Journal of Official Statistics, Vol. 27, No. 1, 2011, pp. 49–63

Assessing Mode Effects in a National Crime Victimization Survey using Structural Equation Models: Social Desirability Bias and Acquiescence

Dirk Heerwegh[1] *and Geert Loosveldt*[2]

The study compared Likert-scale responses obtained in a mixed mode survey using a telephone and a mail national crime victimization survey in Belgium. Theoretically, more socially desirable responses and more acquiescence were expected in the telephone survey. Results showed that, consistent with the social desirability hypothesis, responses were significantly more positive in the telephone survey, but no evidence was found for differences in acquiescence across the survey modes. These results were obtained with structural equation models (SEM). In addition to accounting for differences in sample composition in non-experimental data by including covariates, an SEM also allows dealing with a wide variety of mode effects not usually considered in empirical mixed mode research, such as interaction effects, differential item functioning, and the structure of measurement errors. Analyses detected an interpretable interaction effect between age and survey mode in this study, illustrating the usefulness of the SEM method in mixed mode research.

Key words: Multimode survey; hybrid survey; telephone survey; mail survey.

1. Introduction

While mixed mode surveys are rapidly gaining popularity (de Leeuw 2005), important implications of combining different survey modes are sometimes ignored in nonmethodological research. Theoretical frameworks provide reasons why answers obtained through different survey modes may be different (Dillman 2000). These so-called mode effects may occur because of differences between survey modes in cognitive burden, primacy and recency effects, question order effects, acquiescence, and social desirability (Dillman 2000; Tourangeau et al. 2000; Bowling 2005).

Different survey modes may, however, produce different answers for other reasons than mode effects alone. More specifically, different survey modes may attract different types of respondents because of characteristics like not owning a telephone or not having access to the Internet (noncoverage), or because of a selective nonresponse mechanism (Biemer 2001; Voogt and Saris 2005), such as preference for a certain survey mode (Groves and Kahn 1979). In controlled experiments, it is in principle possible to eliminate the effect of differential sample composition (e.g., by allocating the sample cases to one of the modes

[1] Katholieke Universiteit Leuven, Center for Sociological Research, Parkstraat 45 – bus 3601, 3000 Leuven, Belgium. Email: Dirk.Heerwegh@telenet.be

[2] Katholieke Universiteit Leuven, Center for Sociological Research, Parkstraat 45 – bus 3601, 3000 Leuven, Belgium. Email: Geert.Loosveldt@soc.kuleuven.be

Table 1. Balanced scale used in current analyses

Q1. It is easy for civilians to cooperate with the police force (+)
Q2. Different police forces do not cooperate well with each other (−)
Q3. Police forces have sufficient means (+)
Q4. Police forces are poorly directed (−)
Q5. Police forces work in a professional manner (+)
Q6. Police forces and the judicial system do not work together smoothly (−)

Note. Items listed in same sequence as in the questionnaire. Responses collected on a 4-point scale (original codes were 1 = completely agree, 2 = agree, 3 = disagree, 4 = completely disagree for all items, but the positively worded items, indicated with + in the table, were reverse coded for the current analyses).

after cooperation is secured). In more realistic mixed mode surveys, however, this is not possible, and covariates need to be included in the analysis to control for differences in sample composition.

The current study evaluated responses obtained via a mixed mode crime victimization survey in which a telephone survey (CATI) was combined with a mail and a web survey. The most general research question was whether responses obtained through these modes are comparable. Because the survey contained a balanced response scale (combination of positive and negative items, see Table 1), formatted as a Likert scale, the focus was on the latent variable rather than on the individual survey items. Using a (multiple group) structural equation model has the advantage that random measurement errors ("noise") can be separated from systematic errors (the "signal"). The systematic measurement errors are the possible mode and nonresponse effects.

In order to separate mode and nonresponse effects, and so to reduce the possibility of observing significant differences between survey modes because of differences in sample composition, important control variables were included in the analysis: age, sex, education level, residence type (apartment or not), and having a paid job (yes/no). These are the traditional background characteristics used in weighting adjustment procedures.

To further increase the accuracy of the analysis, acquiescence was also modeled. Likert-type scales like the one used in the current survey are often shown to be susceptible to an acquiescent or agreeing-response bias. Acquiescence is commonly defined as the tendency to agree with items irrespective of their content (Billiet and McClendon 2000; McClendon 1991). Not accounting for acquiescence may lead to biases in the assessment of the invariance of loadings of content factors across groups (Welkenhuysen-Gybels et al. 2003). Since survey modes can be considered "groups," it was advisable to include an acquiescence factor in each mode. In addition, this allowed testing whether the degree of acquiescence was equal across survey modes, and which covariates affected the agreeing-response bias.

2. Theoretical Background and Hypotheses

According to de Leeuw (1992; 2005), one of the most consistent findings in mode comparisons is that self-administered forms of data collection yield better data quality when sensitive questions are asked. The absence of an interviewer eliminates social interaction and therefore reduces respondents' tendency to take into account social norms when responding to survey questions – thereby decreasing social desirability response bias (Bowling 2005, p. 285). The crime victimization survey, from which the data used in the current study originated, included a six-item balanced Likert-type response scale gauging respondent's evaluations of how well the police perform their duties (see Table 1). The desirable response is probably to say that the police are doing a good job. The fact that the study was done on behalf of the police supports this expectation. Consequently, it was expected that the telephone survey would elicit more positive views than the mail and the web survey.

Theoretically, it can be argued that self-administered questionnaires foster less acquiescence than interviewer administered surveys (de Leeuw 2005, p. 245). This is based on the premise that the tendency of respondents to take "mental shortcuts" increases as the allowed amount of time to cognitively process the survey question decreases (de Leeuw 1992). Respondents who take "mental shortcuts," are said to "satisfice" (Krosnick 1991), by which it is meant that they do not (properly) perform all the necessary cognitive steps to answer a survey question – question interpretation, information retrieval, information integration, and mapping the response onto a response category (Tourangeau et al. 2000). Despite the appealing theoretical argument, the literature is not consistent regarding differences in acquiescence across survey modes (Hox et al. 1995; de Leeuw 2005). Consequently, this study held only a relatively weak expectation that the telephone survey would produce more acquiescence than the mail and the web survey.

Still on the issue of acquiescence, the literature suggests that this response style may be associated with age and level of education (Billiet and McClendon 2000). Since satisficing is inversely related with cognitive ability (Krosnick 1991), it is reasonable to assume that old age and lower education are associated with a decreased cognitive ability to process survey questions. This study hence hypothesized that acquiescence would increase with old age and less education. It was further assumed that these relationships would be retrieved in all the considered survey modes.

3. Data

This study used data from a crime victimization survey conducted in Belgium in 2007. The 2007 survey was a "special" survey in the sense that it was a methodological experiment set up to address some of the problems encountered by the regular crime victimization survey (called the "Security Monitor"), which is conducted biannually in the even-numbered years (the last one was conducted in 2006). The Security Monitor is conducted by telephone (CATI), but it was observed that younger respondents were increasingly being left out of this survey – possibly due to increasing numbers of mobile phone only households in the younger parts of the population (to date, no accurate sampling frame of mobile phone numbers exists in Belgium). To address this issue, a mixed mode survey was set up in 2007. The sample was drawn by a commercial fieldwork agency from a registry of postal addresses. For each of these addresses, a telephone number was looked up by matching the addresses with a list containing (landline) phone numbers and addresses. This way, sample cases with and without a landline telephone were identified. Sample cases for which no telephone number could be identified were assigned to the mail survey, while sample cases for which a telephone number was matched were allotted to the CATI survey. At the time, about 85% of Belgian households owned a landline telephone while 15% only owned a mobile phone. To ensure sufficiently large numbers of cases in each of the modes to allow meaningful mode comparisons, this survey was designed as a disproportionately stratified sample with 50% of the cases with known telephone numbers ($n = 3,000$; allocated to the CATI survey) and 50% without known telephone numbers ($n = 3,000$; assigned to the mail survey). In order to maximize the chance of recruiting younger respondents, the mail survey was complemented with a web survey. The web address (URL) along with a unique access code was printed on the invitation and reminder letters sent to the mail survey cases. Where these respondents possessed a telephone (possibly a mobile phone), they were given the opportunity to call a toll-free number to participate in the telephone survey. Only 24 sample cases called this number, of which 12 completed the survey by telephone.

The survey questionnaire normally used in the Security Monitor was overly complex for self-administration. Questions and routings were simplified, and unimode construction principles were fully applied to eliminate or maximally reduce mode effects (Dillman 2000). In total, 156 questions were asked, of which 110 were to be answered by every respondent.

The CATI survey was conducted during March and April of 2007. In total, 1,060 sample cases were interviewed by telephone, accounting for a response rate of 35.23% according to AAPOR response rate definition 1 (AAPOR 2008). The mail (and web) survey was conducted March-May 2007. In total, 3,000 sample cases were invited to participate in the survey. After three contacts (Dillman 2000), 979 completed mail surveys and 132 web surveys were received, accounting for an AAPOR definition 1 response rate of 37.66% (excluding 50 nonexistent addresses).

The web survey resulted in relatively few completed cases, and analyses (not reported here) showed that if the web survey data were left out altogether no different conclusions were reached. There are also indications in the literature that the differences between mail and web are very small and often negligible (Denscombe 2006; for an overview, see de Leeuw 2005). Therefore the data from the web survey were collapsed with those from the mail survey. The mode comparison conducted in this study hence involved a telephone versus a mail (and web) survey.

Analyses on sample composition showed some differences between the two completed samples. No significant differences were found for the distributions of sex, age, and whether or not the respondent holds a paid job (see Table 2). The mail and web surveys, however, did contain significantly more respondents with a lower education level. With respect to residence type (dichotomized variable: apartment vs. not apartment), the distributions were extremely different (see Table 2). This was caused by the matching procedure used by the commercial fieldwork agency. Although the telephone registry included the street address and number, it did not always include the apartment number. As a consequence, sample cases living in an apartment and selected to be included in the sample could not always be matched with a number in the telephone registry. To control for differences between modes, we will use all the variables in Table 2 as covariates in our analysis. By doing so, we can specify and evaluate interaction effects between these variables and mode.

Table 2. Sample characteristics by survey mode

Characteristic	Mail/web	CATI	Statistical test
Sex	$n = 1,111$	$n = 1,060$	
Male	45.18%	45.00%	
Female	54.82%	55.00%	$\chi^2(1) = 0.01; p = 0.93$
Education	$n = 1,086$	$n = 1,060$	
None	8.10%	2.26%	
Level 1	12.15%	11.23%	
Level 2c	10.96%	4.06%	
Level 2b	7.37%	6.42%	
Level 2a	4.60%	8.40%	
Level 3c	8.01%	7.36%	
Level 3b	12.34%	13.77%	
Level 3a	8.93%	11.13%	
Level 4	19.52%	26.13%	
Level 5	8.01%	9.25%	$\chi^2(9) = 96.86; p < 0.0001$
Paid job	$n = 1,111$	$n = 1,060$	
Yes	48.51%	48.02%	
No	51.49%	51.98%	$\chi^2(1) = 0.05; p = 0.82$
Residence type	$n = 1,065$	$n = 1,050$	
Apartment	40.75%	4.29%	
Other	59.25%	95.71%	$\chi^2(1) = 401.32; p < 0.0001$
Age	$n = 1{,}086$	$n = 1{,}060$	
Mean	48.06	49.48	
s.d.	17.89	17.08	$t(2,144) = -1.89; p = 0.06$

Note on education. The levels are ranked from low to high. In the Belgian education system, distinction is made between primary (ages 7–12, Level 1), secondary schooling of the first cycle (ages 13–15, Level 2), secondary schooling of the secondary cycle (ages 16–18, Level 3), higher education outside of universities (Level 4), and higher education within universities (Level 5). Within the secondary schooling system (both cycles), distinction is made between education preparing for skilled manual jobs (Level c), education preparing for more highly skilled manual work and the possibility to participate in higher education – often outside of universities (Level b), and education preparing for higher education in colleges or universities (Level a).

4. Method

In accordance with the guidelines described by Welkenhuysen-Gybels et al. (2003, p. 710), the current study tested several models in a specific sequence. First, a measurement model that included only the latent construct that the indicators were assumed to measure was fit in each group separately. This is a test of an elementary confirmatory factor analysis in which the precise structure of the factor model is specified a priori. Using the items of Table 1, the latent construct or the content factor only model (Model A in Figure 1) can be considered as a general opinion about police forces. In a second step, the response style factor (acquiescence) was included as a second factor in this model (Figure 1, Model B). The loadings of this factor on the items are identical for each item. This factor measures the tendency to agree with positively and negatively worded items. In a previous application of this kind of measurement model the style factor correlated very highly (0.90) with the number of times the respondent agreed to a balanced set of items (positive and negative items about the same topic) (Billiet and McClendon 2000). In general, it can be said that if the model in the second step fits significantly better

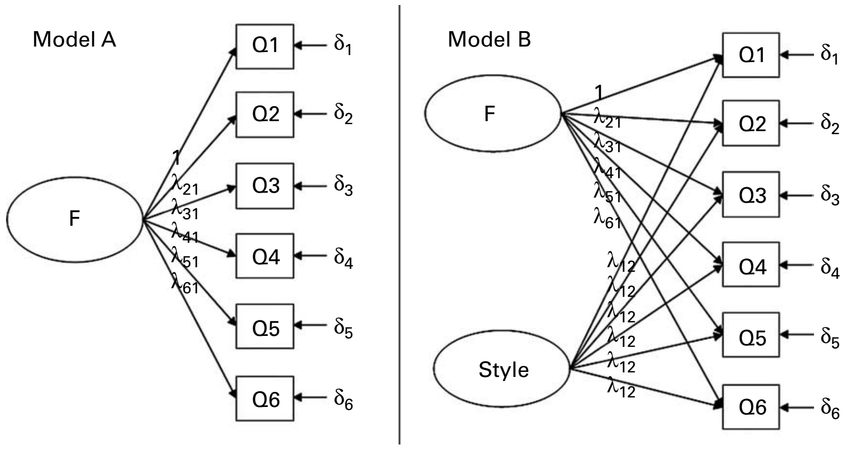

Fig. 1. Content factor only model (Model A), and model including a style factor (Model B)

than the model in Step 1, then the acquiescence factor is needed to make meaningful comparisons across groups (Welkenhuysen-Gybels, Billiet, and Cambré 2003, p. 710). Steps 1 and 2 relate to models tested in the different groups separately and are considered as an assessment of the baseline model (see section 5.1.). The groups in the present application are the survey modes. In a third step a multiple group analysis was conducted in which the parameters in the various groups were simultaneously estimated. This step tested whether the instrument measures the same construct in different modes. This test of construct equivalence compares a model that constrains the factor loadings of the items equal across modes to a model that does not impose any equality constraints with respect to the factor loadings. If the null hypothesis of this test cannot be rejected, the loadings are invariant and the construct is equivalent across groups (van de Vijver and Leung 1997). In general it can be said that if it is necessary to free factor loadings to improve model fit, noninvariance is found for the involved indicators. If factor loadings are equal in both modes (acceptance of imposed equality restrictions across modes), we obtain construct equivalence in the modes (see Section 5.2., assessment of construct equivalence).

Because the current study included covariates to account for differences in sample composition, these covariates were included before latent means were compared across the survey modes. Including covariates effectively modifies the comparison of the latent means into a comparison of the intercepts of the latent variables. In this analysis, all covariates were allowed to affect the content factor, but only age, education and sex were allowed to influence the style factor. The reason was that only variables measuring "cognitive ability" should influence the style factor, so it made little sense to let the "paid job" and the "residence type" variables affect the style factor. There are no mode effects when a model holds with equality restrictions across modes on all parameters (see Section 5.3.: assessment of mode effects).

Model testing was performed using Mplus version 4.0 (Muthén and Muthén 1998-2007). Because the dependent variables were ordinal, the Robust Weighted Least Squares estimator was used (Jöreskog 2005). The variables were declared ordinal in the Mplus program file in accordance with the program guidelines (Muthén and Muthén 1998-2007).

5. Results

5.1. Assessment of the Baseline Model

The first step in the testing strategy is the determination of the baseline model for each group separately. Our hypothesized baseline model is the model with the content factor and the response style factor (Model B in Figure 1). The first analysis concerned whether or not the style factor needed to be included in this model. Therefore a measurement model was tested in each group separately without a style factor (Model A in Figure 1). Then, the same measurement model was tested with a style factor added (Model B in Figure 1) and it was determined whether the fit of the model including the style factor was better than that of the model without the style factor.

To scale the content factor F, λ_{11} was set to 1 (implicating that a higher latent variable score referred to a more positive opinion about the functioning of the police), the other factor loadings were estimated freely. To identify Model B shown in Figure 1, the covariance between the content factor and the style factor was fixed at zero. In accordance with common practice, the factor loadings on the style factor were set equal across all items (λ_{12} in Figure 1) because it was assumed that all items were influenced by the style factor to the same degree since they all shared the same response format. There is no reason to assume that any of the items would be differently affected by acquiescence (Billiet and McClendon 2000; Welkenhuysen-Gybels et al. 2003).

Table 3 shows the fit statistics for both Models A and B in both survey modes separately. It can be seen that Model A had a relatively poor fit in the telephone survey mode with an RMSEA value of just over 0.08 and a CFI value below 0.90. Including the style factor significantly improved model fit as indicated by the chi-squared difference test (the DIFFTEST procedure as described in Muthén & Muthén (1998–2007) was used to calculate the chi-squared difference test. All other chi-squared difference tests reported in this text were also calculated using this procedure). Model A had also a very poor fit in the mail and web survey mode, with an RMSEA of 0.123 and a CFI of 0.836. Similarly to the telephone survey mode, model fit significantly improved when including the style factor. However, even with the style factor included, model fit was not very satisfactory. Because including covariates normally increases the amount of information and generally leads to better model fit, rather than rejecting these models, they were developed further.

In the next step of the assessment of the baseline model the nature of the response style factor was determined. Even though it was previously called an acquiescence style factor,

Table 3. Fit statistics for the content factor model only (Model A) and the model including the style factor (Model B) per survey mode

	Model A			Model B			Chi-squared difference test	
Mode	χ^2	RMSEA	CFI	χ^2	RMSEA	CFI	$\Delta\chi^2$	p value
CATI $n = 548$	36.38	0.081	0.888	21.78	0.062	0.942	10.96	0.0009
Mail/web $n = 431$	60.33	0.123	0.836	29.41	0.086	0.930	22.15	< 0.0001

it could not be ruled out that it was another response style such as a middle-alternative effect (Billiet and McClendon 2000, p. 621). Evidence that this style factor was acquiescence could be obtained by correlating that factor with a variable that counted the number of times the respondent agreed with each of the six statements in the scale (this is called "scoring for acquiescence" and is a procedure described in detail by Billiet and McClendon 2000). The correlations between the style factor and the count variable are very high and significant at the level 0.0001: 0.782 in the CATI mode and 0.836 in the mail/web mode. These correlations suggest that the style factor indeed represented acquiescence. For that matter, it can be noted that the style factor was positively correlated with the count variable in both groups. It is therefore unlikely that the factor represents a response-order effect. In visual survey modes (such as the mail and web surveys in the current study), a primacy effect is indistinguishable from acquiescence if the first response option listed is the "completely agree" category (which was the case in the current study). The positive correlation between the style factor and the count variable could hence just as well point to a primacy effect. However, since the opposite response-order effect – recency – is expected in auditory survey modes (Krosnick and Alwin 1987), a *negative* correlation between the style factor and the count variable should be found in the telephone mode if the style factor represented a response-order effect. As this was not the case, it was concluded with relative certainty that the style factor represented acquiescence.

5.2. Assessment of the Construct Equivalence

Do we have the same measurement model in each mode? This is the core question of the assessment of construct equivalence or measurement invariance between modes. For this analysis, covariates were included. The model allowed an effect of all covariates (age, sex, education, residence type, and employment status) on the content factor. In contrast, only age, sex and education were allowed to exert an influence on the style factor (see Figure 2). Age was a continuous variable, while a dummy variable Male represented sex (1 = male, 0 = female). Education level was represented by two dummy variables, Educ1 (education Levels 2c through 3b; see Table 2) and Educ2 (Levels 3a, 4, and 5). The reference category contained the respondents with no education or only Level 1 education. The dummy variable Job reflected whether or not the respondent held a paid job (1 = yes, 0 = no). Residence type was reflected in the dummy variable Aprtm (1 = apartment, 0 = other).

A first invariance test pertained to configural invariance (Byrne 1998). This means that no equality constraints are imposed on the parameters and only the number of factors and the pattern of factor loadings are the same across groups. Because no equality constraints are imposed the factor structure for each group must be similar but not identical. As expected the fit improved when covariates were added to Model B in Table 3: Mail/web: RMSEA = 0.039, CFI = 0.945; CATI: RMSEA = 0.044, CFI = 0.901. With these values of the fit statistics one can conclude that the same configural model fit reasonably well to the data in each of the groups. When Model C with the response style factor and the content factor (see Figure 2) was simultaneously fit in both groups and no restrictions were placed on any of the parameters, RMSEA equaled 0.042 and the CFI value was 0.936. This provided further evidence of configural equivalence in each mode.

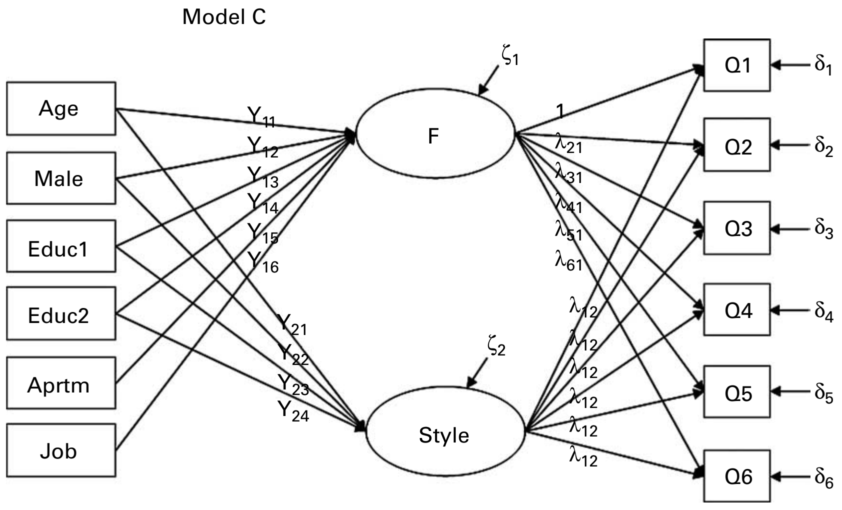

Fig. 2. Measurement model and structural model with covariates

Factorial invariance was tested next. The factorial invariance refers to the extent to which a factor model that is assumed to hold in a general population also holds in subpopulations. Tests for invariance require the imposition of equality constraints across groups. Model D imposed the restriction that all loadings (and thresholds) on the content factor as well as on the style factor were equal across the modes. Comparison of RMSEA and CFI values across Models C and D revealed no deterioration of model fit. Model D actually obtained better RMSEA and CFI values (0.039 and 0.942, respectively) than Model C. The more formal chi-squared difference test revealed no significant fit deterioration either ($\Delta\chi^2(6) = 3.79$, $p = 0.71$). It was therefore concluded that factorial invariance was present across the modes. There are no mode effects on the measurement model. This means that the same measurement model with a content factor and a response style factor holds in both modes. In the next section more specific mode effects are evaluated.

5.3. Assessment of Mode Effects

To test for more specific mode effects in the data, additional restrictions were placed on the model. The theoretical reasoning was that if the data were not influenced by mode effects at all, both survey modes should yield the same results: similar attitudes toward the police should be reported (as evidenced by equal intercepts of the content factor), a comparable level of acquiescence should be observed in both survey modes (indicated by equal intercepts of the style factor), and the effects of the covariates exerted on the content and on the style factor should be similar across survey modes. If these conditions were met, it could be concluded that no mode effects were present at all, as none of the model parameters would deviate across the survey modes. Model E placed these restrictions onto the model: all factor intercepts and all effect parameters (gammas) were constrained to be equal across the survey modes. Constraining the gammas to be equal across survey modes does not imply that the sample composition was constrained to be equal across the survey modes. These restrictions only express that the effect of the covariates on the factors is constrained to be equal across survey modes. Not surprisingly, Model E had a poor fit, with an RMSEA value of 0.066 and a CFI of 0.855. The chi-squared difference test showed a significant decrease in model fit compared to model D ($\Delta\chi^2(6) = 42.13$, $p = 0.00$).

In order to improve model fit one can undo some of the equality restrictions. One can use Modification Index (MI) values to identify these restrictions. The largest MI value (45.78) was associated with the intercept of the content factor. Freeing this model parameter produced the better-fitting model F (RMSEA = 0.046; CFI = 0.915). The model output showed that the content factor had a significantly higher intercept in the telephone than in the mail survey. This indicated that controlling for the covariates and acquiescence, respondents interviewed by telephone gave more positive evaluations of the functioning of the police. This finding was consistent with the social desirability response bias hypothesis formulated earlier. So the fact that the equality restriction on the intercept of the content factor is not possible is informative about the mode effect on the general opinion about police.

Even though Model F exhibited a relatively good fit, the model still fits significantly worse than model D ($\Delta\chi^2(7) = 21.63$, $p = 0.00$). Inspection of the MI values revealed that one more parameter had a relatively high MI value (11.93). This was a gamma parameter reflecting the effect of age on the content factor (γ_{11}). Freeing this parameter produced the well-fitting Model G (RMSEA = 0.041; CFI = 0.933). The model output showed that age had a significant (positive) effect on the evaluation of the police in the mail survey, whereas no such effect was present in the telephone survey. In the telephone survey, age apparently did not influence the evaluation of the police at all. This finding pointed to an interaction effect between a survey mode effect and a background characteristic. To illustrate the finding, the expected scores on the content factor (evaluation of the police) were calculated for a 20-year-old and for a 60-year-old in both survey modes. While calculating these expected factor scores, the remaining covariates were set to 0. Hence, these values are valid for females (male = 0) of the lowest education group (educ1 = 0 and educ2 = 0), who do not live in an apartment (apartment = 0), and who do not have a paid job (job = 0). The estimated scores on the content factor are shown in Figure 3. The figure clearly shows a positive effect of age on the content factor in the mail survey, while this effect was absent in the telephone survey (the slight decrease in the telephone group was not statistically significant).

It can be seen from the figure that for some reason, the mode effect was more pronounced for younger respondents than for older respondents. One possible interpretation is as follows. If it is assumed that the positive relationship between age and the evaluation of the police exists in the population, the mail survey correctly revealed this relationship. The question then becomes why this relationship was not present in the telephone survey. It is possible that the relationship disappeared in the telephone survey because the effect was overwhelmed by the social desirability response bias: if everyone (old and young) gave a positive evaluation of the police, a ceiling effect might have occurred causing the existing relationship to disappear. Alternative interpretations could be offered, but the main conclusion of the analysis was that mode effects need not be restricted to simple main effects as is often assumed in studies on mixed mode surveys (by e.g., only investigating the averages across the groups).

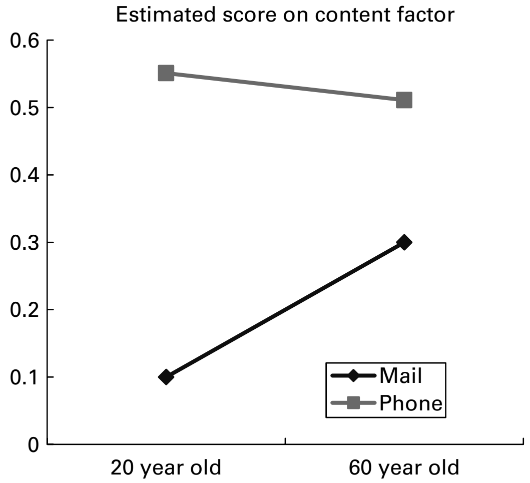

Fig. 3. Estimated factor scores for a 20 year old and a 60 year old, in both survey modes

One more technical note, it emerged that while Model G fit the data relatively well, the chi-squared difference test suggested that Model G fit the data less well than Model D ($\Delta\chi^2(7) = 14.62$, $p = 0.04$). However, ΔCFI equaled -0.009, which is smaller than the cutoff value of -0.01 proposed by Cheung and Rensvold (2002, p. 251), suggesting that Model G fit the data just as well as Model D did. In addition, the model output did not contain any MIs of substantial value. Therefore, Model G was accepted as the final model. Table 4 shows the parameter values of Model G. Except for the intercept of the factor concerning the opinion about police and the effect of age on this factor (γ_{11}), all parameters are restricted to be equal across modes. This pattern of equality constraints and free parameters makes an assessment of mode effects possible.

6. Discussion and Conclusion

This study sets out to investigate mode effects between a mail (and web) survey on the one hand and a telephone (CATI) survey on the other. Because this mixed mode survey was not a fully controlled methodological experiment in which nonresponse was randomly distributed across survey modes, control variables were included in the models to avoid confounding mode effects with differences in nonresponse bias across the survey modes. Care was taken to include as many relevant control variables as possible.

Besides controlling for relevant covariates, the present study also included a response style. Since the item battery on which the analyses were performed was balanced and responses were obtained on a continuum ranging from completely agree to completely disagree, acquiescence was the most obvious response style to include in the model.

Because of the interaction with an interviewer, it was hypothesized that the telephone survey mode would induce a social desirability response bias. More specifically, more positive evaluations of the police (the topic of the item battery) were expected in that mode as compared with the mail (and web) survey, controlling for sample composition and acquiescence. Regarding the response style, it was hypothesized that the telephone survey would engender more acquiescence mainly because the pace of the interview is dictated by

Table 4. Parameter values for Model G (all parameters constrained to be equal across groups except for the content Factor intercept, and gamma 11)

	Mail/web			CATI		
Parameter	Param. value	t−value	Stand. par.	Param. value	t−value	Stand. par.
λ_{11}	1.00 (fixed)	–	0.34	1.00 (fixed)	–	0.27
λ_{21}	−1.73	−6.24***	−0.58	−1.73	−6.24***	−0.50
λ_{31}	0.80	5.08***	0.27	0.80	5.08***	0.24
λ_{41}	−1.76	−6.89***	−0.59	−1.76	−6.89***	−0.52
λ_{51}	1.65	7.24***	0.54	1.65	7.24***	0.56
λ_{61}	−1.76	−6.39***	−0.59	−1.76	−6.39***	−0.52
λ_{12}	1.00 (fixed)	–	0.26	1.00 (fixed)	–	0.26
γ_{11}	0.01	3.24**	0.24	−0.00	−1.20	−0.07
γ_{12}	0.05	1.60	0.07	0.05	1.60	0.07
γ_{13}	0.00	0.04	0.00	0.00	0.04	−0.00
γ_{14}	−0.03	−0.72	−0.04	−0.03	−0.72	−0.04
γ_{15}	0.03	0.75	0.04	0.03	0.75	−0.02
γ_{16}	−0.05	−1.75	−0.08	−0.05	−1.75	−0.08
γ_{21}	0.00	3.69***	0.24	0.00	3.69***	0.25
γ_{22}	0.00	0.04	0.00	0.00	0.04	0.00
γ_{23}	−0.05	−0.94	−0.09	−0.05	−0.94	−0.09
γ_{24}	−0.10	−1.93	−0.18	−0.10	−1.93	−0.19
Intercepts						
Content F	0.00	0.00	0.00	0.57	4.00***	1.70
Style F	0.00	0.00	0.00	0.00	0.00	0.00

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Stand. par. = standardized parameter value.

the interviewer – who may have a tendency to keep that pace rather high. This would make it more difficult for people to perform the necessary cognitive process to answer the survey questions, resulting in "satisficing" behavior such as simply agreeing with the statements read out by the interviewer (acquiescence). It was further hypothesized that acquiescence would depend on age and education level of the respondent, in both the mail (and web) survey and the telephone survey.

Using structural equation modeling, it was found that including the acquiescence factor significantly improved model fit. This implied that this response style should be included before conducting multiple group comparisons, which is in line with findings from previous studies (Welkenhuysen-Gybels et al. 2003). The multiple group comparisons showed that not only did factorial invariance hold for the content factor, but also the loadings of the response style factor were equal across groups. Holding all model parameters equal across survey modes led to poor model fit, indicating that the results in both modes were not identical and some mode effects were present in the data. Further analyses revealed that the intercept of the content factor could not be held equal across the survey modes: the telephone mode generated significantly more positive evaluations of the police, while controlling for sample composition and response style. This finding provided strong support for the social desirability hypothesis.

In addition, it was found that the across group equality restriction of one effect parameter needed to be relaxed. In the mail survey, older respondents held more positive views toward the police, while no such effect was found in the telephone survey. This hints at possible interaction effects between survey mode switches and background characteristics of respondents. While the literature on mode effects so far has concentrated mainly on differences in means across the survey modes (and thereby only on main effects), the present study shows that interaction effects should not be neglected. It makes theoretical sense to include interaction effects because switching from one survey mode to another, more difficult survey mode might for instance be disproportionately more challenging for less educated or older respondents than for highly educated young respondents. Such effects can be picked up by modeling interactions.

In summary, this mode comparison suggests that in comparison to a mail (and web) survey, a telephone survey may upwardly bias the responses because of a social desirability response bias, thereby potentially camouflaging existing relationships between background variables and the factor scores because of interaction effects between survey mode changes and background characteristics of respondents. In contrast to the former conclusion, the latter is relatively new and makes an important addition to the literature, which so far has not paid much attention to interaction effects.

The analysis also revealed that the intercept of the style factor could be held equal across groups. This rejected the hypothesis of more acquiescence in the telephone survey than in the mail (and web) survey. Though rather surprising, this does echo results from various other published mode comparisons (de Leeuw 1992; Fricker et al. 2005).

The present study used structural equation models to detect different types of mode effects simultaneously (social desirability and acquiescence). As far as the authors know, this is unprecedented in the mixed mode survey research field. Because multi-group structural equation modeling allows researchers to detect the presence of mode effects while controlling for differences in sample composition and response styles, using structural equation modeling could make important additions to our knowledge concerning mode effects. Moreover, since mode effects might be associated with any of the many parameters available in the structural equation model, much more detailed understanding about mode effects could be gained from these models than from more traditional analysis techniques. Assessment of equality restrictions across modes is the point of special interest. Different intercepts point to main effects of switching from one mode to another, while different effect parameters signal interaction effects between background characteristics and the mode switch. The present study provided illustrations of these types of mode effects. Other mode effects are also theoretically possible, although they did not occur in the present study: a different factor loading indicates *differential item functioning*, meaning that an indicator operates differently in one mode as compared to another. And different error covariances might suggest important differences in the structure of measurement errors across the survey modes. It is hoped that the current study stimulates researchers to use this method to investigate mode effects in both experimental and non-experimental data. At the same time, the structural equations method should be subjected to rigorous tests (e.g., by conducting simulation studies) to fully grasp the potentials as well as the possible shortcomings of this method in the domain of mixed mode surveys. This constitutes an important research agenda for the mixed mode research community over the next years.

7. References

AAPOR (2008). Standard definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 5th Edition. Lenexa, KS: AAPOR: AAPOR http://www.aapor.org/uploads/Standard_Definitions_07_08_Final.pdf

Biemer, P.P. (2001). Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. Journal of Official Statistics, 17, 295–320.

Billiet, J.B. and McClendon, M.J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. Structural Equation Modeling, 7, 608–628.

Bowling, A. (2005). Mode of Questionnaire Administration Can Have Serious Effects on Data Quality. Journal of Public Health, 27, 281–291.

Byrne, B.M. (1998). Structural Equation Modeling With LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming. Mahwah, N.J. Erlbaum.

Cheung, G.W. and Rensvold, R.B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. Structural Equation Modeling, 9, 233–255.

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

de Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. Journal of Official Statistics, 21, 233–255.

Denscombe, M. (2006). Web-Based Questionnaires and the Mode Effect. An Evaluation Based on Completion Rates and Data Contents of Near-Identical Questionnaires Delivered in Different Modes. Social Science Computer Review, 24, 246–254.

Dillman, D.A. (2000). Mail and Internet Surveys: the Tailored Design Method. New York, NY: John Wiley.

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. Public Opinion Quarterly, 69, 370–392.

Groves, R.M. and Kahn, R.L. (1979). Surveys by Telephone: a National Comparison With Personal Interviews. New York, NY: Academic Press.

Hox, J., de Leeuw, E.D., and Vorst, H. (1995). Survey Participation As Reasoned Action; a Behavioral Paradigm for Survey Nonresponse? Bulletin de Méthodologie Sociologique, 48, 52–67.

Jöreskog, K.G. (2005). Structural Equation Modeling With Ordinal Variables Using LISREL. Retrieved March 10, from http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf.

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

Krosnick, J.A. and Alwin, D.F. (1987). An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. Public Opinion Quarterly, 51, 201–219.

McClendon, M.J. (1991). Acquiescence and Recency Response-Order Effects in Survey Interviews. Sociological Methods and Research, 20, 60–103.

Muthén, L.K. and Muthén, B.O. (1998-2007). Mplus. Statistical Analysis With Latent Variables. User's Guide, (Fourth Edition). Los Angeles, CA: Muthén & Muthén.

Tourangeau, R., Rips, L.J., and Rasinski, K.A. (2000). The Psychology of Survey Response. Cambridge: Cambridge University Press.

van de Vijver, F. and Leung, K. (1997). Methods and Data Analysis for Cross-cultural Research. Thousand Oaks, CA: Sage.

Voogt, R.J.J. and Saris, W.E. (2005). Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects. Journal of Official Statistics, 21, 367–387.

Welkenhuysen-Gybels, J., Billiet, J., and Cambré, B. (2003). Adjustment for Acquiescence in the Assessment of the Construct Equivalence of Likert-Type Score Items. Journal of Cross-Cultural Psychology, 34, 702–722.

Received April 2009
Revised August 2010