., 1986). One hypothesis is that people opt to respond DK at first because this provides an easy short-cut which reduces cognitive effort, but that they can give well-considered substantive answers if pressed (Krosnick 2002; Krosnick et al. 2002). On the other hand, it may be that the pressure to answer generates measurement error, by leading the individual to give a hasty response different from the 'true' answer they would have supplied after careful consideration (Schuman and Presser 1981; Sanchez and Morchio 1992).

We say that probing has a *measurement effect* if responses obtained after probing have different properties as measures of the constructs of interest than do responses which did

not require probing. This is the case if probed answers are less accurate, but also if they are more accurate or just behave differently from unprobed answers (these possibilities can be explored in a multivariate setting, where we define a measurement effect more precisely in Section 3.3.2). In standard survey analysis, probed and unprobed answers would be treated in exactly the same way. This is appropriate if there is no measurement effect, but not otherwise. The key question of interest is then whether probing has *any* measurement effect. If it does not, it may be judged appropriate. If it does, probing might be deemed inappropriate, because it could mean replacing the 'error of omission' of the initial DK with the 'error of commission' of answers with lower or different quality (Beatty et al., 1998).

In this paper we seek to assess the measurement effects of probing. We propose a model designed to uncover these effects, and use it to analyse data from an experiment to compare the DK-neutral and the probing approaches. The study was conducted using the European Social Survey (ESS) Innovation Sample and forms part of an ongoing programme of work to develop the methodology of the ESS, one of the leading sources of cross-national survey data on public attitudes and opinions (Jowell et al. 2007; Stoop et al. 2010, chap. 3).

Previous studies provide both experimental and observational evidence on the effects of DK-discouragement; see, for example, McLendon and Alwin (1993), Beatty and Herrmann (2002), Krosnick (2002), Krosnick et al. (2002), Schaeffer and Presser (2003), Alwin (2007, sec. 9.5), and references therein. This literature has focussed especially on attitude items but has also considered other kinds of questions, such as ones on political knowledge. One broad finding is that probing can lead to substantial reduction in the DK rate (Schuman and Presser, 1981). More contested is what the evidence reveals about the effect of probing on measurement quality and the implications for the preferred way of asking questions.

We contend that a key problem in establishing measurement effects of probing is that they may be confounded with a *selection effect* of which respondents do and do not require probing. The main source of evidence on measurement effects consists of comparisons of answers from two subsamples: immediate responders who responded without probing and responders who answered only after probing. The problem is that differences between them in the distributions of the responses could occur either because of measurement effects or because the two subsamples of respondents differ in what is being measured. Disentangling these two possible explanations is the methodological challenge addressed in this paper.

Based on the evidence of only univariate comparisons between immediate and probed responses, measurement and selection effects are entirely confounded. We suggest, however, that this confounding can be disentangled in the case of multi-item scales, where the items are regarded as multiple imperfect indicators of latent constructs such as attitudes. The dependence of the indicators on the latent attitudes is represented by a measurement model. We define measurement effects of probing as differences in the measurement models which apply to immediate responses and to probed responses. We can also examine the possibility of a 'carry-over' effect where probing on an item affects the measurement of subsequent items. We further suppose that DK responses to different items are also generated from an underlying latent response propensity. A selection effect is then defined as an association between the latent attitudes and the response propensity. We shall show how, given data on multiple items and a suitable latent variable modelling framework, it is possible to identify and estimate such measurement effects, whilst allowing for such selection effects.

Although the literature on the effects of different ways of handling DK responses is large, we are not aware that there has been any other attempt to separate selection and measurement effects. McLendon and Alwin (1993) comes perhaps the closest. They stress (p. 460) that they 'assume that any effects' of different DK-discouragement treatments 'come about because the composition of the sample included in the analysis changes' and remark that 'this assumption is much different than another possibility, which is that the same individuals might respond either more reliably or less reliably' to questioning with different treatments. This may be interpreted as recognition of the confounding issue. They also use a latent variable measurement model, testing for differences in the parameters of the models fitted to respondents under DK-encouraging and DK-neutral treatments. They do not attempt, however, to represent selection effects in the model.

Literatures on two related topics are also relevant to our work (here we cite some key references and recent reviews; more information can be found in the references therein). The first of them is the parallel question on *unit nonresponse*, that is, whether there are differences in measurement between respondents with different levels of willingness to answer a survey at all. Here willingness or reluctance may be measured by paradata indicators such as the number of nonresponse follow-ups required to obtain an interview (Olson et al. 2008; Kaminska et al. 2010; Hox et al. 2012; Olson 2013), in panel surveys by

whether a respondent dropped out in later waves (Roberts et al., 2014), or (less directly) by different levels of financial incentives offered to the respondents (Singer and Ye 2013; Grauenhorst et al. 2015; Medway and Tourangeau 2015).

The overall conclusions from this research are mixed. Some studies conclude that reluctant responders give poorer-quality answers in at least some contexts, while others find no evidence of such differences. Methodologically, it is again usually not possible to fully separate effects of measurement and selection (an exception are questions where the true values are known; see Olson 2013). Latent variable models which aim to do this have, to our knowledge, been used by only two studies in this context (Hox et al. 2012 for different numbers of follow-ups, and Medway 2012 for different levels of incentives). We discuss them further in Section 3.3.4, where we compare them with our models for probed responses.

The second related topic is the effect of the mode of data collection (face to face, telephone, mail or online) on measurement quality, a question of key interest in mixed-mode surveys (see e.g. Hox et al. 2015). Here the comparison of interest is different from ours (different modes rather than probing or not probing), but there are methodological similarities in that several studies have used latent variable models which are partially analogous to our models for probing. These methods are also discussed in Section 3.3.4.

Our probing study and initial analyses of it are described in Section 2. The general latent variable model which allows modelling of multi-item data with probing is introduced in Section 3, and used to analyse the data from the probing study in Section 4. Conclusions and implications of the results for survey practice are discussed in Section 5.

2 THE PROBING STUDY

To test the effect of probing on response we fielded eight survey items on omnibus surveys in three European countries, and compared the responses received with and without probing DK answers. Data were collected via face-to-face interviews in 2013 in Bulgaria, Hungary and Portugal. The respondents were selected from the general population in each country using quota sampling on the basis of gender, age and region. Around 1500 individuals were interviewed in each country, and the combined sample has 4470 respondents.

The study was done in three countries in order to examine also how effects of probing may vary between countries. The possibility of such cross-national variation in measurement effects is an important concern for cross-national surveys such as the ESS. It is, however, not the focus of this paper. To maximize the power of the analyses, we will mostly use the combined data from the three countries together. We thus treat them in effect as a single non-probability sample of subjects for an observational study where the focus is on comparing probed and unprobed responses. Variation between the countries is examined briefly through country-specific analyses, which are reported at the end of the data analysis in Section 4, and in the supplementary materials.

Because of the quota sampling design no design weights were available, and although post-stratification weights had been calculated they were also omitted and we use unweighted estimates throughout. The results presented here are thus not intended as estimates of the distributions of the attitude variables which the items are designed to measure, among the populations of the three countries. Rather than on these attitudes, the focus of the study is on the comparison of the measurement properties of probed and unprobed responses. Even though the data are not a probability sample, we have no obvious reason to think that the conclusions that we obtain from it about the measurement effects of probing would be substantially different from what they would be in the broader population.

Each respondent was randomly assigned to either a treatment (probing) group (75% of the sample in each country) or a control group (25%). Both groups were asked identical survey questions. The questions did not offer an explicit DK response option, but a respondent could still give such a response unprompted. If a respondent in the control group gave a DK response, this was accepted and treated as nonresponse, and the interviewer moved on to the next question. In contrast, if a respondent in the treatment group gave a DK response, the interviewer would probe them by reading out a probing statement and then repeating the original question. The probing statement was worded as 'We are interested in your views. If you are not sure please give the answer that comes closest to what you think'. Respondents were probed only once for each item; if they still gave a DK response, this was accepted and the interviewer moved on. The same procedure was repeated for every item for which the respondent gave an initial DK response. Another type of nonresponse which could occur was an immediate refusal to answer a question. Such refusals were not

Table 1: Wordings (in English) of the eight survey items included in the probing study, and short labels which will be used to refer to the items in the rest of the article.

Label	*Question wording*
Pseudo-knowledge items:	
'Of every 100 people of working age in [country] how many would you say...'	
PK1	'...are unemployed and looking for work?'
PK2	'...are long-term sick or disabled?'
PK3	'...do not have enough money for basic necessities?'
Response options:	0–4 (coded as 1), 5–9 (2), 10–14 (3), 15–19 (4), 20–24 (5), 25–29 (6), 30–34 (7), 35–39 (8), 40–44 (9), 45–49 (10), 50 or more (11).
Attitude items:	
'[...] Please say how much you agree or disagree with each of the following statements about people in [country].'	
AD1	'Most unemployed people do not really try to find a job.'
AD2	'Many people with very low incomes get less benefit than' they are legally entitled to.'
AD3	'Many people manage to obtain benefits and services to which they are not entitled.'
AD4	'There are insufficient benefits in [country] to help the people who are in real need.'
AD5	'Employees often pretend they are sick in order to stay at home.'
Response options:	Agree strongly (1), Agree (2), Neither agree nor disagree (3), Disagree (4), Disagree strongly (5).

probed. They were rare, amounting only to 0.1% of the responses across the eight items.

To ensure that the distinction between the probed and control groups was maintained, interviewers were carefully briefed as to the purpose of the study and the importance of probing for a response only when instructed. The eight items were fielded as close as possible to the start of the omnibus questionnaire to avoid responses being contaminated by previous items or usual omnibus probing procedures.

The survey items were taken from a module of questions measuring 'Welfare attitudes in a changing Europe' which was fielded in Round 4 of the European Social Survey (ESS) in 2008–9 (Svalfors et al., 2008) and which is included again (with some changes) in Round 8 in 2016–17. The wordings of the items and their response options are shown in Table 1, along with our labels for the items. The items were fielded in the same order and using the

same translations as in the ESS, with only the exception that a pre-emptive probe worded as 'if you are not sure please give your best guess', which was included for items *PK1* and *PK3* in the ESS, was omitted here. More information on ESS Round 4, including the source questionnaire and translations, can be found at `www.europeansocialsurvey.org`.

The items form two separate multi-item scales, allowing us to test the effect of probing on two different types of survey questions. The first scale of three items (*PK1*–*PK3*) is designed to measure a respondent's view on the level of need within society. Here the interest lies as much in people's perceptions of reality as in whether the answers are factually correct. We therefore refer to these items as *pseudo-knowledge* items. The second battery of five questions (*AD1*–*AD5*) are more standard attitudinal items, using a common five-point agree–disagree response scale which includes a middle option of 'Neither agree nor disagree'. The original questionnaire design team's expectation was that these items would measure two related aspects of attitudes towards welfare provision and its target beneficiaries, with *AD1*, *AD3* and *AD5* serving as measures of perceived abuse of the welfare system, and *AD2* and *AD4* of perceived sufficiency of welfare benefits. Our analysis suggests that in fact *AD3* appears to relate to both of these attitudes, so we will later use models where this item serves as a measure of both of them.

We treat DK answers as nonresponse rather than a meaningful statement of the absence of knowledge. The reason for doing so for the pseudo-knowledge items is that, as they require personal guesses only, any respondent should be able to produce one (this is perhaps clearer from the wording of the probe than the question itself). For the attitude items, a DK answer could also express a respondent's genuine lack of knowledge. In practice, however, DK responses to agree-disagree questions of this type are most commonly treated as nonresponse, and this is also what we will do here.

Table 2 shows the percentage of responses which were not DK or refusals for each of the items in the control and probed groups. Even without probing, these response rates were relatively high. For example, the rate for item *PK1* was 94.2% among the control group, whilst among the probed group 93.7% responded immediately and a further 3.5% when probed. Nevertheless, levels of item nonresponse are still non-negligible, especially in their cumulative effect over the course of the items. In the treatment group, 21.5% of the respondents gave an initial DK response to at least one item and were thus probed at least

Table 2: Response rates for the eight items considered in the probing study.

Item:	*PK1*	*PK2*	*PK3*	*AD1*	*AD2*	*AD3*	*AD4*	*AD5*
Control group	94.2	87.8	95.6	98.0	95.8	96.4	98.3	95.5
Probing group:								
Immediate response	93.7	88.1	95.5	98.2	95.9	96.0	97.9	95.1
Probed response	3.5	6.0	2.0	1.1	2.2	2.3	1.0	2.5
No response	2.8	5.8	2.5	0.7	1.9	1.7	1.1	2.3
Previously probed	*0*	*2.4*	*11.1*	*14.5*	*14.4*	*15.6*	*17.9*	*16.7*

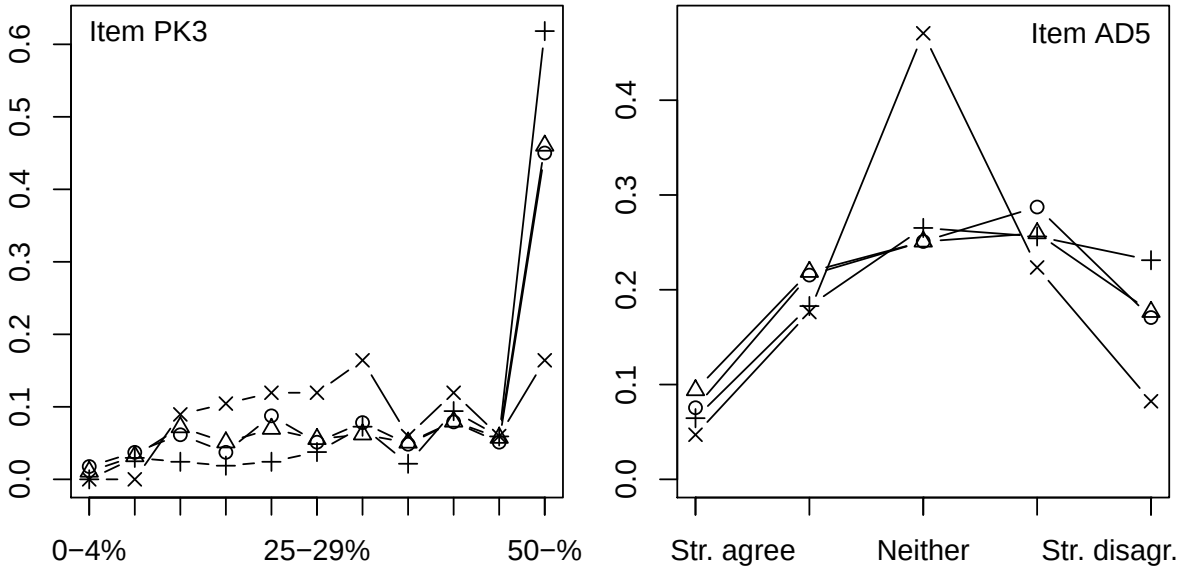
NOTE: The first three rows show percentages of non-DK responses. 'No response' shows the combined percentage in the probing group of final DK responses and the very small number who gave an immediate refusal and were thus not probed. 'Previously probed' shows the percentage of respondents in the probing group who responded to an item immediately but who had been probed for any item before it in the study. The percentages are calculated for the combined sample of 4470 respondents from three countries.

once. There is then scope for the probing to have an effect.

Probing did reduce levels of item nonresponse markedly. Around half of those in the treatment group who initially answered DK gave a substantive response after being probed. This pattern is observed for both pseudo-knowledge and attitude items. Overall, around 690 answers — 2.6% of the total responses for these eight items in the treatment group — were obtained after probing.

Comparing response rates between the probing group and the control group, who were never probed, allows us to look for evidence of a *carry-over* effect, that is whether probing influences the propensity to respond DK to subsequent items. If carry-over was present, we would expect response rates in the control group to diverge from those of immediate responders in the treatment group as the interview progresses and the treatment group adapt their behaviour after having been previously probed. In Table 2, the 'Previously probed' row represents respondents who could be subject to this effect. It shows the percentages of respondents in the treatment group who gave an immediate answer to an item but who had been probed for at least one previous item. For *AD5*, for example, this proportion is 16.7%, representing 17.6% (16.7/95.1) of the immediate responses to this item. However, there is no evidence of carry-over here: the proportions of respondents who gave immediate responses were not significantly different between the control and treatment groups for any item in the study.

Figure 1: Sample distributions of responses to one pseudo-knowledge item (PK3) and one attitude item (AD5) in the probing study. The horizontal axis of each plot spans the response options of the item, and the vertical axis their proportions in the data. The distributions are shown separately for the control group (∘) and, in the probing group, for immediate responses for respondents who had not (△; 'un-probed') and had (+; 'pre-probed') been probed for any previous item, and for probed responses (×).



Our central concern in this paper is that, in addition to influencing the response rate to an item, probing may also affect the nature of responses given. As an initial assessment of this, Figure 1 shows the observed distributions of responses to items *PK3* and *AD5*, i.e. the last items of the two batteries (the distributions of the other items are given in the supplementary materials). They are shown separately for four different types of response: (i) immediate responses in the control group, (ii) immediate responses from members of the treatment group with no previous probing experience ('un-probed'), (iii) immediate responses from members of the treatment group who have been probed on previous items ('pre-probed'), and (iv) responses obtained only after probing on that item ('probed').

There is very little difference between the control and un-probed responses. This is as we would expect given the randomization of respondents to the control or treatment groups. There is also relatively little difference between the un-probed and pre-probed responses, suggesting that there is no marked carry-over effect of probing on response patterns. This is the case even for the last item *AD5* where the numbers of pre-probed respondents, and thus the scope for carry-over effects, is greatest.

The main point of comparison is between the probed responses and the other three types of response. For every item, there are clear differences here. For example, for *PK3* the probed group is less likely to select categories at the extreme ends of the response scale. The same is true for the attitude item *AD5*, for which probed respondents are less likely to say 'strongly agree' or 'strongly disagree', and much more likely to choose the middle 'neither agree nor disagree' category. This middle option is the modal response for the probed group across all five of the *AD* items.

There are, broadly, two possible and opposite interpretations for these observed differences between responses obtained with and without probing, in particular the greater tendency for probed respondents to give answers nearer the middle of the response scale. On the one hand, this could be taken as evidence that probing pressures individuals into giving an unconsidered response, reflecting a greater degree of equivocation or guessing. This would be a measurement effect of probing on the answers given. If this was the case, probing would simply be serving to introduce measurement error rather than providing meaningful additional information. On the other hand, the observed differences may also be explained by the claim that probed respondents are simply expressing genuinely-held attitudes that are different (i.e. more 'moderate') from those of respondents who did not require probing. This would be a selection effect of which kinds of respondents are more likely to require (and respond to) probing. In this case the probed responses would provide equally useful data as immediate responses, and probing would in fact help to rectify the underrepresentation of such moderate attitudes among the immediate responders.

On the basis of these univariate results alone, it is not possible to distinguish between these two interpretations. To separate them, we must make use of the multivariate nature of the data — in essence, to look also at what answers a respondent gives to other items in the same battery. Table 3 shows descriptive statistics of this kind. It gives the sample correlations between pairs of the items, separately by how the answers were obtained. For example, among those respondents who gave an unprobed response to both items, the correlation between *PK1* and *PK2* was 0.44 in both the control and probing groups, but it was 0.42 when both responses were given after probing and 0.30 when *PK1* was unprobed and *PK2* was probed. If probed and unprobed responses behaved in exactly the same way, we would expect all of these correlations to be similar. This is not the case here.

Table 3: Sample correlations between pairs of items in the probing study, separately for observations with different combinations of probed and unprobed answers to the items.

	Group:				
	Control:	Probing:			
Items	Unprobed – Unprobed	Unprobed – Unprobed	Probed – Probed	Unprobed – Probed	Probed – Unprobed
Pseudo-knowledge items:					
PK1-PK2	.44	.44	.42	.30	.39
PK1-PK3	.49	.53	.25	.37	.32
PK2-PK3	.37	.37	.38	.26	.27
Attitude items:					
AD1-AD3	.24	.23	−.24	.29	.02
AD1-AD5	.40	.40	.40	.27	−.23
AD3-AD5	.21	.22	.17	.10	.31
AD1-AD2	−.04	.01	−.35	.01	−.07
AD1-AD4	−.11	−.04	−.58	−.17	.12
AD2-AD5	.03	.00	.57	−.09	.24
AD4-AD5	−.01	−.07	.44	.02	.29
AD2-AD3	.20	.24	.35	.37	.04
AD2-AD4	.41	.42	.71	−.44	.03
AD3-AD4	.12	.17	.37	−.45	.33

NOTE: The pairs of attitude items which are not shown in a bold font are ones where the latent variable measurement model that we use later would imply a low correlation.

For the pseudo-knowledge items, correlations which involve probed responses are mostly somewhat smaller. Larger differences are observed for the attitude items, where some correlations involving probed answers are substantially smaller or larger (and some with different signs) than the ones between unprobed answers. This suggests that at least for these items probing may affect the survey responses.

From these results it is still not possible, however, to draw firm conclusions about possible measurement effects. The correlations in Table 3 may also be affected by selection effects between probed and unprobed respondents. The sample sizes for the correlations involving probed responses are also fairly small (between 9 and 171; they are listed in the supplementary materials), so the majority of the differences within the rows of Table 3 are not individually significant. To remedy these difficulties and to obtain a more parsimonious

and powerful analysis, we need to analyse all the items and types of response in a battery of questions together, using a model for their joint distribution. A general modelling framework for doing this is introduced next.

3 A GENERAL MODEL FOR SURVEY RESPONSES WITH PROBING

3.1 The variables and the joint model

We will first specify models for an individual respondent, so a respondent subscript is omitted from the notation for now. We shall later, in Section 3.4, assume independence between individuals and construct a joint model from the product of the individual-level models developed here.

Let $\mathbf{Y}_C = (Y_1, \ldots, Y_p)'$ denote the values of p survey questions ('items') for a respondent, with the item subscript $j = 1, \ldots, p$ indicating also the order in which the questions were asked. The items are regarded as measures of $q \geq 1$ latent variables $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)'$. Also observed are covariates $\mathbf{X} = (X_1, \ldots, X_r)'$ which are treated as potential predictors of $\boldsymbol{\eta}$ and of the indicators $\mathbf{M}$ defined below. We assume that $\mathbf{X}$ is fully observed.

Suppose that probing of initial nonresponse was used for at least some respondents. Let $\mathbf{M} = (M_1, \ldots, M_p)'$ indicate whether and when a respondent answered the items, such that

$$M_j = \begin{cases} 1 & \text{if } Y_j \text{ was immediately recorded, without probing,} \\ 2 & \text{if } Y_j \text{ was recorded after probing for it,} \\ 3 & \text{if } Y_j \text{ is missing even after probing for it, and} \\ 4 & \text{if } Y_j \text{ is missing and the respondent was not probed for it,} \end{cases}$$

for each item $j = 1, \ldots, p$. In our probing experiment there were two experimental groups, identified here with an indicator variable T. In the control group ($T = 0$), probing was never used, so all 'Don't know' (DK) responses and refusals for item j are coded as $M_j = 4$. In the treatment (probing) group ($T = 1$), initial DK responses were probed and thus lead to $M_j = 2$ or $M_j = 3$, while immediate refusals were left unprobed and coded as $M_j = 4$.

Let $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ denote the vectors of the observed and missing elements of $\mathbf{Y}_C$ respectively. The observed distribution which carries the information on the model parameters is

$$p(\mathbf{Y}, \mathbf{M}|\mathbf{X}) = \int\int p(\mathbf{Y}_C, \mathbf{M}, \boldsymbol{\eta}|\mathbf{X})\, d\boldsymbol{\eta}\, d\mathbf{Y}_{mis} \tag{1}$$

where $p(\cdot|\cdot)$ denotes a conditional probability or density function, and the integrals are over the ranges of the corresponding variables. In our probing study, this distribution can in principle depend also on the experimental group T. We omit T from the notation, however, justifying this with the randomization and other arguments which are discussed below.

We decompose the distribution under the integral in (1) as

$$p(\mathbf{Y}_C, \mathbf{M}, \boldsymbol{\eta}|\mathbf{X}) = p(\boldsymbol{\eta}|\mathbf{X})\, p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X})\, p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X}), \tag{2}$$

and refer to the three distributions on the right-hand side of (2) as the *structural model*, *response model* and *measurement model* respectively. The structural model is typically of most interest in substantive survey analysis. Here, however, our focus is on methodological research questions about the effects of probing, as discussed in Sections 1 and 2. These can be formulated as questions about the response and measurement models. In Section 3.2 below we define the specific versions of the three submodels that we will consider, and then state in Section 3.3 how these models relate to the questions of interest. In that discussion we will also refer to Figure 2 which represents the joint model as a path diagram.

3.2 Specification of the three submodels

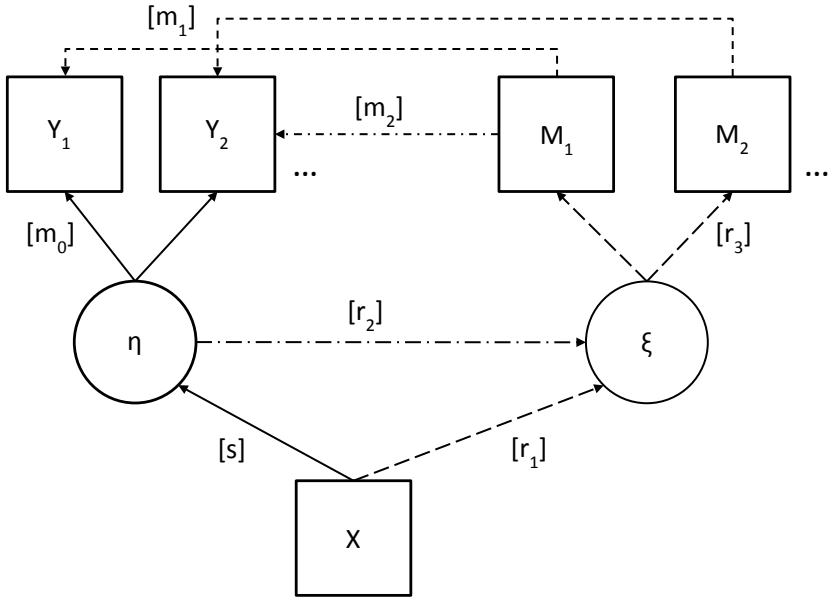
3.2.1 The structural model

The latent variables $\boldsymbol{\eta}$ represent the constructs which a survey aims to measure, and the structural model $p(\boldsymbol{\eta}|\mathbf{X})$ specifies how their distribution depends on observed covariates $\mathbf{X}$. In general, $\boldsymbol{\eta}$ may be taken to be continuous or categorical variables. In our analyses they will be continuous and normally distributed, and the structural model is the linear model

$$\boldsymbol{\eta} \sim N(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\mathbf{X}, \boldsymbol{\Sigma}) \tag{3}$$

with parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Sigma}$. In the probing experiment, $\boldsymbol{\eta}$ should be conditionally independent of the experimental group T given $\mathbf{X}$ because of the randomization, justifying

Figure 2: Path diagram for the model for multi-item survey responses with probing which is defined in Section 3. For simplicity of presentation, this shows only one attitude latent variable (η) and one response propensity latent variable (ξ), and only two items (Y_1, Y_2) with their response indicators (M_1, M_2).



the omission of T from $p(\boldsymbol{\eta}|\mathbf{X})$. This and other conditional independencies implied by the randomization were also checked in the analysis of the data, and found to be satisfied.

3.2.2 The response model

The model $p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X})$ specifies the probabilities of different patterns $\mathbf{M} = (M_1, \ldots, M_p)'$ of responding, i.e. whether a respondent answers the items immediately, after probing, or not at all. This model should allow a sufficiently flexible specification of the associations between the M_j for different items j. We obtain this by introducing a second set of latent variables $\boldsymbol{\xi}$ which induce the joint distribution of $\mathbf{M}$. Specifically, we consider models where $\boldsymbol{\xi} = \xi$ is a univariate categorical variable with K categories $\xi = 1, \ldots, K$, leading to *latent class models* for $\mathbf{M}$. With this device, the model is specified as $p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X}) = \sum_{k=1}^{K} p(\mathbf{M}|\xi = k, \boldsymbol{\eta}, \mathbf{X}) p(\xi = k|\boldsymbol{\eta}, \mathbf{X})$. We further assume that, given ξ, the M_j are conditionally independent of each other and of $(\boldsymbol{\eta}, \mathbf{X})$, so that $p(\mathbf{M}|\xi = k, \boldsymbol{\eta}, \mathbf{X}) = \prod_{j=1}^{p} p(M_j|\xi = k)$.

For the implementation of the estimation it will be convenient to re-code the information in M_j in two binary variables R_{1j} and R_{2j}. Here R_{1j} is an indicator of an immediate

(unprobed) response, which is 1 if $M_j = 1$ and 0 otherwise, while R_{2j} is an indicator of a response after probing, i.e. equal to 1 if $M_j = 2$, 0 if $M_j = 3$ and missing (undefined) otherwise. Denoting a missing value by NA, the values of (R_{1j}, R_{2j}) which correspond to $M_j = 1, 2, 3, 4$ are thus $(1, NA)$, $(0, 1)$, $(0, 0)$ and $(0, NA)$ respectively. We can then write

$$\prod_{j=1}^{p} p(M_j|\xi = k) = \left[\prod_{j=1}^{p} p(R_{1j}|\xi = k)\right] \left[\prod_{j\in\mathcal{R}_2} p(R_{2j}|\xi = k)\right] \tag{4}$$

where $\mathcal{R}_2$ denotes the indices of the items for which R_{2j} is not missing. Here R_{1j} exist for all items and respondents, and because of randomization their distribution conditional on ξ should be the same in both the treatment ($T = 1$) and control groups ($T = 0$) of our experiment. All respondents in both groups thus contribute information about the whole of the product term for R_{1j} in (4). On the other hand, only instances of probing give information about the last product term in (4), and in particular the whole term is omitted in the control group. This also implies that here $p(R_{2j}|\xi)$ really means only $p(R_{2j}|R_{1j} = 0, \xi, T = 1)$, since R_{2j} does not exist for other combinations of R_{1j} and T.

With these choices, the response model is specified as

$$p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X}) = \sum_{k=1}^{K} \left[\prod_{j=1}^{p} p(R_{1j}|\xi = k)\right] \left[\prod_{j\in\mathcal{R}_2} p(R_{2j}|\xi = k)\right] p(\xi = k|\boldsymbol{\eta}, \mathbf{X}). \tag{5}$$

Larger values of K, the number of latent classes, yield more flexible models for the conditional association structure of $\mathbf{M}$. The parameters to be estimated are the probabilities $p(R_{1j} = 1|\xi = k)$ and $p(R_{2j} = 1|\xi = k)$ for $j = 1, \ldots, p$ and $k = 1, \ldots, K$, plus the parameters of the model for $p(\xi|\boldsymbol{\eta}, \mathbf{X})$, which we specify as a multinomial logistic model.

3.2.3 The measurement model

The model $p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X})$ specifies how a respondent's answers $\mathbf{Y}_C$ to the survey items are associated with the latent variables $\boldsymbol{\eta}$ that the items aim to measure, and how these measurement properties may also depend on the types of response $\mathbf{M}$ and the covariates $\mathbf{X}$. We will throughout make the common assumption that responses to different items are conditionally independent, i.e. that $p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X}) = \prod_{j=1}^{p} p(Y_j|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X})$. The joint measurement model is then specified by the univariate models for the individual items.

The vector $\mathbf{M}$ carries the information about when a respondent is probed during the survey. We will assume that the measurement model of an item may depend on this 'probing history' for that item and those that were asked before it, but not on probing for later items. This still leaves a potentially large number of distinct probing histories, many of which may be rarely observed. To simplify the models, we group the histories for each item j into three types: (1) 'unprobed' responses where the respondent has not been probed for that item or any before it ($M_j = 1$ and $M_k = 1$ or 4 for all $k < j$), (2) 'pre-probed' responses where the item is not probed but the respondent has received at least one probe before ($M_j = 1$ and $M_k = 2$ or 3 for at least one $k < j$), and (3) 'probed' responses where the item itself is answered after probing ($M_j = 2$). This the same grouping that we used earlier, in Figure 1. Here 'before' includes all items which are asked in the survey before item j, including ones which are not intended as measures of the same latent variables.

After this grouping, we also modify the notation to write the measurement models as $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X})$, where $h = 1, 2, 3$ indicate the models for an unprobed, pre-probed and probed response respectively. This notation gives the impression of a situation where there are three distinct versions of each item ($Y_j^{(1)}, Y_j^{(2)}$ and $Y_j^{(3)}$), of which a respondent answers at most one. This notion turns out to be a helpful one, both for describing hypotheses about the measurement effects of probing and for implementation of the estimation of the models. The measurement model can then be expressed as

$$p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X}) = \prod_{j=1}^{p} \prod_{h=1,2,3} p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X}) \tag{6}$$

with the convention that $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X})$ for the value of h which corresponds to the respondent's value of $\mathbf{M}$, and $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X}) = 1$ otherwise. If Y_j is missing, all three terms for it are set to 1 and thus omitted from (6). This achieves the integration over missing responses $\mathbf{Y}_{mis}$ in (1), so (6) is also the measurement model for the observed $\mathbf{Y}$. We again omit the experimental group T from this notation, because $p(Y_j^{(1)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(1)}|\boldsymbol{\eta}, \mathbf{X}, T)$ due to randomization, and $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X})$ for $h = 2, 3$ really mean only $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X}, T = 1)$.

The specification of the univariate measurement models $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X})$ can employ conventional forms for latent variable models, with the specific choice depending on the types of the items. In our analyses in Section 4 we use two kinds of measurement models. The first,

which is used for continuous items, is the linear (factor analysis) measurement model

$$Y_j^{(h)} \sim N(\tau_j^{(h)} + \boldsymbol{\lambda}_{xj}^{(h)\prime}\mathbf{X} + \boldsymbol{\lambda}_j^{(h)}(\mathbf{X})'\boldsymbol{\eta},\, \delta_j^{(h)}) \tag{7}$$

for each combination of $j = 1, \ldots, p$ and $h = 1, 2, 3$. The second is used when $Y_j^{(h)}$ is treated as a categorical variable with L_j ordered categories. The measurement model is then an ordinal logistic model where $Y_j^{(h)}$ is multinomially distributed with

$$\log\left[\frac{p(Y_j^{(h)} \leq l|\boldsymbol{\eta}, \mathbf{X})}{p(Y_j^{(h)} > l|\boldsymbol{\eta}, \mathbf{X})}\right] = \tau_{jl}^{(h)} - \boldsymbol{\lambda}_{xj}^{(h)\prime}\mathbf{X} - \boldsymbol{\lambda}_j^{(h)}(\mathbf{X})'\boldsymbol{\eta} \tag{8}$$

for $j = 1, \ldots, p$, $h = 1, 2, 3$ and $l = 1, \ldots, L_j - 1$. Here $\tau_j^{(h)}$, $\tau_{jl}^{(h)}$ $\boldsymbol{\lambda}_{xj}^{(h)}$, $\boldsymbol{\lambda}_j^{(h)}(\mathbf{X})$ and $\delta_j^{(h)}$ are model parameters, and the notation $\boldsymbol{\lambda}_j^{(h)}(\mathbf{X})$ indicates an interaction where the coefficients of $\boldsymbol{\eta}$ may depend on $\mathbf{X}$. In most of our analyses we will in fact consider only measurement models which do not depend on $\mathbf{X}$ at all, so will set $\boldsymbol{\lambda}_{xj}^{(h)} = \mathbf{0}$ and $\boldsymbol{\lambda}_j^{(h)}(\mathbf{X}) = \boldsymbol{\lambda}_j^{(h)}$.

When the latent $\boldsymbol{\eta}$ have a continuous (here normal) distribution, some constraints on the parameters of the structural and/or measurement models are needed to identify the scale of $\boldsymbol{\eta}$ and the remaining model parameters. We do this by fixing in (3) the intercepts $\boldsymbol{\beta}_0 = \mathbf{0}$ and the residual variances (diagonal elements of $\boldsymbol{\Sigma}$) at 1.

3.3 How does the model represent the research questions?

As defined in the previous section, we specify the joint model (1)–(2) with the structural model $p(\boldsymbol{\eta}|\mathbf{X})$ given by the linear model (3), the response model $p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X})$ by the latent class model (5), and the measurement model $p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X})$ by (6) with the item-specific measurement models given by models such as (7) or (8). In this section we discuss how different parts of this correspond to different quantities of interest, and use them to define selection and measurement effects and our research questions. The joint model is represented graphically in Figure 2, with parts of this diagram labelled with letters which are also referred to in the discussion below.

The left-hand side of Figure 2 represents a latent variable model where all the item responses are of one (unprobed) type. In the diagram, $[s]$ stands for the structural model and $[m_0]$ for the measurement model for the responses. Standard survey analysis would include

only these models. They can be used to address the kinds of substantive questions which motivate survey research in the first place, especially questions about the latent attitude variables and how they depend on covariates. Here, however, our focus is on methodological research questions about the effects of probing on the survey responses. These are captured by the other parts of the model.

3.3.1 Response model and selection effects

Our specification of $p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X})$ with the aid of an additional latent variables $\boldsymbol{\xi}$ is inspired by the use of similar approaches for the modelling of item nonresponse in multi-item survey questions (see e.g. Knott et al. 1990, O'Muircheartaigh and Moustaki 1999, and Holman and Glas 2005). There each response indicator M_j would have only two possible values (response or no response), while here we add a third one for a response given after probing.

Here $\boldsymbol{\xi}$ is primarily a convenient device for inducing a flexible family of joint distributions for $\mathbf{M}$, and need not have any other purpose. We may, however, also choose to give a substantive interpretation for $\boldsymbol{\xi}$, as characteristics of the survey respondents which are related to their responding behaviour. The levels of our scalar categorical ξ could then be thought of as *latent response classes* of the individuals.

The response model is represented by the right-hand side of Figure 2. The arrows labelled $[r_3]$ indicate the models for the probabilities of different response types M_j (i.e. R_{1j} and R_{2j} in (5)) given the response classes of ξ. The arrows $[r_1]$ and $[r_2]$ indicate the dependence of the response classes on the latent attitude variables $\boldsymbol{\eta}$ and the covariates $\mathbf{X}$ respectively. These can be used to examine questions such as whether the probabilities of an immediate answer, or of giving an answer if prompted, depend on respondent characteristics.

Consider first a model with no covariates, and the conditional distribution $p(\xi|\boldsymbol{\eta})$ of ξ given $\boldsymbol{\eta}$ only. We say there is a *selection effect* in responses to probing if $p(\xi|\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$, i.e. if there is an association between the latent attitudes $\boldsymbol{\eta}$ and the response propensities ξ. Next, if covariates $\mathbf{X}$ are included, we can write $p(\xi|\boldsymbol{\eta}) = \int p(\boldsymbol{\eta}|\mathbf{X})p(\xi|\boldsymbol{\eta}, \mathbf{X})p(\mathbf{X})\, d\mathbf{X}$, where $p(\mathbf{X})$ denotes the distribution of the covariates. This shows that a selection effect can arise in two ways: if $p(\xi|\boldsymbol{\eta}, \mathbf{X})$ depends on $\boldsymbol{\eta}$ even given $\mathbf{X}$, or if $p(\xi|\boldsymbol{\eta}, \mathbf{X}) = p(\xi|\mathbf{X})$ but some covariates $\mathbf{X}$ have non-zero effects in both $p(\boldsymbol{\eta}|\mathbf{X})$ and $p(\xi|\mathbf{X})$. In the latter case (where

$[r_2]$ in Figure 2 is absent), the joint model (1) simplifies to

$$p(\mathbf{Y}, \mathbf{M}|\mathbf{X}) = p(\mathbf{M}|\mathbf{X}) \times p(\mathbf{Y}|\mathbf{M}, \mathbf{X}) = p(\mathbf{M}|\mathbf{X}) \times \iint p(\boldsymbol{\eta}|\mathbf{X})\, p(\mathbf{Y}_C|\mathbf{M}, \boldsymbol{\eta}, \mathbf{X})\; d\boldsymbol{\eta}\, d\mathbf{Y}_{mis} \quad (9)$$

where $p(\mathbf{M}|\mathbf{X}) = p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X})$ is given by (5) with $\boldsymbol{\eta}$ omitted. Likelihood-based estimation and inference can then be done separately for $p(\mathbf{M}|\mathbf{X})$ and for the rest of the model (if their parameters are also fully distinct). A response model of this kind is *ignorable* for likelihood inference, in the same sense in which this term is used for models for missing data (see e.g. Little and Rubin 2002; the condition $p(\mathbf{M}|\boldsymbol{\eta}, \mathbf{X}) = p(\mathbf{M}|\mathbf{X})$ is analogous to Missingness at Random). If, on the other hand, $p(\xi|\boldsymbol{\eta}, \mathbf{X})$ depends on $\boldsymbol{\eta}$, the response model is *non-ignorable* and needs to be included in the estimation of the joint model (1)–(2) in order to get valid estimates for the whole model.

Both ignorable and non-ignorable response models are identifiable within our model for multiple items (whereas neither of them could be identified, separately from models with measurement effects, from analyses of one item at a time). When the goal is to examine measurement effects of probing, we may in fact be indifferent to the exact form of the response model we use, as long as it allows for selection effects in an adequate way. Response and structural models which include sufficient shared covariates to render the responding ignorable are computationally convenient, because then $\mathbf{M}$ functions simply as additional covariates in the measurement model and estimation of $p(\mathbf{M}|\mathbf{X})$ can even be omitted if it is of no separate interest. However, we could also use a non-ignorable model with fewer covariates, including $p(\xi|\boldsymbol{\eta})$ with none of them, since this also allows for the association between $\boldsymbol{\eta}$ and ξ. In the analysis of the data in Section 4 we will use and compare both of these possibilities.

3.3.2 Measurement model and measurement effects

Once possible selection effects are allowed for through the response model, the measurement model can be used to examine if the measurement characteristics of the survey items are affected by probing of DK answers. We say that there is a *measurement effect* of probing on an item Y_j if the measurement model of the item depends on the respondent's probing history $\mathbf{M}$. In Figure 2 these effects are represented by $[m_1]$, indicating the effect of probing on an item itself, and $[m_2]$, indicating carry-over effects that probing on previous items

may have on responses to a later item. We operationalise these effects by comparing the unprobed ($Y_j^{(1)}$), pre-probed ($Y_j^{(2)}$) and probed ($Y_j^{(3)}$) responses to each item, as defined in Section 3.2.3, and test the following hypotheses:

$$\textit{no direct probing effect}: \quad p(Y_j^{(1)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(3)}|\boldsymbol{\eta}, \mathbf{X}); \tag{10}$$

$$\textit{no carryover}: \quad p(Y_j^{(1)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(2)}|\boldsymbol{\eta}, \mathbf{X}). \tag{11}$$

If both of these are true, there is no measurement effect on the item (to the extent captured by these possibilities), and if this is true for all of the items, probing has no measurement effect of the whole scale $\mathbf{Y}_C$. A measurement effect thus exists if, for at least one item, the relationship between an item response and the latent $\boldsymbol{\eta}$ that the item is designed to measure depends also on how the response was obtained. This is the case if probed responses are poorer measures of (i.e. have a weaker association with) $\boldsymbol{\eta}$ than are unprobed responses, but also if probed answers are better or just behave differently as measures of $\boldsymbol{\eta}$.

An extreme version of a measurement effect occurs if probing yields only random responses which are not related to the latent constructs at all, i.e.

$$\textit{probed answers do not measure } \boldsymbol{\eta}: \quad p(Y_j^{(3)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(3)}|\mathbf{X}); \tag{12}$$

$$\textit{pre-probed answers do not measure } \boldsymbol{\eta}: \quad p(Y_j^{(2)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(2)}|\mathbf{X}). \tag{13}$$

A measurement model $p(Y_j^{(h)}|\boldsymbol{\eta}, \mathbf{X})$ may also depend on covariates $\mathbf{X}$. This would mean *non-equivalence of measurement* where the answers to a survey item are affected not only by the latent attitudes $\boldsymbol{\eta}$ but also by other characteristics of the respondent. This possibility is not represented in Figure 2, and we will consider it only briefly. We assume throughout that $p(Y_j^{(1)}|\boldsymbol{\eta}, \mathbf{X}) = p(Y_j^{(1)}|\boldsymbol{\eta})$ for all j, i.e. that there is full equivalence in the unprobed responses (this assumption is separate from the questions on probing considered here). We will, however, briefly examine the possibility of non-equivalence in the probed and pre-probed responses. It would be present, for example, if probing had a different measurement effect for respondents with different levels of education. Such differences can be identified within the general model, although the amount of information available for estimating them is likely to be limited in typical probing data.

3.3.3 The role of the control group

In the ESS probing study, 25% of the respondents were randomly assigned to a control group and never probed. The main purpose of including this group was to examine the possibility of carry-over in response rates, that is whether some respondents would be more likely to give an immediate response after they had received a probe earlier in the survey. If this was the case, we would have expected to see higher immediate response rates in the probing group than in the control group for the items which were asked later in the sequence. As seen in Table 2, however, there was no evidence of this kind of carry-over.

The control group was not essential for the analyses with the general model in Figure 2, in that the model is identified with the treatment group only. Data from the control group were still included in the estimation. They contribute information on the parts of the models which are not directly related to the probing, i.e. all but the models for R_{2j} in (4) and for $Y_j^{(2)}$ and $Y_j^{(3)}$ in (6). They also allowed us to test and confirm the hypotheses implied by randomization, e.g. that $p(\boldsymbol{\eta}|\mathbf{X})$ and $p(Y_j^{(1)}|\boldsymbol{\eta})$ were the same between the treatment and control groups. This provided additional reassurance that the estimates of those parts of the model were not affected by the probing in the treatment group.

3.3.4 Comparable models in related literature

As discussed in Section 1, latent variable models for multiple items which are partially analogous to our models have been applied to two other questions in survey methodology. Here we return to outline these connections, now that our models have been defined above. The two related areas are comparisons of respondents with different levels of propensity to *unit* nonresponse (Hox et al. 2012; Medway 2012), and of different modes of data collection (a larger literature, starting with de Leeuw 1992; an excellent review is given by Hox et al. 2015, and additional references by Klausch et al. 2013). The same type of modelling can be used in both of these contexts. The aim is to compare two or more groups, defined by a measure of response propensity in the first case, and by survey mode in the second. A latent variable model is then specified in such a way that both the distributions of the latent variables and the measurement model of the items given the latent variables may depend also on the group. Models like this are known as *multigroup* latent variable models.

This research has commonly employed multigroup factor analysis (Jöreskog 1971; see Hox et al. 2015 for a summary of how it is used in this context), but models for categorical items could also be used (for current overviews of multigroup modelling, see e.g. Kankaraš et al. 2011a, 2011b).

There is said to be a measurement effect of the group if the parameters of the measurement model for some or all of the items depend on the group. For example, in a mode comparison study this would mean that some survey questions work differently depending on which mode was used to ask them. These effects can be examined by comparing models fitted with and without such group differences. At the same time, we should also allow for selection effects, i.e. that respondents who, say, choose to answer in one mode may be systematically different from those who answer in another mode. Such effects may be assumed to be absent if respondents are assigned to the groups at random, but otherwise they should be allowed for. This can be done in multigroup latent variable modelling by letting the distributions of the latent variables depend on the group and/or on other covariates. Both measurement and selection effects can then be estimated, although only under the identification condition that a sufficient number of measurement parameters do not depend on the group (i.e. that there are in fact no measurement effects for these parameters).

Our models also involve possibly different measurement models in different circumstances. In contrast to multigroup modelling, however, in our analysis the 'group' is the probing status of an individual item. This is a characteristic of an item because the same respondent may answer one item unprobed and another only after probing. Our analysis thus draws on variation in types of response both within and between respondents. A by-product of this situation is that in these models it is possible to allow for measurement effects for all the items, without further identification constraints on the measurement parameters.

3.4 Estimation of the joint model

Let $\boldsymbol{\theta}$ denote all the parameters of the model (1)–(2). We have data on n respondents, assumed to be independent. Adding a respondent subscript i, the log-likelihood for $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\boldsymbol{\eta}|\mathbf{X}_i)\, p(\mathbf{M}_i|\boldsymbol{\eta}, \mathbf{X}_i)\, p(\mathbf{Y}_i|\mathbf{M}_i, \boldsymbol{\eta}, \mathbf{X}_i)\, d\boldsymbol{\eta}. \tag{14}$$

Maximum likelihood estimates of the parameters are obtained by maximizing $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In our analyses, this was done using the Mplus software (Muthén and Muthén, 2010). Our research questions correspond to constraints on elements of $\boldsymbol{\theta}$, and these can be tested with standard likelihood ratio tests of nested models.

In survey terms, the use of the log-likelihood (14) effectively assumes that the sampling scheme can be ignored. More generally, it could be replaced with a pseudo log-likelihood to allow for survey weights and complex sampling designs in estimation of standard errors, and in modifications to the likelihood ratio tests (see e.g. Skinner et al. 1989 for an overview of such methods). This was not done in the analysis of our probing study; survey weights were omitted as discussed in Section 2, and other features of the sampling designs were also ignored.

4 ANALYSIS OF THE PROBING DATA

4.1 The pseudo-knowledge items

The models defined in Section 3 were used to analyse the data from the probing study introduced in Section 2. The three pseudo-knowledge items (*PK1–PK3* in Table 1) are considered first, followed by the five attitude items in Section 4.2. In both cases the covariates $\mathbf{X}$ are age (in years), sex, and education (as highest level obtained, coded as Primary or less, Secondary, or University/College education). As discussed in Section 2, we will mostly use the data for all three countries combined. This means that the models are estimated under the assumption of measurement equivalence across the countries. The conclusions about the effects of probing are mostly unaffected if this assumption is relaxed by carrying out the analyses separately for each country. These analyses are discussed briefly for both types of items at the end of Section 4.2.

We treat the responses to the pseudo-knowledge items as continuous and use the linear measurement model (7). The continuous latent variable $\boldsymbol{\eta} = \eta$ is taken to be univariate (this produces a saturated measurement model for the three items). An individual with a large value of η will tend to guess that large proportions of people in his or her country are unemployed, long-term sick and lacking money for basic necessities.

We first took a simplifying preliminary step to improve the numerical stability of the joint modelling (the same was done in the analyses of the attitude items). This involved fitting the response model (5) first separately and with the latent class variable ξ alone, without conditioning on η and $\mathbf{X}$. The fitted probabilities $p(R_{1j}|\xi)$ and $p(R_{2j}|\xi)$ were then taken from this model and fixed at these values in subsequent estimation of the rest of the joint model. With the three items, the largest number of latent classes which can be considered is $K = 3$. This reproduces the observed distribution of the response indicators essentially perfectly and much better (according to both AIC and BIC model assessment statistics) than a two-class model, so the three-class model is used in the following analyses.

Next, the first question of the main analysis was whether there was a selection effect, as defined in Section 3.3.1. Here there is evidence of such an effect: when the joint model (1)–(2) was fitted without covariates, the hypothesis that ξ and η are independent was rejected, with $p < 0.001$ for the likelihood ratio test. The direction of this association is such that large values of η are associated with higher probabilities to be a nonresponder. When, however, the covariates $\mathbf{X}$ were included, the hypothesis that $p(\xi|\eta, \mathbf{X}) = p(\xi|\mathbf{X})$ was not rejected ($p = 0.81$). This means that there is a selection effect of which respondents tend to require probing, but that this effect is accounted for by the covariates (especially education, which is the strongest predictor of both η and ξ). The joint distribution then simplifies to the ignorable form (9), where the selection effect is allowed for by controlling for $\mathbf{X}$, and the model $p(\mathbf{Y}|\mathbf{M}, \mathbf{X})$ for the answers can be estimated separately from the response model $p(\mathbf{M}|\mathbf{X})$. This is what we do in the analyses presented below. Towards the end of this section, however, we will briefly return to the non-ignorable model without covariates, to assess whether this affects our main conclusions about the measurement effects.

Table 4 shows the estimated response model. Here the joint distribution of the response indicators $\mathbf{M}$ is a mixture of distributions for three latent classes of individuals with different responding tendencies. The upper part of the table shows, for these three classes, the estimated response probabilities derived from the model, expressed as probabilities of an immediate response, response after probing, and final nonresponse (these apply to a respondent who may be probed after initial nonresponse; if not, the complement of the probability of immediate response represents final nonresponse). It can be seen that a respondent in the first class is highly likely to give an immediate response to each item,

in the second class essentially certain to give an immediate response or respond if probed, and in the third very likely to not respond even if probed. We label the three classes as the 'Ready responders', 'Reluctant responders' and 'Nonresponders' respectively.

The lower part of Table 4 summarises the estimates of $p(\xi|\mathbf{X})$, as fitted probabilities of the latent classes given two ages and the three levels of education for men (the probabilities for women are very similar and not significantly different). In all cases the majority of individuals are ready responders, but there is some variation in the probabilities given the covariates. In particular, older age is associated with significantly higher relative odds of being a reluctant rather than ready responder, and respondents who have at most primary education are significantly more likely to be reluctant or nonresponders than are respondents with higher levels of education. There are no significant covariate effects on being a reluctant responder vs. nonresponder, perhaps in part because of the much smaller number of observations which contribute information about this contrast.

We now turn to the combined structural and measurement model $p(\mathbf{Y}|\mathbf{M},\mathbf{X})$, considering for the moment only models where the measurement model does not depend on the covariates $\mathbf{X}$. With this and the other assumptions discussed above, the model reduces to

$$p(\mathbf{Y}|\mathbf{M},\mathbf{X}) = \int \left[\prod_{j=1}^{p}\prod_{h=1,2,3} p(Y_j^{(h)}|\eta)\right] p(\eta|\mathbf{X})\, d\eta \tag{15}$$

where $p(\eta|\mathbf{X}) \sim N(\beta_0 + \boldsymbol{\beta}_1'\mathbf{X}, \sigma^2)$ and $p(Y_j^{(h)}|\eta) \sim N(\tau_j^{(h)} + \lambda_j^{(h)}\eta, \delta_j^{(h)})$, obtained as special cases of (3) and (7); for identification, $\beta_0 = 1$ and $\sigma^2 = 1$ are fixed. This is a standard factor analysis model with covariates $\mathbf{X}$ and with different probing histories $\mathbf{M}$ incorporated by treating the unprobed, pre-probed and probed responses (i.e. $Y_j^{(1)}$, $Y_j^{(2)}$, and $Y_j^{(3)}$) for each item notionally as distinct items of which at most one is observed.

Table 5 shows the estimated model. Its lower part gives the structural model, which shows that younger respondents, women, and less highly educated respondents tend to perceive that higher proportions of people are in need of welfare benefits (these associations are significant). Our main interest, however, is in possible differences between the measurement models of the different types of responses, which are shown in the upper part of the table.

We note first that hypotheses (12) and (13), that the pre-probed or probed responses are entirely unassociated with the latent variable η (i.e. that the factor loadings $\lambda_2^{(2)} = \lambda_3^{(2)} = 0$

Table 4: A three-class latent class model for response patterns for the three pseudo-knowledge items in the probing study, conditional on respondent's age, sex and education.

		Responding class:		
		'Ready responders'	'Reluctant responders'	'Nonresponders'
Probabilities of different types of response:				
Response	Item			
Immediate response	*PK1*	.98	.44	.21
	PK2	.93	.23	.06
	PK3	.99	.61	.27
Probed response*	*PK1*	.02	.53	.05
	PK2	.04	.77	.00
	PK3	.01	.38	.00
No response*	*PK1*	.00	.03	.75
	PK2	.03	.00	.94
	PK3	.00	.01	.73
Probabilities of the responding classes, given education and selected ages†:				
Education	Age			
Primary or less	30	.91	.04	.05
	60	.87	.07	.06
Secondary	30	.97	.02	.02
	60	.95	.03	.02
College or university	30	.97	.01	.02
	60	.96	.02	.02

NOTE: $*$ = For respondents who may be probed. For those who are never probed, the sum of these probabilities is the probability of no response. † = For men; probabilities for women are not significantly different.

or $\lambda_1^{(3)} = \lambda_2^{(3)} = \lambda_3^{(3)} = 0$ respectively), are both rejected for all of the items. This suggests that these types of responses do carry information on the attitude that is being measured. However, their estimated factor loadings are for all items different from (and here smaller than) those of the unprobed responses. The estimated measurement intercepts are also smaller for probed responses, suggesting that, for any given true level of the attitude, probed answers tend to be lower than unprobed ones. To test these differences, we consider first hypothesis (10) that the unprobed and probed responses have the same measurement models. This is strongly rejected when it is applied to all three items jointly. There is thus evidence that the measurement properties of probed responses differ from those of unprobed

Table 5: Estimated parameters of a one-factor factor analysis model for the three pseudo-knowledge items, conditional on respondent's age, sex and education.

Measurement model:

	Intercept ($\tau_j^{(h)}$)			Factor loading ($\lambda_j^{(h)}$)			Residual s.d. ($\sqrt{\delta_j^{(h)}}$)		
Item:	Unpr.	Pre-pr.	Probed	Unpr.	Pre-pr.	Probed	Unpr.	Pre-pr.	Probed
PK1	8.47		8.14	2.34		1.62	1.85		2.24
PK2	5.29	5.14	4.49	1.54	0.83	1.15	2.33	2.74	2.26
PK3	9.12	9.67	7.84	1.98	1.63	1.17	2.33	1.82	2.29

Structural model — coefficients ($\boldsymbol{\beta}_1$) of covariates (with standard errors):

		Education	
Age	Sex: Woman	Secondary	University/College
−.003 (.001)	.23 (.04)	−.42 (.05)	−.70 (.06)

NOTE: For each item, all parameters of the measurement model are estimated separately for unprobed, pre-probed and probed responses. For *PK1*, none of the measurement parameters are significantly different between unprobed and probed responses (according to a likelihood ratio test, at 5% level of significance). For each of *PK2* and *PK3*, factor loadings of pre-pobed and probed responses are not significantly different. *PK1* is the first probed item in the survey, so no responses to it are pre-probed. The structural model for the latent factor is a linear model with intercept fixed at 0 and residual variance at 1.

responses for at least some of the items. Tests of individual items further indicate that (10) is not rejected for the first item (*PK1*) but is rejected for the other two items.

Hypothesis (11) claims that unprobed and pre-probed responses have the same measurement models. This is the hypothesis of no carry-over measurement effect from previous probing. It too is rejected for *PK2* and *PK3* (there are no pre-probed responses to the first item *PK1*). In fact, for these two items the factor loadings (but not other measurement parameters) of the pre-probed responses are not significantly different from those of the probed responses, although this may be largely due to lower power of the test, as these estimates are based on relatively little data.

These conclusions were derived from models where selection effects were accounted for by covariates in the structural and response models. As discussed earlier, this can also be achieved with a model which omits the covariates but includes a direct association between ξ and η. If this is done, both the patterns of the values of the measurement parameters and conclusions about hypotheses (10)–(13) about them are the same as the ones discussed above (with the one difference that the hypothesis of equivalence of probed and unprobed responses to *PK1* is now also rejected, with $p = 0.020$).

We also examined the possibility of non-equivalence of the measurement models with respect to covariates, i.e. that $\mathbf{X}$ may have a direct effect on Y_j (so that $\boldsymbol{\lambda}_{xj}^{(h)} \neq \mathbf{0}$ in (7)) or modify the association between Y_j and $(\mathbf{M}, \mathbf{X})$ (so that $\boldsymbol{\lambda}_j^{(h)}(\mathbf{X})$ depends on $\mathbf{X}$). We explored this for the probed and pre-probed responses. Such models are computationally demanding and the data do not provide much information for detecting non-equivalence of this kind. So we discuss these analyses only briefly, mainly to illustrate the meaning of such non-equivalence models. One difference which was statistically significant was that having university education had a positive direct effect on probed responses to item *PK3*. This means that for respondents with this level of education, the expected values of probed responses (given the latent η) were not smaller than those of unprobed and pre-probed responses, in the same way they are in the equivalence model (see Table 5). This difference might arise, for example, if the pattern that probed responses generally avoided very high answers did not occur in the same way for the most educated respondents.

4.2 The attitude items

The five attitude items (*AD1–AD5* in Table 1) were analysed in a broadly similar way. The latent response class model was again examined first. We again selected a three-class model, which was preferred by the BIC statistic (AIC would prefer at least five classes). The interpretation of the response classes was similar to those for the pseudo-knowledge items (estimates of the fitted model are not shown). For the items, a model with two correlated continuous latent factors $\boldsymbol{\eta} = (\eta_1, \eta_2)$ was used, after initial analysis indicated that a one-factor model clearly did not fit the data well. The measurement structure was such that items *AD1*, *AD3* and *AD5* were measures of (i.e. had non-zero loadings for) factor η_1, and *AD2*, *AD3* and *AD4* were measures of η_2. Here η_1 may interpreted as a measure of perceived abuse of the welfare system (with high values indicating a more positive attitude towards the beneficiaries in this respect), and η_2 as perceived sufficiency of welfare benefits (with high values indicating the view that current benefits are sufficient). The measurement model for each item was the ordinal logistic measurement model (8) with no effects of the covariates $\mathbf{X}$, i.e.

$$\log\left[\frac{p(Y_j^{(h)} \leq l|\boldsymbol{\eta})}{p(Y_j^{(h)} > l|\boldsymbol{\eta})}\right] = \tau_{jl}^{(h)} - \lambda_{j1}^{(h)}\eta_1 - \lambda_{j2}^{(h)}\eta_2 \tag{16}$$

Table 6: Estimated parameters for a latent variable model for the five attitude items, conditional on respondent's age, sex and education.

Measurement model:									
	$p(Y_j = 3 \mid \boldsymbol{\eta} = \mathbf{0})$			Loading of factor 1 ($\lambda_{j1}^{(h)}$)			Loading of factor 2 ($\lambda_{j2}^{(h)}$)		
Item:	Unpr.	Pre-pr.	Probed	Unpr.	Pre-pr.	Probed	Unpr.	Pre-pr.	Probed
AD1	.30	.36	.58	1.79	2.22	(0.19)			
AD5	.35	.34	.55	1.52	1.40	1.13			
AD3	.18	.19	.36	0.82	0.88	1.05	0.91	0.51	(0.76)
AD2	.11	.04	.52				1.95	(3.37)	(−0.92)
AD4	.05	.06	.53				1.86	1.33	(−1.93)

Structural model — coefficients of covariates (with standard errors):

			Education	
	Age	Sex: Woman	Secondary	University/College
Factor 1:	.000 (.001)	.11 (.04)	.10 (.05)	.23 (.07)
Factor 2:	.001 (.001)	−.06 (.04)	.17 (.05)	.16 (.07)
Covariance between the factors:		−.08 (.03)		

NOTE: The model has two latent factors, for which higher values correspond to higher probabilities of disagreement with the items. The measurement models are ordinal logistic models, with the parameters estimated separately for unprobed, pre-probed and probed responses. $p(Y_j = 3|\boldsymbol{\eta} = \mathbf{0})$ is the probability of responding 'Neither agree nor disagree' given both factors at 0. In the measurement model, the loadings in parentheses are not significantly different from 0 (at 5% level of significance). The structural model for the two latent factors is a bivariate linear model with intercepts fixed at 0 and residual variances at 1.

for $j = 1, \ldots, 5$, $h = 1, 2, 3$ and $l = 1, \ldots, 4$, with $\lambda_{jk}^{(h)} = 0$ when item j was not treated as a measure of factor η_k. A key reason for considering the ordinal model for these items was that it allows the probabilities of individual response categories to be modelled more flexibly than would be possible with a linear measurement model. In particular, we will later highlight estimated probabilities for the middle response category, which was previously seen to be relatively more common among probed responses.

There is little evidence of a selection effect for these items. The association between ξ and $\boldsymbol{\eta}$ was not significant when the covariates were omitted ($p = 0.43$) or when they were included ($p = 0.53$). The only covariate effect which was significant in both the response and structural models was a contrast between those with secondary education and the other two education levels, but this did not induce a strong association between ξ and $\boldsymbol{\eta}$. We will again first use models which include the covariates (but not $\boldsymbol{\eta}$ in the response model) to allow for what selection effect there is, and will consider some alternatives later.

The lower part of Table 6 shows the estimated structural model for the two factors measured by the attitude items. The factors are significantly (although weakly) negatively correlated, indicating that individuals who have positive attitudes toward welfare recipients also tend to feel that current benefits are insufficient. The only significant covariate effects are that women and more educated individuals tend to have more positive attitudes toward recipients, and that the more educated tend to perceive current benefits as more sufficient.

Measurement effects of probing on the attitude items are revealed by the estimated measurement model, shown in the upper part of Table 6. Full equivalence of measurement parameters is rejected for the joint hypothesis for all items and types of response together, as well as for each pair of two types of responses and for every item individually. The only one of these comparisons which is of borderline significance is that of equal loadings between unprobed and pre-probed responses (with $p = 0.054$). The estimated intercept terms for these types of responses ($\tau_{jl}^{(h)}$ for $h = 1, 2$) are also much more similar to each other than they are for probed responses. This is reflected in the estimated response probabilities, of which the probabilities of the middle category ($Y_j = 3$) are shown in Table 6 (at $\boldsymbol{\eta} = \mathbf{0}$, approximately the predicted value for $\boldsymbol{\eta}$ for a low-educated man at any age). Hypotheses of zero factor loadings are also rejected for each type of response. However, the likelihood ratio test statistic for this hypothesis is much smaller for the probed responses (test statistic of 27) than for pre-probed ones (305). Furthermore, zero loadings of probed responses are not rejected for several items individually, so we cannot even reject the hypothesis that these probed responses are not associated with the latent factors at all (although this may be due in part to low precision of the estimates from small numbers of probed responses).

For the attitude items there is thus a relatively small difference between unprobed and pre-probed responses, and a substantial difference between them and probed responses. For every item, the probability of the middle response category is much higher for probed than for other types of responses, even after we condition on the latent variables $\boldsymbol{\eta}$. This agrees with observations from the univariate analyses, like those in Figure 1. The results of the model-based analysis thus suggest that those observed differences were not due to selection effects of who the probed respondents are (which appear to be weak in any case) but to measurement effects of probing.

As an added check, we again did the analysis also using models which omitted the covari-

ates and included instead a non-ignorable response model with a direct association between $\boldsymbol{\eta}$ and ξ (even though this was not actually significant). The conclusions about the measurement effects of probing were unchanged from the ones reported above. Finally, we also tried a non-ignorable response model where $p(\xi|\boldsymbol{\eta})$ included also quadratic effects of η_1 and η_2. This was motivated by the observation that probed responses to these items are very often in the middle of the response scale, which could indicate a selection effect where such responders are more likely to have intermediate levels of the attitudes. The quadratic effect in the association between $\boldsymbol{\eta}$ and ξ is indeed clearest for the probability of being a reluctant responder, which is largest when a respondent has an average level of η_1 and a low or high level of η_2. However, this non-ignorable response model is only borderline significant compared to the non-ignorable model ($p = 0.059$), and even if the quadratic effects were included, conclusions about the measurement effects were again unchanged.

We have also fitted models for Bulgaria, Hungary and Portugal separately. This allows us to examine cross-national variation in the results, although with more uncertainty because the country-specific sample sizes are correspondingly smaller, especially for the relatively rare pre-probed and probed responses. We used the same model specifications as for the combined data (with covariates and ignorable response models), and focus on the estimated measurement models which describe the measurement effects of probing. Their parameter estimates are shown in the supplementary materials.

For the pseudo-knowledge items, the hypothesis of zero loadings for pre-probed and probed responses is rejected also for each country separately. The measurement intercepts are again smaller for probed than for unprobed responses, but the differences in the loading parameters are not consistently in one direction. The sizes of these differences are roughly similar in each country. However, they reach statistical significance only for Portugal (where the number of probed responses is largest) but not for Bulgaria or Hungary, so the conclusion of a significant measurement effect of probing which holds for the pooled sample is reached only for Portugal in the country-specific analyses. For the attitude items, in contrast, the hypothesis of no overall measurement effect is rejected for all countries. For the probed responses, none of their measurement loadings are significantly different from zero when tested individually (although the joint hypothesis that they are all zero is rejected for Bulgaria and Hungary). The measurement intercepts are again such that the

middle response option is for most items much more common for probed responses. These results for the attitude items are similar to the ones for the pooled sample.

Comparing, finally, the two types of items in the study (and drawing on both the pooled and country-specific analyses), the measurement effects of probing appear to be substantially larger for the attitude items than for the pseudo-knowledge items. For the attitude items, there is even an indication that probed responses to several items were not far from being unassociated with underlying attitudes. This is perhaps not surprising, given the different styles of the questions. Probing DK responses to the pseudo-knowledge items might encourage respondents to give answers which are comparable to those obtained without probing. Many responses to these questions, even those by immediate responders, are likely to be based on guesses or vague perceptions and probing may encourage people to return similar guesses that they were initially reluctant to express for fear of giving the 'wrong' answer. Probing DK responses to the more straightforward attitudinal items, on the other hand, could be more likely to result in poorly considered responses which differ from more firmly grounded answers given by immediate responders.

5 CONCLUSIONS AND IMPLICATIONS FOR SURVEY PRACTICE

For survey practitioners, the question of interest on probing DK responses is the operational one of whether it would or would not be a good idea to ask questions with probing. The current approach taken in many large attitudinal surveys, including the ESS, is not to probe. However, practice varies across surveys and survey houses. The work reported in this article was designed to help answer this question. We examined the effects of probing on item-level response rates and response properties using freshly collected data and a modelling framework developed for the purpose.

We proposed a latent variable model which combines models for whether and when a respondent answers the survey questions, how the answers relate to the constructs that the questions are designed to measure, and how the attitudes and responding behaviour depend on each other and on explanatory variables. This framework allows us to disentangle the

selection and measurement effects of probing and to test hypotheses about them. While the full model was constructed specifically to address these questions on probing, it draws on and combines related tools such as multigroup models for assessing measurement equivalence and latent variable models for non-ignorable nonresponse. The same ideas can also be applied, given appropriate data, to other comparable methodological questions, such as the effects of different question formats or response modes on multi-item survey data.

The potential benefit of probing is that it can reduce the undesirable effects of item nonresponse. First, it reduces the proportion of missing data whenever it converts initial nonresponse to a substantive answer. In our study, and in broad agreement with previous research, this effect was substantial, with around half of initial nonrespondents giving an answer when probed. Second, and depending on the relative differences between the immediate respondents, probed respondents and final nonrespondents, probing could also decrease systematic differences between the missing and observed data and thus reduce potential nonresponse biases.

The unavoidable cost of probing is that, by introducing additional follow-up questions within the survey, it necessarily increases the length (and thus cost) of the interview and may impose additional burden and fatigue on the respondents. These increases are likely to be non-trivial. Even in our small study with just eight possibly probed items, over 20% of respondents received at least one probe. Persistent use of probing in a longer survey could raise tensions when the respondents have otherwise been assured that they need not answer questions when they cannot or do not want to do so. It may be desirable to exercise restraint in the amount of probing and limit it at most to selected questions only.

A different cost of probing arises when it has a measurement effect. A clear finding from our analyses was that there was such an effect for most of the items in the study. In other words, the observed differences in the distributions of unprobed and probed responses were not explainable solely by the selection effect that the answers came from systematically different individuals, but also (and more) by the measurement effect that the two types of responses had different properties as measures of the latent constructs.

The exact nature and level of the measurement effects of probing are likely to vary between different types of survey items. In our study these effects were generally larger for agree–disagree attitude items than for pseudo-knowledge items. The measurement effects could

also be affected by question wording and response options. Here a particularly clear finding was that the neutral middle option was chosen very often as a probed response. This would not happen for questions which did not have a middle response option, but probing for them might simply mean forcing a random choice between the other response options instead.

Measurement effects could be allowed for in the analysis by employing models, such as the ones used here, which explicitly allow for different measurement properties for probed and unprobed answers. This, however, is not a practicable approach for everyday survey analysis, where instead probed responses would be treated just like unprobed ones, regardless of any measurement differences. This would risk introducing bias to the results of the analysis. It would, however, also reduce the rate of item nonresponse to the extent the probing does, and might not have a substantially deleterious effect on the estimates of interest, if the overall rate of nonresponse was low or the measurement effects were small. Nevertheless, it seems inadvisable as a general approach, because introducing additional measurement variation with unpredictable consequences is unappealing, and because the effects of the probed responses on the estimates will be known only after the interview costs of probing have already been incurred.

If there is evidence that probing leads to measurement effects, as was the case for the attitudinal and pseudo-knowledge items considered in our study, we would thus recommend against the use of probing of DK responses. However, the effect of probing, and therefore recommendations as to its use, could be different in other situations, for example for behavioural questions or ones on topics such as income where nonresponse rates are particularly high. Probing such items may reduce nonresponse without leading to significant measurement effects, though it is also possible that measurement effects may be even higher than for attitudinal items. Further research on the effects on probing in other contexts would thus be desirable.

If probing is not used, this leaves us to try to minimise item nonresponse and its consequences in other ways, including good design of questionnaires and interviews, and methods of analysis which use all the observed data while allowing for the nonresponse — approaches which should be used in any case, even if we did also employ probing.

Finally, these findings may also have implications for other probing-like approaches in surveys. A common example is a 'pre-emptive probe', that is a phrase such as 'if you are

not sure please give your best guess' which is included as part of the question stem for everyone. This has the advantage that all respondents receive the same stimulus which conveys the expectations of the interviewer and the encouragement to give an answer. At the same time, it is conceivable that the pre-emptive probe could also affect the responses of some respondents, in ways that could be comparable to the effects of a probe for initial nonresponse. We did not consider this question, but a comparable study with pre-emptive probes could be conducted and analyzed using the methods proposed here.

References

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken: Wiley.

Beatty, P. and D. Herrmann (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. M. Groves, D. Dillman, L. Eltinge, and R. Little (Eds.), *Survey Nonresponse*, pp. 71–85. New York: Wiley.

Beatty, P., D. Herrmann, C. Puskar, and J. Kerwin (1998). "Don't Know" responses in surveys: Is what I know what you want to know and do I want you to know it? *Memory 6*, 407–426.

Bishop, G. F., A. J. Tuchfarber, and R. W. Oldendick (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly 50*, 240–250.

Bradburn, N. (1978). Respondent burden. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 35–40.

Converse, P. E. (1964). The nature of belief systems in mass publics. *Critical Review 18*, 1–74.

de Leeuw, E. D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. Amasterdam: TT-Publikaties.

Grauenhorst, T., M. Blohm, and A. Koch (2015). Respondent incentives in a national face-to-face survey: Do they affect response quality? *Field Methods 28*, 266–283.

Holman, R. and G. A. W. Glas (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology 58*, 1–17.

Hox, J., E. D. de Leeuw, and H.-T. Chang (2012). Nonresponse versus measurement error: Are reluctant respondents worth pursuing? *Bulletin of Sociological Methodology/Bulletin de Methodologie Sociologique 113*, 5–19.

Hox, J. J., E. D. De Leeuw, and E. A. O. Zijlmans (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology 6*.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika 36*, 109–133.

Jowell, R., M. Kaase, R. Fitzgerald, and G. Eva (2007). The European Social Survey as a measurement model. In R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (Eds.), *Measuring Attitudes Cross-nationally: Lessons from the European Social Survey*. London: Sage.

Kaminska, O., A. L. McCutcheon, and J. Billiet (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly 74*, 956–984.

Kankaraš, M., G. Moors, and J. K. Vermunt (2011a). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, and J. Billiet (Eds.), *Cross-Cultural Analysis: Methods and Applications*, pp. 359–384. New York: Routledge.

Kankaraš, M., J. K. Vermunt, and G. Moors (2011b). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and Research*, 279–310.

Klausch, T., J. J. Hox, and B. Schouten (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research 42*, 227–263.

Knott, M., M. T. Albanese, and J. Galbraith (1990). Scoring attitudes to abortion. *The Statistician 40*, 217–223.

Krosnick, J. (2002). The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. In R. M. Groves, D. Dillman, L. Eltinge, and R. Little (Eds.), *Survey Nonresponse*, pp. 87–100. New York: Wiley.

Krosnick, J. A., A. L. Holbrook, M. K. Berent, R. T. Carson, W. M. Hanemann, R. J. Kopp, R. C. Mitchell, P. A. Ruud, V. K. Smith, W. R. Moody, M. C. Green, and M. Conaway (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly 66*, 371–403.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Hoboken, NJ: Wiley.

Luskin, R. C. and J. G. Bullock (2011). "Don't know" means "Don't know": DK responses and the public's level of political knowledge. *The Journal of Politics 73*, 547–557.

McLendon, M. J. and D. F. Alwin (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods and Research 21*, 438–464.

Medway, R. (2012). *Beyond Response Rates: The Effect of Prepaid Incentives on Measurement Error*. PhD thesis, University of Maryland, College Park, MD.

Medway, R. L. and R. Tourangeau (2015). Response quality in telephone surveys: Do prepaid cash incentives make a difference? *Public Opinion Quarterly 79*, 524–543.

Muthén, L. K. and B. O. Muthén (2010). *Mplus User's Guide (Sixth Edition)*. Los Angeles, CA: Muthén & Muthén.

Olson, K. (2013). Do non-response follow-ups improve or reduce data quality?: A review of the existing literature. *Journal of the Royal Statistical Society A 176*, 129–145.

Olson, K., C. Feng, and L. Witt (2008). When do nonresponse follow-ups improve or reduce data quality? A meta-analysis and review of the existing literature. Paper presented at the International Total Survey Error Workshop, June 1-4, 2008, Research Triangle Park, NC.

O'Muircheartaigh, C. and I. Moustaki (1999). Symmetric pattern models: A latent variable approach to itemnon-response in attitude scales. *Journal of the Royal Statistical Society A 162*(2), 177–194.

Roberts, C., N. Allum, and P. Sturgis (2014). Non-response and measurement error in online panels based on probability samples– are efforts to recruit reluctant panelists worth it? In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective*, pp. 337–362. Chichester: Wiley.

Sanchez, M. E. and G. Morchio (1992). Probing "don't know" answers: Effects on survey estimates and variable relationships. *Public Opinion Quarterly 56*, 454–474.

Schaeffer, N. C. and S. Presser (2003). The science of asking questions. *Annual Review of Sociology 29*, 65–88.

Schuman, H. and S. Presser (1981). *Questions & Answers in Attitude Surveys*. New York: Academic Press.

Singer, E. and C. Ye (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science 645*, 112–141.

Skinner, C. J., D. Holt, and T. M. F. Smith (Eds.) (1989). *Analysis of Complex Surveys*. New York: Wiley.

Stoop, I., J. Billiet, A. Koch, and R. Fitzgerald (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Chichester: Wiley.

Svalfors, S., W. van Oorschot, P. Taylor-Gooby, C. Staerklé, J. Goul Andersen, J. Hills, and S. Mau (2008). Welfare attitudes in a changing Europe. `http://www.europeansocialsurvey.org/docs/round4/questionnaire/ESS4_final_welfare_module_template.pdf`.