Statistics Netherlands

Discussion Paper

The impact of Survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey

2014 | 16

Dirkjan Beukenhorst
Bart Buelens
Frank Engelen
Jan van der Laan
Vivian Meertens
Barry Schouten

23-05-2014

THE IMPACT OF SURVEY ITEM CHARACTERISTICS ON MODE-SPECIFIC MEASUREMENT BIAS IN THE CRIME VICTIMISATION SURVEY

Summary: The interaction between survey item characteristics and survey mode is investigated in terms of measurement bias. For this purpose a typology of survey items was constructed. Mode-specific measurement bias was estimated using a large-scale mixed-mode experiment linked to the Crime Victimisation Survey 2011. This experiment consisted of a randomized allocation of sample persons to the four survey modes web, paper, telephone and face-to-face. A multi-level model is used to explain the mode-specific measurement bias for a set of approximately 125 items from the Crime Victimisation Survey.

Keywords: Mode effects, Measurement effects, Mixed-mode surveys.

1. Introduction

This paper is the last in a series of three discussion papers that analyze and discuss a large-scale mixed-mode experiment conducted in 2011. The experiment was designed and conducted as a part of the project Mode Effects in Social Surveys (in Dutch: Mode Effecten in Persoonsstatistieken or MEPS). The project was initiated after finding large mode effects in the Dutch Crime Victimization Survey (CVS), see Buelens and Van den Brakel (2011). Although the project focuses mainly on the CVS, questionnaire sections form the Dutch Labour Force Survey (LFS) and the European Social Survey (ESS) were also included. A detailed background and motivation of this project is given by Schouten (2010). An overview of all the project analyses and papers is given in Schouten and Klausch (2011).

The MEPS project had three main goals:

1. To estimate mode-specific selection and measurement bias for a number of key statistics from the LFS and the CVS.
2. To better understand the relation between type of person, nature of survey question, mode-dependent nonresponse and answering behaviour.
3. To give recommendations to improve mixed-mode survey methodology (data collection strategies, nonresponse adjustment methods and questionnaire design) in the CVS and LFS.

The first two discussion papers dealt with the first and third goal, respectively. The first discussion paper (Buelens, Van der Laan, Schouten, Van den Brakel, Burger and Klausch 2012) developed a mathematical framework for the mode effect components: mode-specific coverage bias, mode-specific nonresponse bias and mode-specific measurement bias. It also constructed estimators for the three components and applied the estimators to key survey variables from the CVS and LFS. Detailed descriptions of these estimates can be found in Buelens, Van der Laan and Schouten (2012) and Schouten, Van den Brakel, Buelens, Van der Laan and Klausch (2013). In this paper, a similar decomposition of mode effects is made for all

CVS survey items. In the second discussion paper (Schouten, Cobben, Van der Laan and Arends-Toth 2014) the focus was on the third goal by linking mode effects to contact effort and interviewer performance.

The current paper focusses on the second main goal: the relation between type of person, type of survey item and mode-specific measurement bias. In this paper coverage and nonresponse bias are not considered.

Why is it meaningful to look at mode-specific measurement bias as a function of type of person and type of survey item? Mode-specific measurement bias can be a major component to mode effects; for the CVS survey items it indeed turned out to be the dominant component. In order to achieve comparability over time and between population subgroups that respond to different modes, it is imperative that mode effects can be kept at the smallest possible level. The key to minimal measurement bias lies in questionnaire design for multiple modes. However, even with careful questionnaire design, mode-specific measurement biases occur. In order to be better able to predict the risk of such biases in advance and to be better able to perform questionnaire design to multiple modes, it would be very helpful to know how such biases relate to the type of person and type of survey item. If such dependencies can be detected and explained, then future questionnaire design may profit from this knowledge and be more effective in ensuring comparability.

The quest to understand and explain measurement bias, and in particular mode-specific measurement bias, is not new. The problem of mode-specific measurement bias is several decades old, but regained new impetus due to the rise of web surveys. There is a vast literature about mode differences and implications for survey statistics. Overviews and general discussions of mode effects are presented in De Leeuw (2005), Jäckle, Roberts and Lynn (2010), Dex and Gumy (2011) and Klausch, Hox and Schouten (2013a). Specific discussions of the impact of the survey mode on the answering process can be found in De Leeuw (1992), Holbrook, Green and Krosnick (2003), Fricker, Galesic, Tourangeau and Yan (2005), Greene, Speizer and Wiitala (2008), Chang and Krosnick (2009) and Klausch, Hox and Schouten (2013b). However, most of the existing literature based analyses and conclusions on measurement biases that were confounded with selection biases; they were not or only partially adjusted for different selections of respondents. In the present study this confoundedness is overcome by the special experimental design.

Even with adjusted estimates of measurement biases, the quest to understand them is complex. First it needs to be decided what is meant by type of survey item, i.e. a typology or classification of items along various dimensions must be made. In this paper such a typology is proposed and applied to the CVS survey items. This typology shows a strong resemblance to Saris and Gallhofer (2007) and Campanelli et al (2011). A number of additions, adaptations and omissions were made in order to make the typology fit for an investigation of the impact of the survey mode and in order to avoid characteristics that turned out rare in the CVS. The resulting typology consists of a set of 26 characteristics.

The dependencies between measurement bias, type of person and type of survey item are modelled by multilevel models. The survey items form a level within individuals, so that it is possible to investigate between survey item variation and between individual variation. The estimated mode-specific measurement biases are modelled and analyzed on two levels of aggregation: the item level and the category level. Models for the item level relate the total absolute change in the distribution of a categorical CVS survey item over its categories to the characteristics of the item. In the analysis of category level mode-specific measurement bias, the analysis goes a step further and investigates response styles. Many authors (e.g.

Tourangeau and Rasinski 1988, Krosnick 1991, Greenleaf 1992, Billiet and McGlendon 2000, Baumgartner and Steenkamp 2001, Heerwegh and Loosveldt 2011, Lynn and Kaminska 2012 and Aichholzer 2013) have discussed response styles, or, more generally, deficiencies in answering behaviour that persist throughout a large part of the questionnaire. Respondents are believed to provide answers to a survey item following four cognitive steps: Interpretation, Information retrieval, Judgment and Reporting. For various reasons, consciously or unconsciously, respondents may circumvent one of the steps or may move quickly through one or more steps. If a respondent repeatedly shows this behaviour, then one speaks of a response style. Well-known examples are social desirable answering, acquiescence, disacquiescence, primacy, recency, straightlining, non-differentiation and telescoping. A larger class of styles is satisficing, which consists of all styles where respondents circumvent the first three steps and move to the Reporting step directly. Many of the response styles are conjectured to relate to the survey mode. Five response styles are considered for the CVS: don't know (DK) answers, primacy, recency, extreme response style and straightlining. The proportion of DK answers in the CVS is considerable and is conjectured to cluster within individuals. Extreme response style and straightlining are forms of satisficing behaviour, i.e. behaviour where respondents shortcut the answering process. Primacy and recency result from a mixture of satisficing behaviour and memory effects. The response styles are conjectured to be mode-specific and to play an important role in the CVS. The analysis on the category level is similar to the analyses in Billiet and McGlendon (2000), Heerwegh and Loosveldt (2011), Lynn and Kaminska (2012), Bennink, Moors and Gelissen (2013), Cordova Cazar and Powell (2013) and Klausch, Hox and Schouten (2013b), although some of these authors use different statistical approaches like structural equation models in which multiple survey items are linked through latent factors.

It must be noted that the CVS questionnaire consists of a specific choice of survey items. The items are often closely related and are clustered within questionnaire sections. This implies that care is needed when interpreting and generalizing the results from the analyses. Nonetheless, the CVS is a useful instrument to explore the relation between item properties and measurement bias; it was selected because of its wide variety of survey items and its conjectured sensitivity to mode-specific measurement bias.

It must also be noted that this paper is an exploratory analysis. No explicit hypotheses are formed and tested about the size and direction of the impact of the survey item characteristics on mode-specific measurement bias and on mode-specific response styles. Hypotheses are formed about the mode-dependence of response styles but unrelated to survey item characteristics. The evaluation in this paper will be used to form hypotheses that can be tested in future studies.

In section 2, first, the typology of survey items is presented and motivated. In section 3, the multilevel model is described. The typology and model are applied to the CVS in section 4. Finally, section 5 ends with a discussion of the findings.

2. A typology of survey items

In this section, a list of survey item characteristics is presented that will be used in the analyses of measurement bias and response styles. A coding scheme was developed with 26 variables characterizing the survey items. This scheme was based on the SQP typology described in Saris and Gallhofer (2007) and Gallhofer, Scherpenzeel and Saris (2007), and on an unpublished paper by Campanelli et al (2011), who made a classification for mixed-mode surveys. Saris and

colleagues developed this typology of survey item characteristics in order to predict in advance the quality of a survey question. Their classification was used as the starting point and supplemented by characteristics from Campanelli et al (2011). Characteristics were omitted in case they showed little or no variation for the CVS questionnaire items or in case reliability between coders was low. The survey items of the CVS questionnaire were coded independently by three of the authors, all questionnaire experts and involved in testing and designing questionnaires at Statistics Netherlands. Two characteristics that are, generally, considered as important in the literature, 'liable to socially desirable answering' (Campanelli et al 2011) and 'liable to be outside knowledge of average respondent and to lead to a do not know answer' (Saris and Gallhofer 2007), were omitted because of low reliability between the three coders.

A survey item is the combination of an introduction, a question, a set of answer options or categories and a context in the questionnaire. The characteristics are assigned to one of these four elements. Table 2.1 contains an overview of the selected characteristic and Appendix A provides the labels that will be used throughout the paper to denote the characteristics. In the following the characteristics are motivated and explained in more detail. Essentially, the characteristics linked to the question apply to the first three cognitive steps in the answering process, Interpretation, Information retrieval and Judgment, whereas the characteristics linked to the answer apply to the last cognitive step, Reporting. The characteristics linked to the introduction may apply to all steps and the context characteristics may influence all four steps.

Table 2.1: Selected survey item characteristics.

Element	*Characteristic*
Introduction	Is an instruction provided?
Question	Concept of question: opinion, knowledge or fact
Question	Complexity - 1: Length of question in words
Question	Complexity - 2: Does question use difficult language?
Question	Complexity - 3: Are there conditions or exceptions in question?
Question	Complexity - 4: Does question require recall?
Question	Complexity - 5: Does question contain a hypothetical setting?
Question	Complexity - 6: Does question require calculations?
Question	Complexity: Number of scores on six complexity dimensions
Question	Complexity: Scores on at least one complexity dimension
Question	May question arouse strong emotions?
Question	Is question multidimensional?
Question	Is question formulated as statement?
Question	Time reference of question: past, present or future
Answer	Number of answer categories
Answer	Is there a mismatch between question and answer options?
Answer	Is scale ordinal?
Answer	If ordinal: Are categories presented as report mark?
Answer	If ordinal: Is range of scale bipolar?
Answer	If ordinal: Is direction of scale positive – negative?
Answer	Is DK explicitly offered as answer category?
Context	Questionnaire section to which item belongs (see Appendix A for overview)
Context	Is item an element of a battery of items?
Context	If item in battery: Relative position in battery
Context	Absolute position in questionnaire

Introduction: Differences may occur between modes if instructions are not consistently given to all respondents or do not contain exactly the same information. For example, interviewers are instructed to read instructions if they think it is necessary. If the same instruction is available in the web or mail versions of the questionnaire, then it is up to the respondents to read them or not.

Question: Six characteristics linked to the question have been selected: concept, complexity, emotional content, dimensionality, formulation and time reference. The complexity of the question is divided into six dimensions but also an aggregated characteristic is derived.
The concept of the question is a choice between attitude, knowledge and fact. Questions about attitudes, and to a lesser extent, about knowledge are conjectured to be more sensitive to the presence of an interviewer than factual questions.
The complexity of a question may influence the extent to which respondents need assistance but may also determine the amount of time they need to process the question. For these reasons it is suspected that the mode has a stronger impact on complex questions than on easy questions. Defining what is a complex question is, however, not straightforward and for this reason it was decided, following the literature, to distinguish a number of dimensions: the length of a sentence (counted in number of words), the use of difficult words, the use of conditions or exceptions, the use of a hypothetical setting, the requirement to recall past events, and the need to perform calculations. The dimensions are treated as separate characteristics but also summary measures are derived: the number of dimensions on which the question is complex, and an 0-1 indicator for complexity on at least one dimension.
When a question has an emotional content, then respondents may be reluctant or, on the contrary, be very eager to answer it. Interviewers may mitigate this effect by motivating to answer, or checking if the correctness of the answer is threatened. An example is forward or backward telescoping: When people have been victim of a crime, then they may place this victimization further in the past than the reference period or, on the contrary, report it as happened in the reference period although it happened before the reference period.
Multidimensionality of a question is, generally, considered bad questionnaire design; a question consists of at least two subquestions and this multidimensionality may confuse a respondent. This confusion may again be mitigated by the amount of time the respondent has to answer the question and also by the interviewer who may be aware of this from experience.
The formulation of a question is a choice between a statement and any other form. There has been a strong debate about the use of questions posed as statements in the literature and the answering process of statements is conjectured to be affected by the mode (see discussion in Fowler 1995, Saris, Krosnick, Revilla and Shaeffer 2010 and Ye, Fulton and Tourangeau 2011).
The final question characteristic, the time reference of the question, is a choice between past, present and future. It overlaps partially with the complexity dimension on the need for recall of past events. However, it is introduced as a separate dimension as also expectations about the future may complicate the answering process. Retrospectively, a separate complexity dimension for a reference to the future may have been a more logical choice, but during coding it was not included in summary measures for complexity. It is, therefore, treated as a separate characteristic.

Answer: For the answer seven characteristics have been selected; the number of categories, a mismatch between question and answer options, the measurement level of the answer, when the measurement level is ordinal three further classifications (use of report marks, range of the scale and direction of the scale), and the use of an explicit DK answer category.
The number of answer categories complicates the reporting of an answer. The greater the number of categories, the more likely it is that a respondent will pick the first category that

seems to apply when reading the answer categories (a primacy effect), or, conversely, to report one of the last categories that seems to apply when being read the answer categories (a recency effect).
A mismatch between a question and the answer options is considered bad questionnaire design: it occurs when the response options do not correspond to the question. The impact of a mismatch is conjectured to be stronger when the answer options are clearly presented or are being read to the respondent.
A number of answer characteristics is linked to the measurement level of the survey item. The measurement level itself is a choice between nominal and ordinal. Other levels (e.g. continuous, discrete, interval) exist but are not used in the CVS. When the level is ordinal, then three more indicators are constructed. The first is whether categories do not have clear labels and are presented as report marks. This type of labelling prevents recency and primacy effects and is more adapted to multi-mode surveys. Similarly, the visual presentation of the categories (horizontal or vertical) may play a role. However, in the CVS all scales are presented horizontally, so that this characteristic was not added. The second indicator is whether the range of the scale has a single pole (unipolar) or has two poles (bipolar). The third indicator is whether the answer categories are ordered from negative to positive or from positive to negative. The direction of the scale may interact with primacy and recency effects and some surveys use randomization of the direction of items for this reason.
The final characteristic, the availability of a "do not know" (DK) answer category, is considered influential in the literature. Various mechanisms may enforce or weaken each other. First, respondents may feel pressured to give a substantive answer and tend to give such an answer, even if they do not have an opinion (Beatty and Hermann 2002). This tendency will be strong when an interviewer administers the questionnaire. In modes without interviewers respondents might feel less pressure to give substantive answers and will admit more frankly they have no opinion. In this respect mode effects are expected for questions on topics that many respondents have no opinion about. Second, respondents may feel reluctant to give a truthful substantive answer. This holds especially if an interviewer is present and the answer may be socially undesirable. Under such circumstances a respondent can easily revert to a DK (Beatty and Hermann 2002). This mechanism operates to a lesser extent in modes without interviewers. Third, a lack of motivation is suspected to have a greater effect when no interviewer is present to encourage a substantive answer. These mechanisms are expected to create differences in mode effects, if the DK answer is presented in an identical way in all modes. The presentation of DK answer categories has received a lot of interest for that reason, e.g. Couper (2008).

Context: Three simple characteristics are linked to the survey item: the topic of the questionnaire section to which the item belongs, the presentation of the item as part of a grid or battery of multiple items, and the position of the item in the questionnaire. In case the item is element of a grid or battery also the local position may play a role. The presentation and position in a grid or battery and the questionnaire section form the local context of the item, whereas the position is part of the global context. Clearly, the global context of the item is also formed by the content of all preceding questionnaire sections, but it is hard or impossible to translate that full context to simple indicators. The context of the item has a complex impact on the answering process. The most obvious impact is that on motivation and concentration, which is conjectured to be mitigated by the interviewer but also by the possibility to complete the questionnaire in steps and at a self-selected point in time.

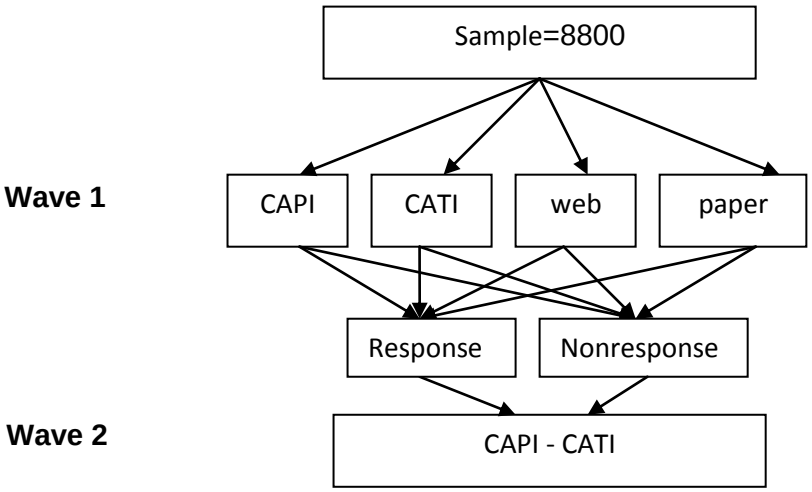
3. Estimating mode-specific measurement bias

This section presents the multilevel model that is used to separate the impact of survey item characteristics and individual characteristics on measurement bias. The model is applied in section 4 to the CVS. Before moving to the model, the underlying experimental design is briefly described. Details can be found in Buelens et al. (2012) and Schouten et al (2013).

3.1 The experimental design

The experiment consisted of two waves. In the first wave, 8800 sample units were randomly assigned to one of the four survey modes, CAPI (Computer Assisted Personal Interviewing), CATI (Computer Assisted Telephone Interviewing), web or paper. The data collection strategies for the four modes equalled the standard strategies at Statistics Netherlands, e.g. length of data collection period and number of visits/calls/reminders. The full sample, excluding administrative errors and some exceptional nonresponse types like language problems, is approached once more in the second wave. Approximately 80% of the sample persons are administered by face-to-face in the second wave. The remaining 20% of the sample persons were interviewed through CATI to reduce administration costs. When a sample person had a registered telephone number, then the allocation to CATI or CAPI was random with probabilities 70% to CAPI and 30% to CATI. When a sample person did not have a registered number, then he/she was always allocated to CAPI. The 70%-30% CAPI-CATI distribution was chosen such that anticipated mode effects between CATI and CAPI are much smaller than the sampling errors. Figure 3.1 presents the design of the experiment.

Figure 3.1: Design of the experiment



The first wave of the experiment is the Crime Victimization Survey (CVS) with two modifications. Part of the modules at the end of the survey questionnaire is replaced by the Labour Force Survey (LFS) module for employment status and by two sets of four questions from the European Social Survey (ESS).

The second wave of the experiment employs a new questionnaire, consisting of:

- A repetition of the key statistics from the CVS
- General attitudes towards safety and politics

- General attitudes towards surveys
- Evaluation of survey participation in wave 1 and survey design features like the advance letter and the interviewer
- Evaluation of the CVS questionnaire (wave 1 respondents only)
- Access to web and mode preferences

The repeated CVS items and the general attitudes on safety and politics and surveys are used to adjust for mode-specific coverage and nonresponse bias, collectively forming mode-specific selection bias. The remaining mode-specific measurement biases are always defined relative to CAPI, i.e. they must be interpreted as the average difference in answer between a survey mode and CAPI given response to CAPI. The conclusions from the biases can, therefore, not be extrapolated to mode-specific measurement biases for all persons in the population. However, in practice these extrapolations have little relevance as no survey has a 100% response. The experiment was designed from the assumption that CAPI has the smallest selection bias to the population.

3.2 Estimating mode-specific measurement bias at the category level

In this study, a different method is used than in Buelens et al. (2012) and Schouten et al. (2013). In those papers the measurement bias was estimated by weighting the response to each mode in the first wave to the response of the second wave. The advantage of this method is that the coverage and selection bias can be estimated. The disadvantage is, however, that it is not possible to directly investigate the effect of person or question characteristics on the measurement bias. Therefore, a different approach is taken.

A regression model is used in which the answer to a survey item of the first wave is a dependent variable and the second wave variables and available register variables are independent variables. The mode of wave 1 is included as a separate independent variable. The wave 2 variables and register variables are included to adjust for selection differences between the modes. Under assumptions that resemble those made in Buelens et al. (2012) and Schouten et al. (2013), the regression coefficients for the modes reflect the mode-specific measurement bias.

Since all CVS survey items are categorical (nominal or ordinal), every category k of each question j is modelled separately using logistic regression. Each answer y_{ijk} of respondent i can take on the values 1 or 0 depending on whether the respondent has chosen option k of question j or not. Each respondent also has a vector of explanatory variables $\mathbf{x}_i$ obtained from the second wave or from administrative sources. The model that is estimated can be written as

$$\begin{aligned} \text{logit}(\ p_{ijk}\ ) &= \mathbf{x}_i'\boldsymbol{\beta}_{jk} + \mathbf{m}_i'\boldsymbol{\delta}_{jk}, \\ y_{ijk} &\sim \text{Binom}(\ p_{ijk}\ ), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_{jk}$ is the vector of coefficients, m_i a vector with three dummy elements for each of the modes (not equal to CAPI) and $\boldsymbol{\delta}_{jk}$ the corresponding coefficients.

In case the coefficients for mode ($\boldsymbol{\delta}_{jk}$) are significant, there is a significant mode-specific measurement bias; the size of the coefficients indicates the size of the bias. However, to aid interpretation these coefficients will be converted to mode averages. For all wave 2 respondents predictions are calculated for their CAPI, telephone, web and paper response respectively:

$$\hat{p}_{ijk}^{m} = \frac{\exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{jk} + \hat{\delta}_{jkm}\right)}{1 + \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{jk} + \hat{\delta}_{jkm}\right)}. \qquad (2)$$

An average of these gives the fraction of respondents that selects answer option *k* in mode *m* on question *j*

$$\hat{p}_{jk}^{m} = \frac{1}{n}\sum_{i} \hat{p}_{ijk}^{m}, \qquad (3)$$

and $\sum_k \hat{p}_{jk}^{m} = 1$.

The measurement error for mode m at category *k* of item *j* is estimated by

$$\hat{\mu}_{jk}^{m} = \hat{p}_{jk}^{m} - \hat{p}_{jk}^{CAPI}. \qquad (4)$$

The main difference between the approach described here and the approach in Buelens et al. (2012) and Schouten et al. (2013) lies in the selection of persons that responded in both wave. The logistic regression coefficients for the mode apply to this subset and not to persons that respond to face-to-face as is the case in the Buelens et al. (2012) and Schouten et al. (2013) studies. The estimated measurement biases were compared to those in these papers and were very close and not significantly different for virtually all items.

3.3 Estimating mode-specific response style propensities

In section 4, a series of response styles is investigated. In all cases, the response styles are constructed from the survey item category level and are based on the 0-1 event that the respondent chose a specific answer category or chose an answer out of a specific set of categories. For example, for DK-answering the 0-1 event is that the respondent selected the DK answer category.

Since the interest is in the effect of question characteristics and the effect of personal characteristics, model (1) is adapted in three ways. First, for each person *i* and question *j* a 0-1 indicator y_{ij} is defined as the event that a certain response style was observed for person *i* at question *j*. Second, question characteristics z_j are added as explanatory variables. Third, since response styles exist only if they persist throughout a significant part of the questionnaire, the indicators y_{ij} are modelled simultaneously and random effects are added for both person and survey item in order to model remaining between variation. The model then becomes

$$\begin{aligned} \text{logit}(p_{ij}) &= \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\gamma} + \mathbf{m}_i'\boldsymbol{\delta} + e_j + e_i \\ y_{ij} &\sim \text{Binom}(p_{ij}) \\ e_j &\sim \text{Normal}(0, \sigma_j) \\ e_i &\sim \text{Normal}(0, \sigma_i). \end{aligned} \qquad (5)$$

3.4 Estimating mode-specific measurement bias at the item level

The measurement biases estimated in section 3.2 apply to categories and not to survey items. In order to study measurement bias at the item level, in relation to characteristics of the items, it is necessary to also define a measurement bias at the item level. Considering item *j* and

answer categories $k = 1, \ldots, c_j$, the item level measurement bias for mode m (CATI, paper or web) is defined as

$$\mathrm{M}_j^m = \frac{1}{2}\sum_{k=1}^{c_j} \left|\hat{\mu}_{jk}^m\right|, \qquad (6)$$

with $\hat{\mu}_{jk}^m$ as defined by (5).

The item level measurement bias M_j^m can be interpreted as being the fraction of respondents selecting a different answer category for item j in mode m, compared to the CAPI mode. Selection effects deliver no contribution as they have been accounted for through the modelling approach.

The definition of M_j^m is explained by observing that

$$\sum_{k=1}^{c_j} \hat{\mu}_{jk}^m = \sum_{k=1}^{c_j} (\hat{p}_{jk}^m - \hat{p}_{jk}^{CAPI}) = \sum_{k=1}^{c_j} \hat{p}_{jk}^m - \sum_{k=1}^{c_j} \hat{p}_{jk}^{CAPI} = 1 - 1 = 0, \qquad (7)$$

basically meaning that measurement bias across categories compensate each other, and, hence, sum to zero. Consequently, in absolute value, the measurement bias amounts to twice the percentage point of respondents selecting a different category in the alternative mode compared to CAPI. This motivates the definition of M_j^m.

It became apparent that a major contribution to the M_j^m is often delivered by a shift towards the category "don't know". This occurs particularly when the "don't know" option is not offered explicitly in the CAPI mode, while it is offered explicitly in the self-administered modes. In paper, not answering a question is considered a "don't know" answer, which is possible for any item. In order to differentiate between shifts towards the "don't know" category, and between true categories, the item level measurement bias is split into two components,

$$\mathrm{M}_j^m = \mathrm{M}_{j,dk}^m + \mathrm{M}_{j,edk}^m, \qquad (8)$$

where the first term is the contribution due to differences with respect to the *don't know* (*dk*) category,

$$\mathrm{M}_{j,dk}^m = \left|\hat{p}_{j(k=dk)}^m - \hat{p}_{j(k=dk)}^{CAPI}\right|, \qquad (9)$$

and the second component is the contribution due to selecting different answer categories *exclusive don't know* (*edk*), defined as

$$\mathrm{M}_{j,edk}^m = \mathrm{M}_j^m - \mathrm{M}_{j,dk}^m. \qquad (10)$$

The item level measurement bias is used to estimate the measurement bias of groups of items, generally grouped according to some item characteristics. The measurement bias for a group G of items is estimated using the available items belonging to that group, simply by averaging,

$$\hat{M}_G^m = \frac{1}{n_G}\sum_{j=1}^{n_G} M_j^m, \qquad (11)$$

with variance estimate

$$\hat{V}(\hat{M}_G^m) = \frac{1}{(n_G - 1)}\sum_{j=1}^{n_G} (M_j^m - \hat{M}_G^m)^2. \qquad (12)$$

These equations ignore the fact that the M_j^m themselves are estimates. Here they are assumed to be known exactly. Hence, the true variance will be somewhat larger than the estimate given here. However, the uncertainty due to having a limited set of questions available is much larger than the uncertainty in the measurement bias estimates of the specific items.

4. Application to the Crime Victimisation Survey

This section consists of three parts. First, the survey item taxonomy of section 2 is applied to the CVS. Next, the item-level measurement bias as defined in section 3.3 is derived and explored using a subset of the survey item characteristics. Finally, an analysis is conducted on five response styles.

4.1 The CVS questionnaire profile

The wave 1 questionnaire consisted of 246 survey items that were taken from the CVS and are then followed by 8 items from the European Social Survey (ESS) on Civic duty and police, a section about employment status taken from the Labour Force Survey (LFS) and a section on demographics and educational level of the sampled person and his/her household. In the analysis, only the CVS and ESS items will be considered, a total of 254 items. Of these 254 items, 126 were presented to all respondents i.e. non-nested. The remaining 128 items nested items mostly consisted of follow-up questions in the questionnaire sections on Victimisation and Reporting to the police. Table 4.1.1 shows the number of non-nested items per questionnaire section. Of the 126 non-nested items , 99 are ordinal, 4 are nominal and 23 are yes-no questions.

Table 4.1.1: Overview of numbers of non-nested items per questionnaire section in wave 1 and numbers of repeated items in wave 2.

Section	*Non-nested items*	
	wave 1	*wave 2*
Living conditions neighbourhood	16	10
Neighbourhood problems	26	8
Feeling safe	5	2
Victimisation	14	7
Reporting to police	0	0
Satisfaction with police contact	5	3
Police performance neighbourhood	13	1
Police performance in general	9	0
Municipality performance	6	0
Prevention I	12	0
Prevention II	5	0
Unsafe places	7	0
Civic duty and police	8	0
Total	126	31

The wave 1 questionnaire consisted of ten batteries consisting of 25, 11, 9, 7 (two times), 6, 5 (two times), 4 (three times) and 3 items. A set of three or four items may not be considered to be a battery. The items are, however, presented in series and have exactly the same set of answer categories. For this reason, they were included.

Table 4.1.2 presents frequencies for most of the item characteristics. Two characteristics are missing: number of categories (ncategories) and aggregated complexity level (complexity 1 and complexity2). The 23 yes-no items obviously have two categories. Of the remaining 103 items, 98 items have five answer categories and five have six or more categories. Concerning aggregated complexity, 55% of the items is coded as complex on at least one feature, and 18% is coded as complex on three or more features (out of seven possible features). Three characteristics turn out to be very rare in the CVS: question may evoke emotion, question is multi-dimensional and answer categories are presented as report marks. These three characteristics are ignored in the analyses for the CVS. Furthermore, the CVS does not have items that ask for knowledge and only three items that refer to the future. The category future of the time reference variable is collapsed with category present.

Table 4.1.2: Frequency distributions for selected item characteristics. Properties position and direction apply to ordinal items only.

Property	*Frequency*	*Property*	*Frequency*
Concept	27% fact, 73% opinion	Formulation	33% is statement
Length	24% has >25 words	Mismatch	8% has mismatch
Language	29% difficult	Resp scale	79% ordinal
Conditional	25% with condition	Report marks	2% as report marks
Memory	17% requires memory	Position	52% unipolar
Hypothetical	10% hypothtcl setting	Direction	62% neg to pos
Calculation	16% needs calculation	Battery	71% in battery
Time ref	83% present, 17% past	Instruction	15% modes differ
Emotion	2% evokes emotion	DK available	79% modes differ
Dimensional	2% multidimensional		

In the analyses survey item characteristics are included univariately, because of the relatively small number of non-nested survey items, except for questionnaire section which is sometimes added to other characteristics. However, clearly characteristics may be collinear in general, or, more specifically, in the CVS. A full elaboration of the associations between all selected characteristics would, however, go beyond the scope of this paper and needs to be investigated in future research. A principal component analysis on all variables in table 4.1.2 plus number of categories gives four factors with eigenvalues above 1 which explain 72% of the variance. The first factor explains 45% of the variance and shows large negative or positive loadings for concept, memory required, calculations required, response scale, number of answer categories, item in battery, DK available, time reference and time reference. Two additional observations are useful to mention for the CVS questionnaire. There is no strong association between availability of a DK answer category and any of the other characteristics. There is a small positive correlation between number of complex item features and absolute position in the questionnaire; items that are positioned towards the end are more complex.

4.2 Explaining measurement bias at the item level

The item level measurement bias is estimated for various item characteristics, thereby using the available CVS items and their estimated measurement bias. The results are shown graphically in the figures in Appendix B. The error bars indicate plus/minus one standard error, obtained as the square root of the variance defined above. Since the number of items available is rather limited, interactions between two or more item characteristics cannot sensibly be analyzed.

A number of observations is made from the figures. The first figure shows measurement bias for items according to availability of don't know (DK) as an explicit choice. Not surprisingly, there is a shift towards DK for the questions where it is offered (it never or hardly ever is in CAPI and CATI). The bias not due to DK is larger for questions where DK is offered in some modes, and this applies to all modes.

Regarding the concept questions ask for, there is a larger shift to the (DK) category for opinions than for facts. The measurement bias excluding DK is larger for opinions than for facts, in all three modes, but it is smallest for CATI.

The complexity of items is considered through the question length (number of words), difficulty of the language, use of conditions, memory use required, use of hypothetical constructs, and whether calculations are needed or not. The characteristics complexity1 and complexity2 summarize these aspects. There is a smaller shift to DK for longer questions, put in difficult language, using conditions, or requiring calculations. This seems counterintuitive as one might expect an escape to DK for hard questions. The contrary seems to be true. As far as the non-DK measurement bias is concerned, CATI is most often closest to the CAPI reference mode, and web the most extreme. Again, generally complex questions seem to have smaller bias, although the complexity of the language, and whether the question is a hypothesis, do not make any difference within a given mode.

All but one of the nominal scale questions are in fact yes/no questions. The measurement bias for these does not vary much between modes, and is always smaller than for items with ordinal scales. The latter have smallest bias in CATI and largest in web.

The same findings apply to the characteristic indicating whether items occur in a battery or not. The direction and the range of answer categories do not affect the magnitude of the measurement bias a lot. CATI bias is somewhat lower than web and paper.

Questions formulated as statements give rise to a larger measurement bias, both exclusive of DK and because of DK, in all three modes.

Finally, questions referring to the past have smaller non-DK bias, and are comparable between modes. Questions about the present have larger bias, specifically in web.

4.3 Explaining measurement bias at the category level

In this section, mode-specific measurement bias is detailed to the level of categories of survey items. More specifically, the response to wave 1 and wave 2 of the experiment is used to test the presence of mode-specific response styles. In the first subsection, the identification of such styles is discussed. In the subsequent subsections three such styles are investigated.

4.3.1 Mode-specific response styles in the Crime Victimisation Survey

The experimental data do not allow for an identification of all possible response styles, since these would require careful randomization of the order and form of survey items. Furthermore, the experiment only allows to distinguish mode-specific styles, e.g. if respondents provide the same socially desirable answer to all modes, then the impact cannot be measured or estimated. The data do, however, allow for an analysis of some mode-specific response styles, as the total sample was randomly allocated to each of the four survey modes and strong auxiliary information is available from wave 2. The wave 2 variables ensure that the mode-specific selection of respondents can be neutralized to the extent needed here. It is believed that remaining selection biases are small relative to standard errors of the mode-specific measurement bias estimates.

In the following subsections, the occurrence of five styles is evaluated: don't know (DK) answers, primacy, recency, extreme response style and straightlining. The amount of DK answers is considerable for the non-interviewer modes and has the potential to bias statistics based on the CVS. Primacy and recency are two styles that are very similar in nature and correspond to choosing one of the first or last answer categories. The extreme response style refers to respondents consistently choosing one of the extreme categories on ordinal rating scale questions. Straightlining occurs when a respondent sticks to the same answer for multiple items in the same questionnaire block. This behaviour is mostly associated with rating scale questions in grids or batteries.

In all cases, 0-1 indicators are formed for each question and respondent, and modelled using the multi-level approach of section 3. Hence, survey items form a level within respondents. Respondents are selected when they participated in both wave 1 and wave 2, and mode-specific selection effects are adjusted using all wave 2 variables. The adjusted mode effect represents the mode-specific measurement effect and is always defined relative to CAPI, i.e. the mode-specific measurement effect describes the difference in answer between a survey mode and CAPI given a response to CAPI. In the interpretation of results it is important to realize that they apply to persons that respond to a CAPI survey and cannot be extrapolated to the full population.

The multilevel model allows for the inclusion of both survey item random effects and individual random effects. The survey item random effects indicate the extent to which the behaviour varies between items, whereas the individual random effect corresponds to variation between different respondents. The variation between items is then modelled using the survey item characteristics of section 2. The variation between persons is not modelled except for DK answers, since the various styles are relatively rare and sample sizes are limited.

With the analysis of response styles comes one conceptual question: How to deal with simultaneously occurring response styles in the CVS? The response styles can be investigated separately but they obviously do not occur in isolation. A respondent may, for example, provide more DK answers, have a tendency to choose the last answer category, and the last answer category may often be the extreme endpoint of a rating scale. When response styles occur simultaneously, then they may mask each other to some extent. In the example: When respondents in one mode choose the DK answer very frequently, and, when they do not give a DK answer, choose one of the last answer categories, then it appears as if recency is absent while in fact it occurs for all substantive (non-DK) answers. This implies that response styles need a hierarchy and such a hierarchy should naturally come from the cognitive steps (Interpretation, Information retrieval, Judgment and Reporting) in the answering process. Such

a hierarchy is, however, not at all straightforward. A DK answer may occur because the respondent does not understand the question, is not willing to spend cognitive effort in retrieving the answer or simply does not know the answer, is not able to put all pieces of the answer together in his mind, or does not understand the answer categories. Hence, the DK answer may arise from any of the four steps and may thus precede or follow the other response styles. The response styles primacy, recency and straightlining all seem to bypass the steps Information retrieval and Judgment, while extreme response style seems to take place at the last step, Reporting. Since DK answering does not have a unique hierarchy to the other styles, it is decided not to remove records with a DK answer but in all instances evaluate response styles using all responses.

4.3.2 Don't know answers

In the CVS, DK answers play a prominent role. They are not explicitly offered for CATI or CAPI, but are offered for web and paper for 79% of the items. In the web and paper modes DK answers are selected frequently. There are questions where approximately 30% of the answers consist of 'don't know' in these modes. Therefore, the effect of question and personal characteristics on the probability of answering 'don't know' will be investigated. Since DK answers are very rarely selected in CATI and CAPI, these modes are removed from the analyses.

First, it is investigated which of the question properties has the strongest effect on the probability of answering 'don't know'. Therefore, for each property a model of the following form is fitted and estimated

$$\text{logit}(\ p_{ij}\ ) = gender + age + property + mode + e_j + e_i\ , \qquad (13)$$

where *property* is the question property under investigation. Gender and age are included in the model to correct for selection and coverage effects. Table 4.3.1 shows for each of the models the AIC. A lower AIC indicates a better model fit given the number of parameters used in the model.

Table 4.3.1: AIC values for model (13) for each of the question properties. The AIC are sorted in increasing order. The model without a question property is labelled the base model.

Pos.	*Question property*	*AIC*	*Pos.*	*Question property*	*AIC*
1	section	63915	12	in battery	64043
2	DKavailable	63995	13	instruction	64051
3	range	64011	14	formulation	64056
4	ncategories	64018	15	cmplxty conditional	64059
5	direction	64021	16	question number	64071
6	cmplxty memory	64021	17	cmplxty hypothetical	64071
7	respscale1	64022	18	complexity 1	64074
8	respscale2	64022	19	*base model*	64075
9	complexity 2	64024	20	cmplxty language	64075
10	concept	64034	21	mismatch	64077
11	cmplxty calculations	64036			

Most question properties explain some of the variation in 'don't know'. Appendix C shows the regression coefficients for the first three properties and the base model. Looking at the large value of the variance coefficient of *question* in the base model, a large spread in the probability

to answer 'don't know' can be observed over the different questions. The table also shows that persons with an age over 65 and females have a higher probability of answering 'don't know'. Furthermore, respondents in web have a lower probability of answering 'don't know'.

Tables C.2 to C.4 show the models for *questionnaire section*, *don't know explicit*, and *position* respectively. From table C.2, it can be seen that the section on victimisation shows little 'don't know', while the three sections on police and municipality performance show a large amount of 'don't know'. Also the amount of remaining variability between the questions has dropped from 2.61 to 0.51, which indicates that a large amount of the original variability is explained by the questionnaire section. In the analysis, the position of the item in the questionnaire was not included. However, from the regression coefficients for the questionnaire sections it can be concluded that there is no clear pattern about an increase or decrease of DK answers during the interview. Table C.3 shows that explicitly offering 'don't know' increases the probability of a 'don't know' answer. In case of web, this is obvious, as it is not possible to choose 'don't know' if it is not offered. In case of paper a respondent can still choose to not fill in a question if 'don't know' is not explicitly offered (this is coded as 'don't know'). A model with the interaction between *mode* and *don't know explicitly offered* confirmed that the probability of 'don't know' is higher for paper when 'don't know' is not explicitly offered. Table C.4 shows that questions that have a unipolar range have a smaller probability of 'don't know' than questions with a bipolar range. Questions with a range *Other* (which are mainly yes/no questions) show the least amount of 'don't know'.

The three sections that show a large amount of 'don't know' are all quite similar: all questions have a bipolar scale, and 'don't know' is offered in all questions. Both of these properties were also found to have a strong effect. Also on other properties (measured and also non-measured such as subject) the questions are similar. This makes it difficult to decide which of the properties actually cause the 'don't knows'. In order to be on the safe side, it was decided to first explain most of the 'don't know' with the questionnaire sections. After that it was checked if there are any remaining properties that predict 'don't know'. Therefore, models of the following form are estimated

$$\text{logit}(\ p_{ij}\ ) = gender + age + section + property + mode + e_j + e_i\ , \qquad (14)$$

where property is the question property under investigation. A likelihood ratio test is then applied to test if this model is significantly better than the model without the property (but with gender, age, block and mode) . Table 4.3.2 shows the results.

Table 4.3.2: p-values for question property in model (14).

Pos.	*Question property*	*p-value*	*Pos.*	*Question property*	*p-value*
1	cmplxty calculations	0.001	11	response scale 2	0.353
2	cmplxty conditional	0.001	12	direction	0.486
3	question number	0.103	13	complexity 1	0.497
4	number of categories	0.136	14	formulation	0.512
5	concept	0.155	15	complexity 2	0.606
6	range	0.187	16	mismatch	0.633
7	don't know explicit	0.213	17	in battery	0.730
8	response scale	0.236	18	cmplxty hypothetical	0.836
9	cmplxty memory	0.271	19	cmplxty language	0.931
10	instruction	0.288			

What can be noted is that when correcting for section, range and don't know explicit are no longer significant, which indicates that these variables are correlated and from this experiment it cannot be decided which of these actually affects 'don't know'. Two properties still show a significant effect when correcting for section: complexity calculations and complexity conditional. The models for these two properties are shown in tables C.5 and C.6 respectively. These tables show that questions involving calculations or conditional statements show more 'don't know'.

4.3.3 Primacy and recency

Primacy and recency correspond to the tendency to choose one of the first and one of the last answer categories, respectively, and are seen as signs of satisficing. Primacy is hypothesized to be a response style that is specific to non-interviewer modes, while recency is thought to be more frequent in interviewer modes. In this section, two hypotheses are tested: 1) Does primacy occur more frequently for non-interviewer modes?, and 2) Does recency occur more frequently for interviewer modes? No further distinction is made between the interviewer modes CATI and CAPI and between the non-interviewer modes paper and web.

In the CVS all items were selected that are part of a battery of items. This led to a selection of 90 items from ten questionnaire sections. All selected items are ordinal and have five categories. In paper and web, the items are presented horizontally as grid questions. A 0-1 indicator for primacy was derived per item and per respondent as the indicator that the respondent chose the first answer category of the item. Similarly, an indicator for recency was derived for the last answer category. If an explicit DK category was presented in paper or web, then the category before the DK answer was selected as the last answer category. This choice was made because recency is believed to be specific to interviewer modes, where the DK answer is never explicitly offered. It is important to note that primacy and recency are known to be strongest when the number of categories is large. In the CVS, most survey items had a relatively small number of categories, so that only the first or last answer category was picked rather than a few categories at the beginning or end. This choice is analogous to the analysis by Lynn and Kaminska (2012) who also picked only one category. It must, however, be conjectured beforehand that primacy and recency effects will be relatively modest in the CVS.

In order to adjust for selection effects, primacy and recency profiles were also derived from wave 2. For both styles, two profiles were made, one based on the repeated CVS items and one based on the repeated CVS items plus the extra attitudinal questions in wave 2. The first profile equals the proportion of repeated CVS items for which the respondent showed the style. The second profile is the same but computes the proportion on CVS repeated items and the attitudes. Both profiles take values in between 0 and 1, where 0 means that the respondent never showed the particular style and 1 means that the respondent always showed the particular style. Since the analysis essentially estimates the probabilities that an arbitrary respondent has the response style when answering an arbitrary question, it is sufficient to control for mode-specific selection on the wave 2 response style profiles. One may, however, assume that the response styles are not constant, i.e. respondents may show the behaviour at one interview but not at the other. Hence, each respondent has a primacy and a recency response style probability, rather than a fixed style. For the analysis presented here, it means that the mode-specific selection effects based on wave 2 are still controlled for but that the association between mode and profile is harder to distinguish.

For primacy and recency, the analysis is restricted to two question properties: the questionnaire section and the absolute position in the questionnaire. Since primacy/recency is

thought to be a behaviour that stretches over multiple similar questions, the characteristics of single survey items are not included in the models. Individual characteristics are also not considered, apart from the wave 2 primacy/recency profiles.

Following notation in section 3, the models that are estimated have the form

$$\text{logit}(\ p_{ij}\ ) = \ profile\ _i + \ property\ _j + mode\ _i + e_j + e_i\ , \qquad (15)$$

where profile is either the profile based on repeated CVS items or the profile based on the extended set of items in wave 2, and where property is either questionnaire section or absolute position.

Table 4.3.3 presents the regression coefficients for the model for primacy and recency without survey item properties. Both for primacy and recency the model coefficient for interviewer mode is significant at the 5% level. The direction is the same; respondents show more primacy and recency in the interviewer modes. For primacy this means that the hypothesis that non-interviewer modes show more primacy is not confirmed. The lack of a primacy effect for non-interviewer modes might be the result of the horizontal presentation of the questions and/or the relatively small number of answer categories. Nonetheless, it is remarkable that the effect is present for interviewer modes. For recency, the finding confirms the hypothesis of a recency effect for interviewer modes.

Table 4.3.3: Regression coefficients for the primacy and recency model without survey item characteristics. Mode-specific selection is controlled for using profile based on CVS repeated items (CVS) and extended set of items (All).'' and '**' represent 5% and 1% significance levels, respectively.*

		Primacy		*Recency*	
		CVS	*All*	*CVS*	*All*
Std dev random effect	individual	0.54	0.54	0.36	0.36
	item	1.34	1.34	3.18	3.18
Intercept		-3.39**	-3.36**	-3.44**	-3.44**
Interviewer mode		0.07*	0.07*	0.51**	0.50**
Profile		2.83**	3.91**	1.90**	1.96**

Table 4.3.4: Regression coefficients for the recency model with absolute position added as main effect and as interaction with mode.'' and '**' represent 5% and 1% significance levels, respectively.*

		Absolute position	
		Main	*Interaction*
Standard deviation random effect	individual	0.54	0.54
	item	2.35	2.35
Intercept		-0.60	-0.73 **
Interviewer mode		0.51 **	0.71 **
abspos		-0.02 **	-0.02 **
abspos * interview mode			-0.003 **
Profile		1.90 **	1.89 **

Since the primacy hypothesis is not confirmed, the analysis is detailed only for recency. The estimated variances of the individual and survey item random effects show that there is relatively little between person variation but a considerable between item variation. This was

to be expected as the selected set of CVS items contains a wide range of different questions with different directions. It is interesting to investigate to what extent this between variation can be explained by the questionnaire section and item position in the questionnaire. Table 4.3.3 shows that there is no difference in adjustment for mode-specific selection between the profiles based on repeated CVS items and all wave 2 items. In the following, the CVS-based profile is used for this reason.

Table 4.3.4 shows the model coefficients for recency when item position is added as a main effect and as an interaction effect with mode. As expected the variance of the survey item random effect decreased with respect to the model without item position and the variance of the individual random effect did not change. The absolute position of the item ranges from 1 to 254 and has a significant effect on the probability of choosing the first answer category. The closer the item is to the end of the questionnaire, the less likely it becomes that the last answer category is selected. The small negative interaction effect with interview mode indicates that the recency effect gets weaker during the course of the interview. In fact, around item position 170 the recency effect shifts sign and the probability of choosing the last category is larger for the non-interviewer modes.

Table 4.3.5: Regression coefficients for questionnaire section in the recency model.'' and '**' represent 5% and 1% significance levels, respectively.*

Living conditions	0	Municipality performance	0.14
Neighbourhood problems	4.47 **	Prevention I	1.02 *
Feeling safe	3.08 **	Prevention II	-15.19
Police performance – ngh	-0.08	Unsafe places	-15.19
Police performance – gen	-0.11	Civic duty and police	0.64

To get more insight into the recency effect, item position is replaced in the model by the questionnaire section of the item. The regression coefficients for the intercept, mode and profile are almost the same as for the model with item position as a main effect. Table 3.4.5 contains the regression coefficients of the ten sections, in the order in which they appear in the questionnaire. The first section, Living conditions neighbourhood, is taken as the reference section. The coefficients show that especially the second and third sections have larger proportions of answers to the last category, which explains the decrease that was observed for item position. The model with questionnaire section as interaction effect with interview mode is not shown here, but estimates positive interaction regression coefficients for almost all sections. The interviewer modes lead to a recency effect throughout the questionnaire which is larger at the second and third sections.

4.3.4 Extreme response style

The extreme response style refers to respondents that always choose one of the extreme answer categories in ordinal rating scale questions and is conjectured to occur more in non-interviewer modes based on the literature (e.g. Greenleaf 1992 and Aichholzer 2013). In this section, one hypothesis is tested: Does an extreme response style occur more frequently for non-interviewer modes? Again, no further distinction is made between the interviewer modes CATI and CAPI and between the non-interviewer modes paper and web.

For the extreme response style a 0-1 indicator was derived per item and respondent as the indicator that the respondent selected either the first or the last answer category before the DK

answer category (if available). The same 90 CVS items were selected as for the primacy-recency evaluation. The other items are mostly yes-no questions or are not nominal.

In order to adjust for selection effects an extreme response style profile was derived from wave 2. Analogous to the primacy and recency response styles, two profiles were made, one based on the repeated CVS items and one based on the repeated CVS items plus the extra attitudinal questions in wave 2.

Table 4.3.6 contains the regression coefficients for the extreme response style model without characteristics of survey items. Again, the item-level variation is much larger than the person-level variation. So there is no indication that the extreme response style clusters strongly within individuals like the DK answers. The positive regression coefficient for the interviewer mode shows that the hypothesis is rejected. This result comes as no surprise given the stronger "primacy" and "recency" on interviewer modes. However, some care is needed with this conclusion as DK answers occur very frequently in the non-interviewer modes and it is unclear what these persons would have answered had they not been offered the option of a DK answer. For the extreme response style, the wave 2 profile based on all items is a stronger predictor than the profile based on the repeated items only. This profile is selected in the following.

Table 4.3.6: Regression coefficients for the extreme response style model without survey item characteristics. Mode-specific selection is controlled for using profile based on CVS repeated items (CVS) and extended set of items (All).'' and '**' represent 5% and 1% significance levels, respectively.*

		Selected items profile	
		CVS	*All*
Standard deviation random effect	individual	0.31	0.30
	item	1.90	1.90
Intercept		-1.79 **	-1.82 **
Interviewer mode		0.40 **	0.39 **
Profile		1.06 **	1.54 **

Although the hypothesis is not confirmed, the analysis is detailed for questionnaire section. Table 4.3.7 contains the regression coefficients when questionnaire section is added. The item-level variation is strongly reduced as a consequence, the standard deviation drops from 1.44 to 0.54, while the person-level variation is not affected. The first questionnaire section, Living conditions, is the benchmark section and has a zero regression coefficient by definition. The coefficients show that the first three sections and sections Prevention I and Civic duty and police show the largest proportion of extreme answers. The first two sections also have the largest numbers of survey items.

Table 4.3.7: Regression coefficients for questionnaire section in the extreme response style model.'' and '**' represent 5% and 1% significance levels, respectively.*

Living conditions	0.00	Municipality performance	-0.97 **
Neighbourhood problems	3.06 **	Prevention I	2.62 **
Feeling safe	1.28 **	Prevention II	-1.38 **
Police performance – ngh	-0.68 **	Unsafe places	-1.80 **
Police performance – gen	-1.24 **	Civic duty and police	0.29

4.3.5 Straightlining

Straightlining is another form of satisficing that occurs when respondents stick to the same answer category throughout questionnaire batteries. It is conjectured that straightlining is more frequent in non-interviewer modes and CATI as opposed to CAPI and may become stronger during the course of the interview and in batteries of questions. So, again two questons are asked: 1) Do non-interviewer modes show more straightlining?, and 2) Does CATI show more straightlining than CAPI?

The analysis for straightlining is performed in a similar way as for primacy and recency, but now the 0-1 indicator for a survey item is equal to 1 when the answer to that item is the same as the answer to the preceding survey item. The number of items that are modelled, therefore, reduces to 90 – 10 = 80 items, as the first items in each battery are ignored. Furthermore, the analysis is performed for interviewer modes versus non-interviewer modes, and for all modes. The absolute item position is again added as a property to the multilevel model.

As for the primacy and recency models, the adjustment for mode-specific selection is done using two individual profile variables based on wave 2. The first profile variables measure the proportion of answers that are the same as preceding answers for CVS repeated items. The second profile again adds the attitudinal questionnaire sections from wave 2. Analogous to the primacy and recency models it was found that the CVS-based profile performs best. In the following this profile is used to adjust for mode-specific selection.

Table 4.3.8 contains the regression coefficients in the straightlining model for interviewer mode. The regression coefficient for interviewer mode is not significantly different from 0. Hence, there is no indication of differences in straightlining behaviour for interviewer versus non-interviewer modes. Again the variation between items is much larger than the variation between individuals.

Table 4.3.9 contains the regression coefficients in the straightlining model for all modes separately without item position, with item position as a main effect and with item position as interaction with mode. Without the interaction between mode and item position, CATI and paper have a significant difference in straightlining behaviour with respect to CAPI. Both modes show more straightlining. Surprisingly, web shows the same amount of straightlining over the total interview. However, when the interaction with mode is added the sign for paper changes and web also has a negative impact on straightlining. The full model indicates that, at the start of the questionnaire, respondents to the non-interviewer modes show less straightlining than the interviewer modes, but as the interview progresses show more and more straightlining. Around position 75 the estimated straightlining probability changes sign for paper and becomes positive. For web, this occurs around position 138. For web the straightlining probability is always larger than for CAPI and decreases very gradually during the interview.

Table 4.3.8: Regression coefficients for the straightlining model for interviewer mode.'' and '**' represent 5% and 1% significance levels, respectively.*

		Straightlining
Std dev random effect	individual	0.30
	item	2.61
Intercept		-1.22 **
Interviewer mode		-0.02
Profile		1.11 **

Table 4.3.9: Regression coefficients for the straightlining model for all modes without absolute position (empty), with absolute position as main effect and as interaction with mode.'' and '**' represent 5% and 1% significance levels, respectively.*

		Absolute item position		
		Empty	*Main*	*Interaction*
Std dev random effect	individual	0.30	0.30	0.30
	item	2.61	2.59	2.60
Intercept		-1.28 **	-0.95 *	-0.85
CATI		0.13 **	0.13 **	0.17 **
Paper		0.11 **	0.11 **	-0.13 **
Web		0.00	0.00	-0.29 **
Abspos			-0.002	-0.003
Abspos * CATI				-0.0003 **
Abspos * paper				0.002 **
Abspos * web				0.002 **
Profile		1.09 **	1.09 **	1.10 **

Summarizing, the analyses confirm the hypothesis that CATI shows more straightlining than CAPI. They do not confirm that non-interviewer modes show more straightlining throughout the interview but as the interview progresses straightlining becomes stronger and eventually is more frequent than for CAPI. However, again some care is needed as survey items and batteries were not randomly ordered in the questionnaire. Table 4.3.10 presents the regression coefficients for questionnaire section, when item position is replaced by questionnaire section. There is no clear pattern. The second, sixth and ninth sections show stronger straightlining, while the seventh section shows less straightlining at the 1% significance level.

Table 4.3.10: Regression coefficients for questionnaire section in the straightlining model.'' and '**' represent 5% and 1% significance levels, respectively.*

Living conditions	0.00	Municipality performance	0.63 **
Neighbourhood problems	0.79 **	Prevention I	-0.40 **
Feeling safe	-0.09 *	Prevention II	0.01
Police performance – ngh	0.07 *	Unsafe places	0.56 **
Police performance – gen	0.20 **	Civic duty and police	-0.06

5. Discussion

In this paper, mode-specific measurement bias in the Crime Victimisation Survey (CVS) was analyzed at the survey item level and at the survey item category level and was related to survey item characteristics. The characteristics were taken from a typology of survey items that was derived from commonly used typologies in the literature. The mode-specific measurement bias was estimated relative to the face-to-face survey mode by adjusting for mode-specific selection bias. The adjustment for selection effects is based on repeated CVS survey items in a follow-up wave. The CVS is a very useful instrument for the analysis of the impact of mode on measurement as it has a wide range of survey items and is conjectured to be subject to various mode-specific response styles. Although the number of items that are posed to all respondents is limited, the CVS indeed proved useful for the analysis of the impact of the survey mode.

First, the typology of survey items turned out to be meaningful and useful. The construction of the typology itself required instructive preparatory discussions on the most important survey item properties. The typology is in a sense eclectic and adapts commonly used typologies to the type of surveys that are conducted by Statistics Netherlands. Given the limited number and range of survey items in the CVS, in 2014 the typology will be supplemented and adapted and will be applied to a wide set of surveys. The application is conducted in order to find out whether the survey item characteristics can be used to form questionnaire profiles, and whether the resulting profiles are predictive of mode-specific measurement bias and explain differences found in in parallel runs of old and new survey designs. For the LFS and CVS it was concluded that measurement effects dominate differences between modes after regular weighting adjustment. It may be conjectured that for other, similar surveys this should hold as well.

Don't know (DK) answers are a prominent source of differences between survey modes. In the CVS, DK answers were not explicitly offered in the interviewer-assisted modes. In the non-interviewer modes they were explicitly offered for around 80% of the items that are posed to all respondents. The item level investigation of the measurement bias showed that DK answers indeed make up a considerable part of the total difference between modes. The item level analysis did not reveal strong predictors for the occurrence of DK answers from the typology due to the limited number of survey items. However, some relations were found and confirmed by the category level analysis. The category level analysis showed large differences between questionnaire sections and a relatively large between individual variation. The latter implies that indeed DK answers cluster within individuals.

The analysis of the other response styles, primacy, recency, extreme response style and straightlining, only partially confirmed hypotheses in the literature. Respondents do not show primacy effects for the non-interviewer modes, but recency effects occur more often for interviewer modes. Respondents also show more extreme response style in the interviewer modes. Hence, the interviewer modes show more differentiation of respondent answers over item categories. This finding does not clearly confirm hypotheses in the literature. Straightlining occurs more often in telephone interviews and increases in frequency for non-interviewer modes during the questionnaire. Taken together with the DK answers, the various response styles support the observed large differences between the modes.

When generalizing results to other surveys, care is needed. First, the set of survey items in the CVS is limited, but more importantly is not a random, arbitrary set of items. The results in this paper cannot be applied to surveys with completely different profiles of survey item characteristics. Second, simultaneously occurring response styles may mask each other; A respondent that provides a lot of DK answers and that otherwise chooses the first answer category may not seem to show a primacy effect. In order to investigate response styles, it is imperative that a hierarchy is constructed based on the cognitive steps in the answering process. However, constructing such a hierarchy between response styles is by no means straightforward and is a topic for future theory building and research. Still, it is believed that some of the findings translate to other, similar surveys. The presentation and analysis of DK answers needs careful consideration; it plays a dominant role in differences between modes. Although they are strongly confounded with the order of the CVS questionnaire sections, the traces of non-differentiation and straightlining are likely to apply to similar surveys as well.. The results indicate that rating scale questions lead to a different mix of straightlining and non-differentation in different survey modes, which affects especially the reliability of underlying scales and factors. These effects should also be accounted for in questionnaire design. Overall, these results indicate that questionnaire design in surveys using multiple modes needs special

attention and might be a key to minimise measurement differences between modes. Further research of survey item characteristics and response styles may lead to more fundamental considerations when designing questions of mixed mode surveys, i.e., particular survey questions might be less suitable to certain modes while other are portable across all modes.

The analysis presented in this paper was restricted to hypotheses about mode-specific response styles, that are put forward by many authors in the literature. Although the CVS has a limited, specific set of survey items, the data from the mixed-mode experiment allow for more in-depth analyses of differences between modes. Some of the analyses should be detailed and elaborated through measurement error models assuming latent factors. The reader is referred to Klausch, Hox and Schouten (2013b) for the application of such models to the CVS.

6. References

Aichholzer, J. (2013), Intra-individual variation of extreme response style in mixed-mode panel studies, Social Science Research, 42 (3), 957 – 970.

Baumgartner, H., Steenkamp, J.E.M. (2001), Response styles in marketing research: a cross—national investigation, Journal of Marketing Research, 28, 143 – 156.

Beatty, P., Hermann, D. (2002), To answer or not to answer: Decision processes related to survey item nonresponse, pages71 – 87 In R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little (eds), *Survey Nonresponse*, Wiley: New York, USA.

Bennink, M., Moors, G., Gelissen, J. (2013). Exploring response differences between face-to-face and web surveys: A qualitative comparative analysis of the Dutch European Values Survey 2008, Field Methods, 25 (4), 319 – 338.

Billiet, J. B., McClendon, M.J. (2000), Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items, Structural Equation Modeling: A Multidisciplinary Journal 7(4), 608–628.

Buelens, B., Van den Brakel, J. (2011), Inference in surveys with sequential mixed-mode data collection, Discussion paper 201121, Statistics Netherlands, Heerlen, The Netherlands.

Buelens, B., Van der Laan, J., Schouten, B. (2012), MEPS – decomposition of mode effects for CVS and LFS, Research paper, BPA PPM-2012-02-25-BBUS-DLAN-BSTN, Statistics Netherlands, Den Haag, The Netherlands.

Buelens, B., Van der Laan, J., Schouten, B., Van den Brakel, J., Burger, J., Klausch, T. (2012), Disentangling mode-specific selection and measurement bias in social surveys, Discussion paper 201211, Statistics Netherlands, Den Haag, The Netherlands.

Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., Hope, S., Blake, M., Gray, M. (2011), A classification of question characteristics relevant to measurement (error) and consequently important for mixed mode questionnaire design, Paper presented at the Royal Statistical Society, October 11, London, UK.

Chang, L., Krosnick, J.A. (2009), National Surveys Via RDD Telephone Interviewing Versus the Internet, Public Opinion Quarterly 73(4), 641–678.

Córdova Cazar, A.L., Powell, R. (2013), Examining item nonresponse through paradata and respondent characteristics. A multilevel approach. Paper presented at the 68th AAPOR conference, May 16 – 19, Boston, USA.

Couper, M., (2008), Designing Effective Web Surveys, Cambridge U.P.

De Leeuw, E. (1992), Data Quality in Mail, Telephone, and Face to Face surveys, Amsterdam: TT-Publicaties.

De Leeuw, E. (2005), To mix or not to mix? Data collection modes in surveys, Journal of Official Statistics, 21, 1 – 23.

Dex, S., Gumy, J. (2011), On the experience and evidence about mixing modes of data collection in large-scale surveys where the web is used as one of the modes in data collection, National Centre for Research Methods Review paper, National Centre for Research Methods, UK.

Fowler, F.J. (1995), Improving Survey Questions: Design and Evaluation. Applied Social research Methods Series, 38. Sage Publications.

Fricker, S., Galesic, M., Tourangeau, R., Yan, T. (2005), An Experimental Comparison of Web and Telephone Surveys, Public Opinon Quarterly 69(3), 370–392.

Gallhofer, I., Scherpenzeel, A., Saris, W.E. (2007), The code-book for the SQP program, available at www.sqp.nl.

Greene, J., Speizer, H.,Wiitala, Y. (2008), Telephone and Web: Mixed-Mode Challenge, Health Services Research 41 (1p1), 230–248.

Greenleaf, E.A. (1992), Measuring extreme response style, Public Opinion Quarterly, 56 (3), 328 – 351.

Heerwegh, D., Loosveldt, G. (2011), Assessing Mode Effects in a National Crime Victimization Survey using Structural Equation Models: Social Desirability Bias and Acquiescence, Journal of Official Statistics, 27(1), 49 – 63.

Holbrook, A. L., Green, M.C., Krosnick, J.A. (2003), Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias, Public Opinion Quarterly 67(1), 79–125.

Jäckle, A., Roberts, C., Lynn, P. (2010), Assessing the effect of data collection mode on measurement, International Statistical Review 78, 3 – 20.

Klausch, T., Hox, J., Schouten, B. (2013a), Assessing the mode-dependency of sample selectivity across the survey response process, Discussion paper 201303, Statistics Netherlands, Den Haag, The Netherlands.

Klausch, L.T., Hox, J., Schouten, B. (2013b), Measurement effects of survey mode on the equivalence of ordinal rating scale questions, Sociological Methods and Research,42 (3), 227 – 263.

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

Lynn, P., Kaminska, O. (2012), The impact of mobile phones on survey measurement error, Public Opinion Quarterly, 77 (2), 586 – 605.

Saris, W.E., Gallhofer, I. (2007), Estimation of the effects of measurement characteristics on the quality of survey questions, Survey Research Methods, 1, 29 – 43.

Saris, W.E., Krosnick, J., Revilla, M., Shaeffer, E. (2010), Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options, Survey Research Methods,4,1, 61-79.

Schouten, B. (2010), Mode effecten in persoonsstatistieken. Een mixed-mode experiment op de iVM 2011, Project Initiation Document, BPA DMV-2010-09-02-BSTN, CBS, Den Haag, The Netherlands.

Schouten, B., Brakel, J. van den, Buelens, B., Laan, J. van der, Klausch, L.T. (2013), Disentangling mode-specific selection and measurement bias in social surveys, Social Science Research, 42, 1555 – 1570.

Schouten, B., Cobben, F., Van der Laan, J., Arends-Toth, J. (2014), The impact of contact effort and interviewer performance on mode-specific nonresponse and measurement bias, Discussion paper 201405, Statistics Netherlands, Den Haag, The Netherlands.

Schouten, B., Klausch, T. (2011), Mode effects in social surveys. MEPS analysis plan, Research paper, BPA DMV-2011-09-25-BSTN-LKAH, Statistics Netherlands, Den Haag, The Netherlands.

Schwarz, N. (1999), Self-Reports. How the questions shape the Answers, American Psychologist 54, 2, 93-105.

Tourangeau, R., Couper, M.. Conrad, F. (2004), Spacing, position and order: Interpretive heuristics for visual features of survey questions, Public Opinion Quarterly 68, 368-393.

Tourangeau, R., Couper, M., Conrad, F. (2007), Color, labels and interpretive heuristics for response scales, Public Opinion Quarterly 71, 91-112.

Tourangeau, R., Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 103, 299-314.

Ye, C. J. Fulton, Tourangeau, R. (2011), More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice, Public Opinion Quarterly, 75, 2, 349-365.

Appendix A: Labels of selected survey item characteristics

Label	*Property*
section	Questionnaire section to which item belongs: Living conditions neighbourhood, Neighbourhood problems, Feeling safe, Victimisation, Reporting to police, Satisfaction with police contact, Police performance neighbourhood, Police performance general Municipality performance, Unsafe places, Prevention I, Prevention II, Civic duty and police.
concept	Question asks for opinion, knowledge, fact
complexity - length	Yes-no question longer than 25 words
complexity - language	Yes-no difficult language in question
complexity - conditional	Yes-no question conditional
complexity - memory	Yes-no questions demands memory
complexity - hypothetical	Yes-no question contains hypothetical situation
complexity - calculations	Yes-no question requires calculations
complex1	Yes-no complex for at least one feature
complex2	Not complex, 1 or 2 complex features, 3 or more complex features
emotional	Yes-no question may evoke emotion
dimension	Yes-no question is multidimensional
mismatch	Yes-no mismatch between question and answer categories
formulation	Yes-no question formulated as statement
respscale1	Ordinal, nominal
respscale2	Ordinal, nominal, yes-no question
reportmark	Yes-no answer categories presented as report mark
ncategories	Number of categories 1 - 2, 3 - 5, 6 or more
range	Bipolar, unipolar
direction	Positive-negative, negative-positive
inbattery	Yes-no item in battery of items
relpos	Relative position in battery
abspos	Absolute position in questionnaire
instruction	Yes-no modes have the same instructions
DKavailable	Yes-no DK offered in non-interviewer modes
timeref	Past, present, future

7. Appendix B: Item level measurement bias

The item level bias, as defined in section 3.2, is estimated for each mode and each survey item characteristic. The figures B.1 to B.19 present the estimates per characteristic including and excluding the DK category.

Figure B.1: Item level bias for DK availability.

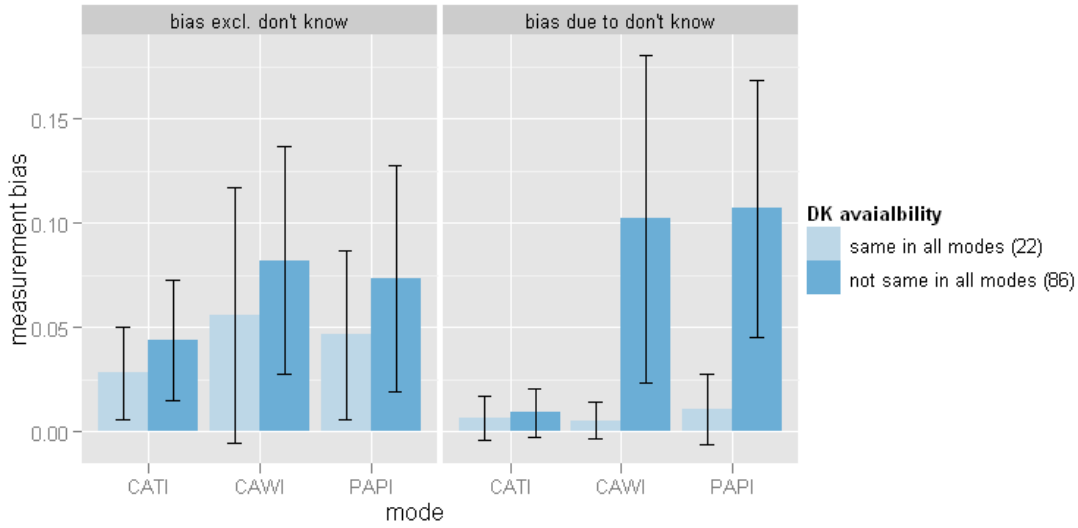


Figure B.2: Item level bias for concept.

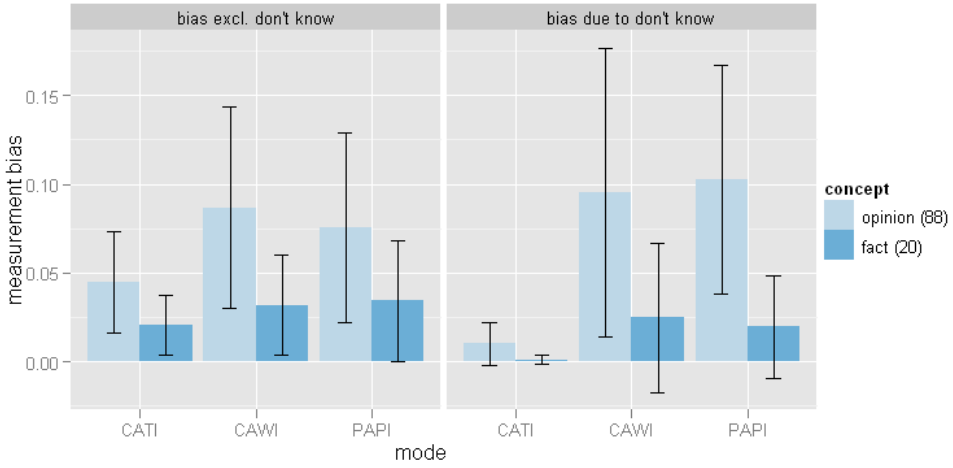


Figure B.3: Item level bias for complexity - length.

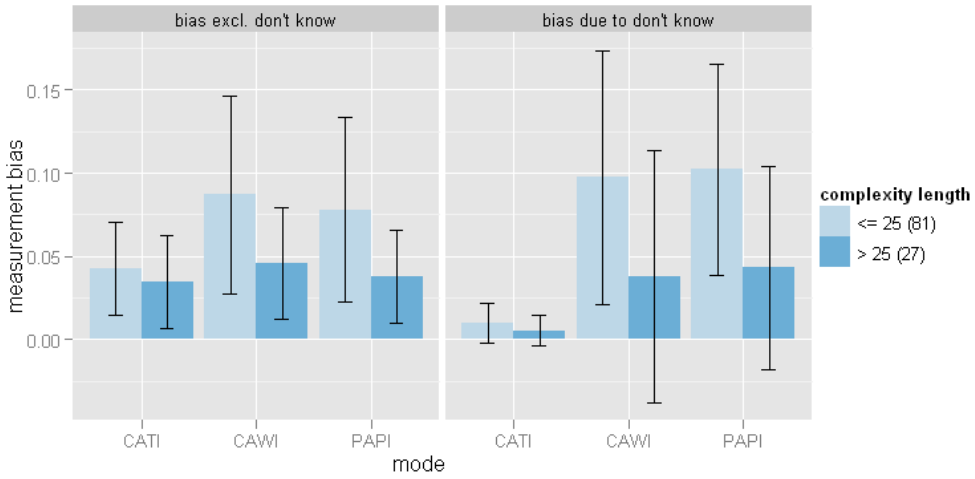


Figure B.4: Item level bias for complexity - language.

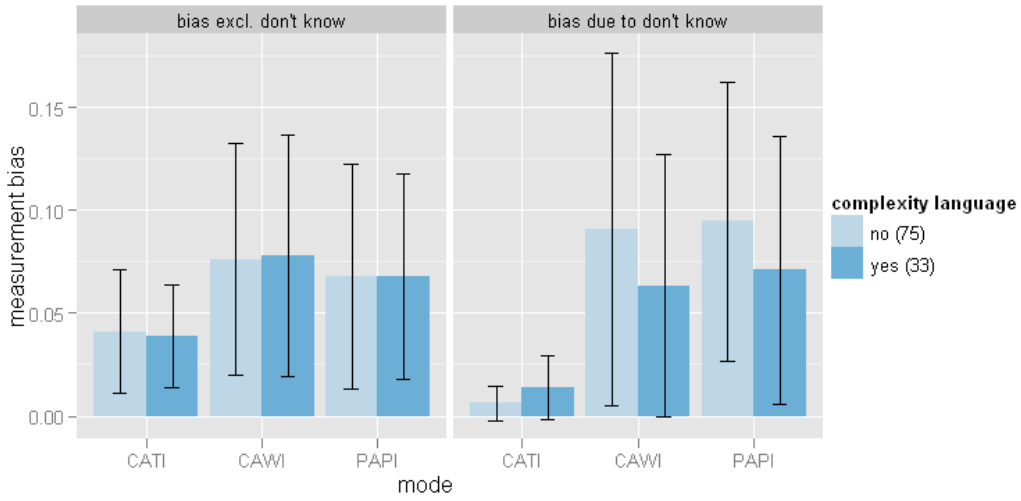


Figure B.5: Item level bias for complexity - conditional.

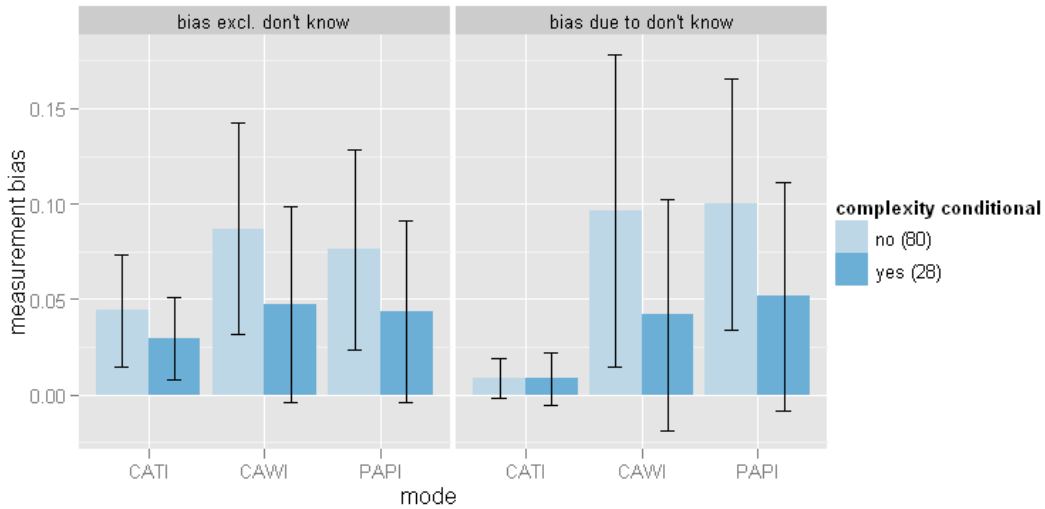


Figure B.6: Item level bias for complexity - memory.

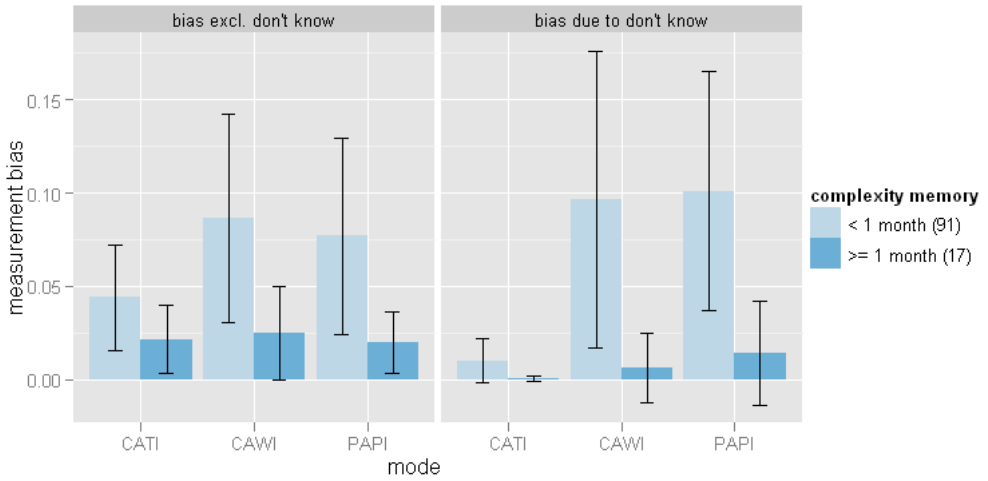


Figure B.7: Item level bias for complexity - hypothetical.

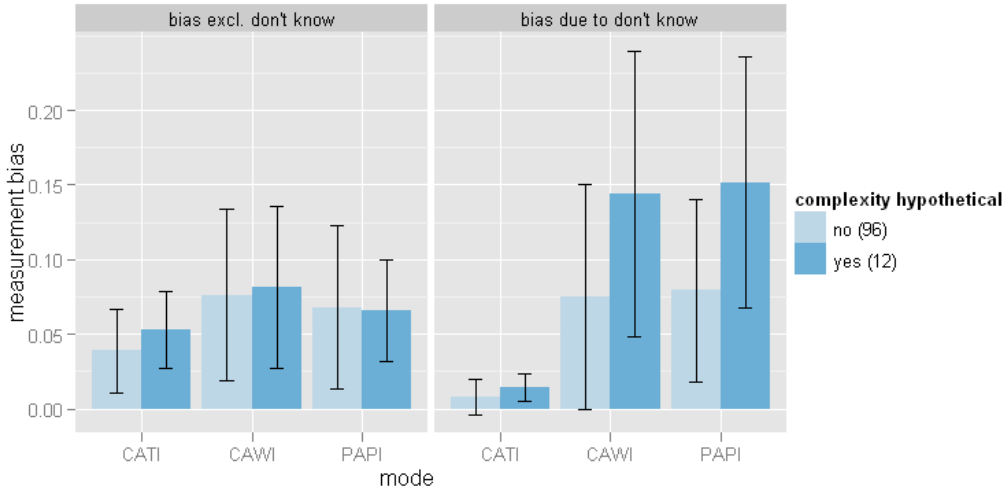


Figure B.8: Item level bias for complexity - calculations.

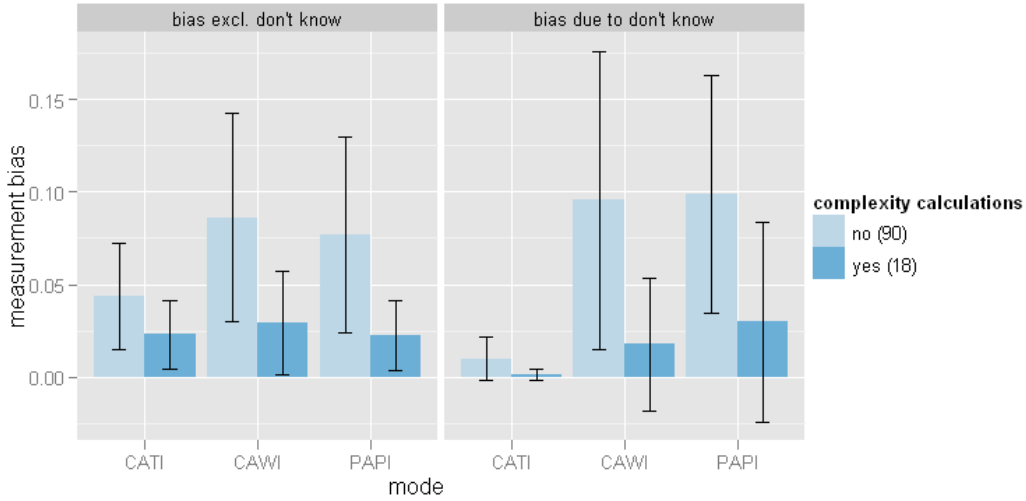


Figure B.9: Item level bias for aggregated complexity indicator complexity1.

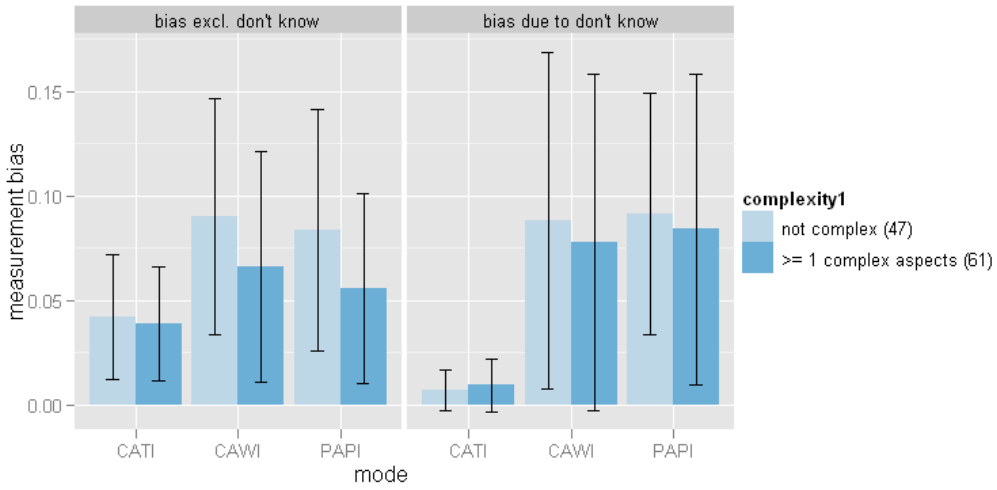


Figure B.10: Item level bias for aggregated complexity indicator complexity2.

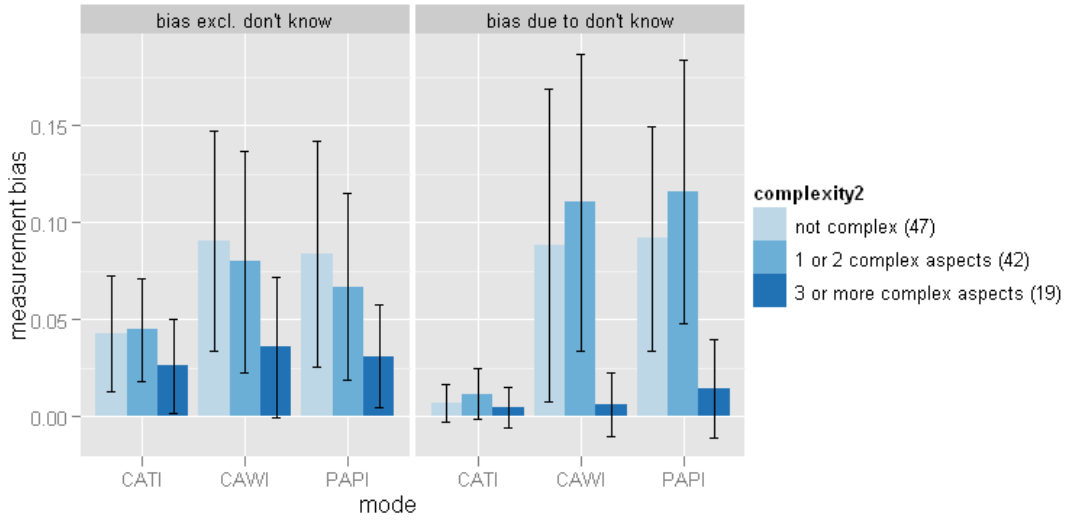


Figure B.11: Item level bias for response scale.

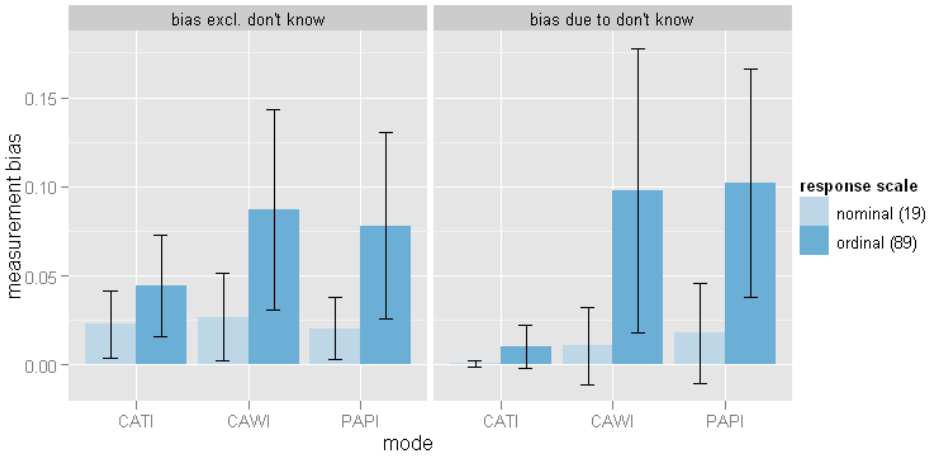


Figure B.12: Item level bias for response scale plus yes-no questions..

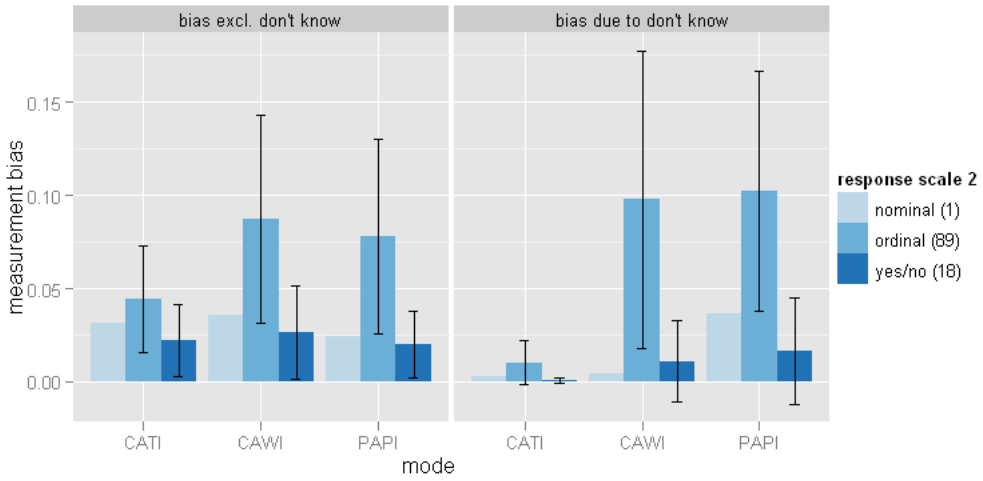


Figure B.13: Item level bias for 0-1 battery indicator..

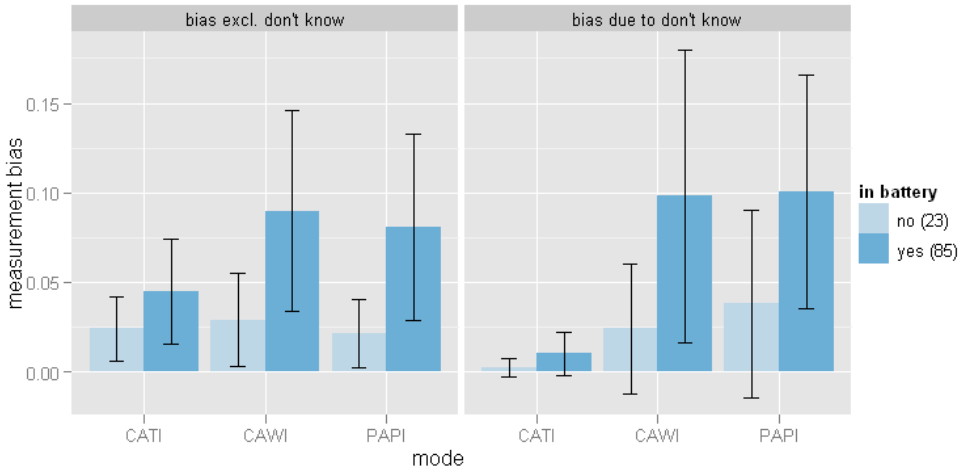


Figure B.14: Item level bias for number of categories.

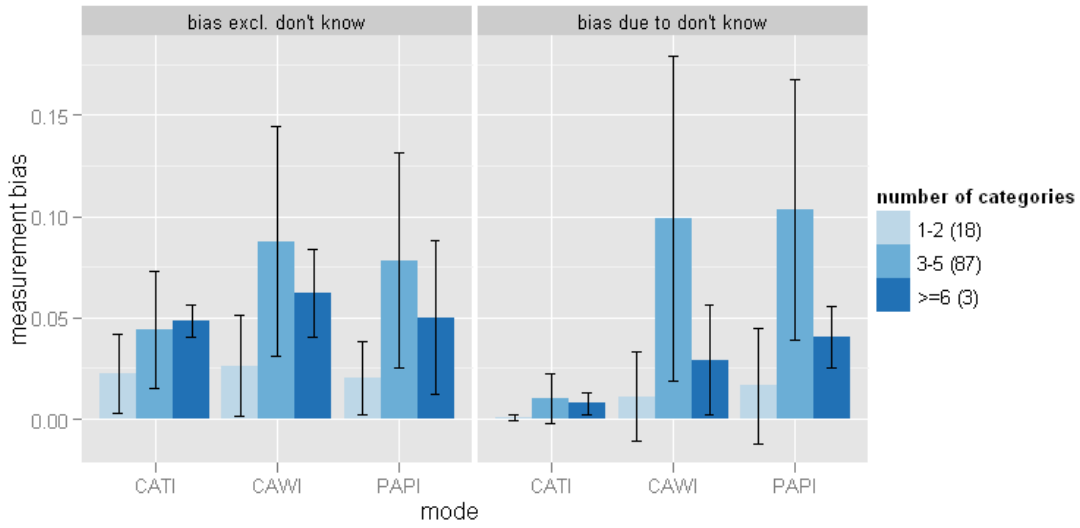


Figure B.15: Item level bias for direction.

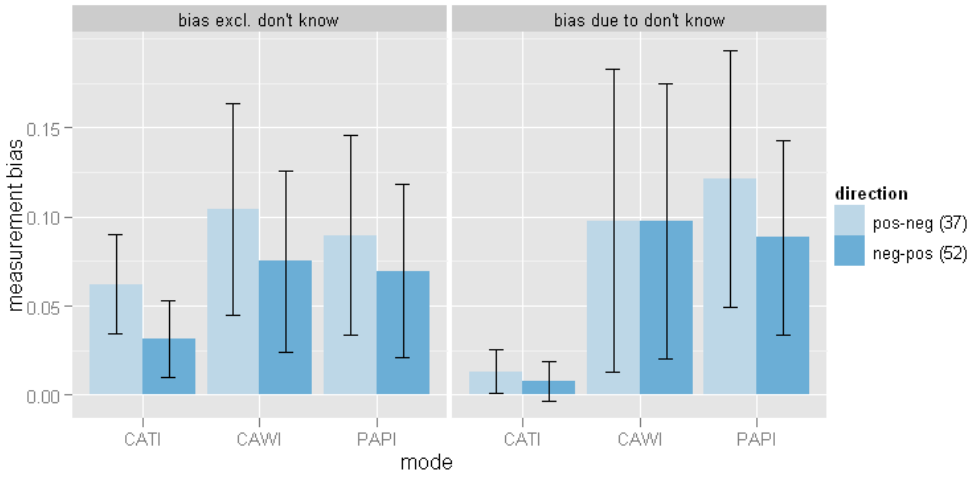


Figure B.16: Item level bias for range.

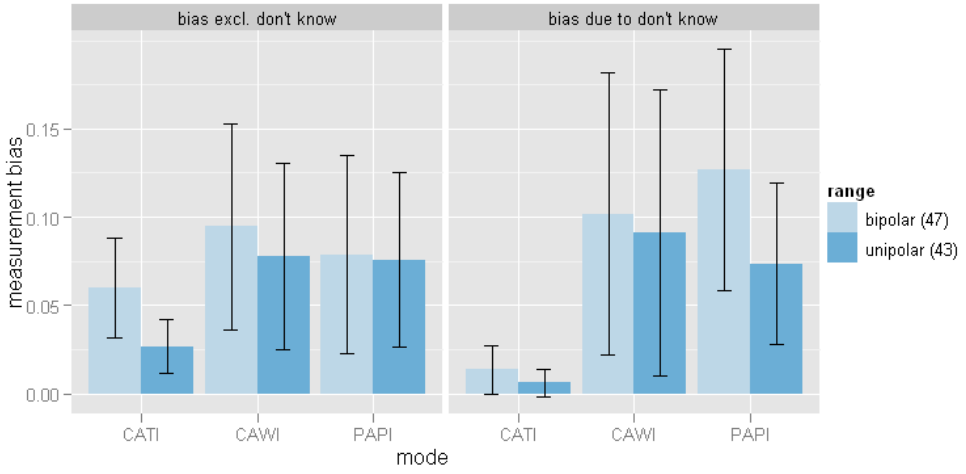


Figure B.17: Item level bias for formulation.

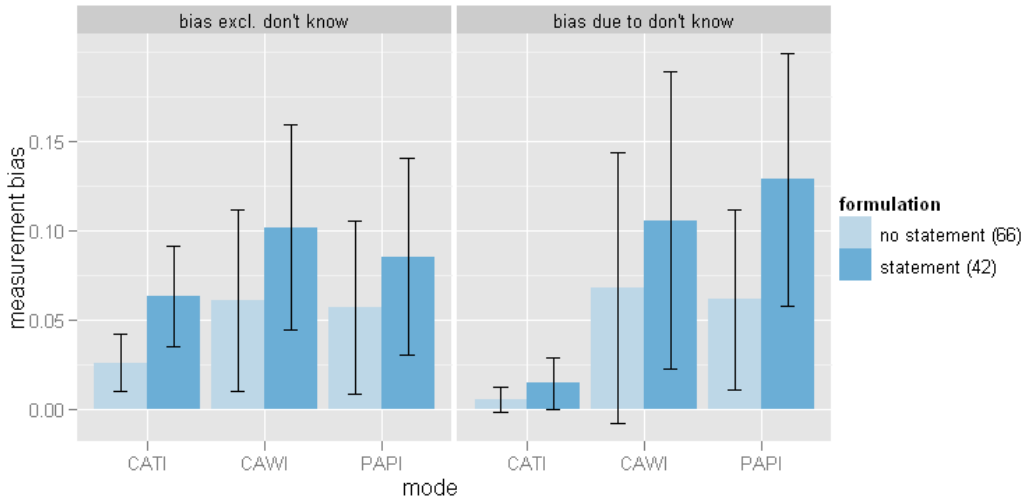


Figure B.18: Item level bias for mismatch.

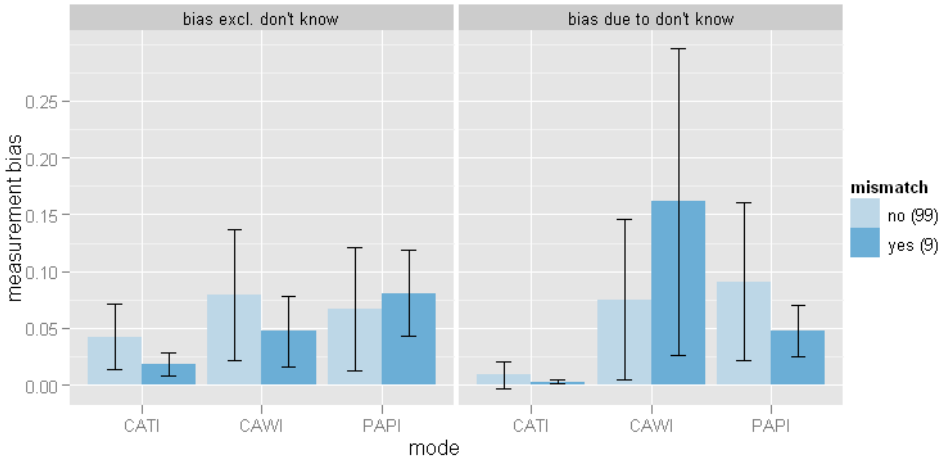
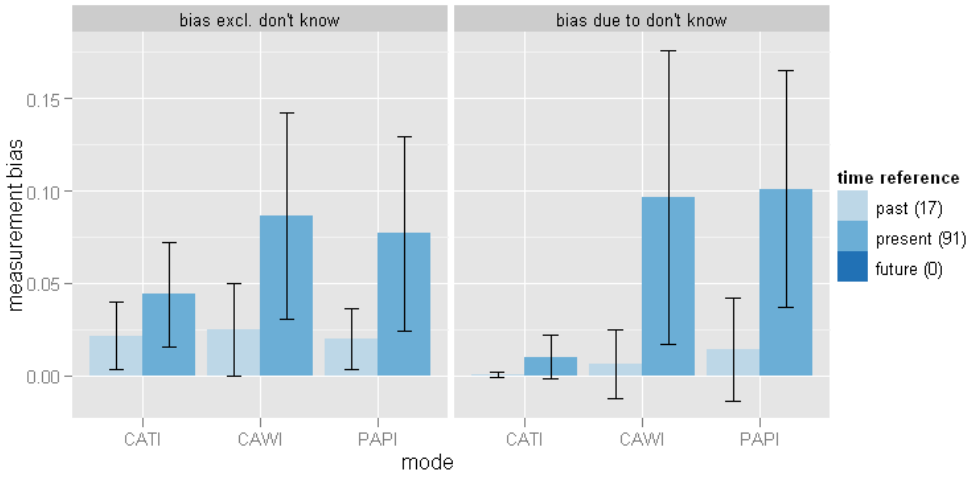


Figure B.19: Item level bias for time reference.



8. Appendix C: Regression coefficients for DK multilevel models

Table C.1: Regression coefficients for the base model: gender + age + mode + e_j + e_i

Random effects

Group	Variance
Person	2.14
Question	2.61

Fixed effects

Coefficient	Estimate	Std.Err.	Sig.
(Intercept)	-3.91	0.23	***
Gender=Male	0		
Gender=Female	0.39	0.09	***
Age=[0,25)	0		
Age=[25,35)	0.28	0.20	
Age=[35,45)	0.16	0.19	
Age=[45,55)	-0.20	0.19	
Age=[55,65)	-0.07	0.19	
Age=[65,75)	0.56	0.20	**
Age=[75,200)	1.07	0.23	***
Mode=Paper	0		
Mode=Web	0.13	0.09	

Table C.2: Regression coefficients for the model: gender + age + mode + section + e_j + e_i

Random effects			
Group	Variance		
Person	2.14		
Question	0.51		
Fixed effects			
Coefficient	Estimate	Std. Err.	Sig.
(Intercept)	-4.54	0.25	***
Gender=Male	0		
Gender=Female	0.39	0.09	***
Age=[0,25)	0		
Age=[25,35)	0.28	0.20	
Age=[35,45)	0.16	0.19	
Age=[45,55)	-0.20	0.19	
Age=[55,65)	-0.07	0.19	
Age=[65,75)	0.56	0.20	**
Age=[75,200)	0.107	0.23	***
Section=Liveability neighbourhood	0		
Section=Perception neighbourhood problems	0.96	0.23	***
Section=Feeling safe	0.98	0.37	**
Section=Victimisation	-2.29	0.29	***

	Section=Last police contact	-0.44	0.76	
	Section=Police performance - neighbourhood	2.57	0.27	***
	Section=Police performance - general	2.57	0.30	***
	Section=Municipality performance	1.87	0.35	***
	Section=Prevention I	0.03	0.41	
	Section=Prevention II	0.56	0.37	
	Section=Unsafe places	1.31	0.33	***
	Section=Civic duty and police	-0.89	0.32	**
	Mode=Paper	0		
	Mode=Web	0.13	0.09	

Table C.3: Regression coefficients for the model:
gender + age + mode + don't know explicit + e_j + e_i

Random effects			
Group	Variance		
Person	2.14		
Question	1.27		
Fixed effects			
Coefficient	Estimate	Std. Err.	Sig.
(Intercept)	-6.15	0.29	***
Gender=Male	0		
Gender=Female	0.39	0.09	***
Age=[0,25)	0		
Age=[25,35)	0.28	0.20	
Age=[35,45)	0.16	0.19	
Age=[45,55)	-0.20	0.19	
Age=[55,65)	-0.07	0.19	
Age=[65,75)	0.56	0.20	**
Age=[75,200)	1.07	0.23	***
Mode=Paper	0		
Mode=Web	0.13	0.09	
Don't know explicit=no	0		
Don't know explicit=yes	2.86	0.27	

Table C.4: Regression coefficients for the model:
gender + age + mode + range + e_j + e_i

Random effects			
Group	Variance		
Person	2.14		
Question	1.44		
Fixed effects			
Coefficient	Estimate	Std. Err.	Sig
(Intercept)	-3.04	0.25	***
Gender=Male	0		
Gender=Female	0.39	0.09	***
Age=[0,25)	0		

	Age=[25,35)	0.28	0.20	
	Age=[35,45)	0.16	0.19	
	Age=[45,55)	-0.20	0.19	
	Age=[55,65)	-0.07	0.19	
	Age=[65,75)	0.56	0.20	**
	Age=[75,200)	1.07	0.23	***
	Mode=Paper	0		
	Mode=Web	0.13	0.09	
	Range=Bipolar	0		
	Range=Unipolar	-0.87	0.25	***
	Range=Other	-3.29	0.35	***

Table C.5: Regression coefficients for the model:
gender + age + section + mode + complexity calculations + e_j + e_i

Random effects			
Group	Variance		
Person	2.14		
Question	0.46		
Fixed effects			
Coefficient	Estimate	Std. Err.	Sig
(Intercept)	-4.54	0.24	***
Gender=Male	0		
Gender=Female	0.39	0.09	***
Age=[0,25)	0		
Age=[25,35)	0.28	0.20	
Age=[35,45)	0.16	0.19	
Age=[45,55)	-0.20	0.19	
Age=[55,65)	-0.07	0.19	
Age=[65,75)	0.565	0.20	**
Age=[75,200)	1.07	0.23	***
Section=Liveability neighbourhood	0		
Section=Perception neighbourhood problems	0.96	0.22	***
Section=Feeling safe	-0.32	0.53	
Section=Victimisation	-4.40	0.69	***
Section=Last police contact	-2.56	0.96	**
Section=Police performance - neighbourhood	2.57	0.26	***
Section=Police performance - general	2.57	0.29	***
Section=Municipality performance	1.87	0.33	***
Section=Prevention I	0.04	0.39	
Section=Prevention II	0.56	0.36	
Section=Unsafe places	1.31	0.31	***
Section=Civic duty and police	-0.89	0.31	**
Mode=Paper	0		
Mode=Web	0.13	0.09	
Complexity Calculations=No	0		
Complexity Calculations=Yes	2.12	0.63	***

Table C.6: Regression coefficients for the model:
gender + age + section + mode + complexity conditional + e_j + e_i

Random effects				
	Group	Variance		
	veilignr	2.14		
	vraag	0.46		
Fixed effects				
	Coefficient	Estimate	Std. Err.	Sig
	(Intercept)	-4.54	0.24	***
	Gender=Male	0		
	Gender=Female	0.39	0.09	***
	Age=[0,25)	0		
	Age=[25,35)	0.28	0.20	
	Age=[35,45)	0.16	0.19	
	Age=[45,55)	-0.20	0.19	
	Age=[55,65)	-0.07	0.19	
	Age=[65,75)	0.56	0.20	**
	Age=[75,200)	1.07	0.23	***
	Section=Liveability neighbourhood	0		
	Section=Perception neighbourhood problems	0.92	0.22	***
	Section=Feeling safe	0.35	0.40	
	Section=Victimisation	-3.32	0.42	***
	Section=Last police contact	-1.48	0.79	.
	Section=Police performance - neighbourhood	2.57	0.26	***
	Section=Police performance - general	2.57	0.29	***
	Section=Municipality performance	0.83	0.46	.
	Section=Prevention I	0.04	0.39	
	Section=Prevention II	0.35	0.36	
	Section=Unsafe places	1.31	0.31	***
	Section=Civic duty and police	-1.28	0.33	***
	Mode=Paper	0		
	Mode=Web	0.13	0.09	
	Complexity Conditional=No	0		
	Complexity Conditional=Yes	1.04	0.32	**

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
empty cell	Not applicable
2013–2014	2013 to 2014 inclusive
2013/2014	Average for 2013 to 2014 inclusive
2013/'14	Crop year, financial year, school year, etc., beginning in 2013 and ending in 2014
2011/'12–2013/'14	Crop year, financial year, etc., 2011/'12 to 2013/'14 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia)
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

60083201416 X-10