Journal of Official Statistics, Vol. 20, No. 1, 2004, pp. 77–89

# A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers

*Jan Pickery*[1] *and Geert Loosveldt*[2]

In this article we go further along the path of multilevel analysis of respondent and interviewer effects on survey data. With a simultaneous analysis of five data quality indicators we show how a multivariate multilevel model can make a further contribution in survey data quality research. Furthermore we demonstrate a practical application. We show how the results of a (multivariate) multilevel analysis can be used to identify exceptional interviewers.

*Key words:* Survey data quality; multilevel analysis; multivariate model.

## 1. Introduction

It has become common practice to analyse interviewer effects on survey data with multilevel techniques. In a multilevel analysis the respondents constitute the first level and the interviewers the second level. The variance of a dependent respondent variable is divided into a respondent and an interviewer part and variables of both levels are used to explain both variances. Hox et al. (1991) and Hox (1994) provide an introduction and a discussion of the appropriateness of this technique. Typical examples can be found in Campanelli and O'Muircheartaigh (1999), Van Tilburg (1998) and Marsden (2003).

The literature about interviewer effects is rather comprehensive. Most research findings show small interviewer variances (intra-class correlation coefficients below 0.02) (Groves 1989) and the effects of interviewer characteristics are not univocal (Hox et al. 1991). The present article can be situated in this research tradition. We will use a multivariate multilevel model to examine interviewer effects on several data quality indicators simultaneously. We will not try to explain these interviewer effects. We want to show how the results of the multilevel analysis can be used to identify exceptional interviewers. The interviewer residuals, produced by a multilevel analysis, can be used for this purpose. With "exceptional" we mean that they register unusual response patterns compared to other interviewers. The identification of exceptional interviewers is another application of the analysis of interviewer effects with multilevel models, which is very

[1] Administration of Planning and Statistics, Flemish Administration, Boudewijn 30, 1000 Brussels, Belgium. Email: jan.pickery@azf.vlaanderen.be

[2] Katholieke Universiteit Leuven, Department of Sociology, E. Van Evenstraat 2B, 3000 Leuven, Belgium. Email: geert.loosveldt@soc.kuleuven.ac.be

**Acknowledgments:** Jan Pickery did most of this research while being a Postdoctoral Researcher funded by the Fund for Scientific Research - Flanders (FWO).
We would like to thank the two anonymous reviewers and the associate editor for their helpful comments.

useful to evaluate interviewer performance. This is another application of the analysis of interviewer effects with multilevel models.

This article complements an article published previously (Pickery and Loosveldt 2001). In that article we analysed interviewer effects on five different kinds of item nonresponse and examined which nonresponse was subject to interviewer effects. In the analysis in this article some variables are the same and others are similar, but the model is different, as well as the results we present. We already know that the variables we examine are subject to interviewer effects. Consequently the analysis we present in this article is supposed to answer two other research questions. Firstly, are interviewer effects on various data quality indicators correlated? Secondly, can we identify generally exceptional interviewers?

## 2. Data and Variables in the Analysis

Data come from the survey "Cultural Shifts in Flanders: Survey 2000." This survey was organized by the Flemish Administration, Department for Planning and Statistics (Ministerie van de Vlaamse Gemeenschap 2001). The survey is rather general, with questions on time spending, social relations and values like ethnocentrism, utilitarianism, and political powerlessness. In the survey 1,345 respondents (aged 16 to 85) were interviewed by 85 interviewers. In our analysis we take a closer look at five data quality indicators that also appeared to be subject to interviewer effects in a preliminary analysis (Loosveldt and Carton 2001, pp. 28–32): item nonresponse to the income question, "no opinion" answers to items of various scales, "don't know" answers to items of other scales, the use of "don't know" answers to knowledge questions and the use of extreme response categories to items of various scales.

The respondent's income was asked for both using one open and one closed question. If the respondent did not answer the open question, the closed question had to be asked: 11% of the respondents did not answer either of the questions. They were coded 1 on the first response variable (income). This item nonresponse probably comprises various kinds of response behaviour like refusal and ignorance.

For three five-point scales with a total of 25 items on political powerlessness, way of life/sense of public responsibility and environmentalism, the no opinion alternative was mentioned in the question but not on the show card. More than 92% of the respondents never made use of this alternative. The remaining 8% are grouped together and are coded 1 on the second response variable (no opinion-items). For a scale with 18 items on trust in public organisations the "don't know" answer was also mentioned to the respondents but not included on the show card. 18% of the respondents used this answer at least once. They were coded 1 on the third response variable (don't know items). Krosnick (1991) argues that answering "don't know" or "no opinion" is a form of satisficing. Satisficing occurs when the respondent is not motivated to expend the mental effort necessary to generate optimal answers. The "no opinion" answer is an acceptable answer, but it is the result of a weak cognitive process. Interviewers have been found to have an effect on this process as well (Pickery and Loosveldt 1998).

We also examined the "don't know" answer to knowledge questions. The respondents were asked to give an explanation of four abbreviations relating to waste treatment and the environment. Here the "don't know" answer is used very frequently. Only 7% percent of the respondents did not use it. They have the value 0 on the fourth response variable (don't know knowledge). For knowledge questions the "don't know" response is acceptable and plausible, but an interviewer effect on its occurrence also indicates dissimilarities during the interviewing process.

Finally we examined the use of extreme response categories with regard to 25 items of the three scales that also defined the second dependent variable ("no opinion" answer). These were five-point scales (from totally agree to totally disagree) and 25% of the respondents never used the first or the last category. The other 75% were coded 1 on the last dependent variable (extreme response categories). Scales like this are used to obtain information about the direction and the strength of the respondent's opinion. One can expect that even respondents without a pronounced opinion will use the extreme response options now and then, especially when different scales are considered. If they do not, they probably do not expend the effort required. As this is a form of satisficing, the interviewer can affect it the same way as in the case of the use of the "no opinion" answer.

Independent respondent variables in our model are sex, education, and age. Except for age these variables are dummies: female and low education and high education (with mean education being the reference category). Age is centred around the grand mean.

In this analysis we are not particularly interested in explaining item nonresponse or the use of the extreme response categories. We want to look for correlations between the interviewer effects and identify exceptional interviewers. Consequently it is not a major problem that our models are not fully specified. We include and control for the basic respondent characteristics, but we are not interested in a full model with all the relevant respondent characteristics, neither do we include interviewer variables in the analysis. Given our research question, it is also justifiable to simplify the analysis by dichotomising all the dependent variables and including the same independent variables for all dependent variables, although both are technically not necessary.

## 3. The Model

As already mentioned, we selected five variables that refer to the data quality of the survey. Since we dichotomised all dependent variables a multilevel logistic model is appropriate. Goldstein (1995, pp. 77–111) discusses multilevel models for discrete response data. Random coefficients can be included in models with dichotomous outcomes by linearizing the model by a Taylor series expansion. In such a linearized model an approximation of the nonlinear component for each unit can be obtained using the fixed part predictor or adding the estimated level 2 residuals to that predictor. The former is called marginal quasi-likelihood (MQL), the latter penalized or predictive quasi-likelihood (PQL). PQL estimation in combination with a second order Taylor series expansion (involving the second derivative) produces the most accurate results (Goldstein 1995, pp. 99–101). MQL and PQL are for instance available in MLwin. A disadvantage of quasi-likelihood estimation is that the log likelihood value is only approximative. A likelihood-ratio test cannot be used. An alternative to these quasi-likelihood procedures is direct integration of the nonlinear likelihood. This is for example available in MIXNO. With direct integration the likelihood-ratio test can be used.

*Table 1. Example data matrix*

| Respondent | Interviewer | Female | Nonresponse income question | No opinion items |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 |
| 3 | 2 | 0 | 1 | 1 |
| 4 | 2 | 0 | 0 | 0 |

We have five dependent variables. To analyse all the dependent variables in one model a variable level has to be created below the respondent level; the respondents then constitute the second level and the interviewers the third. Furthermore dummy variables have to be created for the dependent variables. We can illustrate this with a small example. Consider a dataset with four respondents, two interviewers, one independent variable (female) and two dependent variables (nonresponse income question and no opinion items). This dataset is represented in Table 1.

*Table 2. Reorganized example data matrix*

| Respondent | Interviewer | Female | Response variable | Income dummy | No opinion dummy |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 | 1 |
| 2 | 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 0 | 0 | 1 |
| 3 | 2 | 0 | 1 | 1 | 0 |
| 3 | 2 | 0 | 1 | 0 | 1 |
| 4 | 2 | 0 | 0 | 1 | 0 |
| 4 | 2 | 0 | 0 | 0 | 1 |

In Table 2 the reorganized data matrix suited to a multilevel multivariate analysis is shown. The original variables are copied twice and the new response variable alternates between nonresponse income question and no opinion items. Furthermore two dummy variables are created for the respective dependent variables.

For this dataset three level equations can be defined for each dependent variable. These equations are then multiplied by the dummy variables and summed into one model. In such a three-level model we obtain separate fixed parameter estimates for all dependent variables. Moreover we obtain variance and covariance estimates both at the respondent (second) and the interviewer (third) level. At the first level (variable) no (co)variances are estimated. That level exists solely to fit the multivariate structure. A discussion of the model can be found in Goldstein (1995, pp. 69–71).

## 4. Results of the Analysis

We analysed the five dependent variables including the independent variables sex, education (one and two dummies, respectively) and age in the model. Although the model does not necessarily require so, we included the same independent variables for all five dependent variables. The results of this multivariate multilevel logistic model fitted to our data are reported in Table 3. The model was fitted in MLWiN, using second order PQL estimation.

As mentioned before, we obtain separate fixed effects for the dependent variables. The table shows that the effects of the independent variables vary considerably. There is for example no independent variable that makes a significant contribution to explaining the use of extreme response categories. For the use of the "don't know" answer to 18 items of a scale, on the other hand, significant effects are found for female, high education, and age. Women have a larger probability to use this response alternative as well as older respondents, but for highly educated respondents this probability is smaller. For nonresponse to the income question there are also significant effects of sex (a larger probability for women), education and age (a smaller probability for less educated and older respondents). Moreover women have a larger probability to use the "no opinion" answer to items of a scale and the "don't know" answer to knowledge questions. For highly educated respondents this latter probability is smaller.

The random effects are more interesting for us. In the linearized multilevel model, the variance parameter at the first level is a dispersion parameter. It is a proportionality parameter for binomial variance. It is constrained to be 1 and therefore there are no standard errors. The constrained parameter can be released to check the binomial assumption. In the multivariate model this assumption check caused estimation problems, but univariate models showed that the assumptions of binomial variance are reasonable. At the interviewer level the variances can be interpreted as in a standard multilevel model. We find interviewer variances for all dependent variables. For the random part we do not use the star notation to denote the significance of the parameters, since a WALD test is not valid (Longford 1999). Such a test lacks power (Berkhof and Snijders 2001).

Table 3. Results of the multivariate analysis of the data quality indicators

*Fixed effects*

| | Income question | | No opinion items | | Don't know items | |
|---|---|---|---|---|---|---|
| | param. | s.e. | param. | s.e. | param. | s.e. |
| constant | −1.584 | 0.304 ** | −3.424 | 0.365 ** | −2.919 | 0.276 ** |
| female | 0.461 | 0.195 * | 0.689 | 0.224 ** | 0.618 | 0.157 ** |
| low education | −0.549 | 0.263 * | 0.482 | 0.263 | −0.061 | 0.185 |
| high education | −0.168 | 0.233 | −0.615 | 0.341 | −0.894 | 0.230 ** |
| age | −0.020 | 0.006 ** | 0.007 | 0.007 | 0.024 | 0.005 ** |
| | Don't know knowledge | | Extreme response categories | | | |
| | param. | s.e. | param. | s.e. | | |
| constant | 2.615 | 0.396 ** | 0.792 | 0.199 ** | | |
| female | 1.165 | 0.277 ** | −0.207 | 0.128 | | |
| low education | −0.516 | 0.329 | 0.224 | 0.164 | | |
| high education | −0.849 | 0.301 ** | −0.026 | 0.158 | | |
| age | 0.005 | 0.008 | 0.008 | 0.004 | | |

*Table 3. Continued*
*Random effects*

| | Param. | s.e. | Correlation |
|---|---|---|---|
| Interviewer level | | | |
| $\sigma^2_{u_{01}}$ | 0.887 | 0.281 | |
| $\sigma_{u_{0201}}$ | 0.249 | 0.197 | 0.338 |
| $\sigma^2_{u_{02}}$ | 0.613 | 0.245 | |
| $\sigma_{u_{0301}}$ | 0.032 | 0.172 | 0.040 |
| $\sigma_{u_{0302}}$ | 0.574 | 0.187 | 0.865 |
| $\sigma^2_{u_{03}}$ | 0.719 | 0.200 | |
| $\sigma_{u_{0401}}$ | −0.093 | 0.220 | −0.113 |
| $\sigma_{u_{0402}}$ | −0.213 | 0.206 | −0.309 |
| $\sigma_{u_{0403}}$ | −0.185 | 0.186 | −0.248 |
| $\sigma^2_{u_{04}}$ | 0.775 | 0.321 | |
| $\sigma_{u_{0501}}$ | 0.003 | 0.101 | 0.011 |
| $\sigma_{u_{0502}}$ | 0.118 | 0.096 | 0.435 |
| $\sigma_{u_{0503}}$ | 0.074 | 0.085 | 0.252 |
| $\sigma_{u_{0504}}$ | 0.036 | 0.108 | 0.119 |
| $\sigma^2_{u_{05}}$ | 0.119 | 0.069 | |

| | Param. | s.e. |
|---|---|---|
| Respondent level | | |
| $\sigma^2_{r_{01}}$ | 1.000 | |
| $\sigma_{r_{0201}}$ | 0.164 | 0.027 |
| $\sigma^2_{r_{02}}$ | 1.000 | |
| $\sigma_{r_{0301}}$ | 0.124 | 0.027 |
| $\sigma_{r_{0302}}$ | 0.353 | 0.023 |
| $\sigma^2_{r_{03}}$ | 1.000 | |
| $\sigma_{r_{0401}}$ | 0.037 | 0.028 |
| $\sigma_{r_{0402}}$ | 0.059 | 0.028 |
| $\sigma_{r_{0403}}$ | 0.014 | 0.028 |
| $\sigma^2_{r_{04}}$ | 1.000 | |
| $\sigma_{r_{0501}}$ | 0.021 | 0.028 |
| $\sigma_{r_{0502}}$ | 0.013 | 0.028 |
| $\sigma_{r_{0503}}$ | −0.007 | 0.028 |
| $\sigma_{r_{0504}}$ | −0.016 | 0.028 |
| $\sigma^2_{r_{05}}$ | 1.000 | |

$*p < 0.05$ $**p < 0.01$

In this table we denote the variance terms with the subscripts *r* and *u* and additional number subscripts and not with the variable names. That would only further complicate the table. *r* denotes the (co)variance terms for the respondent, *u* the (co)variance terms for the interviewer. The first number of the subscript notation denotes the independent variable (always the constant, 0); the second the dependent variable following the order in Table 3 (1 = income question, 2 = no opinion items, 3 = don't know items, 4 = don't know knowledge, 5 = extreme response categories). So, for example, $\sigma^2_{u_{02}}$ is the variance at the interviewer level for the constant in the no opinion - items equation, $\sigma_{r_{0301}}$ is the covariance of the respondent residuals for the constant in the income question and don't know items equations.

As mentioned before a log likelihood test is also not available. But most interviewer variances (except from the last one - extreme response categories) are much larger than their standard errors, which is still an indication that they are substantial. We will mainly discuss their sizes.

In a logistic model the intra-class correlation can be calculated in different ways. In the threshold model formulation, an underlying latent variable is assumed and the variance at the first level is defined as 3.290 ( $= \pi^2/3$) (see Snijders and Bosker 1999, p. 224). Following this approach the remaining intra class correlation for the income question is estimated as 0.21 ( $= 0.887/(0.887 + 3.290)$). The other estimates are 0.16 for "no opinion" items, 0.18 for "don't know" items, 0.19 for "don't know" knowledge and 0.04 for extreme response categories. Since values above 0.15 can be considered large in survey research and since we control at least for the basic socio-demographics, we can conclude that the interviewers definitely have an effect on four of the five data quality indicators.

The interviewer variances can be represented graphically, using the interviewer residuals. The multivariate model produces interviewer residuals for the five dependent variables. These residuals, however, are sample estimates with a degree of uncertainty. They have (comparative) standard errors that depend on the number of respondents for the interviewers involved and the between and within variation (Goldstein and Thomas 1996, p. 161). The residuals can be represented graphically (with the rank on the x-axis, so that they go up from the lowest to the highest) and the uncertainty can be depicted by an error bar. That error bar displays the confidence interval. With the residuals and the confidence interval, interviewers can be compared to the general mean. The graphical representation of the residuals with the [± 1.96 s.e.] confidence intervals shows the significant non-overlap with the general mean. Figure 1 reports such a representation for no opinion - items and Figure 2 displays this picture for "don't know" items. So the figures exhibit 85 interviewer residuals and 85 tests for these residuals. Both figures show several interviewers with intercepts significantly varying from the general mean. For "no opinion"

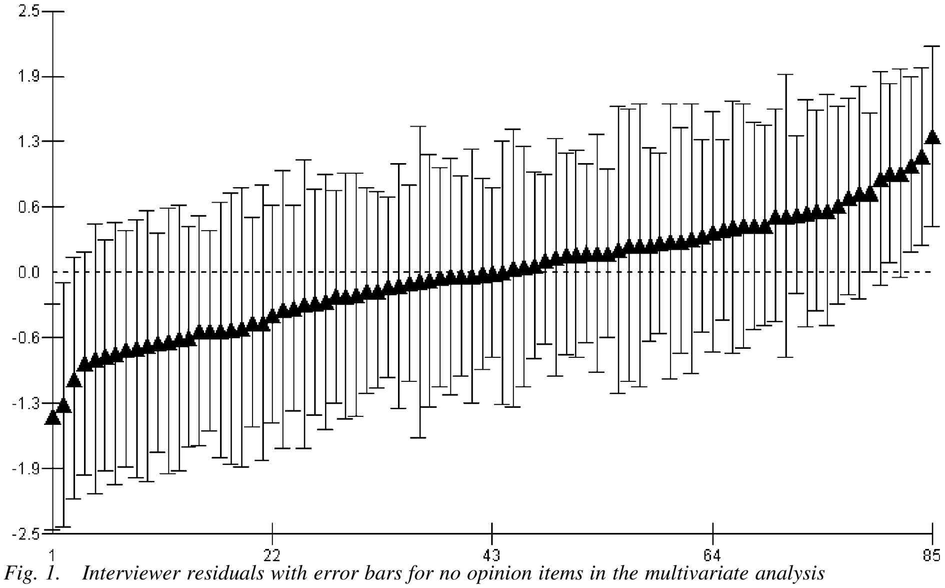

*Fig. 1. Interviewer residuals with error bars for no opinion items in the multivariate analysis*

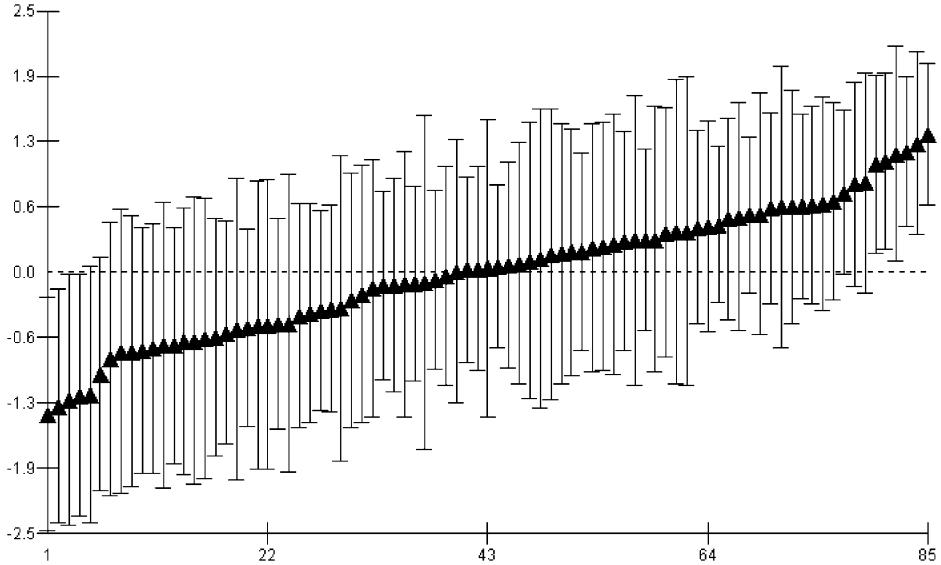

*Fig. 2. Interviewer residuals with error bars for don't know items in the multivariate analysis*

items two interviewers have significantly lower residuals and four interviewers have significantly higher residuals. For "don't know" items these numbers are four and six, respectively. We chose to display the residuals for "no opinion" items and "don't know" items, because we also display the high correlation between these residuals in the following figure. The residual plots for the other dependent variables show a similar pattern, except from the one for extreme response categories, which shows no interviewer with a residual which differs significantly from the general mean.

It should be noted that these figures involve multiple testing. When performing 85 tests, it is normal to find a number of interviewers with significantly varying intercepts (residuals that are significantly different from 0). Moreover these interviewers are not "outliers." They are extreme but they do not violate the assumption of a normal distribution.

Apart from variances at both levels we also obtain covariances (and correlations). These are the covariances after controlling for the independent variables and the variance structure at the other level. The variances and covariances at the respondent level are those for the transformed outcome variables, for which the variances are identical to one. As a consequence these respondent covariances can be interpreted as correlations. The highest correlation is the one between the chance to use the "no opinion" answer to different items and the chance to use the "don't know" answer to another set of items. This correlation amounts to 0.353. This is not surprising since they relate to similar response behaviour. Both can be labelled as a form of satisficing. Both response behaviours correlate also with the nonresponse to the income question: correlations of 0.164 and 0.124 for the "no opinion" and the "don't know" answers, respectively. Often all are considered as nonresponse and they certainly relate to a respondent reluctance to answer. So a correlation could be expected. It is rather moderate, however. Apart from these three, there are no substantive correlations at the respondent level between the various data quality indicators. The avoidance of extreme response categories can for example be labelled as satisficing as well. One could expect a negative correlation between this dependent variable and the use of the "no opinion" and the "don't know" answer. But the data do not support this hypothesis. These correlations at the respondent level can as well be calculated without partialling out the respondent variables and even without taking the multilevel structure into account. In those calculations the same three correlations stand out. The most important correlation ("no opinion" items – "don't know" items) is higher in the multilevel model without respondent variables (0.424). Both the other correlations are however smaller (0.123 and 0.108 for income – "no opinion" items and income – "don't know" items, respectively).

Also in the random part at the interviewer level correlations are estimated, controlling for the correlation at the respondent level and for all the respondent variables in the model. The largest correlation is found for the interviewer effects on the "no opinion" and "don't know" items questions: a correlation of 0.865. As mentioned before the two responses are very similar. They are both forms of satisficing. Apparently the interviewer effects for both are similar as well. Interviewers who obtain more "no opinion" answers also tend to obtain more "don't know" answers. This analysis shows that the argumentation that links answering "don't know" to satisficing is too limited when it grounds its explanation only on respondent behaviour. The interviewers have to be taken into account. The correlation between the interviewer effects can be shown by plotting both residuals jointly in a two-dimensional figure. Figure 3 displays the level two residuals for the two variables (without the error bars).

The figure clearly represents the positive correlation between the two residuals and thus the relation between the two interviewer effects. Interviewers who obtain more "no opinion" answers also tend to obtain more "don't know" answers. It may be noted that the correlation between both interviewer residuals is not exactly the same as the one reported in Table 3. The variance covariance matrix at the third level is estimated as part of the whole model and accounts for all other parameters, e.g., the variance covariance matrix

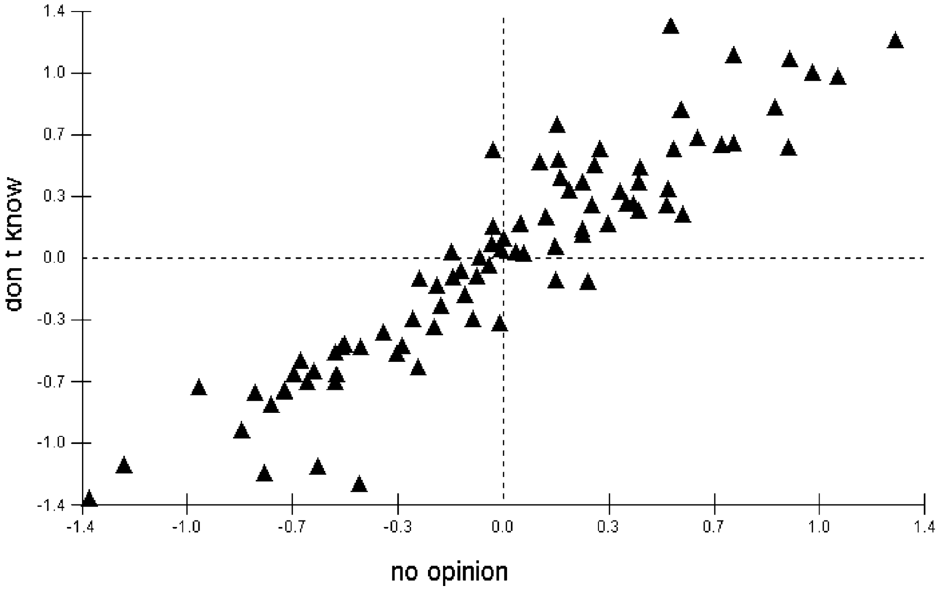

*Fig. 3. Interviewer residuals for no opinion items and don't know items in the multivariate analysis*

at the second level. The interviewer residuals are estimated or predicted afterwards on the basis of the observed data and the parameters in the model.

Table 3 also reports the nine other correlations between the interviewer effects. They are all much smaller, although one is still higher than the highest correlation at the respondent level, the correlation of the interviewer effects on the use of the no opinion answer and on the use of extreme response categories. But although it is not a valid test, it is worth noting that the covariance, corresponding with this correlation is not much larger than its standard error.

These correlations tell us something about the structure of the interviewer effects. The interviewer residuals can also be used to identify exceptional interviewers. In Figure 3 the most exceptional interviewers are the ones who are the furthest away from the central point (0,0). A distance measure could identify these interviewers. It is also possible to combine the five interviewer residuals. One can conceive these residuals as stemming from a five-dimensional space and calculate the distance to the central point in that space (0,0,0,0,0). The interviewers with the largest distance diverge the most from the central mean. Since the central point is a vector of zeros the Euclidean distance for each interviewer is simply the sum of the squared residuals. When calculating this distance measure, one can standardise the residuals in order to account for the different spread of the residuals. Using standardised residuals, every dependent variable has the same weight, whereas in our analysis the variable with the largest interviewer variance (income) will have the largest effect on this distance. The simplest standardisation only involves a division by the standard deviation of the residuals (Snijders and Bosker 1999, p. 132). Goldstein (1995, pp. 41–42) shows however that different standardisations apply depending on the purposes: model diagnostics or comparison of level two units. We tried several calculations and the order of the distance measures was always practically the same, especially for the extreme interviewers. We calculated this Euclidean distance for the raw residuals. The resulting distance measure has a mean of 1.406 with a standard deviation of 1.481 and a median of 1.013. We plotted the distance measures of all interviewers by their rank to see the pattern. Figure 4 displays this picture.

The figure shows smoothly increasing distance measures for the first 70 interviewers. Afterwards the distances increase more rapidly. At the far right-hand side of the figure three interviewers have a markedly larger distance measure than the other interviewers. Given the pattern of the figure they are outliers and they can clearly be labelled as exceptional in the analysis of the data quality indicators. We identified these interviewers to have a closer look at the answers they registered. The pattern of the interviewer with the largest distance measure (interviewer 85) is very remarkable: he or she has no item nonresponse on the income question, none of his or her 40 respondents ever used the “no opinion” or “don’t know” answer to the presented items. On the other hand all his or her respondents answered “don’t know” at least once to the knowledge questions. So for this interviewer there is actually no respondent variance on these dependent variables. Although no nonresponse is of course better than nonresponse, this finding is at least suspicious. Interviewer 84 stands out because of the “don’t know” knowledge question. Only 2 out of his or her 10 respondents used the “don’t know” answer at least once, which is very small compared to the general mean of 93%. Interviewer 83 reported nonresponse to the income question for 18 out of his or her 29 respondents, which is more than 62% and very large compared to the general mean of 11%. Furthermore this interviewer has

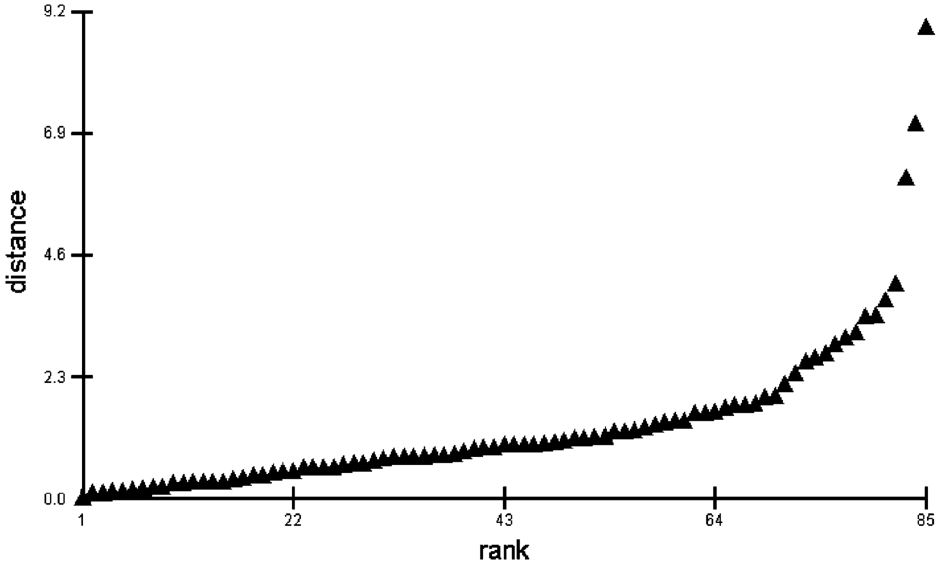

*Fig. 4. Interviewer distance measures by rank*

also no respondent variation on the “don’t know” items (no respondent with that answer) and the “don’t know” knowledge (all respondents have this answer) questions. Interviewer 82 is not a similar outlier, but his or her response pattern is also exceptional and resembles that of the first interviewer: no respondent variation on the first four variables. For the following interviewers we find similar patterns (scarce respondent variation and/or extremely high or extremely low nonresponse to one of the questions), but the patterns become less clear-cut (as expected, given the lower distance scores). This return to the original data shows that the distance measure that came out of the multivariate multilevel analysis serves well to trace interviewers with remarkable and (rather) unlikely response patterns. The survey researcher interested in the quality of the data should evaluate and control the interviews accomplished by these interviewers.

## 5. Conclusions and Discussion

In this article we showed how respondent and interviewer effects on various data quality indicators could be analysed simultaneously with a multivariate multilevel model. In such an analysis we obtain separate fixed parameter estimates alike in a standard multilevel analysis. Furthermore the model allows for an estimation of covariances at the interviewer and the respondent level. These can be used for example to examine nonresponse patterns and relations between interviewer effects.

Secondly we demonstrated a practical application. We showed how the results of the multivariate multilevel analysis could be used to identify exceptional interviewers, interviewers who have exceptional measurements for data quality indicators in the survey. The multilevel analysis produces interviewer residuals and interviewer specific intercepts (posterior means) that can be evaluated in this respect. For an interviewer who interviewed a (very) large group of respondents, these posterior means will be practically equal to the intercept of a separate regression equation for that interviewer. However for most interviewers in this survey these posterior means are pushed a bit towards the general mean (shrinkage to the mean). They can be a rather conservative appraisal of group differences (Snijders and Bosker 1999, p. 59). On the other hand the shrinkage expresses the lack of information in small groups and takes the overall population value into account (Goldstein 1995, p. 24). In that sense it is worth noting that we found that interviewer 85 and interviewer 82 registered similar response patterns. The latter is however no outlier in the distance figure, because he or she interviewed only 20 respondents compared to the 40 respondents of the former. This is a clear indication of the sensitivity of the residuals (and distances calculated out of them) to the group size.

Moreover these posterior means are sample estimates with an amount of uncertainty. They have standard errors that depend on the number of respondents for the interviewer involved and the between and within-interviewer variation. Using these residuals and their standard errors one can compare the interviewers to a general mean and to one another. Sometimes a comparison to one another will be more informative. After all, the general mean is not an objective value, indicating good data quality. When a survey researcher wants to set up a new survey and has to decide whether or not to hire interviewers again, he or she will rather compare the interviewers with one another. We certainly managed to identify interviewers with exceptional response patterns. But if we compared these interviewers to other interviewers for one of the data quality indicators, the comparison would probably show that some of the interviewers we denoted as exceptional do not differ significantly from interviewers we did not give a similar mark.

The approach used in this article relates to the differences between a fixed and a random effects model. Snijders and Bosker (1999, pp. 43–44) discuss these differences. If the researcher considers the interviewers of the survey as a fixed group and he or she wishes to draw conclusions for each (or several) of these interviewers a fixed effects model is conceptually more appropriate. In the random effects model the interviewers of the sample are regarded as a sample from a (real or hypothetical) population and the conclusions pertain to this population. In that case it is of greater relevance to search for interviewer variables that account for the interviewer variance than to identify the particular interviewers who are responsible for it. This information can then be used when training interviewers or when selecting new interviewers. But a fixed effects model is not very comfortable when a lot of interviewers are involved and the interviewer residuals will not be estimated accurately when the number of respondents per interviewer is small. Furthermore focussing on exceptional interviewers might reveal cheating, which is probably not correlated with collectable interviewer characteristics like background characteristics, experience or even interviewer style. The generalisability to the hypothetical population of interviewers is then of lesser importance than the evaluation of the data quality of the survey at hand. The research question of the identification of exceptional interviewers conceptually calls for a fixed effects model. But in a survey with a lot of interviewers and (mostly) a small number of respondents per interviewer the multilevel model proves useful. The return to the original data also showed that the residuals produced by the multilevel analysis are certainly meaningful.

The multivariate multilevel model is even more flexible than the one we presented. The included independent respondent (and interviewer) variables do not have to be the same for all dependent variables. Furthermore continuous dependent variables can be combined with categorical ones. We think that this multivariate multilevel model can have an important contribution in survey data quality research.

## 6. References

Berkhof, J. and Snijders, T.A.B. (2001). Variance Component Testing in Multilevel Models. Journal of Educational and Behavioral Statistics, 26, 133–152.

Campanelli, P. and O'Muircheartaigh, C. (1999). Interviewers, Interviewer Continuity, and Panel Survey Nonresponse. Quality and Quantity, 33, 59–76.

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

Goldstein, H. and Thomas, S. (1996). Using Examination Results as Indicators of School and College Performance. Journal of the Royal Statistical Society, Series A., 159, 149–163.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley.

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 Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds). Measurement Errors in Surveys, New York: Wiley.

Krosnick, J.A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213–236.

Longford, N.T. (1999). Standard Errors in Multilevel Analysis. Multilevel Modelling Newsletter, 11, 10–13.

Loosveldt, G. and Carton, A. (2001). Kwaliteitsevaluatie van surveys. Een toepassing op de surveys naar culturele verschuivingen in Vlaanderen. In Vlaanderen gepeild. De Vlaamse overheid en burgeronderzoek 2001, ed. Ministerie van de Vlaamse Gemeenschap, Brussels: Administratie Planning en Statistiek [In Dutch]

Marsden, P.V. (2003). Interviewer Effects in Measuring Network Size Using a Single Name Generator. Social Networks, 25, 1–16.

Ministerie van de Vlaamse Gemeenschap (2001). Vlaanderen gepeild. De Vlaamse overheid en burgeronderzoek 2001. Brussels: Administratie Planning en Statistiek [In Dutch]

Pickery, J. and Loosveldt, G. (1998). The Impact of Respondent and Interviewer Characteristics on the Number of "No Opinion" Answers. A Multilevel Model for Count Data. Quality and Quantity, 32, 31–45.

Pickery, J. and Loosveldt, G. (2001). An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse. Journal of Official Statistics, 17, 337–350.

Snijders, T.A.B. and Bosker, R.J. (1999). Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling. Newbury Park - London: Sage.

Van Tilburg, T. (1998). Interviewer Effects in the Measurement of Personal Network Size. A Nonexperimental Study. Sociological Methods and Research, 26, 300–328.

Received June 2002
Revised June 2003