Survey Methodology, December 2000
Vol. 26, No. 2, pp. 189-198
Statistics Canada, Catalogue No. 12-001

Modeling Interviewer Effects in Panel Surveys: An Application

Jan Pickery and Geert Loosveldt [1]

Abstract

In this paper we will combine two applications of multilevel models. The multilevel model is suitable to analyze interviewer effects on survey data. It can also be used to analyze longitudinal - "repeated measurements" - data. We will analyze a data quality indicator of panel data that come from the Belgian Election Studies. These panel data consist of only two waves. The respondents that cooperated twice are for the most part not interviewed by the same interviewers. This results in a complex data structure with measurements nested in respondents, and respondents nested in interviewers, but without an overall hierarchical nesting structure: cross-classification. This complicated data structure will be analyzed in two different ways: an analysis of all respondents and an analysis of only those who are interviewed twice by the same interviewer. The results of these different analyses will be compared. We conclude that the multilevel cross-classified model is a very flexible and useful tool to analyze interviewer effects in panel surveys.

Key Words: Multilevel models; Cross-classifications; Panel surveys; Interviewer effects; Don't know answer.

1. Introduction

In this paper we analyze the effect of respondent and interviewer characteristics on the number of "don't know" answers in two waves of the panel survey from the Belgian Election Studies. We use different multilevel models for a subset of the dataset and for the entire dataset. The main purpose of the article is to illustrate how interviewer effects in a panel survey can be analyzed using multilevel models.

A multilevel or hierarchical model is an appropriate tool to analyze data with nested structures, *e.g.*, pupils nested in schools or patients in hospitals. A multilevel model can include variables of the different levels of nesting, but it also takes account of the variability associated with each level. The typical quality of the models is not the functional form relating the variables of the different levels, but rather a more sophisticated treatment of the error structure (DiPrete and Forristal 1994, 334). In education research for instance a multilevel model can account for variation between schools and variation between pupils. Moreover the model tries to replace this variance attributed to both levels by variables of either level. These models are described in various textbooks like Bryk and Raudenbush (1992), Goldstein (1995), Kreft and de Leeuw (1998) and Snijders and Bosker (1999).

Multilevel or hierarchical models also offer the best possibilities to analyze interviewer effects on survey data (Hox 1994). A hierarchical model is the best tool to tackle the "respondents nested within interviewers" - design. Other statistical techniques require mutual independence of interviewer and respondent characteristics, which is - most of the time - not the case because of the hierarchical structure of the data. In a multilevel model both the regression coefficients and the variance components are conditional on the explanatory variables in the model, which is a useful property if there is no complete orthogonalization of interviewer and respondent variables (Hox 1994, 307). When respondents are not randomly assigned to interviewers, respondent and interviewer characteristics can become confounded since respondents from a specific area will most likely be interviewed by interviewers from the same area. In such a situation, if the relevant respondent variables are put in the multilevel model, interviewers are equalized by statistical means. For that reason the assumptions of an analysis of interviewer effects with a multilevel model are more realistic than those of an ANOVA or ANCOVA. Furthermore the hierarchical model allows estimation of both the interviewer variance and the effects of explanatory variables measured at the interviewer and the respondent model. This possibility of replacing variance attributed to respondents/ interviewers with the effects of respondent/interviewer characteristics allows for wider generalizations.

The multilevel model can also fruitfully be used to analyze longitudinal - "repeated measurements" - data (see *e.g.*, Goldstein 1995, 87-95; Snijders 1996 and Yang and Goldstein 1996). There are alternatives to analyze the "measurements nested in individuals" - design, but multilevel analysis has some clear advantages. Using a hierarchical model, it is feasible to handle unbalanced designs - not all individuals have the same number of measurements - and quite easy to incorporate changing covariates. Besides, the model allows more nesting levels. The individuals can be nested in another higher level unit.

We will analyze respondent and interviewer effects on the number of "don't know" answers on a series of questions regarding political parties in a panel survey. We have

1. Jan Pickery and Geert Loosveldt, Department of Sociology, University of Leuven, E. Van Evenstraat 2B, 3000 Leuven, Belgium. E-mail: jan.pickery@soc.kuleuven.ac.be; geert.loosveldt@soc.kuleuven.ac.be.

measurements (wave 1 and wave 2) nested in respondents (the longitudinal design) and respondents nested in interviewers. Our panel data consist of two waves. During the second wave the respondents are for the most part not interviewed by the same interviewers. The purely hierarchical nesting has broken down and a more complex data structure is the result. To handle this data structure it is necessary to conceive the measurements as being nested into two different classification structures: measurements nested in respondents and in interviewers. This is called a cross-classified design, because the nesting of the levels is not purely hierarchical.

In this article we'll start with a simple data structure and the appropriate model. Afterwards the model will become more complex. We'll perform two analyses. In our first analysis we work only with the respondents who are interviewed twice by the same interviewers. Afterwards we will analyze all the respondents, including those who were interviewed only once. In the first analysis the purely hierarchical structure remains intact. The model is a "simple" three level one: measurements nested in respondents nested in interviewers. Analysis 2 sets up a cross-classified model. In that model the measurements are classified by respondents and interviewers.

The next section reflects on the nature of our dependent variable, the "don't know" answer, and the way to analyze it. In the third section we'll describe our data in detail to clarify the complex structure. The following section (4) treats in brief the different models that we will combine. Section 5 presents the variables in our analysis. In sections 6 and 7 we discuss the setup of our 2 different models and report the results of the analyses. Section 8 concludes the article.

2. Answering "Don't Know"

It has become generally acknowledged that the use of a "don't know" or a "no opinion" filter increases the proportion of respondents who give this answer, and that the increase itself is a function of the nature of the filter used (Schuman and Presser 1981, 143). Krosnick argues that answering "don't know" is one form of satisficing. Satisficing occurs when a respondent is not motivated to expend the mental effort necessary to generate optimal answers. A "no opinion" answer is an acceptable answer but it is the result of a "weak" cognitive process. Satisficing is a function of task difficulty, and the respondent's lack of knowledge, ability and motivation. This theoretical reasoning is consistent with the finding that offering a "don't know" response option increases the proportion of respondents who select it, particularly among respondents with little formal education and people who consider an issue to be less personally important. (Krosnick 1991). Following this argumentation, answering "don't know" is mainly explained by respondent characteristics that can be related to the cognitive aspect of answering questions. Previous research points us to the following characteristics of interest: education (*e.g.*, Sudman and Bradburn 1974), age (see *e.g.*, Groves 1989, 441-443), sex (*e.g.*, Hox, de Leeuw and Kreft 1991), and a measure of involvement or interest in the subject (*e.g.*, Groves 1989, 419).

However answering questions is not only a cognitive process of the respondent but it is also a communicative process (Schwarz and Sudman 1995). Within this process the interviewer plays an important role. There is a lot of literature about the interviewer as a source of survey measurement error (Groves 1989). The main idea is that interviewers are not "neutral" collectors of data but that they can influence the respondents' answers. Item nonresponse too is subject to interviewer effects as has been shown long ago by *e.g.*, Hanson and Marks (1958) and Bailar, Bailey and Stevens (1977). A social scientist interested in the explanation of "don't know" answers should therefore include respondents and interviewers in the analysis. The number of "don't know" answers will be the dependent variable of our analyses.

3. Description of the Data Structure

After the 1991 General Election in Belgium a national survey was set up in which 4,544 face to face interviews were conducted in the three Regions in the early months of 1992. A two-stage self-weighting sample (see *e.g.*, Särndal, Swensson and Wretman 1992, 141-144) was used. The sample was representative for the population of 18-74 years old (ISPO/PIOP 1995). In this article we will use the data from the Flemish region, which cover 2,691 Flemish respondents, interviewed by 163 interviewers (Carton, Swyngedouw, Billiet and Beerten 1993). After the 1995 Elections a similar survey was set up. Due to budgetary constraints the sample had to be smaller for the second wave. So the 2,691 respondents were used as a group to sample from and, in second order, there had to be new respondents to compensate for the aging of the youngest cohort from 1991. Finally 2,099 respondents were interviewed by 167 interviewers. This sample contained 1,762 panel respondents and 337 new respondents (see Beerten, Billiet, Carton and Swyngedouw 1997 for a detailed technical report of the sample plan). Only 55 of the interviewers of the first wave collaborated again. So there were 112 new interviewers in the second wave.

This gives us a dataset with 3028 respondents (2,691 + 337) and 275 interviewers (163 + 112). For 1,762 respondents we have a measurement in both waves, for the rest (1,266) there is only one measurement. The structure of the dataset can be represented in a table similar to table 1 (see also Goldstein 1995, 114). Each x in the table reflects an observation. The complete dataset contains 4,790 observations ((1,762 × 2) + 1,266). Each type of respondent in the table represents a possible occurrence in the dataset.

Table 1
A Representation of the Dataset

	Respondent Type 1		Respondent Type 2		Respondent Type 3	Respondent Type 4	
Wave	1	2	1	2	1	2	N (Interviewers)
Interviewer Type A	x	x	x			x	47
Interviewer Type B				x	x		8
Interviewer Type C			x		x		108
Interviewer Type D				x		x	112
N (Respondents)	374		1,388		929	337	

This table illustrates that we have three kinds of respondents: panel respondents who are interviewed twice by the same interviewer (Type 1), respondents who cooperated twice but were interviewed by different interviewers (Type 2) and respondents who are only interviewed once (Type 3 and 4). Our 2 different analyses are based on these different types of respondents. In Analysis 1 we'll look at the respondents whose situation corresponds with that of Type 1. Only 374 Respondents satisfy this condition. They were interviewed twice by the same interviewer. Analysis 2 takes all the 3,028 respondents into account (Respondents 1 to 4 of the table).

Furthermore the table shows that we can also discern different interviewers: interviewers who collaborated twice (Type A and B) and interviewers who collaborated only the first wave (Type C) or only the second wave (Type D). The interviewers of Type B collaborated in both waves, but never interviewed the same respondents twice (unlike the interviewers of Type A).

To analyze this complex data structure we will combine three different models that are presented in the next section.

4. A Short Description of the Different Multilevel Models Used in the Analyses

4.1 The General Multilevel Model

The first model we need is the general multilevel model, which has the following form:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + e_{ij}, \tag{1}$$

$$\beta_{0j} = \beta_0 + u_{0j},\ \beta_{1j} = \beta_1 + u_{1j} \tag{2}$$

or

$$\beta_{0j} = \beta_0 + \gamma_{01} z_{1j} + u_{0j} \text{ and } \beta_{1j} = \beta_1 + \gamma_{11} z_{1j} + u_{1j}. \tag{3}$$

Subscript i refers to the level 1 unit and subscript j to the level unit 2. In our situation level 1 indicates the respondent and level 2 the interviewer. So the response variable Y of respondent i, interviewed by interviewer j is dependent on the x variable of that respondent. This relationship looks like an ordinary regression model but the parameters of the model are interviewer specific. The β's differ from interviewer to interviewer. For each β, there is an interviewer residual (u_{0j} or u_{1j}). The β's can also be made dependent on higher level variables (interviewer characteristics), allowing for generalization across interviewers. We have one second level variable z. Substituting (3) into (1) results in the following overall model:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \gamma_{01} z_{1j} + \gamma_{11} z_{1j} x_{1ij} + u_{1j} x_{1ij} + u_{0j} + e_{ij}. \tag{4}$$

Of course more x and z variables can be included in these relationships. We assume that the residuals u_{0j}, u_{1j} and e_{ij} have means 0 given the values of the explanatory variables z and x. Furthermore it is assumed that the level 1 residuals (e_{ij}) are independent. The level 2 residuals (u_{0j} and u_{1j}) are assumed to be independent from e_{ij} and to have a joint multivariate normal distribution with covariance matrix Ω. They don't have to be independent from each other. Usually they are correlated.

4.2 The Multilevel Model for Longitudinal Analysis

The second model we need is the longitudinal multilevel model. In an analysis of a "repeated measurements" - design with a hierarchical model, the measurements are considered to be the first level and the individual the second. Most of the time the individual units will be persons, but of course they can be other units, like *e.g.*, schools or countries. In our analysis the individuals are the respondents. The analysis tries to estimate a growth curve on the base of the different measurements and to compare differences in curves given individual characteristics. Each observed value is made conditional upon the time of measurement - which can be a measure of time, but also age - and possible transformations of this measurement. Usually the curve is assumed to be a polynomial, which has the following form:

$$Y_{it} = \pi_{0i} + \pi_{1i} t + \pi_{2i} t^2 + \ldots + \pi_{ki} t^k + e_{it}. \tag{5}$$

Y_{it} is the observed value for respondent i on moment t, t can be time of measurement or age. $\pi_{hi} (h = 0..k)$ are the trajectory parameters or growth parameters for subject i, k is the degree of the polynomial. In a simple case k has the value 1 and then there is a linear curve. If there are m moments of measurement, a polynomial with degree of $m - 1$ will result in an exact reproduction of the curve. Of course it is more interesting to use a polynomial with a lower degree if that yields a satisfactory reproduction of the curve. You can test whether the model with degree $k + 1$ results in a significant improvement compared to the model with degree k.

The growth parameters have also a subscript for the individual (respondent). The model states that these parameters differ from individual to individual. The second part of the model defines these parameters:

$$\pi_{0i} = \pi_0 + r_{0i} \quad (6)$$

or

$$\pi_{0i} = \pi_0 + \beta_{01} x_{1i} + r_{0i}. \quad (7)$$

The individual parameter equals a general parameter (π_0) + an individual residual (r_{0i}). By the inclusion of individual characteristics (x) it may be possible to reduce the individual specific part, thus generalizing across respondents. In line with (1) and (2) we chose x to denote the individual (respondent) characteristics. But it is worth mentioning that in this model the x variables are higher level (level 2) variables. The individual characteristics can be fixed (the same for all moments of measurement) or varying.

4.3 Cross-Classified Models

The third model we will use is the cross-classified model. Not all data structures are purely hierarchical. Units may be classified along more than one dimension (see Goldstein 1995, 113-116). For example students can be classified by the school they go to and by the neighborhood they live in. In our example measurements are classified by respondents and by interviewers. A cross-classified model has the following form (subscripts j_1 and j_2 refer to the 2 different classification structures):

$$Y_{ij_1j_2} = \beta_{0j_1j_2} + \beta_{1j_1j_2} x_{1ij_1j_2} + e_{ij_1j_2}, \quad (8)$$

$$\beta_{0j_1j_2} = \beta_0 + u_{0j_1} + u_{0j_2} \text{ and } \beta_{1j_1j_2} = \beta_{1j_1j_2} + u_{1j_1} + u_{1j_2}. \quad (9)$$

Equation (9) can be reformulated the same way as equation (3).

$Y_{ij_1j_2}$ is the observed value for individual i, classified by j_1 and j_2. In our case: the observed value for measurement i on respondent j_1, interviewed by interviewer j_2. The parameters associated with the independent variable x have a residual for both classifying structures. For this model the additional assumption is made that the residuals of the different classifying structures (in our case: the respondent and interviewer residuals) are mutually independent (u_{0j_1} and u_{1j_1} versus u_{0j_2} and u_{1j_2}).

Raudenbush (1993) discusses this kind of models and the use of the EM algorithm to estimate them. Rasbash and Goldstein (1994) and Goldstein (1995, 123-124) show how these models can be specified and estimated using a purely hierarchical formulation and (consequently) standard multilevel software. The way to do this is to specify one of the classifications as a standard hierarchical one, then define a dummy for each unit of the other classification, specify that each of these dummy variables has a random coefficient at the higher level and constrain the resulting sets of variances to be equal.

In section 6 and 7 we'll use these 3 different models. They can all be implemented in MLn/MLwiN, software for multilevel modeling. Firstly we take a closer look at the variables we will use in the analysis.

5. Variables in our Analysis

One of the more difficult tasks during the interview of the election study was rating six parties on different 11-point scales. Three scales were presented to the respondents: catholicism, economic liberalism and federalism. An explicit "don't know" filter was included in the question, but it was not mentioned on the card with the alternatives given to the respondent. The entire question is included in the Appendix. We expected a considerable number of "don't know" answers because of the degree of complexity of the task. The explicit filter was expected to raise that number as well (see *e.g.*, Schuman and Presser 1981).

In the first wave the average number turned out to run up to more than 4 "don't know" answers per respondent. Almost 20% of the respondents made use of this possibility at least 9 times out of the 18. If we consider only the panel respondents the mean number is a bit lower (3.8). This is not surprising since we could expect that "multi-users" of the "don't know" answer would be underrepresented in the second wave because of lack of interest in the subject of the survey and/or difficulties in answering the questions. In the second wave the overall mean is 3.6 and the mean for the panel respondents 3.4. The numbers for the respondents that were interviewed twice by the same interviewer are 3.9 and 4.2 respectively. There is no explanation why the number of "don't know" answers during the second wave is higher than the overall mean for these respondents.

At the measurement level we'll use the year of the interview as indication of time of measurement. We've recoded this variable, so time has the value 0 for the first wave and 3 for the second wave.

At the respondent level we have 3 independent variables: sex (0 = man, 1 = woman), completed education (0 = low, 1 = high) and the extent to which the respondents follow political news in the press (press: 1 = (almost) always - 5 never). The first 2 variables are constant for the 2 times of measurement. The third is a time-varying covariate and the question phrasing also slightly changed for the second survey. The two different questions are also included in the appendix. The dissimilarity in the phrasing induces an additional difficulty in setting up the model. The way to handle such a variable is to standardize it (mean 0, variance 1) for each time of measurement and (afterwards) to ascribe the value 0 to the time of measurement when the question wasn't asked. The reference value for those variables is their mean (see Snijders 1996, 422). This gives us 2 variables: press1 for the first occasion and press2 for the second. The former has the value 0 for all respondents for the second measurement and the latter for the first measurement. We don't take up the respondent's age in the model, since this variable would correlate too much with the time measurement at the occasion level.

In order not to complicate the analysis too much we don't take up interviewer variables. We just assume there is an interviewer effect, without trying to explain that effect in terms of interviewer characteristics.

6. Analysis 1: Respondent Type 1 of Table 1

The first analysis considers only those respondents who are interviewed twice by the same interviewer (cfr. Type 1 from Table 1). This analysis requires a "simple" three level model: measurements nested in respondents nested in interviewers. The hierarchical structure is unambiguous. This model is similar to the example in chapter 8 in the Bryk and Raudenbush book (1992). In that example the authors analyze the progress in academic achievement of students in schools.

Our dependent variable is the number of "don't knows"s for respondent i on moment t, interviewed by $j\,(Y_{tij})$. We have only 2 measurements so the degree of the polynomial cannot exceed 1. This results in the following level 1 equation:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\,\text{YEAR} + e_{tij}.$$

Our time variable (t) is the year of the interview which has the value 0 (1992) or 3 (1995). We will test whether π_{1ij} is significant. If not, this leaves us a null model or "naive" model (see Snijders 1996, 411), in which the number of "don't know" answers doesn't change over time, and both measurements can be considered as retests of the same constant value. The coefficients in the level 1 equation are respondent and interviewer specific.

At the respondent level we'll include 3 variables: sex, education and the 2 press variables. So our level 2 equation contains 4 variables:

$$\pi_{0ij} = \pi_{0j} + \beta_{01j}\,\text{SEX}_{\text{i}} + \beta_{02j}\,\text{EDUCATION}_{\text{i}} + \beta_{03j}\,\text{PRESS1}_{\text{i}} + \beta_{04j}\,\text{PRESS2}_{\text{i}} + r_{0ij}.$$

If the parameter estimate associated with year is significant we'll have a similar equation for π_{1ij}.

At the third level (interviewer) we won't include anymore variables, but we will fit a random intercept and random slopes. So we have the following level 3 equations:

$$\pi_{0j} = \pi_0 + u_{0j} \quad \text{and} \quad \beta_{01j} = \beta_{01} + u_{01j}, \ldots$$

Implementing these model specifications in MLn gives us the following results.

Model a in the table is the null model. This is a model without independent variables, neither at the measurement level, nor at the respondent level. In this model there is no evolution in the number of "don't know" answers. But the variance of the dependent variable is divided in a measurement part, a respondent part and an interviewer part. All variances are significant. This indicates that there is between wave variation, that some respondents use the answer more than others and that some interviewers will get more "don't know" answers than other interviewers.

Table 2
Analysis of Respondents who were Interviewed Twice by the Same Interviewers (s.e. in brackets)

	model a	model b	model c	model d
Fixed				
Measurement level				
constant	4.136 (0.322)	4.028 (0.358)	3.749 (0.442)	3.754 (0.523)
year		0.072 (0.089)		
Respondent level				
sex			2.393 (0.434)	2.458 (0.414)
education			-1.675 (0.425)	-1.778 (0.446)
press1			0.911 (0.263)	0.887 (0.233)
press2			1.483 (0.236)	1.426 (0.234)
Random				
Interviewer level				
$\sigma^2_{\text{constant}}$	2.249 (1.040)	2.251 (1.043)	2.666 (0.969)	6.090 (2.109)
$\sigma^2_{\text{education/constant}}$				-4.099 (1.816)
$\sigma^2_{\text{education}}$				1.396 (1.819)
Respondent level				
$\sigma^2_{\text{constant}}$	14.470 (1.714)	14.480 (1.714)	8.939 (1.308)	8.692 (1.332)
Measurement level				
σ^2_e	13.320 (0.974)	13.300 (0.974)	13.270 (0.969)	13.250 (0.969)
-2 LL	4,519.35	4,518.62	4,414.52	4,395.68
Δdf^*		1	4	6

Note: * compared to model a

The inclusion of the variable YEAR does not provide a better fit of the model. The decrease in deviance (-2 Log L) is not significant, neither is the parameter of the variable significant (model b). We can conclude that there is no significant overall evolution in the number of "don't know"s. We can go on with a model without the time variable.

The respondent variables do result in a considerable improvement of fit of the model. The decrease of the -2 Log L value is large and clearly significant ($p < 0.001$). According to the analysis women use the "don't know" alternative more than men do and highly educated respondents less than respondents with lower education. Following the political news in the press reduces your chance to answer "don't know". Both press1 and press2 are significant (model c). The inclusion of respondent variables also results in a substantial decrease of the variance at the respondent level.

We also tried to fit random slopes at the interviewer level (model d). Our analysis showed some variation in the parameter associated with the respondent's education. This is the only independent variable with a varying coefficient at the third level. $\sigma^2_{\text{education}}$ is not significant, but there is an important covariance between the residual for the constant and the residual for education ($\sigma^2_{\text{education/constant}} = -4.099$). The covariance is negative, indicating that interviewers with a higher constant have a smaller coefficient for education. Since the fixed parameter for education is negative, it will be even more negative for those interviewers, thus having a larger absolute value. Hence for interviewers who stimulate more "don't know" answers, the difference between less educated respondents and more educated respondents will be larger. In model d the value of $\sigma^2_{\text{constant}}$ at the interviewer level has increased considerably, compared to model c. In this model the variance at the interviewer level is dependent on the values of the explanatory variable education and it will be larger for zero values of education. That is another interpretation of model d: the variance between interviewers is much higher for lower educated respondents than for higher educated respondents. This model with a more complex variance structure at level 3 has a better fit than the previous models.

When including YEAR in model c or model d, it turned out to be not significant either. Also in our final models there is no evidence for an evolution in the number of "don't know"s between the two waves. All models prove a significant interviewer effect. But the relative size of the variance shows that there is more variation between respondents than between interviewers.

7. Analysis 2: All Respondents

In this analysis we look at all the respondents: the panel respondents that were interviewed twice by the same interviewer, the other panel respondents and those who were interviewed only once. This second analysis breaks down the hierarchical structure. Measurements are still nested in respondents and respondents are still nested in interviewers. But there is no overall hierarchical structure, since the interviewer can (and most of the time will) change between the two waves (see section 3). Our dependent variable is still the number of "don't know"s of respondent i interviewed by j on moment $t\,(Y_{tij})$. But the model has changed. The level 1 equation hasn't:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\text{YEAR} + e_{tij}.$$

In this notation we use π, since the level 1 model is also a growth curve. But this equation matches the level 1 model of the cross-classified model (equation (8), section 4.3). Furthermore we still use i for the respondent and j for the interviewer. But it is important to notice that this is not the same model as the one of analysis 1. These subscripts correspond to the j_1 and j_2 of equations (8) and (9).

There is no "real" third level. To fit the cross-classified model in MLn, we have to define a third level, but conceptually the respondent and interviewer are at the same level in this model. This leads to the following level 2 equation:

$$\begin{aligned}\pi_{0ij} &= \pi_0 + \beta_{01}\,\text{SEX}_{\text{i}} + \beta_{02}\,\text{EDUCATION}_{\text{i}} \\ &\quad + \beta_{03}\,\text{PRESS1}_{\text{i}} + \beta_{04}\,\text{PRESS2}_{\text{i}} + r_{0i} + r_{0j}.\end{aligned}$$

The interviewer specific part (r_{0j}) is included in the second level, so there is no interaction between the interviewer variance and the respondent variables. This is the main difference with analysis 1.

A cross-classified model requires enormous computations. We have 3,026 respondents and 275 interviewers. This would mean 275 dummies with all varying coefficients at the artificial third level. Up till now it is impossible to fit such a model. The storage required by the worksheet is far too large (see Goldstein 1995, 118 and Rasbash and Woodhouse 1996, 85-86 for details). It is possible to reduce these storage requirements and improve the speed of model estimation by dividing the dataset in subsets in which the cross-classification implies fewer cells. In our case we look for separate groups of measurements that are classified by fewer respondents and interviewers. The analysis of 1 group of 1,000 measurements classified by 500 respondents and 100 interviewers is computationally more demanding than the analysis of a dataset consisting of 10 groups of 100 measurements each, classified by 50 respondents and 10 interviewers. Sometimes it is worth omitting some of the observations (measurements in combinations of respondents and interviewers that hardly occur) to make the partitioning more efficient.

MLn/MLwiN provides some procedures (via the commands XSEArch and BXSEArch) that are designed for that partition (Rasbash and Woodhouse 1996, 89-93). We used the BXSEArch command. The command starts an enhanced procedure, which attempts to provide the maximum separation with the minimum deletion of data. We started with

4,790 measurements, 3,026 respondents and 275 interviewers. After omitting the observations indicated by the BXSEarch command, we're left with 4,597 measurements on 3,026 respondents interviewed by 275 interviewers. No higher level units (respondents nor interviewers) are left out. The procedure resulted in 7 partitions with a maximum of 44 cells in the cross-classification of respondents and interviewers. The model converged sufficiently fast when implied this way.

The results of the analysis are reported in Table 3.

Table 3
Analysis of all the Respondents (s.e. in brackets)

	model a	model b	model c
Fixed			
Measurement level			
constant	3.894 (0.136)	3.967 (0.155)	3.864 (0.165)
year		-0.053 (0.055)	
Respondent level			
sex			1.808 (0.153)
education			-1.914 (0.148)
press1			1.185 (0.090)
press2			1.197 (0.102)
Random			
Level 2			
Interviewer			
$\sigma^2_{constant}$	2.777 (0.373)	2.716 (0.368)	2.844 (0.363)
Respondent			
$\sigma^2_{constant}$	11.810 (0.635)	11.800 (0.635)	7.017 (0.527)
Measurement level			
σ^2_e	13.130 (0.475)	13.150 (0.476)	13.460 (0.480)
-2 LL	27,717.1	27,716.3	27,042.1
Δdf*			

Note: * compared to model a

This table looks very much the same as Table 2 but there is an important difference. In the random part we marked level 2 – interviewer and respondent to make clear that the interviewers do not constitute a third level in this analysis.

Model a is the null model: no explanatory variables, but the variance of the dependent variable separated in a measurement part, a respondent part and an interviewer part. There is a significant interviewer variance. Thus in this design we again have evidence for an interviewer effect. You have to be careful about the interpretation of the relative sizes of the variances if one classification has far fewer units than the other (Goldstein 1995, 117-118). It's not fully correct to state that there's 5 times as much variation between respondents than between interviewers, but again there is much more variability between respondents than between interviewers.

In the next model (model b) we've included the time variable (YEAR). Again this variable turns out to be not significant and its inclusion does not provide a better fit of the model. Again we can conclude that there is no significant overall evolution in "don't know" answers over time.

Model c is the model with the respondent variables. They are all significant and this model has a far better fit than the previous ones. The substantive interpretation of the parameters is the same as in analysis 1. Women use the "don't know" answer more than men and a higher education results in less "don't know"s. The extent to which the respondents follow the political news in the press is also a predictor of the use of the "don't know" answer. The less they follow politics the more they answer "don't know".

8. Conclusion and Discussion

The general conclusions of this article are methodological as well as substantive.

Our analysis confirms previous research findings about the use of the "don't know" answer. It is related to the respondent's education, sex and a measure of involvement or interest in the subject. Furthermore it is likely to diverge from interviewer to interviewer. All our analyses showed a significant interviewer effect. We did not find a significant evolution in the use of the "don't know" answer over time in the two waves of the survey. The interviewer effects prove that the "don't know" response alternative is not merely a result of the respondent answering the questions. It stresses the necessity of an interviewer training, which includes instructions on how to ask difficult questions and how to deal with "don't know" answers.

As in most panel surveys, the nonresponse in the second wave of this panel survey was not totally random. It is related to the respondent's living arrangement, his or her political interest and a few socio-demographic variables (Loosveldt and Carton 1997). This selective dropout puts limits to the generalizability of the results concerning the evolution in the dependent variable, but our analyses did not show a general evolution in the use of the "don't know" answer anyway. An impact of selective nonresponse in the second wave on the size of the interviewer effect is not unlikely either as interactions between the respondent characteristics and the interviewer effects are possible, as analysis 1 showed. But it is unlikely that this will affect the substantive conclusions about the interviewer effects. Given the results of analysis 1 and the conclusions in the Loosveldt and Carton paper, one could even expect that the interviewer effect in analysis 2 and consequently the overall interviewer effect might be somewhat underestimated. Loosveldt and Carton (1997, 1021) show that lower educated respondents are more likely to drop out of the survey than higher educated respondents and analysis 1 showed that the interviewer variance is higher for lower educated respondents.

The methodological conclusions consider the use of the different models to analyze interviewer effects in panel surveys. The analyses presented in this paper show that quite complex designs with complicated data structures can be analyzed by specifying the appropriate multilevel model.

The first model (Analysis 1) only suits in a tiny number of cases. It is not so common to ascribe the same interviewers to the same respondents for different waves of a panel survey, neither is it always feasible.

The second model (Analysis 2) is an appropriate tool but can require enormous computations. MLn is quite powerful and helps to decrease the storage requirements, at the cost of a small loss of information. Besides, the second model has its limitations too. Using this method it is not possible to model interactions between respondent variables and interviewer variance, as we did in the first analysis, or between respondent and interviewer variables. However the analysis showed that this model could be a very useful and flexible tool. The cross-classified model is also suitable when the number of measurements increases. A panel survey with 3 or 4 or even more waves, where some interviewers are retained and some are new at each occasion would require exactly the same analysis. The multilevel model also knows how to handle respondents for whom 1 or more measurements are missing, as our analysis showed. The pliability of this model outweighs the impossibility to include respondent – interviewer interactions in the model. That would be feasible when analyzing each wave of the panel survey separately. But those analyses could not model a possible evolution in the dependent variable, which is another important advantage of the joint analysis of all waves of the panel.

Acknowledgements

We thank the ISPO-PIOP Centre for Electoral Research for providing us with the data. Jacques Billiet, Marc Swyngedouw, Ann Carton and Roeland Beerten originally collected the Flemish data. The ISPO-PIOP is supported by the Federal Services for Technical, Cultural and Scientific Affairs. Neither the original collectors nor the Centre bear any responsibility for the analyses or interpretations presented here. We would also like to thank Jon Rasbash (Institute of Education, University of London) for some very useful comments about the operation of MLn. Finally we thank the referees for their constructive comments and suggestions on an earlier version of the paper.

Appendix 1

The rating question was: "Political parties are said to be "Catholic" or "non-Catholic". Please place the cards of the various parties on card No. 20 at the place that corresponds best to the degree in which the party is "Catholic" or "non-Catholic". If two or more parties are just as "Catholic" or just as "non-Catholic" in your opinion, place the cards on the same square. If you do not know how "Catholic" or "non-Catholic" a party might be, then simply put its card aside."

With the card:

Catholic 0 1 2 3 4 5 6 7 8 9 10 Non-Catholic

The press question was not identical for both surveys. For the first survey the press question was: "How often do you read the political news in the newspaper?"

With the response categories:

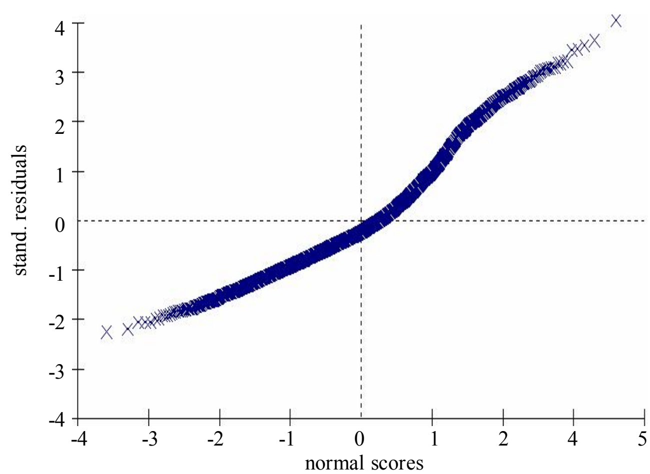
1 = (almost) always, 2 = often, 3 = now and then, 4 = seldom, 5 = never

In the second survey it became: "How often do you follow the political news on the radio, on television or in the paper?"

The response categories remained the same.

Appendix 2

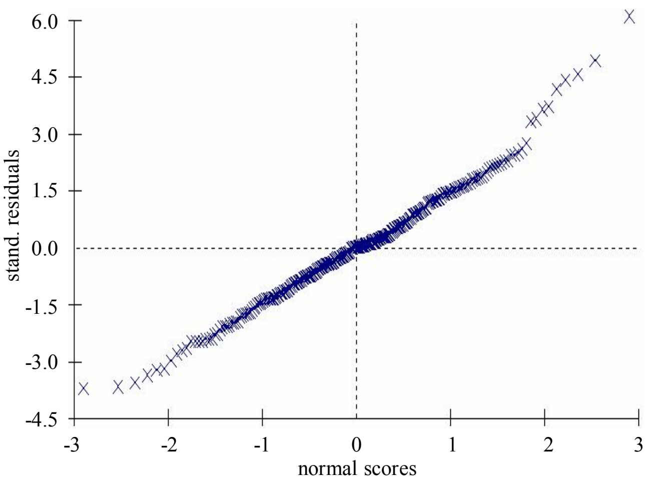
Section 4 set the assumptions of the different models that were used. For the last model the most important assumptions concern the random effects associated with the respondent and the interviewer. The assumption that the $\sigma^2_{\text{constant}}$ values for the respondent and for the interviewer are normally distributed can be assessed by looking at Normal probability plots for the residuals. Graph 1 presents the plot for the standardized respondent residuals and graph 2 the plot for the standardized interviewer residuals.



Graph 1. Standardized respondent residuals by Normal equivalent scores.

In this graph the departures from the diagonal are rather limited and no apparent violation of Normality can be inferred. On the other hand it is worth noting that this graph shows more observations at the upper right hand than at the lower left end.

Graph 2 does not show any clear departures from the diagonal either. But in this graph some outliers draw the attention. Especially the outlier at the upper right hand side of the graph seems to be outside the range of the other interviewer residuals. Moreover in this graph also there are more observations at the upper right hand side than at the lower left end.



Graph 2. Standardized interviewer residuals by Normal equivalent scores.

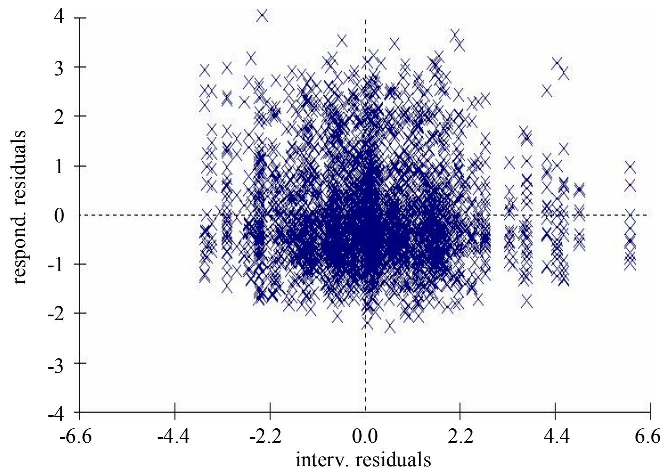
The conclusions from these graphs are as follows: there is nothing clearly wrong with the residuals but the more numerous deviations upwards and the outliers of the interviewer residuals could possibly be further investigated. Efficient techniques for these checks are not yet available for multilevel models (Goldstein 1995 29). But it is of course possible to analyze a dataset without the outliers. That is done in Table 4.

Table 4
Analysis of the Dataset Without the Interviewer Outliers (s.e. in brackets)

Fixed		
Measurement level		
constant	3.853	(0.162)
Respondent level		
sex	1.820	(0.153)
education	-1.929	(0.149)
press1	1.160	(0.090)
press2	1.217	(0.102)
Random		
Level 2		
Interviewer		
$\sigma^2_{constant}$	2.495	(0.333)
Respondent		
$\sigma^2_{constant}$	7.109	(0.530)
Measurement level		
σ^2_e	13.420	(0.481)
-2 LL		26,850.2

For the analysis in Table 4 we excluded two interviewers, the one with the lowest and the one with the highest residual. The coefficients in this table are very similar to those of model c in Table 3. The interviewer variance has decreased a bit, as a result of the exclusion of extremes, but there is no evidence of a considerable impact of the outliers on the results.

The other assumption about the interviewer and respondent random effects is their mutual independence. The interviewer and respondent residuals should not correlate. That is of course more difficult to evaluate since both residuals are connected to their respective units, which do not correspond. You will get 3,028 respondent residuals and 275 interviewer residuals. An indirect check of this assumption is possible by attributing the interviewer residuals to the respondents. This is done in graph 3.



Graph 3. Standardized respondent residuals by standardized interviewer residuals.

In this graph, again the more numerous deviations upwards and the interviewer outliers draw the attention. Apart from that, no pattern can be discerned. Because of the interviewer outliers, there are fewer observations at the right hand side of the graph. But the respondent residuals do not really tend to be smaller if the interviewer residuals are higher. Neither is there any evidence of the opposite.

The check in graph 3 is imperfect as it attributes the interviewer residuals to the respondents. A better alternative might be to fit a more complex model with an interaction term between the two random effects. Goldstein (1995, 119) proposes this model. A test for the model improvement due to the interaction term can give an indication for the presence of a correlation between the residuals. Another alternative is the insertion of an additional level (the region) above interviewers and respondents. That model would include a term for the regional variation, which could cause a correlation between the interviewer and respondent residuals. Snijders and Bosker (1999, 159-160) describe this model. But both models require a different parameterization with various sets of dummies. Their clarification calls for a paper in itself and is consequently outside the scope of this paper.

References

Bailar, B., Bailey, L. and Stevens, J. (1977). Measures of interviewer bias and variance. *Journal of Marketing Research*, 14, 337-343.

Beerten, R., Billiet, J., Carton, A. and Swyngedouw, M. (1997). *1995 General Election Study Flanders-Belgium. Codebook and Questionnaire*. Leuven: ISPO/Departement Sociologie, K.U. Leuven.

Bryk, A.S., and Raudenbush, S. (1992). *Hierarchical Linear Models Applications and Data Analysis Methods*. Newbury Park - London: Sage.

Carton, A., Swyngedouw, M., Billiet, J. and Beerten, R. (1993). *Source Book of the Voters' Study in Connection with the 1991 General Election*. Leuven: Sociologisch Onderzoeksinstituut/ ISPO.

Diprete, T.A., and Forristal, J.D. (1994). Multilevel models: Methods and substance. *Annual Review of Sociology*, 20, 331-357.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold.

Groves, R.M. (1989). *Survey Error and Survey Costs*. New York: John Wiley & Sons, Inc.

Hanson, R.H., and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53, 635-655.

Hox, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

Hox, J.J., De Leeuw, E.D. and Kreft, I.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys*. (Ed. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman). New York: John Wiley & Sons, Inc., 439-461.

ISPO/PIOP (1995). *1991 General Election Study Belgium. Codebook and Questionnaire*. Leuven: ISPO.

Kreft, I.G., and De Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage Publications.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Loosveldt, G., and Carton, A. (1997). Evaluation of nonresponse in the Belgian Election Panel Study '91 - '95. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 1017-1022.

Rasbash, J., and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational Statistics*, 19, 337-350.

Rasbash, J., and Woodhouse, G. (1996). *MLn Command Reference*. Version 1.0a. London: Multilevel Models Project. Institute of Education, University of London.

Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 321-349.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys*: *Experiments on Question Form*, *Wording and Context*. New York: Academic Press.

Schwarz, N., and Sudman, S. (1995). *Answering Questions*. San Francisco: Jossey-Bass.

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, 30, 405-426.

Snijders, T., and Bosker, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Sudman, S., and Bradburn, N. (1974). *Response Effects in Surveys*. Chicago: Aldine.

Yang, M., and Goldstein, H. (1996). Multilevel models for longitudinal data. In *Analysis of Change. Advanced Techniques in Panel Data Analysis*. (Ed. U. Engel and J. Reinecke). Berlijn - New York: Walter de Gruiter, 191-220.