Discussion Paper

Can survey item characteristics relevant to mode-specific measurement error be coded reliably?

2015 | 22

Frank Bais
Barry Schouten
Peter Lugtig
Vera Toepoel
Judit Arends-Tóth
Salima Douhou
Natalia Kieruj
Mattijn Morren
Corrie Vis

Content

1. Introduction 4

2. The item characteristics 7
   - 2.1 Towards a typology of item characteristics 7
   - 2.2 Highlighting the most relevant item characteristics 10

3. Method 14
   - 3.1 Surveys 14
   - 3.2 The allocation of coders 15
   - 3.3 Statistics 16

4. Results 17
   - 4.1 Relative frequencies 17
   - 4.2 Intercoder reliabilities 18
   - 4.3 Explaining low intercoder reliabilities 20

5. Coping with low intercoder reliability 23
   - 5.1 Option 1: Excluding survey items 24
   - 5.2 Option 2: Redefining and refining item characteristics 24
   - 5.3 Option 3: Computerizing the definition and demarcation of item characteristics 25
   - 5.4 Option 4: Using item characteristic scales with multiple applicability categories 25

6. Discussion 27

References 29

Appendix A 32

Appendix B 34

Summary

In multi-mode questionnaire design, usually some consideration is given to mode-specific measurement error. Despite this consideration, however, these measurement effects can be unexpectedly large. For this reason, there is a strong incentive to better predict measurement effects. This may be done by constructing profiles of a questionnaire, in which relevant item characteristics are summarized. For all items of a survey, these item characteristics need to be coded and combined. In this paper, we evaluated a list of item characteristics that literature has reported as relevant to mode-specific measurement error. Most importantly, we evaluated the reliability of the coding of such characteristics. Our results showed that intercoder reliability can be low for the most relevant characteristics. This may be explained by the difficulty of defining the item characteristics and the inherent subjectivity with which these item characteristics are coded. Finally, some suggestions are made for coping with low intercoder reliability.

1. Introduction

In this study, we anticipated the current interest in measurement error that is dependent on the used survey mode. The occurrence and scope of this mode-specific measurement error are partly influenced by the characteristics of the items of the survey (Tourangeau, Rips, and Rasinski 2000). Examples of item characteristics are the content of the question (Gallhofer, Scherpenzeel, and Saris 2007) and the extent to which an item contains an emotional charge or sensitive information (Lensvelt-Mulders 2008). To investigate this relation between mode-specific measurement error and survey item characteristics, we coded the items of general population surveys of Statistics Netherlands and CentERdata on characteristics that are assumed to be influential in evoking mode-specific measurement error. For this purpose, we constructed an item characteristics scheme that is based on the Survey Quality Predictor (SQP) typology of Saris and Gallhofer (2007) and Gallhofer et al. (2007), and on the typology of Campanelli et al. (2011). In order to employ this scheme, however, it is imperative that item characteristics can be coded reliably. If items cannot be coded reliably on their characteristics unambiguously, the relation between item characteristics and mode-specific measurement error cannot be investigated reliably. Therefore, in this paper, we investigated the extent to which the coding of item characteristics can be done reliably.

In current survey practice, a variety of survey modes is used to collect data, including multiple modes within single surveys, which are called mixed-mode surveys (De Leeuw 2005). Combining different modes within surveys has the benefit of increasing the total response rate, as well as reducing survey costs. However, this cost efficiency of mixed-mode designs may come along with potential changes, as answers to the same survey questions that are asked under different modes are not necessarily equivalent (Klausch, Hox, and Schouten 2013). Observed differences in survey outcomes when using different data collection modes are called mode effects (Buelens and Van den Brakel 2011). Different survey modes may produce mode effects, which can be the result of mode-specific measurement effects (Buelens and Van den Brakel 2011), having their impact on survey data quality (Roberts 2007). Mode-specific measurement effects refer to the influence of a survey mode on answers that respondents give (Vannieuwenhuyze, Loosveldt, and Molenberghs 2010) and arise as different modes evoke different kinds of measurement errors while reporting an answer (Buelens and Van den Brakel 2011; Buelens et al. 2012).

Despite this problem of mode-specific measurement effects (in the remainder of this paper simply called ‘measurement effects’), taking into account anticipated mode effects is often done insufficiently, as this is difficult and may cost too much time. Also, users who launch surveys are not always aware of measurement effects due to specific questionnaire characteristics interacting with a mixed mode design. Choosing a survey questionnaire design may have unknown consequences for the quality of the survey questions (Saris and Gallhofer 2007) and it is difficult to come to a design that is optimal for all items, as many items are affected by the survey mode to

different extents. In order to investigate the relation between survey mode, the type of survey item, and measurement effects, Beukenhorst et al. (2013) developed a coding scheme with variables characterizing the survey items of the Crime Victimisation Survey (CVS) 2011. Their coding scheme was based on the typologies of Campanelli et al. (2011), Gallhofer et al. (2007), and Saris and Gallhofer (2007). With such a questionnaire typology, one can try to explore to what extent questionnaire characteristics can explain different answering behaviour in different survey modes.

An example of a characteristic of a survey question is to what extent the question asks for a socially desirable answer (Lensvelt-Mulders 2008). The tendency of a respondent to give a socially desirable answer is relatively larger when an interviewer is administering the interview (Kreuter, Presser, and Tourangeau 2008; Tourangeau and Yan 2007). Surely, it is relatively harder to admit deviating from a social norm towards a present interviewer than anonymously filling out a social norm deviation in a web-administered interview. Consequently, answers to questions that ask for socially desirable answers may differ between interviewer and non-interviewer modes. To put it over-simplified, in the case of social desirability, an answer to such questions in non-interviewer mode may be 'more true' than in interviewer mode, which may be considered a measurement effect. In short, characteristics of survey items are related to the occurrence of mode-specific measurement error. Therefore, it may be useful to construct a typology of item characteristics for complete surveys.

The characteristics of the survey questionnaire and its items are called questionnaire profiles (Beukenhorst et al. 2013), summarizing item characteristics that might lead to undesirable answering behaviour. By constructing typologies for questionnaires, we can investigate to what extent they are able to explain variation in answering behaviour in mixed-mode surveys. Such questionnaire profiles may be helpful in anticipating measurement effects, given successful identification and enough explanatory power. In order to construct questionnaire profiles, survey items need to be coded on their characteristics. By their experiment, Beukenhorst et al. (2013) made a first attempt to characterize a whole survey questionnaire to investigate measurement effects. They concluded that 'measurement effects dominate differences between modes after regular weighting adjustment', but they used only one survey on a specific topic and a restricted selection of items in their study. Hence, the question is to what extent measurement effects may be found for multiple surveys on a broad range of topics and for a large selection of different kind of items. Our study will be a first step toward investigating measurement effects in the mixed-mode context for multiple surveys.

Before we are able to use questionnaire profiles to investigate measurement effects, however, we need to know to what extent items can actually be coded on their characteristics reliably. If multiple item coders would not agree on how to categorize certain items on certain characteristics, one could wonder to what extent complete questionnaire profiles may be constructed at all. For instance, when two coders would disagree on whether an item contains sensitive information, this specific item could not be characterized to the extent to which it contains sensitive information based on the judgment of the two coders. As a consequence of intercoder

disagreement, certain item characteristics may need to be omitted from the typology. Beukenhorst et al. (2013) removed the characteristics sensitive information and centrality, as they evoked too much disagreement. Thus, to be able to construct questionnaire profiles, intercoder agreement in coding the item characteristics is a prerequisite. The current study is about investigating this intercoder agreement on various item characteristics for all items in multiple surveys.

Beukenhorst et al. (2013) executed their experiment by using an item coding scheme that was partly based on the SQP typology of Saris and Gallhofer (2007) and Gallhofer et al. (2007), and on the typology of Campanelli et al. (2011). On the basis of these typologies and extensive discussions of the coders involved, we constructed a questionnaire characteristics scheme consisting of both question and answer characteristics. By coding multiple questionnaires of Statistics Netherlands and CentERdata, we can investigate the intercoder reliability on these characteristics for many items that highly range over various general population topics. In case the intercoder reliability is relatively high on certain characteristics, a questionnaire profile based on these characteristics may be constructed relatively easily. In case the intercoder reliability is relatively low on certain characteristics, we need to explain this low reliability and how to cope with it. For this study, we coded 11 surveys on 15 question characteristics and one answer characteristic to answer three research questions. We 1) investigated the intercoder reliability for each item characteristic over the items of all surveys together; 2) tried to explain potential low intercoder reliability, and; 3) gave suggestions about how to cope with such low reliability.

The motivation for our study and coding scheme is set in the context of multi-mode surveys. Obviously, the scheme may assist any design choice considering measurement error and our findings are not limited to the multi-mode context only. We do concentrate here, however, on the item characteristics that are most relevant to this multi-mode context. From here, we will first motivate the chosen question and answer characteristics in section 2. In section 3, we will present all surveys for which these characteristics are coded and elaborate on the actual coding procedure and the statistics that will be calculated. In section 4, we will present all statistical results of the actual coding experiment and answer research questions 1) and 2). In section 5, we will answer research question 3) and suggest ways of coping with low intercoder reliability. In section 6, we will conclude with a discussion of these results.

2. The item characteristics

In this section, we will first elaborate on a pilot study that we executed and to what changes this resulted for the list of item characteristics that was used for the actual coding study. Second, we will present the final list of item characteristics as used in the current study. And third, we will give a motivation on the item characteristics that evoke measurement effects according to the literature.

2.1 Towards a typology of item characteristics

Saris and Gallhofer (2007) and Gallhofer et al. (2007) created a typology of item characteristics to predict the quality of survey items in terms of validity and reliability. Campanelli et al. (2011) classified item characteristics that are regarded as relevant to mode-specific measurement error and thus to mixed-mode questionnaire design. Based on these typologies, we constructed a list of item characteristics to be used in a pilot study. The pilot study was set up to investigate the factual occurrence of each item characteristic and to check for potential difficulties during the coding process. The list consisted of 28 item characteristics. Expert discussion meetings for the involved coders were planned before the start of the pilot study. During these meetings, the item characteristics were discussed extensively. After these meetings, consensus among the researchers was reached about the exact definitions and accompanying categories of the selected item characteristics. After discussing and defining the item characteristics, the researchers decided to use all the selected characteristics for the pilot study.

In the pilot study, a selection of 31 items of the Dutch Labour Force Survey (LFS) and 50 items of the three LISS core studies 'Personality', 'Politics & Values', and 'Religion & Ethnicity' was coded on its item characteristics by six of the co-authors. We chose these 81 items in such a way to capture as many of the item characteristics as possible. The four chosen surveys differed substantially in topic, so that a relatively broad range of topics and item characteristics was covered. After the pilot study, its evaluation, and several follow-up meetings, the list of item characteristics was finalized for the actual coding study with the adjustments from the pilot study. In total, 29 item characteristics were selected for the actual coding study. Thirteen of these characteristics were considered to be codeable on their true category unambiguously. These 13 characteristics were coded by a single coder and are not taken into consideration for this paper. See table 2.1.1 below for an overview of the 16 item characteristics that are involved in the current study and table 5.4.1 in Appendix A for an overview of these remaining 13 item characteristics.

After the pilot study, a few important changes were made. Based on the results of the pilot study, the item characteristic filter question was split up into two separate item characteristics; one item characteristic that asks whether an item is factually a

filter question (see table 5.4.1); and a second item characteristic that asks whether an item could make the respondent presume that it would be a filter question, regardless of whether the item factually is a filter question (see table 2.1.1). Also, for the item characteristics sensitive information and centrality (see table 2.1.1), the middle option was removed so that only two coding categories remained for these characteristics: Characteristic not applicable and characteristic applicable. This was done because it appeared to be difficult for the coders to choose between two gradual categories of applicability of these characteristics. Finally, a few item characteristics were defined more strictly, as it was not clear for some items what coding category had to be chosen based on their definitions.

2.1.1 Definitions of the item characteristics, their coding numbers and categories, and references

Item characteristic	Definition of the item characteristic as used in the current study	Coding number and categories	References
Content of the question	What kind of topic or aspect is the item about?	1 factual behaviour 2 otherwise factual 3 opinion 4 satisfaction 5 otherwise subjective	Campanelli et al. 2011; Gallhofer et al. 2007; Lozar Manfreda and Vehovar 2002; Saris and Gallhofer 2007; Schonlau et al. 2003
Emotional charge	Does the item contain potentially emotional words or a potentially emotional charge?	0 not applicable / 1 applicable	Lensvelt-Mulders 2008
Sensitive information	Does the item contain sensitive information of some societal, menial or personal kind?	0 not applicable / 1 applicable	Campanelli et al. 2011; Gallhofer et al. 2007; Kreuter et al. 2008; Lensvelt-Mulders 2008; Saris and Gallhofer 2007; Tourangeau and Yan 2007
Presumption of a filter question	Might the respondent be able to presume the item to be a filter question?	0 not applicable / 1 applicable	Bosley et al. 1999; Eckman et al. 2014; Kreuter et al. 2011
Centrality	Does the item go beyond the interest, knowledge or experience of the	0 not applicable / 1 applicable	Gallhofer et al. 2007; Saris and Gallhofer 2007; Van der Zouwen 2000

Item characteristic	Definition of the item characteristic as used in the current study	Coding number and categories	References
	respondent?		
Question complexity 1: Difficult language usage	Does the item contain unknown or difficult words or complex sentences?	0 not applicable / 1 applicable	Beukenhorst et al. 2013; Van der Zouwen 2000
Question complexity 2: Conditions	Does the item contain conditions?	0 not applicable / 1 applicable	Beukenhorst et al. 2013; Van der Zouwen 2000
Question complexity 3: Memory	Does answering require some kind of memory?	0 no memory 1 non-specific memory 2 memory < 1 month ago 3 memory > 1 month ago	Van der Vaart et al. 1995; Van der Zouwen 2000
Question complexity 4: Hypothetical situation	Does the item refer to a concrete, specific hypothetical situation in the future?	0 not applicable / 1 applicable	Van der Zouwen 2000; Van der Zouwen and Dijkstra 1996
Question complexity 5: Calculations	Does answering require the performance of some kind of calculation?	0 not applicable / 1 applicable	Beukenhorst et al. 2013; Van der Zouwen 2000
Question complexity 6: Ambiguity	Does the item contain multiple sub-questions or is the item otherwise potentially confusing?	0 not applicable / 1 applicable	Campanelli et al. 2013; Foddy 1993; Fowler and Mangione 1990; Van der Zouwen 2000
Response complexity	Do the answering options contain unknown or difficult words or complex sentences, or do they require the execution of some kind of performance?	0 not applicable / 1 applicable	Campanelli et al. 2011; Gallhofer et al. 2007; Saris and Gallhofer 2007
Time reference	What time period does the item refer to?	1 past / 2 present / 3 future	Gallhofer et al. 2007; Saris and Gallhofer 2007
Mismatch	Do the question and its answering options match?	0 not applicable / 1 applicable	Beukenhorst et al. 2013; Van der Zouwen 2000
Formulation	Is the item formulated	0 not	Fowler 1995; Gallhofer

Item characteristic	Definition of the item characteristic as used in the current study	Coding number and categories	References
	as a statement?	applicable / 1 applicable	et al. 2007; Saris and Gallhofer 2007; Saris et al. 2010; Ye et al. 2011
Clarification	Does the item contain some kind of clarification?	0 not applicable / 1 applicable	Gallhofer et al. 2007; Saris and Gallhofer 2007; Van der Zouwen 2000

2.2 Highlighting the most relevant item characteristics

From here, we will give a motivation for the inclusion of six specific item characteristics. We will elaborate on these item characteristics specifically, as they are considered particularly influential in evoking measurement effects according to the literature. See table 2.2.1 for examples of items that contain or are related to the most influential item characteristics that we have selected for the current study (see section 3.2).

2.2.1 The most relevant item characteristics with examples of questions containing or relating to these characteristics

Characteristic	Example of a question containing or relating to the characteristic
Content of question: Factual behaviour	”What sport do you practice?”
Content of question: Otherwise factual	“In which year did you enter into employment with your current employer?”
Content of question: Opinions	“What do you think of Mark Rutte?”
Content of question: Satisfaction	”How satisfied are you with the life you lead at the moment?”
Content of question: Otherwise subjective	“How do you feel at the moment?”
Question complexity: Difficult language usage	“How much is the total gross amount that you received in 2007 as WAO, IVA or WGA (preferably as stated on your tax reporting statement)?”
Centrality	“In politics, a distinction is often made between ‘the left’ and ‘the right’. Where would you place yourself on the scale below, where 0 means left and 10 means right?”

Characteristic	Example of a question containing or relating to the characteristic
Sensitive information	"For which party did you vote in the parliamentary elections of 12 September 2012?"
Emotional charge	"Are all your children still alive?"
Presumption of filter question	"Have you ever performed paid work in the past (even if it was only for one or several hours per week or for a brief period)?"
Response complexity	"The rating scale with circles below is used to assess the degree to which people feel connected to other people. Please indicate to what extent you generally feel connected to other people."

Question complexity

A high degree of question difficulty has a negative effect on the quality of the response to that question (Van der Zouwen 2000). In our study, the omnibus item characteristic question complexity consists of six separate characteristics: Difficult language usage, conditions, memory, hypothetical situation, calculations, and ambiguity. According to the cognitive response model (Jenkins and Dillman 1997; Tourangeau et al. 2000), the presence of these characteristics in items may impose difficulty for the respondent in, for instance, understanding the question, or in retrieving or judging relatively complex information, possibly leading to measurement effects.

The characteristic difficult language usage refers to the use of unknown or difficult words or complex sentences within the item (Beukenhorst et al. 2013), possibly having a negative influence on response quality (Van der Zouwen 2000). The characteristic conditions refers to specifically including and/or excluding certain aspects in/from the answer and the characteristic calculations refers to the performance of some kind of mathematical calculation (Beukenhorst et al. 2013). Both characteristics may relate to a relatively high cognitive burden on the respondent while answering a question (Lenzner, Kaczmirek, and Lenzner 2009; Tourangeau et al. 2000; Van der Zouwen 2000). The characteristic hypothetical situation refers to imagining a fictitious or hypothetical situation (Van der Zouwen and Dijkstra 1996). Respondents may have difficulty in accepting the reality of a hypothetical situation or with imagining a situation in the far future (Van der Zouwen 2000).

The characteristic memory refers to retrieving information from the past. Questions requiring information retrieval from the past are retrospective questions that may have a negative effect on response quality (Van der Vaart, Van der Zouwen, and Dijkstra 1995; Van der Zouwen 2000), especially when no recall aiding devices are used (Van der Vaart 1996). The characteristic ambiguity refers to questions that are double barrelled (Bassili and Scott 1996; Campanelli et al. 2011; Foddy 1993; Fowler and Mangione 1990) or otherwise have an unclear meaning of wording (Van der Zouwen 2000). Concerning these six characteristics about question complexity, differences in interviewer-administered versus self-administered survey modes may be expected. In interviewer-administered modes, the respondent can be assisted in

answering a particular question containing some form of complexity. In self-administered modes, however, the respondent does not have this assistance. Respondents can take as much time as they need to understand and answer the particular question (Beukenhorst et al. 2013), but the probability on some form of satisficing may be relatively high in self-administered modes (Krosnick 1991). Therefore, mode-specific measurement effects can be expected.

Centrality

Centrality is particularly about the concept or content of the question. When the item is about a topic that extends beyond the knowledge, experience or interest of the respondent, this is called centrality (Gallhofer et al. 2007; Saris and Gallhofer 2007). This is for instance the case when an item deals with a political or religious topic, which is not 'central' in the life of relatively many respondents. The respondent might be either reluctant or incapable to answer items that are non-central or hardly accessible (Van der Zouwen 2000) to them. The respondent may be assisted or stimulated by the interviewer in interviewer-administered modes concerning such topics, while this assistance or stimulance is less evident in self-administered modes. This difference in interviewer-administered versus self-administered modes makes centrality sensitive to possible measurement effects.

Content of the question

Concerning content of the question, an item may belong to one of the following categories: Factual behaviour, otherwise factual, opinions, satisfaction, or otherwise subjective (Campanelli et al. 2011; Gallhofer et al. 2007; Saris and Gallhofer 2007). Here, otherwise factual refers to items asking for factual data other than factual behaviour. Otherwise subjective refers to items asking for thoughts, feelings or emotions other than opinions or satisfaction of the respondent. We defined factual behaviour and otherwise factual as objective categories that are observable and measureable, as opposed to opinions, satisfaction, and otherwise subjective, which are considered subjective categories. The goal is to distinguish objective versus subjective categories, with the latter categories being more sensitive to the predispositions of the respondent. Especially subjective questions are sensitive to the presence of an interviewer and may be more prone to measurement effects than factual questions (Campanelli et al. 2011; Lozar Manfreda and Vehovar 2002; Schonlau et al. 2003).

Sensitive information

Some items ask for sensitive information that may be perceived as being more or less threatening by respondents (Lensvelt-Mulders 2008). Sensitive questions may be about private, stressful or sacred issues. Answering sensitive questions may evoke emotional responses or the potential fear of stigmatization on the part of the respondent or his social group (Lensvelt-Mulders 2008). In effect, a question is sensitive when it asks respondents to admit that they have violated a social norm (Tourangeau and Yan 2007). This may for instance the case when items ask for information about former or current drug or alcohol use. As a result, respondents might be reluctant to answer the question and may tend to avoid or distort their answer. Interviewer-administered modes may strongly facilitate the tendency to give

socially desirable answers, while this effect will be much less strong in case of self-administered modes. Therefore, this characteristic in particular is sensitive to possible measurement effects due to mode differences and may well evoke socially desirable answering (Campanelli et al. 2011; Kreuter et al. 2008; Tourangeau and Yan 2007).

Emotional charge

This item characteristic is related to the characteristic sensitive information, but is more narrow and specific. In some cases, emotional charge may be considered an intrinsic subcategory of the characteristic sensitive information, potentially evoking strong personal negative emotions (Lensvelt-Mulders 2008). An item contains a potentially emotional charge when it is about for instance a former traumatic experience or another event that the respondent fell victim to. Emotionally charged items and items asking for sensitive information may be distinguished by the idea that the former, in contrast to the latter, will probably be answered candidly. Nevertheless, when a question contains an emotional charge or word, respondents might be either reluctant or very eager to answer it (Beukenhorst et al. 2013). In interviewer-administered modes, the interviewer may mitigate this effect by stimulating the respondent to answer in any case. In self-administered modes, however, there is no interviewer present to regulate potential emotions of the respondent. Thus, measurement effects regarding these mode differences are likely.

Presumption of a filter question

In some surveys more than in others, certain questions may lead to follow up items. These questions are so-called filter questions. Dependent on the content of a question and on the format of asking filter questions, respondents may presume a question to be a filter question (Eckman et al. 2014; Kreuter et al. 2011). When presuming a question to be a filter question, respondents might be motivated to give an answer that avoids them from having to answer follow-up questions (Bosley, Dashen, and Fox 1999). The item characteristic presumption of a filter question was considered a separate characteristic by the involved researchers as a result of a pilot study (see section 2.2). The coders experienced difficulty in distinguishing an item as a factual filter question versus as a question of which the respondent could presume to be a filter question, regardless of whether the question factually is a filter question. Some respondents could avoid a filter question in case they presume a question to be one. It is likely that a respondent's presumption of a filter question may partly be determined by the presence or absence of an interviewer. In mail and web mode, respondents could scroll through the survey to check for follow up questions and filter questions that are repeated later in the survey may be recognized more easily. In personal and telephone mode, respondents do not have the option to scroll through the survey, making filter questions relatively more difficult to detect. Therefore, this item characteristic may be sensitive to measurement effects.

It is important to note, however, that we used the characteristic presumption of a filter question without considering the mode in which surveys were administered. This means that we did not account for possible mode differences concerning visual aspects or scroll through options during the coding process. The benefit of a mode-

free coding process is that items are purely judged on their content, meaning that coding results can be used regardless of the mode in which a survey is executed.

In this section, we described the pilot study that we executed, the list of item characteristics as used in our study, and the item characteristics that evoke measurement effects according to the literature. In the next section, we will elaborate on the selected surveys, the coding procedure and the statistical analyses.

3. Method

In this section, we will first elaborate on the surveys that we used for the study. Second, we will give a short overview of the actual coding procedure. And third, we will elaborate on the statistics that will be calculated to answer our research questions.

3.1 Surveys

This coding research is based on 11 Dutch general population surveys. These are the first wave of the Dutch Labour Force Survey (LFS) administered by Statistics Netherlands and the most recent waves of the ten core studies from the Longitudinal Internet studies for the Social Sciences (LISS) of CentERdata. See table 3.1.1 for an overview of these surveys with a brief description of the topics of their content and the total number of items they contain. In total, the surveys together contain 2470 items of a broad range of topics that covers virtually the whole area of general population statistics. All items of these surveys were coded by a group of survey researchers on all 16 item characteristics. In the following, we will describe the steps of the coding procedure.

3.1.1 Overview of all surveys and a description of their content

Survey (Wave: Number of items)	**Topics of the content**
Labour Force Survey (LFS) (LFS-A: N = 123)	Education; employment and labour
Economic Situation Assets (Wave 3: N = 50)	Income, property and investment
Economic Situation Housing (Wave 6: N = 73)	Housing and household; income, property and investment

Survey (Wave: Number of items)	Topics of the content
Economic Situation Income (Wave 6: N = 286)	Employment, labour and retirement; income, property and investment; social security and welfare
Family and Household (Wave 6: N = 409)	Housing and household; social behaviour
Health (Wave 6: N = 243)	Health and well-being
Personality (Wave 6: N = 200)	Psychology
Politics and Values (Wave 6: N = 148)	Politics; social attitudes and values
Religion and Ethnicity (Wave 6: N = 71)	Religion; social stratification and groupings
Social Integration and Leisure (Wave 6: N = 396)	Communication, language and media; leisure, recreation and culture; social behaviour; travel and transport
Work and Schooling (Wave 6: N = 471)	Education; employment, labour and retirement

3.2 The allocation of coders

The coding procedure consisted of three steps. First, as described in section 2, we set up the list of candidate characteristics based on existing literature. Second, this tentative list was coded on a small but diverse subset of items for executing the pilot study. Based on these coding results, the list was refined and revised. Third, all items of all selected surveys were coded by either two or three coders, depending on the anticipated complexity of the coding task. Throughout these steps, the same group of survey researchers was involved. Altogether, eight researchers from Utrecht University, CentERdata and Statistics Netherlands with knowledge of and experience with survey research were involved in coding the 11 surveys on the final 16 selected item characteristics. All coders were allocated randomly to the surveys, but each coder received a different amount of surveys and survey items to code.

To each survey, two main coders were randomly allocated to code all item characteristics. A third coder was randomly allocated to code only seven specific item characteristics that are assumed to be the most influential in evoking measurement effects. Therefore, we have called these characteristics the 'hard' item characteristics. The hard item characteristics are content of the question, difficult language usage, emotional charge, presumption of a filter question, sensitive information, centrality, and response complexity. For reasons of clarity, we have called the remaining item

characteristics that will be coded by only two coders the 'easy' item characteristics. The easy item characteristics are time reference, conditions, memory, hypothetical situation, calculations, ambiguity, mismatch, formulation, and clarification. All coders were instructed to abide by the agreed definitions and coding categories as strictly as possible during the coding process.

Finally, it is important to note that the researchers coded their allocated survey items in both the pilot study and the actual coding study independently of other coders. This means that they walked through the coding process without communicating with other coders. Also, all researchers coded the surveys and its items throughout their entire coding process consistently. This means that they tried to code all items according to the exact definitions of the item characteristics and its coding categories. Next, we will elaborate on the statistics that will be calculated based on the results of the actual coding study.

3.3 Statistics

First, the relative frequencies for all categories and the intercoder agreement probabilities for all characteristics will be calculated. This will be done in proportions and only for all surveys together, to check each item characteristic on its factual and relative overall occurrence. Second, the intercoder agreement probabilities for the item characteristics that are coded by two or three coders consist of the probability that, respectively, both or all three coders agreed on the coded category of a certain item characteristic over all surveys. Here, the intercoder agreement probability for a specific item characteristic is the number of items for which the coders agreed on the category, divided by the total number of items. These probabilities will directly give an overall indication of the extent to which the item characteristics can be coded reliably.

The intercoder agreement for the easy item characteristics is calculated on the basis of two coders and for the hard item characteristics on the basis of three coders. Therefore, these two different kinds of intercoder agreement are not directly comparable. Here, it seems logical to calculate Fleiss' kappa, which is an indicator of the interrater agreement between multiple coders. Fleiss' kappa incorporates a correction for the degree of agreement that may be expected by chance alone (Fleiss 1971). However, we do not believe that the coding of items by coders involves an element of chance. The coders were instructed on the coding procedure precisely and are assumed to have coded conscientiously and consistently. This means that differences between coders are real differences in the sense that the coders may consider the item characteristics differently for certain items, based on their own perspective. Therefore, we will not use Fleiss' kappa, but instead calculate the fixed probability λ that a coder correctly indicates the true category for an item characteristic. This probability is calculated on the basis of the accompanying intercoder agreement for the concerned item characteristic. Then, the probability λ for an item characteristic is the number of correctly coded items divided by the total number of all items. For this calculation, we assumed that each coder acted

independently and that this probability is the same for each coder. See Table 6 in section 4.2 for the probability λ and its accompanying intercoder agreement for each item characteristic. See Appendix B for an elaboration on the probability λ and table 2 in Appendix B for an overview of specific values for the probability λ and its accompanying intercoder agreement for two or three coders.

4. Results

In this section, we will first give an overview of the relative frequencies of all item characteristics. Second, we will present the intercoder reliabilities for both the hard and easy item characteristics. And third, we will try to explain low intercoder reliability both in general terms and for each concerned item characteristic separately.

4.1 Relative frequencies

Three coders were assigned to each survey, meaning that 33 sets of coding data for 11 surveys were collected. For each survey, this consisted of two sets of coding data for all item characteristics and one set of coding data for only the seven so-called hard item characteristics. For each coding category, we calculated the relative frequencies for all item characteristics. The calculations were done over all surveys, giving an overview of these frequencies for the broad range of all 11 surveys together in proportions. See table 4.1.1 for the overall relative frequencies for the item characteristics with more than two coding categories. Over all surveys, all categories were coded to at least some extent. Factual questions (content of the question), questions for which no memory was needed (memory), and questions about the present (time) were coded most frequently. Questions that ask for a degree of satisfaction (content of the question), questions about events from the past one month (memory), and questions about the future (time) were coded relatively infrequently.

See table 4.1.2 for the item characteristics with only two coding categories. Over all surveys, the category indicating that the characteristic is applicable, was coded to at least some extent for each characteristic. The applicability of an item being formulated as a statement and an item containing some form of clarification were coded most frequently. Complexity of the answering options, questions about a hypothetical situation, ambiguous questions, and questions being a mismatch were coded relatively infrequently. The lowest proportion of 0.02 for questions being a mismatch indicates an applicability of still roughly 40 items of all survey items per coder on average. Because of this substantial amount of items, we decided to include all item characteristics and their coding categories in further analyses.

4.1.1 The relative frequencies of the coding categories for the item characteristics content of the question, memory, and time reference over all surveys (2490 items)

Content of the question	Factual behaviour (1)	Otherwise factual (2)	Opinions (3)	Satisfaction (4)	Otherwise subjective (5)
	0.17	0.59	0.09	0.02	0.12
Memory	No memory (0)	Non-specific memory (1)	Memory < 1 month ago (2)	Memory < 1 month ago (3)	
	0.61	0.12	0.02	0.25	
Time reference	Past (1)	Present (2)	Future (3)		
	0.35	0.62	0.03		

4.1.2 The relative frequencies for the item characteristics with two coding categories over all surveys

Item characteristic	Applicability characteristic	Item characteristic	Applicability characteristic
Conditions	0.14	Difficult language usage	0.19
Hypothetical situation	0.03	Emotional charge	0.12
Calculations	0.20	Presumption of a filter question	0.26
Ambiguity	0.02	Sensitive information	0.25
Mismatch	0.02	Centrality	0.21
Formulation	0.31	Response complexity	0.04

4.2 Intercoder reliabilities

Following this overview of the relative frequencies of the item characteristics over all surveys together, we will now deal with our first research question and present to what extent coding of these item characteristics is actually reliable. As a rule of thumb and for reasons of convenience, we will consider proportions of 0.80 and higher as reasonably high intercoder reliability and proportions of 0.79 and lower as low intercoder reliability. Therefore, we will focus on proportions below 0.80 when we will try to explain potential low intercoder reliability. For clarity reasons, we will

present the intercoder reliabilities for the hard and easy item characteristics separately. See table 4.2.1 for the intercoder reliabilities for the easy item characteristics on the left side and the hard item characteristics on the right side of the table. Regarding the hard item characteristics, see table 4.2.2 for the intercoder reliabilities for the three pairs of coders.

4.2.1 Intercoder reliabilities for the easy and hard item characteristics (and their fixed coder probability λ)

Easy item characteristics	Intercoder reliability	Hard item characteristics	Intercoder reliability
Time reference	0.85 (0.92)	Content of the question (5 categories)	0.56 (0.82)
Conditions	0.89 (0.94)	Content of the question (2 categories)	0.90 (0.97)
Memory	0.85 (0.92)	Difficult language usage	0.61 (0.85)
Hypothetical situation	0.98 (0.99)	Emotional charge	0.75 (0.91)
Calculations	0.94 (0.97)	Presumption of a filter question	0.62 (0.85)
Ambiguity	0.96 (0.98)	Sensitive information	0.53 (0.81)
Mismatch	0.98 (0.99)	Centrality	0.59 (0.84)
Formulation	0.57 (0.68)	Response complexity	0.91 (0.97)
Clarification	0.71 (0.82)		

4.2.2 The intercoder reliabilities for the three pairs of coders for the hard item characteristics.

Item characteristic	Coder 1 vs. coder 2	Coder 1 vs. coder 3	Coder 2 vs. coder 3
Content of the question	0.76	0.65	0.68
Difficult language usage	0.73	0.69	0.81
Emotional charge	0.91	0.83	0.77
Presumption of filter question	0.74	0.74	0.76
Sensitive information	0.74	0.67	0.66
Centrality	0.74	0.70	0.74
Response complexity	0.94	0.94	0.95

Intercoder reliabilities for the easy item characteristics

As can be seen in the left part of table 4.2.1, the intercoder reliabilities for most easy item characteristics were reasonably high, indicating that coding of these item characteristics can be done relatively reliably. For the item characteristics formulation and clarification, however, low intercoder reliabilities were evident. Although formulation and clarification were defined as easy item characteristics and

thus coded by only two coders, coding of these two item characteristics could not be done reliably. This means that coders did often not agree on whether the concerned item was formulated as a question or a statement and whether it contained a clearly present clarification or not.

Intercoder reliabilities for the hard item characteristics

For the item characteristic content of the question, a second kind of intercoder reliability was calculated to investigate to what extent this characteristic could be coded reliably with only an objective and an subjective category. For this specific intercoder reliability, the categories 'factual behaviour' and 'otherwise factual' were merged into one overall objective category and the categories 'opinion', 'satisfaction' and 'otherwise subjective' were merged into one overall subjective category. As can be seen in the right part of table 4.2.1, for the initial item characteristic content of the question, the intercoder reliability was relatively low. For content of the question with merely the objective and subjective category, however, the intercoder reliability was reasonably high. This indicates that this item characteristic could not be coded reliably with five subcategories, but could be coded reliably when only one objective and one subjective category were used. For the items for which no consensus was found, this means that coders usually agreed on whether an item contained either objective or subjective content, but did often not agree on the category within the objective of subjective content.

As can be seen in the right part of table 4.2.1, the intercoder reliabilities for most other hard item characteristics were also relatively low, indicating that coding of these item characteristics cannot be done reliably. For relatively many items, this means that coders did often not agree on when an item contained unknown or difficult words or complex sentences (difficult language usage), when an item was about a topic or contained words that could evoke an emotional reaction (emotional charge), when an item could make respondents presume that follow-up questions might result depending on the answer they would give (presumption of a filter question), when an item asked for some kind of sensitive information so that it may evoke socially desirable answering behaviour (sensitive information), or when an item was difficult to answer as it goes beyond the interest, knowledge or experience of the respondent (centrality). In the following section, we will try to explain low intercoder reliability for the concerned item characteristics.

4.3 Explaining low intercoder reliabilities

Following this overview of the intercoder reliability statistics, we will now deal with our second research question and try to explain the low intercoder reliabilities that we found. Overall, the interaction of two related key factors is probably associated with the obtained low intercoder reliabilities. First, we will briefly discuss these key factors to indicate the difficulty in obtaining reasonably high intercoder reliabilities. Second, with the two key factors in mind, we will discuss the characteristics that had a fixed coder probability λ below the value of 0.90 (see section 3.3 and table 4.2.1 in section 4.2). We do not believe that coders had the same coding probabilities nor

that the correct probabilities are equal for each category, but the criterion allows for a more objective and intuitive decision. See Appendix B and table 2 for a brief explanation. Regarding the hard item characteristics, we will also discuss those characteristics that had an intercoder reliability below the value of 0.80 for at least one of the three pairs of coders (see table 4.2.2).

Key factors associated with low intercoder reliability

We evaluated low intercoder reliability with the survey researchers involved in our study. A first key factor associated with low intercoder reliability is the inherent difficulty with which the item characteristics are defined and demarcated on their categories. Even though the item characteristics are based on existing survey literature and even after extensive discussions with the coders involved, it is difficult for many item characteristics to put concrete boundaries between the categories of a specific item characteristic. For many item characteristics, there is a relatively large grey area between two categories. Hence, it is difficult for the coder to choose between them, no matter how precise the concerning item characteristic has been defined. Also even more specific definitions will leave relatively many items difficult to code. For many item characteristics, this means that many items cannot be coded unambiguously on the basis of their definition and accompanying categories.

As a consequence, a second key factor is the inevitability of a certain extent of personal interpretation from the side of the coders. This means that the coding of surveys by coders is of inherent subjective nature. Even though the item characteristics may be well-defined and well-demarcated, all coders involved have their own life history, personality and current mood, which may all somewhat affect the way a specific item characteristic is interpreted. This will influence the way how certain survey items are coded on this item characteristic. From this point of view, intercoder reliabilities will partly depend on which coders coded the concerned survey. Moreover, it is likely that if the same coder would code the same specific survey for a second time, different coding outcomes will result. As a consequence, somewhat different intercoder reliabilities would emerge. From here, we will integrate these two key factors in a brief discussion about the item characteristics that were coded with low intercoder reliability over all surveys.

Explaining low intercoder reliability

Formulation and clarification

Coders could often not agree on whether an item consisted of a question or a statement. An explanation for this could be that many surveys contain batteries of items with the same response options. These items are often neither direct questions nor full statements, making it difficult for the coder to judge whether the item consists of a statement. Here, it depends on the individual coders and their interpretations how the concerned item is coded for this item characteristic. For many items, coders could also not agree on whether an item contained clarification. This could be explained by the fact that many survey items contain brief examples of what is meant by a concept, remarks about how to fill out the item, or other subordinate clauses. Items contain examples and remarks for a reason, but it may be

unclear to what extent these examples and remarks are full clarifications. This may confuse the coders in their judgment about this item characteristic, resulting in different decisions for different coders.

Content of the question

In particular, coders could often not agree on whether a subjective item was either an opinion or otherwise subjective. A question for which respondents have to state to what extent they agree and which contains the verb 'think' or 'find' logically leads to the coding category opinion. However, when these kind of questions contain verbs like 'believe', 'consider', 'view', 'feel', or 'want' instead, it may become unclear whether the concerned question should be coded as either being an opinion or otherwise subjective. This decision is strongly dependent on which coder is making the judgment, which may partly explain the intercoder disagreement for this item characteristic.

Difficult language usage

It was hard if not impossible for coders to agree on which exact words and phrases to code as difficult language usage. Not only an unrealistically large database of words and phrases that are –if even possible- objectively judged on their difficulty would be needed to secure consensus, the inherent subjectivity of coders in determining what language usage is difficult for the average respondent almost guarantees coding differences between coders. Due to differences in the subjective reference frameworks of coders, this item characteristic cannot be coded reliably.

Emotional charge

Coders could often not agree on whether an item was emotionally charged. A possible explanation is that it may have been tempting for coders to go beyond the demarcation of the agreed definition, as emotions may also be evoked outside the restricted area of personal trauma and victimization. Surely, also words or phrases that are not necessarily about traumatic events may evoke feelings of anxiety or insecurity. It will partly remain a matter of coder subjectivity that determines where the line between traumatic and non-traumatic emotions is drawn. Some coders may have given more room to non-traumatic emotions than others, possibly explaining a relatively low intercoder reliability for this item characteristic over all surveys.

Presumption of a filter question

It was up to the coder to decide whether an average respondent could have the presumption of a filter question concerning a specific item, but this appeared to be difficult. The estimation of this potential presumption for the respondent may not be much more than a rational but subjective guess from the coders. This idea gives this item characteristic a 'dual subjective' nature, with a presumption of the coder about a possible presumption of the respondent. This makes the coding of presumption of a filter question unrealistic and may explain the relatively low intercoder reliability for this item characteristic.

Sensitive information

Coders could often not agree on whether a question asked for sensitive information from the respondent. The broad range of personal, menial and societal topics contains more or less sensitive information to different degrees. Probably, it is difficult for the coder to judge these varying degrees in order to define an item as either sensitive or non-sensitive, making it hard to decide for a consistent demarcation between these two categories. Moreover, all coders have their own personal view, opinion or experience about whether an item would contain sensitive information. In short, this demarcation difficulty and associated subjectivity may explain the relatively low intercoder reliability for this item characteristic.

Centrality

Coders could often not agree on whether an item was a case of centrality. As for the item characteristic difficult language usage, the difficulty in coding centrality for an item may be judging the knowledge, experience or interest of the average respondent. Again, there is no database in which every sort of item content is objectively judged to secure consensus on centrality. Moreover, the inherent subjectivity of coders in determining centrality for an item for the average respondent again almost guarantees coding differences between coders. This item will also not be codeable reliably due to differences in the subjective reference frameworks of coders.

Now that we have tried to explain the resulting low intercoder reliability by the presumed key factors of definition difficulties and inherent coder subjectivity as well as for each item characteristic with a low intercoder reliability separately, we will suggest a few options for coping with low intercoder reliability in constructing questionnaire profiles based on their item characteristics in the following section.

5. Coping with low intercoder reliability

Following this overview of the most likely explanations for the low intercoder reliability that was found, we will now deal with our third research question and suggest four options for coping with low intercoder reliability. These are 1) excluding survey items in constructing questionnaire profiles, 2) redefining and refining the item characteristics for a more strict coding demarcation, 3) computerizing the definition and demarcation of the item characteristics, and 4) using scales consisting of different degrees of applicability of the item characteristics with two categories that are coded by three coders. In this section, we will discuss these four options in some detail.

5.1 Option 1: Excluding survey items

A first option for coping with low intercoder reliability is the most simple and passive one, which is excluding all survey items in constructing questionnaire profiles for which no coding consensus was found for the concerned item characteristic. For instance, when two coders do not agree on whether a certain survey item contains difficult language usage, there is simply no coding consensus for the item characteristic difficult language usage for that specific survey item. Therefore, this specific survey item should not be included in a questionnaire profile for this item characteristic. The advantage of excluding such survey items is the solid and secure foundation on which the questionnaire profile is based for a specific item characteristic for a specific survey, with only items included for which full intercoder consensus is present. The disadvantage of excluding such survey items is that probably relatively many items will have to be excluded before being able to construct the questionnaire profile for the concerned item characteristic and survey. As relatively much information would be lost for constructing the questionnaire profile, this option does not seem to be preferable.

5.2 Option 2: Redefining and refining item characteristics

A second option for coping with low intercoder reliability is to redefine the item characteristics in such a manner that they are conceptually even more narrow and specific than how they were used in the current experiment. For this purpose, all survey items for which low intercoder reliability was evident should be checked on the concerned item characteristic to investigate how the characteristic should be defined more narrow and specific. For instance, let us consider the item characteristic content of the question and the difficulty of distinguishing between the categories opinion and otherwise subjective. Here, it is necessary to check for all items for which low intercoder reliability was evident with a focus on the verbs that are used within the item. Surely, the main verb in an item determines whether the question asks for either an opinion or otherwise subjective. As stated earlier, relatively many items for which low intercoder reliability was found contained ‘believe’, ‘consider’, ‘view’, ‘feel’, or ‘want’ as the main verb. Then, for items containing one of these verbs, it has to be decided whether the item either asks for an opinion or asks for something otherwise subjective for each verb. By refining the definition of item characteristics in this way, coding demarcations will become more strict and intercoder reliability might be improved significantly for the concerned item characteristic. However, this option will not fully account for the inherent coder subjectivity of each coder during the actual coding procedure.

5.3 Option 3: Computerizing the definition and demarcation of item characteristics

To completely avoid the inherent coder subjectivity in the coding procedure, a third option for coping with low intercoder reliability is to computerize the definition and demarcation of item characteristics. By making use of computerized decisions between the different categories of an item characteristic, coder subjectivity is simply no part of the coding process anymore. Here, the definitions of the item characteristics and the demarcations between the categories are programmed by strict rules that cannot be deviated from. Let us consider the example of the item characteristic content of the question for the categories opinion and otherwise subjective again. Here, this would for instance imply that every verb for which no full consensus was evident is programmed to be attributed to either opinion or otherwise subjective. In this way, every verb would be subject to strictly one and only one of both categories. However, before this computerized coding procedure can actually be launched, the same steps from option 2 (see above) will have to be executed. Ironically, human decisions about those strict rules need to be made before they can actually be programmed.

Furthermore, this is just as true for the other item characteristics as it is for content of the question. For instance, let us consider the item characteristics emotional charge and sensitive information. It needs to be decided specifically when the topic or context of the item and the words within an item should be coded as emotionally charged or sensitive. For every specific topic and context and even for every word, strict rules should be made about the item's emotional and sensitive content. Moreover, these decisions and rules also need to distinguish specifically the often subtile differences between emotional charge and sensitive information. Exactly the same is true for, for instance, the item characteristics difficult language usage and centrality. Hence, in fact, the question rises to what extent such strict rules can actually be programmed to a realistic extent at all.

5.4 Option 4: Using item characteristic scales with multiple applicability categories

For a way to avoid redefining and re-demarcating the item characteristics or programming strict rules for the coding procedure, a fourth option for coping with low intercoder reliability is to construct scales with multiple applicability categories for the item characteristics with two categories that are coded by three coders. Let us consider the item characteristic presumption of a filter question here. This characteristic was coded by three coders, meaning that either no, one, two, or three coders indicated its applicability for a certain item. Based on all items for which either no, one, two, or three coders indicated the characteristic's applicability, a questionnaire profile consisting of four respective categories could be constructed. Then, for the items of a survey, the characteristic presumption of a filter question is expressed on a gradual scale with four applicability categories, rather than on a dichotomous scale with only the categories applicable and not applicable. This profile

can be used to investigate to what extent it explains variation in the influence of this item characteristic on evoking measurement effects. For instance, consider items that were coded as presumed to be a filter question by three coders versus two coders. Here, the influence on evoking measurement effects may appear relatively larger for items for which all three coders versus for items for which only two coders presumed them as filter questions. Exactly the same may be true for two coders versus one coder and for one coder versus no coders. In this way, the relative influences of each of these four categories can be compared directly to check for their potential different relations to the occurrence of measurement effects.

To be able to investigate and compare the categories of such an applicability scale, each category should contain enough items to base its profile on. For the current study, we calculated the relative frequencies of each category for all item characteristics with two coding categories that were coded by three coders. As can be seen in table 5.4.1, the applicability of the item characteristics is coded by all three coders for only relatively few items. Hence, it may not be feasible to construct a scale for all four category profiles, as relatively few items may not contain enough power to expose potential measurement effects. Here, an alternative option might be to pool the two categories with two and three coders into a single third category. Then, this third category may contain enough items and will consist of all items that were coded as applicable to the concerned item characteristic by at least two coders.

5.4.1 Relative frequencies of the applicability of the hard item characteristics with two coding categories for the number of coders over all surveys

Item characteristic	No coder (0)	One coder (1)	Two coders (2)	Three coders (3)
Difficult language usage	0.59	0.28	0.11	0.02
Emotional charge	0.73	0.20	0.04	0.02
Presumption of a filter question	0.53	0.25	0.13	0.09
Sensitive information	0.49	0.32	0.14	0.04
Centrality	0.57	0.26	0.15	0.02
Response complexity	0.91	0.06	0.03	0.00

6. Discussion

In this study, we coded all 2470 items of 11 Dutch surveys on 16 item characteristics that are assumed to be relevant in evoking mode-specific measurement error according to the literature. We have investigated to what extent the coding of these item characteristics could be done reliably by multiple coders. In case of reasonably high intercoder reliability, so-called questionnaire profiles may be constructed. Questionnaire profiles summarize the characteristics of the items of a survey. If questionnaire profiles could be identified and would appear to explain variation in answering behaviour on the part of the respondent, they might be helpful in anticipating measurement effects. In case of relatively low intercoder reliability, however, questionnaire profiles cannot be constructed without difficulty. Low intercoder reliability would then need to be explained and suggestions should be made for coping with low intercoder reliability.

We found that item characteristics that are particularly influential in evoking measurement effects could not be coded reliably. For the characteristics content of the question, difficult language usage, emotional charge, sensitive information, presumption of a filter question, and centrality, a relatively low intercoder reliability was found. Surprisingly, also a low intercoder reliability was found for the characteristics formulation and clarification that are not particularly influential in evoking measurement effects per se. In general, the low intercoder reliability may be explained by the difficulty with which the item characteristics had to be defined and by the inherent subjective nature of the coding of survey items by coders. Coders sometimes differed substantially in their relative coding frequencies, depending on the concerned survey and characteristic. Some coders appeared to have the tendency to be generally conservative, while other coders seemed to be generally liberal in indicating the applicability of characteristics. This was especially evident for the characteristics that are particularly relevant to measurement effects. The coders were selected from three different institutions and we believe that they are representative for any set of coders in similar studies and institutions. We consider it unlikely that substantially different coding outcomes will result with another set of coders.

At the start of our study, we distinguished item characteristics that were coded by either two or three coders. In principle, we wanted the characteristics to be coded by two coders, but we assigned a third coder to characteristics that were supposed to be particularly relevant to evoking measurement effects according to the literature. Considering the study results, the intercoder reliability for characteristics coded by three coders was generally lower than for characteristics coded by two coders. However, it is difficult to say to what extent this can be explained by the different degree of difficulty of coding the characteristics versus to what extent this can be attributed to the different number of coders; the particularly relevant characteristics may have been relatively more difficult to code, but it is also obvious that consensus decreases as more coders are involved. First, the fixed intercoder probabilities for

most characteristics coded by three coders were clearly below the value of 0.90 that we set as a minimum as a reasonable intercoder probability, while the fixed intercoder probabilities for most characteristics coded by only two coders were clearly above this value (see table 4.2.1 in section 4.2 and table 2 in Appendix B). Second, for most characteristics coded by three coders, the intercoder reliabilities for all three pairs of coders showed that either one, two or all three pairs of coders had an intercoder reliability below the value of 0.80 that we set as a minimum for reasonable intercoder reliability (see table 4.2.2 in section 4.2). Based on both the intercoder probabilities that are assumed to be fixed and equal for each coder and the intercoder reliabilities for the pairs of coders, this means that characteristics particularly relevant to measurement effects were relatively more difficult to code.

It must be noted that, according to the coders, the occurrence of some characteristics was relatively rare (see table 4.1.2 in section 4.1). The rareness of a characteristic is logically related to the intercoder reliability of a characteristic. For instance, let us consider the characteristic mismatch with an intercoder reliability of 0.98 and a relative frequency of 0.02. This means that, for almost all items, both coders did not indicate its applicability, explaining the high intercoder reliability of 0.98. Thus, for the remaining 0.02 percent of all items, one of the two coders indicated the applicability of the characteristic mismatch and the other coder did not. In fact, there were no items at all for this characteristic for which both coders indicated the applicability. This means that the high intercoder reliability for this characteristic is solely based on the majority of items for which both coders did not indicate the applicability. In short, when a characteristic appears to be rare, a high intercoder reliability is a logical result and may mask a low consensus for those items on the boundary of having the characteristic.

Despite the potential limitations in our study, the results may have far-reaching consequences for the literature on measurement error and survey design features. Although there are obvious associations between question complexity, question centrality, question sensitivity and measurement error, these features are not easily identified; they may lead to inconsistent, weak or even spurious conclusions. To be able to construct questionnaire profiles to investigate their relation to measurement effects, more research needs to be done. Based on the results of our study, questionnaire profiles cannot be constructed without difficulty. This is especially evident for characteristics that are particularly relevant to measurement effects. Four options to cope with low intercoder reliability were suggested: Excluding items for which no consensus was found, redefining the item characteristics, computerizing the item characteristics, and using applicability scales for the item characteristics. Excluding items for which no coder consensus was found and computerizing the item characteristics do not seem to be attractive options to base questionnaire profiles on. The former option would mean a relatively large loss of information and the latter option would be time consuming and still contain a substantial subjective element in deciding on the definitions of the characteristics and the coding rules. In constructing valuable questionnaire profiles, it seems plausible to investigate the items for which no consensus was found. By drawing up an inventory of these items and using the literature, the definitions of characteristics could be complemented and part of these

items may still be coded unambiguously for at least the 'easy' characteristics that did not have a reasonable intercoder reliability. For the characteristics particularly relevant to measurement effects, the applicability scales may also be used for items for which no consensus was found to obtain an indicative questionnaire profile for a survey.

Acknowledgements
We would like to thank Rachel Vis-Visschers for her contribution to the pilot study and her review of the draft paper, Vivian Meertens for her review of the draft paper and Rodinde Pauw for coding the 'easy' item characteristics.

References

Bassili, John N. and B. Stacey Scott 1996. "Response Latency as a Signal to Question Problems in Survey Research." Public Opinion Quarterly 60(3):390-399.

Beukenhorst, Dirkjan, Bart Buelens, Frank Engelen, Jan van der Laan, Vivian Meertens, and Barry Schouten 2013. "The impact of Survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey." Discussion paper 201416. Statistics Netherlands, Den Haag.

Bosley, John, Monica Dashen, and Jean E. Fox 1999. "When should we ask follow-up questions about items in lists?" Proceedings of the Survey Research Methods Section of the American Statistical Association :749-754.

Buelens, Bart and Jan van den Brakel 2011. "Inference in surveys with sequential mixed-mode data collection." Discussion paper 201121. Statistics Netherlands, Heerlen.

Buelens, Bart, Jan van der Laan, Barry Schouten, Jan van den Brakel, Joep Burger, and Thomas Klausch 2012. "Disentangling mode-specific selection and measurement bias in social surveys." Discussion paper 201211. Statistics Netherlands, Den Haag.

Campanelli, Pamela, Gerry Nicolaas, Annette Jäckle, Peter Lynn, Steven Hope, Margaret Blake, and Michelle Gray 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.

De Leeuw, Edith D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." Journal of Official Statistics 21(2):233-255.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser 2014. “Assessing the Mechanisms of Misreporting to Filter Questions in Surveys.” Public Opinion Quarterly 78(3):721-733.

Fleiss, Joseph L. 1971. “Measuring nominal scale agreement among many raters.” Psychological Bulletin 76(5):378-382.

Foddy, William 1993. Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research. Cambridge: Cambridge University Press.Fowler, Floyd J., Jr. 1995. Improving Survey Questions: Design and Evaluation. Applied Social Research Methods Series, 38. Thousand Oaks, CA: Sage Publications.

Fowler, Floyd J., Jr. and Thomas W. Mangione 1990. Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Applied Social Research Methods Series, 18. Newbury Park, CA: Sage Publications.

Gallhofer, Irmtraud N., Annette Scherpenzeel, and Willem E. Saris 2007. “The code-book for the SQP program”, available at (http://sqp.upf.edu).

Klausch, Thomas, Joop J. Hox, and Barry Schouten 2013. “Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions.” Sociological Methods and Research 42(3):227-263.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau 2011. “The Effects of Asking Filter Questions in Interleafed versus Grouped Format.” Sociological Methods and Research 40(1):80-104.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau 2008. “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” Public Opinion Quarterly 72(5):847-865.

Krosnick, Jon A. 1991. “Response strategies for coping with the cognitive demands of attitude measures in surveys.” Applied Cognitive Psychology 5(3): 213-236.

Lensvelt-Mulders, Gerty J. L. M. 2008. “Surveying sensitive topics.” Pp. 461-478 in International Handbook Of Survey Methodology, edited by E. D. de Leeuw, J. J. Hox, and D. A. Dillman. New York: Taylor & Francis, Psychology Press, EAM series.

Lenzner, Timo, Lars Kaczmirek, and Alwine Lenzner 2009. “Cognitive burden of survey questions and response times: A psycholinguistic experiment.” Applied Cognitive Psychology 24(7): 1003-1020.

Lozar Manfreda, Katja and Vasja Vehovar 2002. “Mode effect in web surveys.” In the proceedings from The American Association for Public Opinion Research (AAPOR) 57th Annual Conference, 2002, (http://www.amstat.org/sections/srms/Proceedings/y2002/files/JSM2002-000972.pdf)

Roberts, Caroline 2007. “Mixing modes of data collection in surveys: A methodological review.” ESRC National Centre for Research Methods, NCRM Methods Review Papers, NCRM/008, UK.

Saris, Willem E. and Irmtraud Gallhofer 2007. “Estimation of the effects of measurement characteristics on the quality of survey questions.” Survey Research Methods 1(1):29-43.

Saris, Willem E., Jon A. Krosnick, Melanie Revilla, and Eric M. Shaeffer 2010. “Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options.” Survey Research Methods 4(1):61-79.

Schonlau, Matthias, Kinga Zapert, Lisa Payne Simon, Katherine Sanstad, Sue Marcus, John Adams, Mark Spranca, Hongjun Kan, Rachel Turner, and Sandra Berry. 2004. “A Comparison Between a Propensity Weighted Web Survey and an Identical RDD Survey.” Social Science Computer Review 22(1):128-138.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski 2000. The psychology of survey response. Cambridge: Cambridge University Press.

Tourangeau, Roger and Ting Yan 2007. “Sensitive Questions in Surveys.” Psychological Bulletin 133(5):859-883.

Van der Vaart, Wander 1996. Inquiring into the Past: Data Quality of Responses to Retrospective Questions. PhD dissertation Vrije Universiteit Amsterdam.

Van der Vaart, Wander, Johannes van der Zouwen, and Wil Dijkstra 1995. “Retrospective questions: data quality, task difficulty, and the use of a checklist.” Quality and Quantity 29(3):299-315.

Van der Zouwen, Johannes 2000. “An assessment of the difficulty of questions used in the ISSP-questionnaires, the clarity of their wording and the comparability of the responses.” ZA-Informationen 45:96-114.

Van der Zouwen, Johannes and Wil Dijkstra 1996. “The Impact of the Question on the Interactions in Survey-Interviews.” Paper presented at the Fourth International ISA Conference on Social Science Methodology (Essex ’96), July 1-5, University of Essex, Colchester UK.

Vannieuwenhuyze, Jorre, Geert Loosveldt, and Geert Molenberghs 2010. “A Method for Evaluating Mode Effects in Mixed-mode Surveys.” Public Opinion Quarterly 74(5):1027-1045.

Ye, Cong, Jenna Fulton, and Roger Tourangeau 2011. “More positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice.” Public Opinion Quarterly 75(2):349-365.

Appendix A

Table 1. Definitions of the item characteristics coded by one coder and their coding numbers and categories.

Item Characteristic	Definition	Coding numbers and categories
Number of words	How many words does the text of an item (clarifications included) contain up to the answering categories?	0 <= 25 words 1 > 25 words
Factual filter question	Is the question a factual filter question?	0 no 1 yes
Measurement level	What is the measurement level of the answering categories?	0 closed nominal 1 closed ordinal 2 open numeric 3 open non-numeric
Number of answering options	How many answering options does the item contain?	0 not applicable 1 1 or 2 categories 2 3 to 5 categories 3 6 or more categories
Answer as a mark	Does the question need to be answered as a mark from 0 or 1 to 10?	0 not applicable 1 no 2 yes
Polarity of the scale	Is the polarity of the answering scale unipolar or bipolar?	0 not applicable 1 unipolar 2 bipolar
Balance of the scale	Is the answering scale balanced or unbalanced?	0 not applicable 1 balanced answering scale 2 unbalanced answering scale
Neutrally formulated middle category	Does the answering scale contain a neutrally formulated middle category?	0 not applicable 1 with middle category 2 without middle category
Direction of the scale	What is the direction of the answering scale?	0 not applicable 1 from positive to negative 2 from negative to positive
Labels of the scale	Do the categories of the answering scale contain labels?	0 not applicable 1 no labels 2 partly labeled 3 fully labeled

Item Characteristic	Definition	Coding numbers and categories
'Don't know' explicitly present	Is 'don't know' an explicit answering option?	0 no 1 yes
Item part of a battery	Is the item part of an item battery?	0 no 1 yes
Relative position of item in battery	What is the relative position of the concerned item in the item battery?	

Appendix B

Let us consider the fixed probability that coders correctly indicate the true category for an item characteristic with the two coding categories applicable and not applicable. This probability consists of 1) the probability that the characteristic is applicable to an item and the coders correctly indicate its applicability, and 2) the probability that the characteristic is not applicable to an item and the coders correctly indicate its non-applicability. By combining these two probabilities, we get Formula 1:

$$\lambda^m + (1-\lambda)^m, \tag{1}$$

where λ is the fixed probability that coders correctly indicate the true category for a characteristic and m is the number of coders. By using Formula 1, we can calculate the probability –the intercoder agreement– that m coders indicate a characteristic on the same category for each fixed probability λ. See table 2 for the intercoder reliability for specific values of λ for two and three coders. By means of table 2, we are able to compare the intercoder reliability for two versus three coders to determine that the intercoder reliability decreases relatively faster for three coders versus two coders. For instance, for a fixed coder probability λ of 0.90, the intercoder reliability is 0.82 for two coders, but only 0.73 for three coders. Merely on the basis of the fixed coder probability, we expect the intercoder reliability for item characteristics with two coding categories to be lower for three coders than for two coders.

Table 2. The intercoder reliability based on the fixed true coding probability λ and the number of coders m

λ	m = 2	m = 3
1	1	1
0.95	0.91	0.86
0.90	0.82	0.73
0.85	0.75	0.62
0.80	0.68	0.52
0.75	0.63	0.44
0.70	0.58	0.37
0.65	0.55	0.32
0.60	0.52	0.28
0.55	0.51	0.26
0.50	0.50	0.25

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

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

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information