Article

Sociological Methods & Research
2019, Vol. 48(2) 263-295

DOI: 10.1177/0049124117729692
journals.sagepub.com/home/smr
SAGE

Can Survey Item Characteristics Relevant to Measurement Error Be Coded Reliably? A Case Study on 11 Dutch General Population Surveys

Frank Bais[1], Barry Schouten[2], Peter Lugtig[1], Vera Toepoel[1], Judit Arends-Tòth[2], Salima Douhou[3], Natalia Kieruj[4], Mattijn Morren[2] and Corrie Vis[4]

Abstract

Item characteristics can have a significant effect on survey data quality and may be associated with measurement error. Literature on data quality and measurement error is often inconclusive. This could be because item characteristics used for detecting measurement error are not coded unambiguously. In our study, we use a systematic coding procedure with multiple coders to investigate the extent to which the coding of item characteristics could be done reliably. For this purpose, we constructed an item characteristics scheme that is based on typologies of characteristics. High intercoder reliability indicates a clear relation between item characteristic, item

[1] Utrecht University, Utrecht, the Netherlands
[2] Statistics Netherlands, The Hague, the Netherlands
[3] City University London, London, United Kingdom
[4] CentERdata, Tilburg, the Netherlands

Corresponding Author:
Frank Bais, Utrecht University, PO Box 80140, 3508 TC Utrecht, the Netherlands.
Email: f.bais@uu.nl

content, and measurement error. Our results show that intercoder reliability is often low, especially for item characteristics that are hard to code due to subjectivity. Low intercoder reliability complicates comparisons between studies about item characteristics and measurement error. We give suggestions for coping with low intercoder reliability.

Keywords

intercoder reliability, questionnaire profiles, item characteristics, measurement error, mode-specific measurement error, data quality

Literature shows that item characteristics can have a significant impact on data quality (see Saris and Gallhofer 2007). For example, when a respondent is asked to report on an item containing sensitive information, he or she might have the tendency to answer don't know, to refuse to answer, not to answer at all, or to give an answer that is socially desirable instead of truthful (Campanelli et al. 2011; Kreuter, Presser, and Tourangeau 2008; Lensvelt-Mulders 2008; Schaeffer 2000), resulting in measurement error (Tourangeau and Yan 2007). To be able to investigate the relation between item characteristics and measurement error, an item characteristic should be assigned to a specific item unambiguously. Determining the degree of presence of a characteristic like sensitive information may, however, be less straightforward. For example, an item about emancipation may be sensitive to some people and not to others. A systematic method is needed to reliably measure the characteristic and, hence, assign the characteristic properly to the content of a specific item. An obvious method would be a formal coding procedure in which multiple coders rate the extent of presence or absence of an item characteristic to obtain a formal measure of intercoder reliability.

Coding procedures have already proven to be useful in survey methodology, for example, for the coding of answering behavior and question–answer sequences to identify difficulties with survey questions (Dijkstra 1994; Holbrook, Cho, and Johnson 2006; see Ongena and Dijkstra 2006; Van der Zouwen and Dijkstra 1998; Van der Zouwen and Smit 2004). However, for many item characteristics, there is no clear definition. For example, survey methodology literature provides no definite answer on what makes a question sensitive (Tourangeau, Rips, and Rasinski 2000). Many papers use their own definition (see Bradburn, Sudman, and Associates 1979; Sudman and Bradburn 1982; Tourangeau et al. 2000). In addition, papers code items differently. Fowler and Mangione (1990) coded survey questions on the likeliness that their answers would be "sensitive" or "embarrassing." Cho et al. (2006) define a nonsensitive question as one that would not cause discomfort "for the average respondent"; Kreuter et al. (2008) had respondents rate of 4 items on whether people they know would "report falsely" or "exaggerate their answers" to the items. The result is a diversity of used operationalizations and hence a lack of empirical convergence (Paulhus 2002). A characteristic should be operationalized from a clear definition into specific coding categories, its presence must be determined on the basis of its definition and accompanying coding categories by multiple skilled coders in advance, and the overall coding procedure has to be systematic in terms of adequate code descriptions and a consistent method (Ongena and Dijkstra 2006). Only if these conditions are met, there is a clear relationship between item characteristic and measurement error.

To investigate the relation between survey mode, the type of survey item, and mode-specific measurement error, Beukenhorst et al. (2014) developed a coding scheme with variables characterizing the survey items of the Crime Victimisation Survey. For their study, they selected the question characteristics concept, time reference, question complexity, emotional content, mismatch, formulation, instruction, sensitive information, and centrality (Campanelli et al. 2011; Gallhofer, Scherpenzeel, and Saris 2007; Saris and Gallhofer 2007). They concluded that "measurement effects dominate differences between modes after regular weighting adjustment" (Beukenhorst et al. 2014, p. 25). However, they used only one survey on a specific topic and a restricted selection of items in their study. Beukenhorst et al. (2014) also decided to delete the question characteristics sensitive information and centrality from the analyses, as they evoked too much disagreement among the coders during the coding process. Here, one question is to what extent measurement error may be found for multiple surveys on a broad range of topics and for a large selection of different kind of items, but this can only be done if such a selection of items could be coded on its characteristics along with high intercoder agreement.

In case of high intercoder agreement, we may conclude that the relation between item characteristic and item content can be unambiguous, allowing us to map the role of measurement error within this relation. A way to do this is to construct questionnaire profiles, giving us insight into the complex relation of item characteristic, item content, and measurement error. Such questionnaire profiles may eventually be helpful in anticipating measurement error in designing questionnaires and executing the administration of surveys. As a consequence of intercoder disagreement, certain item characteristics may need to be omitted from the questionnaire profile. Thus, to be able to construct complete questionnaire profiles for whole surveys, intercoder agreement in coding the item characteristics is a prerequisite.

To our knowledge, no research so far has reported a systematic procedure to code many items of multiple surveys on their characteristics by two or more coders to evaluate intercoder reliability. By their experiment, Beukenhorst et al. (2014) made a first attempt to characterize a whole survey questionnaire to investigate mode-specific measurement error by using an item coding scheme that was partly based on the Survey Quality Prediction (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, we constructed a questionnaire characteristics scheme consisting of both question and answer characteristics. By coding 11 questionnaires of Statistics Netherlands and CentERdata, we can investigate the intercoder reliability on these characteristics for 2,470 items that range over various general population topics such as income, education, work, leisure, and personality. In case the intercoder reliability is *high* on certain characteristics, a questionnaire profile based on these characteristics may be constructed relatively easily. In case the intercoder reliability is *low* on certain characteristics, we need to explain this low reliability and how to cope with it. In this study, we (1) investigate the intercoder reliability for each item characteristic over the items of all surveys together, (2) try to explain potential low intercoder reliability, and (3) give suggestions about how to cope with such low reliability.

From here, we first motivate the chosen item characteristics and accompanying literature background in the second section. In the third section, we present all surveys for which these characteristics are coded and elaborate on the actual coding procedure and the statistics that were calculated. In the fourth section, we present all statistical results of the actual coding procedure. In the fifth section, we suggest ways of coping with low intercoder reliability. In the sixth section, we conclude with a discussion of these results.

The Item Characteristics

In this section, we present the list of 16 item characteristics as used in the current study and elaborate on the literature background of these characteristics. Thirteen other item characteristics were considered to be codable on their true category unambiguously. For instance, one such characteristic is the amount of words that the item contains up till the first answering category. Therefore, these 13 characteristics were coded by a single coder and are not taken into consideration for this article (see Table S8 in Online Appendix A for an overview of these item characteristics). Table 1 presents an

Table 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[a]	What kind of topic or aspect is the item about?	1 = factual behavior; 2 = otherwise factual; 3 = opinion; 4 = satisfaction; 5 = otherwise subjective	Campanelli et al. (2011); Gallhofer, Scherpenzeel, and Saris (2007); Lozar Manfreda and Vehovar (2002); Saris and Gallhofer (2007); Schonlau et al. (2004)
Emotional charge[a]	Does the item contain potentially emotional words or a potentially emotional charge?	0 = not applicable/1 = applicable	Lensvelt-Mulders (2008)
Sensitive information[a]	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, Presser, and Tourangeau (2008); Lensvelt-Mulders (2008); Saris and Gallhofer (2007); Tourangeau and Yan (2007)
Presumption of a filter question[a]	Might the respondent be able to presume the item to be a filter question?	0 = not applicable/1 = applicable	Bosley, Dashen, and Fox (1999); Eckman et al. (2014); Kreuter et al. (2011)
Centrality[a]	Does the item go beyond the interest, knowledge or experience of the respondent?	0 = not applicable/1 = applicable	Gallhofer et al. (2007); Saris and Gallhofer (2007); Van der Zouwen (2000)

(continued)

Table 1. (continued)

Item Characteristic	Definition of the Item Characteristic as Used in the Current Study	Coding Number and Categories	References
Question complexity 1: difficult language usage[a]	Does the item contain unknown or difficult words or complex sentences?	0 = not applicable/1 = applicable	Beukenhorst et al. (2014); Van der Zouwen (2000)
Question complexity 2: conditions[b]	Does the item contain conditions?	0 = not applicable/1 = applicable	Beukenhorst et al. (2014); Van der Zouwen (2000)
Question complexity 3: memory[b]	Does answering require some kind of memory?	0 = no memory; 1 = nonspecific memory; 2 = memory < 1 month ago; 3 = memory > 1 month ago	Van der Vaart, Van der Zouwen, and Dijkstra (1995); Van der Zouwen (2000)
Question complexity 4: hypothetical situation[b]	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[b]	Does answering require the performance of some kind of calculation?	0 not applicable/1 applicable	Beukenhorst et al. (2014); Van der Zouwen (2000)
Question complexity 6: ambiguity[b]	Does the item contain multiple subquestions or is the item otherwise potentially confusing?	0 = not applicable/1 = applicable	Campanelli et al. (2011); Foddy (1993); Fowler and Mangione (1990); Van der Zouwen (2000)

(continued)

Table 1. (continued)

Item Characteristic	Definition of the Item Characteristic as Used in the Current Study	Coding Number and Categories	References
Response complexity[a]	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[b]	What time period does the item refer to?	1 = past/2 = present/3 = future	Gallhofer et al. (2007); Saris and Gallhofer (2007)
Mismatch[b]	Do the question and its answering options match?	0 = not applicable/1 = applicable	Beukenhorst et al. (2014); Van der Zouwen (2000)
Formulation[b]	Is the item formulated as a statement?	0 = not applicable/1 = applicable	Fowler (1995); Gallhofer et al. (2007); Saris and Gallhofer (2007); Saris et al. (2010); Ye et al. (2011)
Clarification[b]	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)

Note: see The Allocation of Coders subsection.

[a] The “hard” item characteristics.

[b] The “easy” item characteristics.

overview of the 16 item characteristics and their references that are involved in the current study.

An important note is that we came to our conclusive list of item characteristics based on a pilot study. The pilot study was set up to investigate the actual occurrence of each item characteristic and to check for potential difficulties during the coding process. To read about the execution of the pilot study and about to what changes the pilot study resulted before coming to the conclusive list of item characteristics, see Online Appendix B. From here, we give a motivation for the inclusion of the item characteristics, considering their influence on data quality, in general, and on measurement error. According to the literature, some item characteristics may be sensitive to mode-specific measurement error in particular. Therefore, we finish this section by briefly elaborating on how these characteristics may be associated with mode-specific measurement error.

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 error.

The characteristic *difficult language usage* refers to the use of unknown or difficult words or complex sentences within the item (Beukenhorst et al. 2014), 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 computation (Beukenhorst et al. 2014; Van der Zouwen 2000). 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 barreled (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).

Response Complexity

Response complexity refers to the use of unknown or difficult words or complex sentences within at least one of the answering categories or to the request for the respondent to execute some kind of complex performance, such as moving figures. The number of response categories (Campanelli et al. 2011), the complexity of the response labels (Gallhofer et al. 2007; Saris and Gallhofer 2007), and the amount of information about the response alternatives that has to be stored in short-term memory (Van der Zouwen 2000) can all have their influence on data quality.

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 noncentral or hardly accessible (Van der Zouwen 2000) to them.

Content of the Question

Concerning content of the question, an item is about factual behavior, 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 behavior, and otherwise subjective refers to items asking for thoughts, feelings, or emotions other than opinions or satisfaction of the respondent. We defined factual behavior 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.

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 are 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 or her social group (Lensvelt-Mulders 2008). Tourangeau et al. (2000) define a sensitive question as being experienced as intrusive, involving a threat of disclosure, or to some extent eliciting an answer that is socially undesirable. 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, be 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.

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. 2014).

Presumption of a Filter Question

In some surveys more than in others, certain questions may lead to follow-up items. These questions are the 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 Online Appendix B). The coders experienced difficulty in distinguishing an item as a true filter question versus as a question of which the respondent could presume to be a filter question, regardless of whether the question *is* a true filter question. Some respondents could avoid a filter question in case they presume a question to be one.

The remaining item characteristics that may have their influence on data quality are *time reference*, which refers to whether the item concerns the past, the present, or the future (Gallhofer et al. 2007; Saris and Gallhofer 2007); *mismatch*, which refers to whether the question matches its accompanying answering options (Beukenhorst et al. 2014; Van der Zouwen 2000); *formulation*, which refers to whether the item is formulated as a statement (Gallhofer et al. 2007; Saris and Gallhofer 2007); and *clarification*, which refers to whether the item contains instruction or clarification for the respondent (Gallhofer et al. 2007; Saris and Gallhofer 2007; Van der Zouwen 2000).

The Item Characteristics and Mode-specific Measurement Error

The characteristics can all have their influence on data quality and may be associated with measurement error. This may, however, differ for each survey mode, possibly leading to mode-specific measurement error. Considering *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. 2014), but the probability on some form of satisficing may be relatively high in self-administered modes (Krosnick 1991). Concerning *centrality*, the respondent may be assisted or stimulated by the interviewer in interviewer-administered modes concerning topics that are not central to the respondent, while this assistance or stimulance is less evident in self-administered modes.

Regarding *content of the question*, especially subjective questions are sensitive to the presence of an interviewer and may be more prone to mode-specific measurement error than factual questions (Campanelli et al. 2011; Lozar Manfreda and Vehovar 2002; Schonlau et al. 2004). Considering *sensitive information*, interviewer-administered modes may strongly facilitate the avoidance or distortion of the respondent's answers, while this effect will be much less strong in case of self-administered modes. Therefore, this characteristic, in particular, is sensitive to mode-specific measurement error and may well evoke socially desirable answering (Campanelli et al. 2011; Kreuter et al. 2008; Tourangeau and Yan 2007). In interviewer-administered modes, the interviewer may mitigate the effect of *emotional charge* 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.

Concerning *presumption of a filter question*, respondents may be able to scroll through the survey to check for follow-up questions in mail and web mode. Filter questions that are repeated later in the survey may also be recognized more easily. In personal and telephone mode, however, respondents do not have the option to scroll through the survey, making filter questions relatively more difficult to detect. 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.

Method

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

Surveys

This coding research is based on 11 Dutch general population surveys. These are the first wave of the Dutch Labour Force Survey administered by Statistics Netherlands and the most recent waves of the 10 core studies from the Longitudinal Internet studies for the Social Sciences of CentERdata (see Table 2 for an overview of these surveys with a brief description of the

Table 2. 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 labor
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
Economic situation income (wave 6: $n = 286$)	Employment, labor, and retirement; income, property, and investment; social security and welfare
Family and household (wave 6: $n = 409$)	Housing and household; social behavior
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 behavior; travel and transport
Work and schooling (wave 6: $n = 471$)	Education; employment, labor, and retirement

topics of their content and the total number of items they contain). In total, the surveys together contain 2,470 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 describe the steps of the coding procedure.

The Allocation of Coders

The coding procedure consisted of three steps. First, as described in the second section, 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 appeared to be hard to code during the pilot study. 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 elaborate on the statistics that will be calculated based on the results of the actual coding study.

Statistics

First, the relative frequencies for all categories and intercoder agreement probabilities for all item characteristics were calculated in proportions over all surveys. This was 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 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 was 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's κ, which is an indicator of the interrater agreement between multiple coders. Fleiss's κ 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 considered the item characteristics differently for certain items based on their own perspective. Therefore, we did not use Fleiss's κ but instead calculated the fixed probability λ that a coder correctly indicates the true category for an item characteristic. This probability was 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 the Intercoder Reliabilities subsection for the probability λ and its accompanying intercoder agreement for each item characteristic. See Online Appendix C for an elaboration on the probability λ and Table S9 in Online Appendix C for an overview of specific values for the probability λ and its accompanying intercoder agreement for two or three coders.

Results

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

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 3 for the

Table 3. The Relative Frequencies of the Coding Categories for the Item Characteristics Content of the Question, Memory, and Time Reference over all Surveys (2,490 Items).

Content of the Question	Factual Behavior (1)	Otherwise Factual (2)	Opinion (3)	Satisfaction (4)	Otherwise Subjective (5)
	.17	.59	.09	.02	.12
Memory	No Memory (0)	Nonspecific Memory (1)	Memory < 1 Month Ago (2)	Memory > 1 Month Ago (3)	
	.61	.12	.02	.25	
Time Reference	Past (1)	Present (2)	Future (3)		
	.35	.62	.03		

Table 4. The Relative Frequencies for the Item Characteristics with Two Coding Categories over all Surveys.

Item Characteristic	Applicability Characteristic	Item Characteristic	Applicability Characteristic
Conditions	.14	Difficult language usage	.19
Hypothetical situation	.03	Emotional charge	.12
Calculations	.20	Presumption of a filter question	.26
Ambiguity	.02	Sensitive information	.25
Mismatch	.02	Centrality	.21
Formulation	.31	Response complexity	.04
Clarification	.36		

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 reference) 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 reference) were coded relatively infrequently.

See Table 4 for the item characteristics with only two coding categories. Over all surveys, the category indicating that the characteristic is applicable

Table 5. 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	.85 (.92)	Content of the question (five categories)	.56 (.82)
Conditions	.89 (.94)	Content of the question (two categories)	.90 (.97)
Memory	.85 (.92)	Difficult language usage	.61 (.85)
Hypothetical situation	.98 (.99)	Emotional charge	.75 (.91)
Calculations	.94 (.97)	Presumption of a filter question	.62 (.85)
Ambiguity	.96 (.98)	Sensitive information	.53 (.81)
Mismatch	.98 (.99)	Centrality	.59 (.84)
Formulation	.57 (.68)	Response complexity	.91 (.97)
Clarification	.71 (.82)		

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 .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.

Intercoder Reliabilities

Following this overview of the relative frequencies of the item characteristics over all surveys together, we 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 consider proportions of .80 and higher as reasonably high intercoder reliability and proportions of .79 and lower as low intercoder reliability. Therefore, we focus on proportions below .80 when we try to explain potential low intercoder reliability. For clarity reasons, we present the intercoder reliabilities for the hard and easy item characteristics separately. See Table 5 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

Table 6. 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	.76	.65	.68
Difficult language usage	.73	.69	.81
Emotional charge	.91	.83	.77
Presumption of a filter question	.74	.74	.76
Sensitive information	.74	.67	.66
Centrality	.74	.70	.74
Response complexity	.94	.94	.95

hard item characteristics, see Table 6 for the intercoder reliabilities for the three pairs of coders.

Intercoder reliabilities for the easy item characteristics. As can be seen in the left part of Table 5, 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 2 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 a subjective category. For this specific intercoder reliability, the categories “factual behavior” 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 5, 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 or subjective content.

As can be seen in the right part of Table 5, 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 behavior (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 try to explain low intercoder reliability for the concerned item characteristics.

Explaining Low Intercoder Reliability

Following this overview of the intercoder reliability statistics, we 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 briefly discuss these key factors to indicate the difficulty in obtaining reasonably high intercoder reliabilities. Second, with the two key factors in mind, we discuss the characteristics that had a fixed coder probability λ below the value of .90 (see Statistics subsection and Table 5 in Intercoder Reliabilities subsection). 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 Online Appendix C and Table S9 for a brief explanation). Regarding the hard item characteristics, we also discuss those characteristics that had an intercoder reliability below the value of .80 for at least one of the three pairs of coders (see Table 6).

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 gray 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 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 nontraumatic emotions is drawn. Some coders may have given more room to nontraumatic 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 this presumption for 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 codable 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 suggest a few options for coping with low intercoder reliability in constructing questionnaire profiles based on their item characteristics in the following section.

Coping with Low Intercoder Reliability

Following this overview of the most likely explanations for the low intercoder reliability that was found, we 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 discuss these four options in some detail.

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.

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.

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 subtle 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.

Option 4: Using Item Characteristic Scales with Multiple Applicability Categories

For a way to avoid redefining and redemarcating the item characteristics or programming strict rules for the coding procedure, a fourth option for coping

Table 7. 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	.59	.28	.11	.02
Emotional charge	.73	.20	.04	.02
Presumption of a filter question	.53	.25	.13	.09
Sensitive information	.49	.32	.14	.04
Centrality	.57	.26	.15	.02
Response complexity	.91	.06	.03	.00

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 no, one, two, or three coders indicated its applicability for a certain item. Based on all items for which 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 error. 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 error 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 error.

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 7, 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 error. 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.

Discussion

In this study, we used a systematic coding procedure to code all 2,470 items of 11 Dutch surveys on 16 item characteristics that we expected to be relevant in evoking measurement error according to the literature. We have investigated to what extent the coding of these item characteristics could be done by multiple coders *reliably*. In case of reasonably high intercoder reliability, this would be indicative for an unambiguous relation between item characteristic, item content, and measurement error. Hence, the so-called questionnaire profiles may be constructed, which summarize the characteristics of the items of a survey. If questionnaire profiles could be identified and would appear to be related to varying answering behavior of the part of the respondent, they might be helpful in controlling for measurement error. 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 8 item characteristics 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, which were coded by three coders, a relatively low intercoder reliability was found. Surprisingly, a low intercoder reliability was also found for the characteristics formulation and clarification, which we expected a relatively high intercoder reliability for. 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. 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 from 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 appeared to be hard to code during the pilot study. 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 characteristics coded by three coders may actually 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 .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 5 in Intercoder Reliabilities subsection and Table S9 in Online Appendix C). Second, for most characteristics coded by three coders, the intercoder reliabilities for all three pairs of coders showed that one, two, or all three pairs of coders had an intercoder reliability below the value of .80 that we set as a minimum for reasonable intercoder reliability (see Table 6 in Intercoder Reliabilities subsection). 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 coded by three coders were indeed 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 in Relative Frequencies subsection). 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 .98 and a relative frequency of .02. This means that, for almost all items, both coders did not indicate its applicability, explaining the high intercoder reliability of .98. Thus, for the remaining .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 error, 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 appeared hard to code during the pilot study. 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 hard characteristics consisting of two coding categories, the applicability scales may also be used for items for which no consensus was found to obtain an indicative questionnaire profile for a survey.

Acknowledgment

We would like to thank Rachel Vis-Visschers for her contribution to the pilot study and Rodinde Pauw for coding the remaining item characteristics.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Bassili, John N. and B. Stacey Scott. 1996. "Response Latency as a Signal to Question Problems in Survey Research." *Public Opinion Quarterly* 60:390-99.

Beukenhorst, Dirkjan, Bart Buelens, Frank Engelen, Jan van der Laan, Vivian Meertens, and Barry Schouten. 2014. "The Impact of Survey Item Characteristics on Mode-specific Measurement Bias in the Crime Victimisation Survey." Discussion paper 201416, Statistics Netherlands, Den Haag, the Netherland.

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, Bureau of Labor Statistics, Washington, DC. Alexandria, VA: American Statistical Association, 749-54.

Bradburn, Norman M. and Seymour Sudman, and Associates. 1979. *Improving Interview Method and Questionnaire Design*. San Francisco, CA: Jossey-Bass.

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.

Cho, Young Ik, Anne Fuller, Thom File, Allyson L. Holbrook, and Timothy Johnson. 2006. "Culture and Survey Question Answering: A Behavior Coding Approach." Pp. 4082-89 in *American Statistical Association 2006 Proceedings of the Section on Survey Research Methods*. Washington, DC: American Statistical Association.

Dijkstra, W. 1994. "SEQUENCE: A Program for Analysing Sequential Data." *Bulletin de Méthodologie Sociologique* 43:134-42.

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* 78:721-33.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 76:378-82.

Foddy, William. 1993. *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research*. Cambridge, MA: Cambridge University Press.

Fowler, Floyd, Jr. 1995. *Improving Survey Questions: Design and Evaluation*. Applied Social Research Methods Series, 38. Thousand Oaks, CA: Sage.

Fowler, Floyd, Jr., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing; Minimizing Interviewer-related Error*. Newbury Park, CA: Sage.

Gallhofer, Irmtraud N., Annette Scherpenzeel, and Willem E. Saris. 2007. "The Code-book for the SQP Program." Retrieved October 17, 2017, (http://www.europeansocialsurvey.org/docs/round7/methods/ESS7_sqp_codebook.pdf).

Holbrook, Allyson L., Young Ik Cho, and Timothy Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70:565-95.

Jenkins, Cleo R. and Don A. Dillman. 1997. "Towards a Theory of Self-administered Questionnaire Design." Pp. 165-96 in *Survey Measurement and Process Quality*, edited by L. E. Lyberg, P. P. Biemer, M. Collins, L. Decker, E. D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. New York: Wiley-Interscience.

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* 40:80-104.

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* 72:847-65.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213-36.

Lensvelt-Mulders, Gerty J. L. M. 2008. "Surveying Sensitive Topics." Pp. 461-78 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* 24:1003-20.

LozarManfreda, 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. Retrieved October 17, 2017 (http://ww2.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-000972.pdf).

Ongena, Yfke P. and Wil Dijkstra. 2006. "Methods of Behavior Coding of Survey Interviews." *Journal of Official Statistics* 22:419-51.

Paulhus, Delroy L. 2002. "Socially Desirable Responding: The Evolution of a Construct." Pp. 49-69 in *The Role of Constructs in Psychological and Educational Measurement*, edited by Henry I. Braun, Douglas N. Jackson, and David E. Wiley. Mahwah, NJ: Erlbaum.

Saris, Willem E. and Irmtraud Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 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* 4:61-79.

Schaeffer, Nora C. 2000. “Asking Questions about Threatening Topics: A Selective Overview.” Pp. 105-21 in *The Science of Self-report: Implications for Research and Practice*, edited by Arthur A. Stone, Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. Mahwah, NJ: Erlbaum.

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: 128-38.

Sudman, Seymour and Norman M. Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, MA: Cambridge University Press.

Tourangeau, Roger and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133:859-83.

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

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* 29:299-315.

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.

Van der Zouwen, Johannes and Wil Dijkstra. 1998. “Het Vraaggesprek Onderzocht. Wat Zegt Het Verloop Van De Interactie in Survey-interviews over De Kwaliteit Van De Vraagformulering?” *Sociologische gids* 45:387-403.

Van der Zouwen, Johannes and Johannes H. Smit. 2004. “Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Transcripts of Question-answer Sequences: A Diagnostic Approach.” Pp. 109-30 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jennifer M.

Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York: John Wiley.

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* 75:349-65.

Author Biographies

Frank Bais (Utrecht University, Department of Methodology and Statistics) is a PhD candidate at the Department of Methods and Statistics at Utrecht University since January 2014. The current project is a cooperation between Utrecht University and Statistics Netherlands. His research focuses on characteristics of respondents and questionnaires to gain insight into mode-specific measurement error. Email: f.bais@uu.nl

Barry Schouten (Statistics Netherlands, The Hague) is a senior methodologist at Statistics Netherlands. His research focuses on methods for the reduction and correction of nonresponse and measurement error in surveys. He has several publications in this field. Email: bstn@cbs.nl

Peter Lugtig (Utrecht University, Department of Methodology and Statistics) is an assistant professor in social science research methodology at the Department of Methods and Statistics at Utrecht University. His research focuses on the methodology of panel surveys and statistical modeling of survey data quality. He has recently published articles on panel attrition, dependent interviewing, and data quality in mobile surveys. Email: p.lugtig@uu.nl

Vera Toepoel (Utrecht University, Department of Methodology and Statistics) is an assistant professor at the Department of Methods and Statistics at Utrecht University. Her research focuses on survey methodology. She has several publications in the field of survey research, for instance, the book *Doing Surveys Online* (Sage). E-mail: v.toepoel@uu.nl

Judit Arends-Tòth (Statistics Netherlands, Heerlen) is a survey researcher at Statistics Netherlands at the Department of Methodology and Quality. E-mail: j.arends-toth@cbs.nl

Salima Douhou (City University London) used to be involved in panel data analysis and survey research at CentERdata. Since May 2015, she is research fellow for the European Social Survey at City University London. E-mail: salima.douhou@city.ac.uk

Natalia Kieruj (CentERdata, Tilburg) is a survey researcher for the CentERpanel at CentERdata. Her work includes developing and programming survey research and administering online data collection. Email: n.d.kieruj@uvt.nl

Mattijn Morren (Statistics Netherlands, Heerlen) works at Statistics Netherlands as a statistical researcher. His current work concentrates on health and holiday statistics. Email: m.morren@cbs.nl

Corrie Vis (CentERdata, Tilburg) used to be a senior researcher in survey research at CentERdata until her retirement in the summer of 2015. Email: corrie8327@gmail.com