Discussion paper

Is undesirable answer behaviour consistent across surveys?

Frank Bais
Barry Schouten
Vera Toepoel

December 2019

Abstract

In this study, we investigated to what extent respondent characteristics may be associated with undesirable answer behaviour consistently across surveys. We used respondent data from ten national population surveys of CentERdata and Statistics Netherlands. An adaptation of the robust effect size Cliff's Delta was used to compare average density distributions on the potentially consistent occurrence of answer behaviour across surveys. The results did not show consistent undesirable answer behaviour. Many characteristics' categories were associated with a relatively *higher* occurrence of answer behaviour for some surveys, but a relatively *lower* occurrence for other surveys. We conclude that the occurrence of answer behaviour may be more dependent on the survey and its items than on respondent characteristics. We recommend follow-up research to investigate the relation between item characteristics and answer behaviour.

1. INTRODUCTION

The relation between survey answer behaviour and measurement error has been studied extensively. According to the literature, the occurrence and size of measurement error and hence response data quality can be influenced by both item characteristics (Campanelli et al., 2011; Krosnick, 1991; Yan & Tourangeau, 2007) and respondent characteristics (Olson & Smyth, 2015; Roberts, 2007; Stern et al., 2007). Respondent characteristics can be thought of as fixed tendencies or features of a respondent. Some of these characteristics may lead to undesirable answer behaviour, like satisficing (Holbrook et al., 2003; Kaminska et al., 2010). Satisficing refers to respondents taking short-cuts in the question-answering process and is the outcome of the interaction of motivation, cognitive ability, and question difficulty (Krosnick, 1991, 1999; Krosnick et al., 1996). Both motivation and cognitive ability may be considered characteristics of the respondent. Cognitive ability is relatively constant over time for a specific respondent, while motivation may be a fixed part of the respondent's personality, but also dependent on the survey topic and other external factors.

Socio-demographic respondent characteristics consist of more concrete and straightforward personal characteristics, like gender, age, origin, educational level, and income. These socio-demographics can be used as auxiliary variables to validate survey data, being adopted from official registers (see Bakker, 2012; Scholtus et al., 2015), but may also function as background variables that are collected by the survey administrator when validation is not of primary importance. The background variables may not be free of measurement errors themselves, but these errors are assumed not to relate to survey response behaviour and to be relatively stable through time (Schouten & Calinescu, 2013).

Answer behaviour should be stable and typical for the respondent in order to investigate its relation to respondent characteristics. That is, the behaviour for a specific respondent must be shown *consistently* in order to be typical for that respondent. Here, the term 'consistent' refers to a pattern of answer behaviour that is shown over several moments in time, across multiple surveys. When a respondent only incidentally shows a certain answer behaviour, it is not to say whether this is typical for that specific respondent. For instance, a respondent could fill out a single battery of ten multiple choice items by choosing the very first answering option for each item. It is however not clear to what extent this may be a form of satisficing (Krosnick 1991, 1999; Krosnick et al., 1996), as the answers may just as well be truly applicable to that respondent. In case of consistent answer behaviour, we may connect the behaviour to other

stable characteristics of the same respondent. *In this paper, we investigate the relation between stable respondent characteristics and consistent undesirable answer behaviour.*

For our study, we use data from ten large population surveys administered by CentERdata in the LISS Panel. In section 2 and 3 of this paper, we introduce the answer behaviours and respondent characteristics respectively, and elaborate on their connection to undesirable answer behaviour and measurement error according to the literature. In section 4, we describe the methods that were used to compare the different categories of the respondent characteristics for each answer behaviour across surveys. In section 5, we show all statistical results and give answers to our main research question. In section 6, we conclude with a discussion of these results and make suggestions on how to proceed.

2. ANSWER BEHAVIOURS

In this section, we elaborate on ten relevant answer behaviours, selected from the literature: Avoiding follow-up questions, socially desirable responding, answering 'don't know', answering 'won't tell', acquiescent responding, neutral responding, extreme responding, primacy responding, recency responding, and straightlining. We motivate their inclusion by elaborating on why they may be referred to as undesirable and how they may be related to measurement error.

Avoiding follow-up questions: Filter questions are questions containing answering options that may lead to follow-up questions. Adding a filter may result in more 'don't knows' and affect respondents' attitudes (Bishop et al., 1983). Filter questions can be burdensome by additional instructions and information (Redline & Dillman, 2002) or by including presuppositions that can make the question suggestive towards a particular answering option (Knäuper, 1998). Both cases may possibly lead to measurement error. Respondents may presume a question to be a filter question, depending on the content and the format of the question (Eckman et al., 2014; Kreuter et al., 2011). This presumption may motivate respondents to give an answer that avoids follow-up questions (Bosley et al., 1999). For filter questions in grouped or ensemble format, respondents report higher disorder prevalence rates (Kessler et al., 1998), more mental health service use (Duan et al., 2007), and more times 'yes' triggering follow-up questions (Jensen et al., 1999; Lucas et al., 1999) than for filter questions in interleafed format.

Socially desirable responding: Socially desirable responding refers to the tendency to minimize showing socially *undesirable* behaviour (DeMaio, 1984; Krosnick, 1999; Paulhus, 2002). It can refer to both automatic and deliberate answer behaviour (Andersen & Mayerl, 2019). Questions asking for sensitive information (Johnson & Van de Vijver, 2003; Kreuter et al., 2008; Tourangeau et al., 2000; Tourangeau & Yan, 2007) or less anonymous modes of data collection (Johnson & Van de Vijver, 2003) may evoke such an answer. The degree of socially desirable responding strongly depends on the data collection mode (see Campanelli et al., 2011; Heerwegh & Loosveldt, 2011; Holbrook et al., 2003; Kreuter et al., 2008; Roberts, 2007; Roberts & Jäckle, 2012; Tourangeau & Yan, 2007). See Jann et al. (2019) for references on methods to collect and analyze sensitive data that may evoke socially desirable responding. Social desirability may be a human tendency rather than be particularly dependent on the situation (Paulhus, 2002) or survey (Johnson & Van de Vijver, 2003). This means that respondents who show this behaviour may do so consistently over multiple surveys.

Answering 'don't know' and 'won't tell': The answering options 'don't know' or 'won't tell' are often added to substantive answering categories. A 'don't know' response is relatively more likely to occur when respondents are unknown to a particular topic and as question specificity or the number of response alternatives is large (Leigh & Martin, 1987). Research shows that sensitive questions are likely to receive more refusals, while questions requiring more cognitive effort are likely to receive more don't knows (Shoemaker et al., 2002). Respondents may not give a substantive answer in case they are relatively inexperienced as a respondent (Binswanger et al., 2013), reluctant or lacking motivation to answer (Beatty & Hermann, 2002; Krosnick et al., 2002), or in case items ask for sensitive information (Bradburn et al., 1978; Tourangeau et al., 2000). However, respondents may give an actual answer without knowing the answer or having an opinion (Beatty & Hermann, 2002; Bishop et al., 1986). This implies that a non-response option should only be included when deemed a realistic plausible option (Vis-Visschers et al., 2008). In these situations, relatively lower data quality can be the result, which may be mode-dependent (Fricker et al., 2005; Roberts, 2007).

Acquiescence: Like socially desirable responding, acquiescence may be considered a stable personality tendency (Messick, 1966; Stricker, 1963). Acquiescence is defined as the tendency to answer affirmatively, regardless of the content of the question (Billiet & McClendon, 2000; De Leeuw, 1992; Heerwegh & Loosveldt, 2011; McClendon, 1991). Acquiescence may be mode-dependent (De Rada & Domínguez, 2015) and potentially result in measurement error.

Saris et al. (2010) found relatively lower data quality for agree-disagree rating scales, referring to the tendency to acquiesce in case of agree-disagree items (O'Muircheartaigh et al., 2000; Schaeffer & Presser, 2003). Considering acquiescence at least partly a stable personality tendency rather than particularly survey- or item-dependent, it may be likely that respondents who show this behaviour do so consistently over multiple surveys.

Neutral and extreme responding: Offering a neutral middle option increases the probability that respondents express this response (Kalton et al., 1980), possibly indicating satisficing (Krosnick & Fabrigar, 1997). The middle option is more likely to be chosen in case the answering options are presented more prominently (Tarnai & Dillman, 1992) and when beneficial options are placed first (Stern et al., 2007). O'Muircheartaigh et al. (2000) suggest including middle alternatives to reduce random measurement error in responses. Extreme responding is the tendency to choose extreme answering categories. This tendency may differ intra-individually when survey mode is switched (Aichholzer, 2013) and more generally in terms of relatively more extreme answers in interviewer-administered versus self-administered surveys (De Leeuw, 1992). Studies reported relatively more extreme positive responding in telephone mode (Ye et al., 2011) and in postal surveys versus internet surveys (De Rada & Domínguez, 2015). An increase in extreme responding may refer to stable response behaviour (De Leeuw, 1992). This means that giving extreme responses may partly be a personal tendency instead of necessarily depending on the survey.

Primacy and recency responding: Depending on the order in which answering options are offered, response order effects may occur; an option at the beginning or at the end of a list may be chosen, respectively called a primacy and a recency effect (Krosnick & Alwin, 1987). These effects occur as some respondents do not give equal consideration to all the response alternatives (Krosnick & Alwin, 1987; McClendon, 1991). Primacy effects may be expected in case items are either self-administered or read from a show card and thus presented visually (Galesic et al., 2008; Krosnick, 1991; Krosnick, 1992; Krosnick & Alwin, 1987). Recency effects may be expected in case items are interviewer-administered and thus presented orally (Krosnick, 1991; Krosnick, 1992; Krosnick & Alwin, 1987). Both effects may lead to measurement error (see Galesic et al., 2008).

Straightlining: Questions followed by a common answering scale are often clustered together (Krosnick, 1991), possibly leading respondents to differentiate to a smaller extent between the

questions in their answers (Krosnick & Alwin, 1989). Straightlining, or non-differentiation, refers to giving the same answers to a series of questions arranged in a grid format (De Rada & Domínguez, 2015; Schonlau & Toepoel, 2015). Straightlining seems more common towards the end versus the beginning of a questionnaire (Krosnick, 1991). It tends to increase for respondents who give answers very quickly or 'speed' (Zhang, 2013; Zhang & Conrad, 2013) or had relatively longer panel experience (Schonlau & Toepoel, 2015). Straightlining may partly be dependent on the type of survey topic or question (Schonlau & Toepoel, 2015).

3. RESPONDENT CHARACTERISTICS

To motivate the inclusion of respondent characteristics for this study, we connect the characteristics to category differences and measurement error (see Table 6 in Appendix A for the categories belonging to each characteristic). In this section, we present an overview of the literature for each of the following eight respondent characteristics: Gender, age, education, domestic situation, primary occupation, income, origin, and whether or not borrowing a computer from CentERdata for filling out the surveys (in case a respondent did not have a computer or an internet connection before participating in the panel).

We state hypotheses regarding the relations between respondent characteristics and answer behaviours across surveys only when this could be based on the literature. This means that more relations may be found than we hypothesize. For most relations however, we did not state explicit hypotheses about what to expect for various reasons. First, the many relations between eight characteristics and ten behaviours would result in a complex overview of many hypotheses. Second, many of these relations are missing, unclear or ambiguous according to the literature. And third, evidence about these relations being consistent across multiple surveys is largely unknown. Therefore, most characteristic-behaviour relations in this study are investigated *exploratively*.

Gender:

Women are found to give more 'no opinion'-answers (Pickery & Loosveldt, 1998) and 'don't know'-answers than men (Antoni et al., 2019; O'Muircheartaigh et al., 2000; Schräpler, 2004), possibly referring to a gender difference in cognitive engagement for certain topics (O'Muircheartaigh et al., 2000). Women are also found to have a larger propensity to give affirmative answers (Hox et al., 1991; O'Muircheartaigh et al., 2000) and socially desirable responses than men (Bernardi, 2006). Men are found to have a larger tendency to give extreme

responses (Marshall & Lee, 1998) and to straightline more than women (Zhang & Conrad, 2013). It is likely that men and women differ in various kinds of undesirable answer behaviour. *Across surveys, we hypothesize more acquiescent, socially desirable, and 'don't know'-answers for women, and more extreme answers and straightlining for men.*

Age

Older respondents show less accurate survey answer behaviour than younger respondents (Andrews & Herzog, 1986) and a decline in attitude reliability measurement in the oldest age group of 66 years and older (Alwin & Krosnick, 1991). Two studies found more acquiescence for older than for younger respondents (Meisenberg & Williams, 2008; O'Muircheartaigh et al., 2000), while other studies found the opposite (Hox et al., 1991) or no effect (He et al., 2014). Older respondents are found to give more extreme answers (Greenleaf, 1992; He et al., 2014; Meisenberg & Williams, 2008), including across questionnaires (Kieruj & Moors, 2013), while younger respondents are found to choose relatively more middle or neutral options (He et al., 2014). One study found more straightlining for older respondents (Schonlau & Toepoel, 2015), while another did not find a relation between age and straightlining for respondents who speed (Zhang & Conrad, 2013). Finally, age may be related to non-substantive responses. Older respondents are found to give more 'no opinion'-answers (Pickery & Loosveldt, 1998) or 'don't know'-answers (O'Muircheartaigh et al., 2000) than younger respondents. It seems realistic to expect an association between age and some of the answer behaviours.
Across surveys, we hypothesize more extreme and 'don't know'-answers for older respondents, and more neutral answers and straightlining for younger respondents.

Education

Higher education is associated with stable reliability of attitude measurement (Alwin & Krosnick, 1991) and more accurate answer behaviour (Antoni et al., 2019). Lower educated respondents are found to give more 'no opinion'-answers (Narayan & Krosnick, 1996; Krosnick et al., 2002; Pickery & Loosveldt, 1998) and 'don't know'-answers (O'Muircheartaigh et al., 2000; Schuman & Presser, 1981) than higher educated respondents. Most studies found a negative relation between education and acquiescence (McClendon, 1991; Narayan & Krosnick, 1996; O'Muircheartaigh et al., 2000), although some research did not find a relation (Bachman & O'Malley, 1984; He et al., 2014; Hox et al., 1991). Also a negative relation between education and extreme responding is found (Aichholzer, 2013; Greenleaf, 1992; He et al., 2014; Marin et al., 1992 - but see Bachman & O'Malley (1984b) for different findings),

while mixed results exist concerning choosing middle or neutral options (see Narayan & Krosnick, 1996 versus He et al., 2014). For specific items and topics, more recency responding was found for lower educated respondents (McClendon, 1986, 1991), while the evidence for primacy responding was mixed (see Krosnick & Alwin, 1987 versus McClendon, 1991). Finally, among respondents who speed, more straightlining was found for lower educated respondents (Zhang & Conrad, 2013). Associations between education and some of the selected answer behaviours can be expected.

Across surveys, we hypothesize more extreme, acquiescent, and don't know-answers, and straightlining for lower educated respondents.

Origin

Answer behaviour may be influenced by cultural factors and have substantive culture-specific meaning (Cheung & Rensvold, 2000; Smith, 2004). Cultural differences in response styles may be explained by differences in judgment style (Bachman & O'Malley, 1984a; Hui & Triandis, 1989), language (Bachman & O'Malley, 1984a; Harzing, 2006), and the extent of individualistic versus collectivistic influences in a country (Bernardi, 2006; Chen et al., 1995; Johnson & Van de Vijver, 2003; Marshall & Lee, 1998; Van Herk et al., 2004). Afro-American, Hispanic, and Mediterranean respondents are found to show the most extreme responding (see Bachman & O'Malley, 1984ab; Baumgartner & Steenkamp, 2001; Hui & Triandis, 1989; Marin et al., 1992; Van Herk et al., 2004), Asian respondents to show the least extreme responding, and respondents from North Western America, Australia, and Europe seem to fall in between (Chen et al., 1995; Chun et al., 1974; Dolnicar & Grun, 2007; Watkins & Cheung, 1995; Zax & Takahashi, 1967). But see Marshall & Lee (1998), and Stening & Everett (1984) for contradicting results. In contrast, more neutral responding was found for Asian respondents than for non-Asian respondents (Si & Cullen, 1998; Stening & Everett, 1984), Australian respondents (Dolnicar & Grun, 2007), and North American respondents (Chen et al., 1995; Zax & Takahashi, 1967), and for non-Western immigrants than for Western immigrants and Dutch native citizens (He & Van de Vijver, 2013). Most acquiescence is evident for Hispanic, Mediterranean, and Asian respondents, and non-Western immigrants, while less acquiescence is shown by non-Hispanic whites (Marin et al., 1992), Australians (Watkins & Cheung, 1995), North Western Europeans (Baumgartner & Steenkamp, 2001; Van Herk et al., 2004), and Western immigrants and Dutch native citizens (He & Van de Vijver, 2013). Finally, most socially desirable responses are shown for Afro- and Mexican Americans, Hispanics, and

Asians, while less for US-born and European Americans, non-Hispanic whites, and Mexicans (see Johnson & Van de Vijver, 2003).

Across surveys, we hypothesize more neutral, acquiescent, and socially desirable answers for non-Western respondents.

Income and Primary Occupation

Income and primary occupation have been shown to be related to answer behaviour and measurement error (see Butler et al., 1987; Greenleaf, 1992; Lynn & Kaminska, 2012; McClendon, 1991; Schräpler, 2004). Antoni et al. (2019) found a relation between higher income and less accurate answer behaviour. Greenleaf (1992) found a negative relation for income and extreme responding. McClendon (1991) found a negative association for income with acquiescence, which they explain by a lower status, rather than by limited cognitive sophistication (McClendon, 1991). Respondents may be reluctant to reveal having no or a relatively low paid job or income because of its lower status. Butler et al. (1987) found that individuals who do not work have the tendency to report their health incorrectly, which may be considered a form of socially desirable responding. Schräpler (2004) found that respondents with a higher occupational status show more 'won't tell'-answers on questions about income than respondents with a lower occupational status. He also refers to respondents with a lower occupational status who answer 'don't know' more often (Schräpler, 2004). We adopt the suggestion of Schräpler (2004) to include answering 'don't know' and 'won't tell' as response categories concerning items asking about income. In sum, various answer behaviours are to be expected for the respondent characteristics income and primary occupation.

Across surveys, we hypothesize more extreme, acquiescent, socially desirable, and 'don't know'-answers in case of no paid work or lower income, and more 'won't tell'-answers for both lower and higher income.

Domestic Situation

Literature suggests that factors concerning the household composition or domestic situation may have their influence on answer behaviour and measurement error. Here, three main and interrelated factors are distraction, the presence of others, and multitasking (Holbrook et al., 2003; Kellogg, 2007; Lavrakas et al., 2010; Lynn & Kaminska, 2012; Olson et al., 2018; Schwarz et al., 1991). As a result of distraction or multitasking, respondents are less likely to provide accurate responses in general (Lavrakas et al., 2010; Olson et al., 2018), especially on cognitively demanding items (Lavrakas & the AAPOR Cell Phone Task Force, 2010). Kellogg

(2007) refers to the inherent speed and quality costs of executing two concurrent tasks simultaneously due to its complexity and attention-demanding nature (Kellogg, 2007), possibly enhancing the likelihood of satisficing (Holbrook et al., 2003). Considering the number of people in a household an easily assessed proxy for cognitive ability (see Alwin & Krosnick, 1991), we use this characteristic as an indicator for risk of distraction or multitasking (see Olson et al., 2018; Schwarz et al., 1991).

Being provided a Computer

Respondents who do not own a personal computer were provided a computer from the involved panel administrators to complete the surveys in question (see Schonlau & Toepoel, 2015). One study showed a lower prevalence of speeders among respondents who received a computer from the panel administrators than the respondents who did not (Zhang, 2013; Zhang & Conrad, 2013). A possible explanation is that respondents who are being provided a computer feel more responsible for participating seriously than respondents owning their own device. Another explanation is that respondents with a provided computer may have less experience using the internet and computers, and may therefore need more time and effort to navigate through the survey than respondents who have their own device and hence more experience (Zhang, 2013; Zhang & Conrad, 2013). Both explanations refer to the plausibility of more accurate answer behaviour and hence better survey data quality for respondents being provided a computer.
Across surveys, we hypothesize more straightlining for respondents owning their own device.

4. METHOD

In this section, we discuss the survey data and statistical methods that we use to answer our research question about the relation between respondent characteristics and consistent answer behaviours. First, we introduce the surveys and their topics from the LISS Panel that we used for this study. Second, we explain how we constructed 'behaviour profiles' that we used for all answer behaviours, surveys, respondent characteristics, and their accompanying categories. We also elaborate on how we come to average behaviour occurrences -the expected values- for groups of respondents by means of these behaviour profiles. Third, we describe the non-parametric Cliff's Delta statistic for comparing the behaviour profiles for the various categories to the overall behaviour profiles. Here, the overall behaviour profile consists of the profiles for all categories of a certain characteristic taken together, except for the category profile to which the overall profile is compared. We close the method section by elaborating on the method of

estimating confidence intervals for Cliff's Delta and by stating the statistics to answer our main research question.

4.1 LISS Panel and Surveys

We selected ten Dutch general population surveys that were administered by CentERdata to the same respondents of the Longitudinal Internet studies for the Social Sciences (LISS) Panel. This was done in the time period between June 2012 and December 2013. The surveys were the first wave of the Dutch Labour Force Survey from Statistics Netherlands and nine of the core studies from CentERdata. All surveys were administered in computer-assisted format. These surveys cover a broad range of topics in the field of general population statistics (see Table 1). The data for the eight background variables as presented in section 3 were also provided by CentERdata. The LISS Panel consists of about 7000 individuals from about 4500 households and is based on a probability sample of households. This sample is drawn from the population registry by Statistics Netherlands. All panel members were invited for all surveys included in this study. The number of respondents that filled out a specific survey differed per survey and the number of surveys that respondents filled out varied across respondents. The average number of surveys filled out by a respondent was 8. Altogether, the surveys contain 2074 items that were used to cover the ten possible answer behaviours as presented in section 2.

Before constructing behaviour profiles, the survey data needed to be coded for each behaviour separately. See Table 7 in Appendix B for an overview of the selection of eligible items and the operationalization of the behaviours; see Table 8 in Appendix C for the proportions of eligible items per survey and in total for all behaviours; see Bais et al. (submitted) for an elaboration on the coding procedure and behaviour operationalization. With respect to Bais et al. (submitted), the behaviours socially desirable responding and acquiescence were re-coded for the current study, see Appendix D.

4.2 Behaviour Profiles and Expected Values for Groups of Respondents

In this subsection, we define behaviour profiles and elaborate on the calculation of expected values for groups of respondents. A behaviour profile represents the relative proportions of a specified population group (for instance women) in showing a specified behaviour (for instance answering don't know) at all possible occurrence rates from 0 to 1.

Table 1. Overview of all Surveys, a Description of their Content, and their Response Rate (and the Number of Respondents).

Survey (administration period, nr. of items)	*Topics of the content*	*Response rate (and nr. of respondents)*
Economic Situation Assets (AS) (Jun/Jul '12, i = 50)	Income, property and investment	75.2% (5588)
Family and Household (FA) (Mar/Apr '13, i = 409)	Housing and household; social behaviour	88.8% (5826)
Health (HE) (Nov/Dec '12, i = 243)	Health and well-being	85.4% (5780)
Economic Situation Housing (HO) (Jun/Jul '13, i = 73)	Housing and household; income, property and investment	58.2% (3199)
Economic Situation Income (IN) (Jun/Jul '13, i = 286)	Employment, labour, retirement; income, property, investment; social security, welfare	78.4% (5015)
Personality (PE) (May/Jun '13, i = 200)	Psychology	90.6% (5169)
Politics and Values (PO) (Dec '12/Jan '13, i = 148)	Politics; social attitudes and values	85.7% (5732)
Religion and Ethnicity (RE) (Jan/Feb '13, i = 71)	Religion; social stratification and groupings	88.6% (5908)
Work and Schooling (WO) (Apr/May '13, i = 471)	Education; employment, labour and retirement	86.5% (5585)
Labour Force Survey (LFS) (Dec '13, i = 123)	Education; employment and labour	81.2% (3166)

For each respondent, the number of items for which a specific behaviour is shown and the number of items for which it was possible to show this behaviour are counted. Dividing these two units gives a probability or proportion between 0 and 1 that refers to the expected occurrence of the behaviour for the concerned respondent. As each respondent filled out a delimited number of items from not more than ten surveys, an extent of uncertainty exists concerning the actual occurrence of the behaviour. Moreover, as each respondent filled out a variable number of surveys and items per survey due to filter questions, the actual occurrence of behaviour is indicated with varying uncertainty across different respondents. It is necessary to include these uncertainties in calculating the expected behaviour occurrences. This can be done by calculating the likelihood of behaviour occurrence for the whole probability range from 0 to 1. The occurrence of each behaviour per respondent is then estimated by calculating the likelihood λ of each possible occurrence from 0 to 1 with a step size of 0.01:

$$\lambda_{v,c,r}(p) = \binom{N}{K} p^K (1-p)^{N-K} \text{ with } 0 \leq p \leq 1, \quad (1)$$

where $\lambda_{v,c,r}$ is the likelihood of occurrence of the behaviour for respondent *r* in category *c* of characteristic *v*, *p* is the posterior probability, *N* is the number of actually filled out items for which the behaviour was possible, and *K* is the number of items for which the behaviour actually occurred. In this way, a likelihood curve or likelihood distribution for the probability range from 0 to 1 can be constructed for each respondent's expected behaviour occurrence. We will name this likelihood distribution a *behaviour profile*, as it delineates the expected occurrence across the full potential probability range from 0 to 1 and hence gives visual consideration to the amount of occurrence uncertainty. For a single respondent *r* in category *c* of characteristic *v*, the average or expected value $EV_{v,c,r}$ for the behaviour occurrence can be estimated on the basis of the respondent's behaviour profile and the integral over *p*:

$$EV_{v,c,r} = \int_{p=0}^{1} p\lambda_{v,c,r}(p)dp \tag{2}$$

To be able to compare groups of respondents with different characteristics, for instance men and women, their expected values are useful starting points as estimates of the average behaviour occurrences per group. By considering all respondents who meet the condition for a specific category of a characteristic, the average estimate for this category can be calculated by summing their behaviour profiles and dividing the outcome by the integral over *p*:

$$\bar{\lambda}_{v,c}(p) = \frac{\sum_{r=1}^{R} \lambda_{v,c,r}(p)}{\int_{p=0}^{1} \sum_{r=1}^{R} \lambda_{v,c,r}(p)dp} \tag{3}$$

where $\bar{\lambda}_{v,c}$ is the likelihood of the behaviour occurrence averaged over all concerned respondents in category *c* of characteristic *v*, and *R* is the total number of these respondents. By means of this average behaviour profile, the expected value $EV_{v,c}$ for the behaviour occurrence for this specific category of respondents can be calculated as follows:

$$EV_{v,c} = \int_{p=0}^{1} p\bar{\lambda}_{v,c}(p)dp \tag{4}$$

The expected values of two groups with different characteristics indicate the average behaviour occurrences for the groups as a whole. In this way, an idea is obtained about the difference of the occurrences of certain behaviour (for instance answering don't know) between two groups

(for instance men and women). The next step is to use a solid analysis to compare the behaviour occurrences of two groups.

4.3 Cliff's Delta for Comparing Groups of Respondents

When comparing the expected values of two groups or categories of respondents of a certain characteristic, it is obvious to compare the behaviour profile for one category to the profile for the combined remainder of other categories of the characteristic. For instance, men can be compared to women for 'gender' and respondents with a Dutch background can be compared to respondents with all other backgrounds combined for 'origin'. To test whether the expected value of a specific category differs from the expected value of the combined remainder of categories for a characteristic in terms of effect size, an adaptation of Cliff's Delta (Cliff, 1993, 1996ab) is calculated. Cliff's Delta δ can be used as a robust alternative to using two independent means.

Using Cliff's Delta for the current research asks for an adaptive version of the statistic, as we are not considering data observations but density distributions. See Appendix E for this adaptation for density distributions and for how Cliff's Delta takes into account both the location and the shape of the behaviour profiles in comparing them. See Appendix F for a brief simulation on how Cliff's adapted Delta quickly approaches Cliff's original Delta as the number of eligible items increases and thus the uncertainty around the expected value decreases.

4.4 Confidence Intervals for Cliff's Delta

For each Cliff's Delta, we used confidence intervals to refer to its amount of uncertainty. For a respondent characteristic, each Cliff's Delta is based on the comparison between the profile of a category and the overall profile of the remaining categories taken together. For a confidence interval, we bootstrapped 10000 category profiles and 10000 overall profiles. For each profile, respondents were randomly sampled with replacement and their individual profiles were averaged conform (3). Here, the number of sampled respondents was equal to the number of respondents in the category or overall group respectively. By means of these averaged bootstrap profiles, we calculated 10000 Cliff's Delta's and ranked them from low to high. Because of the large number of Cliff's Delta's in our study, we chose to use 99% confidence intervals. This means that we used the 51st and the 9950th Cliff's Delta in the ranking to construct each confidence interval.

4.5 Statistics

In the results section, we first give several visual examples of behaviour profiles to compare category profiles that are hardly diffused versus category profiles that are obviously diffused from the overall characteristic profile. Second, we calculate the expected values for every characteristic and their separate categories for each behaviour, both overall and for each survey separately. Third, we calculate the Cliff's Delta's for every characteristic category versus the combined remainder of the concerned characteristic for each behaviour, both overall and for each survey separately. Each Cliff's Delta is accompanied by its 99% confidence interval. All calculations were done in the programming language R.

5. RESULTS

In this section, we first present the descriptive statistics. We show examples of diffusion of category profiles around the overall profiles of those categories together. Then, we give the expected values for these category and overall profiles. Second, we answer our main research question. We show the Cliff's Delta's for all surveys together as if they were one large survey, to obtain overall differences for each category profile versus the combined profile of remaining categories for the concerned characteristic. Then, we consider the number of surveys for which such potential difference was found to give an indication about behaviour consistency across surveys.

First, we need to note that respondents varied in the number of surveys they filled out. Some respondents filled out only one or two surveys, while others filled out all or almost all surveys. Behaviour data for *every* survey that the respondent filled out were used for the analyses. For instance, if a respondent filled out the surveys Assets, Health, Housing, Income, and Personality, this respondent is included in the data analyses for all these surveys. Second, respondents are classified in one category of *every* background characteristic that they filled out. This means that a respondent can be male, older than 65 years, highly educated, retired, and Dutch, and is included in the data analyses for all these characteristics. Hence, respondents are included in each survey and characteristic analysis that is applicable to them. This means that we do not consider individual respondents in this study, but that we focus on *groups* of respondents sharing the same characteristic. Therefore, in describing and interpreting our results, the emphasis is placed on the relations between answer behaviour and the characteristics with their categories.

5.1. Diffusion of Category Profiles around their Combined Overall Profiles

For each category of each respondent characteristic, including for the characteristic's categories together, a behaviour profile was constructed by means of (1) and (3). For every possible behaviour occurrence from 0 to 1 with step size interval 0.01, the profile refers to a density outcome. This outcome refers to the chance on the concerned behaviour for that specific occurrence interval. The behaviour profiles for the categories of each characteristic are to a certain extent diffused around their combined overall behaviour profile.

To compare relatively small and large diffusion, let us consider Figure 1. As can be seen in Graphs 1 and 2, the several category profiles are very close to each other and to the combined overall profile. This refers to a relatively *small* diffusion for the respective characteristics 'being provided a computer' and 'origin' for the concerned surveys for the behaviours 'answering won't tell' and 'extreme responding' respectively. On the contrary, Graphs 3 through 6 show relatively *larger* diffusion for the behaviours 'socially desirable responding', 'straightlining', 'answering don't know', and 'neutral responding' respectively. As can be seen from the graphs, the category profiles for the respective characteristics 'primary occupation', 'domestic situation', 'gender', and 'primary occupation' differ relatively more from their overall profile.

5.2. Expected Values

A useful descriptive statistic is the expected behaviour occurrence for the category and overall profiles. The expected occurrence or value for each category and overall profile, thus for any group of respondents, can be constructed by means of (4). The expected value refers to an average occurrence of the concerned behaviour on the basis of the profiles of all group members. To see to what extent the expected values of the various categories of a certain characteristic may differ, let us consider the aforementioned visual examples. See Table 2.

The two examples for which the diffusion was almost zero show expected values that differ relatively *little*. See the expected values for 'answering won't tell' for 'provided a computer', and for 'extreme responding' for 'origin', in the first and second column of Table 2 respectively. On the contrary, the expected values for the other four examples differ relatively *much*, with respective differences of 0.06, 0.12, 0.04, and 0.23. These varying differences are in line with the graphs of Figure 1; in general, behaviour profiles that show relatively little diffusion have resembling expected values, while profiles that show relatively much diffusion have more divergent expected values.

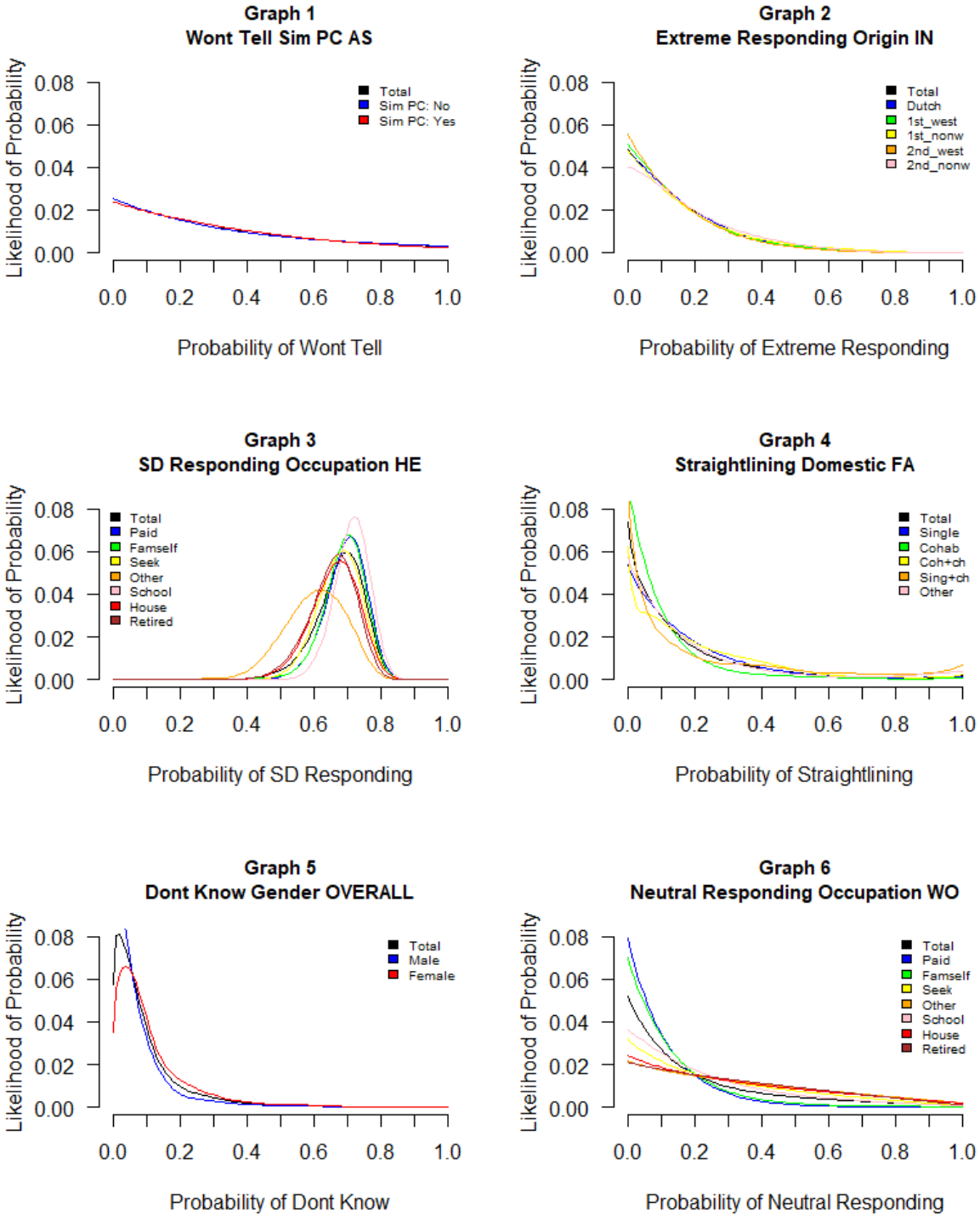

Figure 1. Examples for Little Diffusion (Graphs 1 and 2) versus Much Diffusion (Graphs 3 through 6) between Behaviour Profiles.

In the following parts, we compare the behaviour profiles for each category to the behaviour profile for the combined remainder of the other categories. This is done by means of (9) for the calculation of Cliff's Delta, for all characteristics and behaviours. From here, we first turn to

Table 2. Examples of Expected Values for Various Behaviours, Characteristics, and their Categories.

Behaviour	Answering won't tell	Extreme responding	Socially desirable	Straight-lining	Answering don't know	Neutral responding
Characteristic	Provided a computer	Origin	Age	Domestic situation	Gender	Primary occupation
Survey	Assets	Income	Health	Family	TOTAL	Work
Overall	**0.33**	**0.17**	**0.68**	**0.17**	**0.10**	**0.22**
Category 1 *	0.33	0.17	0.71	0.18	0.08	0.12
Category 2 *	0.32	0.16	0.70	0.12	0.12	0.14
Category 3 *		0.18	0.69	0.21		0.29
Category 4 *		0.15	0.68	0.24		0.35
Category 5 *		0.18	0.66	0.20		0.24
Category 6 *			0.65			0.33
Category 7 *						0.34

* The characteristics in this table have different numbers of categories. Here, the main idea is to illustrate the varying differences between the expected values for the categories of several characteristics. Therefore, we did not explicitly mention the categories of each of the six examples separately.

the Cliff's Delta's for all surveys taken together. For each behaviour, this gives us a global picture about which categories stand out with respect to the others for the various characteristics. Second, we answer our main research question by considering the Cliff's Delta's per survey separately. This gives us an indication about potential consistency across surveys for all behaviours and characteristics.

5.3. Overall Outcomes for Cliff's Delta δ

The overall results for Cliff's Delta δ concern the global picture for groups of respondents with certain characteristics for all surveys taken together. We use the rules that $|\delta| < 0.11$ indicates no effect, $0.11 \leq |\delta| < 0.28$ a small effect, $0.28 \leq |\delta| < 0.43$ a medium effect, and $|\delta| \geq 0.43$ a large effect, as investigated by Vargha & Delaney (2000), see also Goedhart (2016). See Table 3 for the Cliff's Delta's for all surveys taken together with a medium or large effect. From here, we discuss these medium and large effect sizes.

What stands out from Table 3 are the several medium and large effect sizes for the behaviours answering 'don't know' and answering 'won't tell'. Respondents who are young (15-24 years) and/or school-going gave more 'don't know'- and 'won't tell'-answers overall. Respondents with a higher income (2001 EUR or more) gave less 'don't know'-answers overall, while

Table 3. Overall Cliff's Delta (δ) with a Medium or Large Effect (and its 99% Confidence Interval), for the Categories of the Characteristics Age, Primary Occupation, and Income when Applicable, for the Behaviours Avoiding Follow-Up Questions, Socially Desirable Responding, Answering Don't Know, and Answering Won't Tell when Applicable, for All Surveys Taken Together.

	δ avoiding follow-up questions	δ socially desirable responding	δ answering don't know	δ answering won't tell
Age 15_24 yrs			**0.31** * (.25, .36)	**0.33** * (.29, .37)
Occupation Famself	**-0.32** * (-.37, -.26)			
Occupation Other		**-0.40** * (-.45, -.33)		
Occupation School			**0.29** * (.23, .34)	**0.31** * (.26, .35)
Income 2001_3000			**-0.31** * (-.35, -.27)	
Income 3001_more			**-0.34** * (-.41, -.28)	
Income DK			**0.31** * (.19, .42)	**0.28** * (.17, .39)
Income WT				**0.52** # (.45, .58)

* → medium effect; # → large effect

In this table, only the overall |Cliff's Delta's| ≥ 0.28 are shown, as these meet the criterium for a medium or large effect and are discussed in the main text. The overall |Cliff's Delta's| < 0.28 are either not shown or empty cells in this table.

respondents who filled out 'don't know' for income as a background characteristic gave more 'don't know'-answers in the surveys. Respondents who filled out 'don't know' and especially 'won't tell' for income as a background characteristic gave more 'won't tell'-answers overall (see Figure 2 Graph 2). Respondents with an own or family business avoided less follow-up questions. Finally, respondents whose main occupation was 'other' (respondents who are exempted from work seeking, have a work disability, or are doing unpaid or voluntary work) gave fewer socially desirable responses overall (see Figure 2 Graph 1).

A present overall effect size for a certain behaviour and characteristic does not by definition mean a present effect size for various surveys; an overall effect size may exist without effect sizes for any surveys. The opposite may be true as well; an overall effect size may be absent, as positive and negative effects sizes for various surveys cancel each other out. In the following part, we investigate to what extent either positive or negative effect sizes consistently exist across surveys.

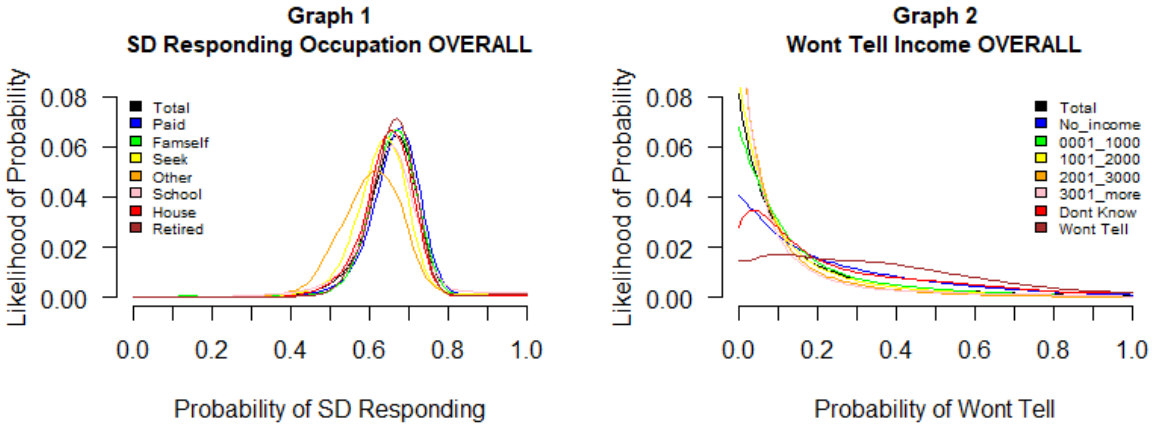

Figure 2. Relatively Less Socially Desirable Responding Overall for Respondents with Occupational Status 'Other' (orange, Graph 1), and Relatively More 'Won't Tell'-Answers Overall for Respondents who filled out 'Don't Know' (red) or 'Won't Tell' (brown) as a Background Characteristic (Graph 2).

5.4. Consistency Outcomes for Cliff's Delta δ

The results for Cliff's Delta δ concern the consistency of groups of respondents with certain characteristics across surveys. To reveal consistency, we considered the number of surveys for which at least a small effect ($|\delta| \geq 0.11$) was the result. If we would consider consistency conservatively, as a small or larger effect for a specific behaviour and characteristic category for *all or almost all* applicable surveys, we would draw the conclusion that there is no consistency to be found. *Strictly, this means that the results do not meet any of our expectations.* See Table 4 containing the categories for which at least two thirds of the applicable surveys showed an either positive or negative effect size: Not one category for a certain characteristic and behaviour shows an effect for all or almost all surveys.

Therefore, for each category, characteristic, and behaviour, we considered the number of surveys for which at least a small either positive or negative effect was found (see Table 5). It is striking that about one third of all cells or category-behaviour pairs showed both positive and negative effects (marked by '2' in Table 5). This means that a certain category may show *more* of a specific behaviour for some surveys, while *less* for other surveys. For instance, consider the category 15-24 years for the behaviour answering 'won't tell' (WT). Here, this age category showed more 'won't tell' than the other categories combined for three surveys, while less 'won't tell' for one other survey. For a more liberal perspective on consistency, we considered all cases for which a more or less substantial number of surveys referred to an indication for

Table 4. Cliff's Delta (δ) (and its 99% Confidence Interval) for the Behaviours Socially Desirable Responding, Answering Don't Know, Answering Won't Tell, Neutral Responding, and Primacy Responding, for the Applicable Categories of All Characteristics, for the Surveys Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), and Labour Force Survey (LF) when Applicable.

	δ AS	δ FA	δ HE	δ HO	δ IN	δ PE	δ PO	δ RE	δ WO	δ LF
Socially Desirable Responding										
Occupation Other	-0.05 (-.09, -.00)	**0.11** ~ (.05, .16)	**-0.51** # (-.56, -.45)		**-0.36** * (-.42, -.29)	**-0.20** ~ (-.27, -.13)	-0.05 (-.12, .01)	0.01 (-.05, .08)	**-0.36** * (-.40, -.31)	**-0.13** ~ (-.18, -.07)
Sim PC no vs yes	-0.02 (-.06, .02)	**-0.24** ~ (-.29, -.18)	**0.24** ~ (.17, .32)		**0.27** ~ (.20, .34)	**0.11** ~ (.03, .18)	**0.19** ~ (.12, .26)	**-0.13** ~ (-.20, -.05)	0.09 (.03, .15)	**0.12** ~ (.05, .19)
Answering 'Don't Know'										
Gender mal vs. fem	**-0.12** ~ (-.15, -.10)	0.04 (.02, .06)		**-0.14** ~ (-.18, -.11)	**-0.17** ~ (-.21, -.14)		**-0.27** ~ (-.30, -.23)	-0.01 (-.02, .00)	**-0.13** ~ (-.16, -.10)	
Age 15_24 yrs	**0.14** ~ (.10, .17)	**-0.12** ~ (-.16, -.09)		0.08 (.01, .16)	**0.43** # (.39, .48)		**0.28** * (.22, .34)	0.05 (.03, .07)	**0.16** ~ (.12, .21)	
Occupation School	**0.15** ~ (.11, .19)	**-0.13** ~ (-.16, -.09)		0.04 (-.04, .12)	**0.43** # (.38, .48)		**0.25** ~ (.19, .32)	0.04 (.02, .07)	**0.20** ~ (.16, .25)	
Income No income	**0.13** ~ (.09, .18)	-0.06 (-.10, -.03)		**0.19** ~ (.11, .27)	**0.27** ~ (.21, .33)		**0.19** ~ (.13, .25)	0.04 (.02, .06)	**0.33** * (.29, .37)	
Income 2001_3000	**-0.12** ~ (-.15, -.09)	-0.02 (-.04, .01)		**-0.14** ~ (-.18, -.10)	**-0.28** * (-.32, -.24)		**-0.25** ~ (-.29, -.21)	-0.02 (-.03, -.01)	**-0.22** ~ (-.26, -.19)	
Income 3001_more	**-0.19** ~ (-.24, -.14)	0.01 (-.04, .06)		**-0.17** ~ (-.24, -.10)	**-0.29** * (-.36, -.21)		**-0.31** * (-.37, -.25)	-0.03 (-.04, -.01)	**-0.20** ~ (-.26, -.15)	
Income DK	**0.15** ~ (.02, .28)	-0.02 (-.10, .07)		**0.25** ~ (.06, .44)	**0.33** * (.20, .47)		**0.23** ~ (.09, .36)	0.05 (-.01, .12)	**0.12** ~ (.01, .22)	
Answering 'Won't Tell'										
Age 15_24 yrs	0.06 (.02, .09)			**0.16** ~ (.10, .21)	**0.24** ~ (.20, .28)				**0.12** ~ (.07, .16)	**-0.14** ~ (-.24, -.04)
Occupation School	0.02 (-.01, .06)			**0.19** ~ (.12, .26)	**0.23** ~ (.19, .28)				**0.15** ~ (.10, .20)	**-0.16** ~ (-.26, -.04)
Income WT	**0.39** * (.32, .46)			**0.42** * (.32, .52)	**0.49** # (.41, .58)				**0.21** ~ (.14, .28)	**0.13** ~ (.05, .22)
Neutral Responding										
Income 3001_more		0.01 (-.05, .07)			**-0.12** ~ (-.18, -.05)	**-0.17** ~ (-.26, -.07)	**-0.16** ~ (-.23, -.09)		**-0.15** ~ (-.21, -.09)	
Primacy Responding										
Occupation School		**-0.39** * (-.43, -.34)	**-0.11** ~ (-.15, -.07)		**-0.33** * (-.38, -.28)	**-0.18** ~ (-.24, -.11)	**-0.11** ~ (-.17, -.05)	-0.09 (-.14, -.04)	**-0.11** ~ (-.15, -.06)	-0.00 (-.10, .10)
Income 2001_3000		-0.02 (-.06, .02)	**0.15** ~ (.12, .18)		**0.16** ~ (.12, .20)	**0.14** ~ (.09, .19)	0.04 (-.01, .08)	**0.15** ~ (.11, .20)	**0.13** ~ (.09, .17)	**-0.14** ~ (-.18, -.10)
Income 3001_more		-0.01 (-.07, .06)	**0.11** ~ (.06, .16)		**0.26** ~ (.19, .33)	**0.18** ~ (.10, .26)	0.02 (-.05, .09)	**0.18** ~ (.11, .24)	**0.20** ~ (.14, .26)	**-0.11** ~ (-.17, .-04)
Origin 2nd Nonw		**-0.26** ~ (-.38, -.13)	**-0.12** ~ (-.24, .00)		**-0.25** ~ (-.40, -.10)	**-0.17** ~ (-.35, .01)	-0.05 (-.20, .10)	**-0.16** ~ (-.28, -.04)	-0.02 (-.13, .09)	**-0.12** ~ (-.31, .06)
Sim PC no vs. yes		**-0.13** ~ (-.19, -.07)	**0.11** ~ (.05, .18)		**0.17** ~ (.11, .24)	**0.11** ~ (.03, .18)	0.01 (-.07, .09)	**0.14** ~ (.07, .20)	**0.22** ~ (.15, .28)	-0.09 (-.16, -.02)

~ → small effect; * → medium effect; # → large effect

consistency. From here, we show the consistency results per behaviour separately. The results that link back to our hypotheses are displayed in italics.

Answering 'don't know'

The most effects were found for the behaviour answering 'don't know' (DK). Respondents who are female, young (15-24 years and school-going), have only followed primary education, have

Table 5. The Categories with either at Least Two Positive *or* Two Negative Effect Sizes Receiving a '1' (Unidirectional Results) and the Categories with at Least One Positive *and* One Negative Effect Size Receiving a '2' (Contrasting Results) for the Behaviours Answering Won't Tell (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Answering Don't Know (DK), Straightlining (ST), Primacy Responding (PR), Recency Responding (RE), Avoiding Follow-Up Questions (FQ), and Socially Desirable Responding (SD).

		5 surveys				7 surveys		8 surveys		9 surveys	
		WT	AC	NE	EX	DK	ST	PR	RE	FQ	SD
Ge	m vs f					1	2	2	2		2
Ag	15-24	2				2	2	1	2	2	2
	25-34				1	2	2	2	2		2
	35-44			2	1	1	1	2		2	2
	45-54					1		1		1	
	55-64					2				2	2
	65+	2		1	1	2	1	2	1	2	2
Ed	primary		1	1		1		2	1	2	2
	vmbo			1		1		2	2	1	2
	havwo							2			2
	mbo										
	hbo		1	1		1		2			1
	wo			1	2	1	1	2	2	1	2
Do	single							1	1		2
	cohab					2	1	2	2	2	2
	coh_ch			2	1	2	1	2		2	2
	sing_ch		1			2		1	2		2
	other							1			2
Oc	paid			2	1	1	2	2	2	2	1
	famauto					2	1		2	1	
	seek							2	2	2	2
	other		1				1	1		2	2
	school	2		2		2	2	1	2	2	2
	house		1	1	1	1		2	1	1	2
	retired	2			1	2	1	2	1	2	2
In	no_inc	1	1			1		2	2	2	2
	-1000					1		2			1
	-2000										
	-3000	1		1		1	1	2	2	1	2
	3001+	1	1	1		1	1	2	2	1	2
	DK	1	1			1	1			1	
	WT	1				1	1			2	
Or	dutch		2			1		1			1
	1_west					1		2	2	2	2
	1_now	1	2			1	1	1	1	2	2
	2_west										
	2_now		2	2		1		1	2	2	2
Si	n vs y		1		1	1	1	2	1	2	2

The empty cells refer to either no effects, or one positive effect, or one negative effect.

housekeeping as their primary occupation, have either no or a lower income (less than 1000 EUR), are from the first generation of non-western immigrants, or filled out 'don't know' or 'won't tell' for the background characteristic 'income', showed more 'don't know'-answers for a substantial amount of surveys. Respondents who are male, higher educated (followed HBO or WO), or have a higher income (2001 EUR or more), showed less 'don't know'-answers. For the characteristic 'income', the overall results are shown in Graph 1 and the consistency results are shown in Graphs 2 through 8 in Figure 3. For the surveys Assets, Housing, Income, Politics, and Work, respondents without an income or who filled out 'don't know' for the background characteristic 'income' showed more 'don't know'-answers (see the blue and red lines respectively for these surveys). Respondents with a higher income (2001 EUR or more) showed less 'don't know'-answers for these surveys (see the orange and pink lines). *Although not across all surveys, the relation between being female, lower education, and a lower income, and saying 'don't know', is in line with our expectations, while the relation between being young and saying 'don't know' is not.*

Answering 'won't tell'

Concerning answering 'won't tell', respondents who are young (15-24 years and school-going), are from the first generation of non-western immigrants, or filled out 'don't know' or 'won't tell' for the background characteristic 'income', showed more 'won't tell'-answers for a substantial number of surveys. Respondents with a higher income (3001 EUR or more) showed less 'won't tell'-answers*, which is in contrast to our expectation. We did not find a consistent relation between lower income and saying 'won't tell'.*

Neutral and extreme responding

Respondents who only followed primary education showed more neutral responding than the other groups for multiple surveys, while respondents with a WO background or a higher income (3001 EUR or more) showed less neutral responding. *We did not find a consistent relation between being young or a non-Western immigrant and giving neutral responses.* Respondents whose main occupation was housekeeping showed more extreme responding for multiple surveys. *We did not find a consistent relation between being male or older, lower education, lower income, or not having paid work and giving extreme responses.*

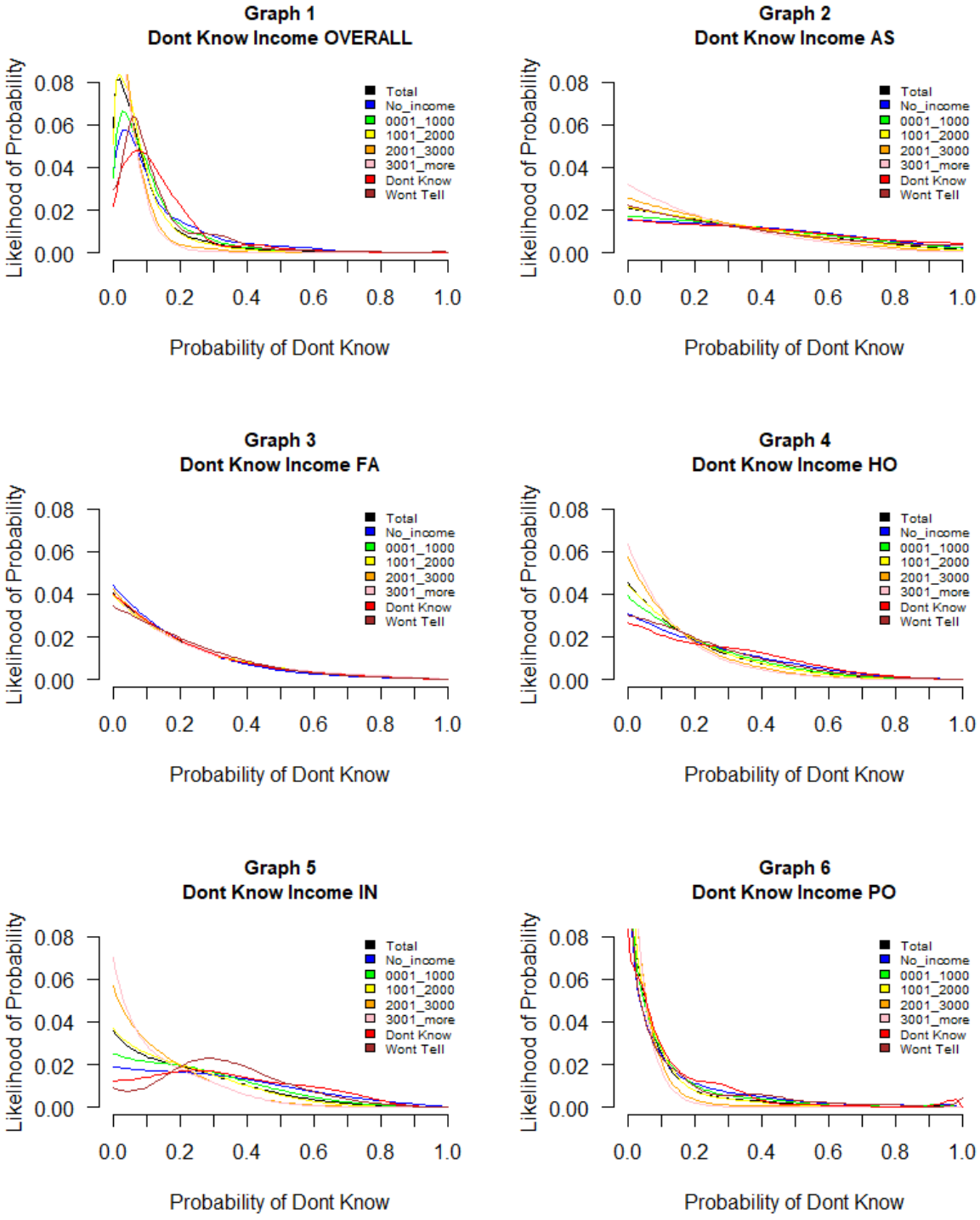

Figure 3. Consistently Relatively *Less* 'Don't Know'-Answers for Respondents without an Income (blue) and who filled out 'Don't Know' as a Background Characteristic (red), and Consistently Relatively *More* 'Don't Know'-Answers for Respondents with an Income between 2001 and 3000 EUR (orange) and above 3000 EUR (pink) for 5 Out of 7 Surveys: Assets (AS, Graph 2), Housing (HO, Graph 4), Income (IN, Graph 5), Politics and Values (PO, Graph 6), and Work and Schooling (WO, Graph 8).

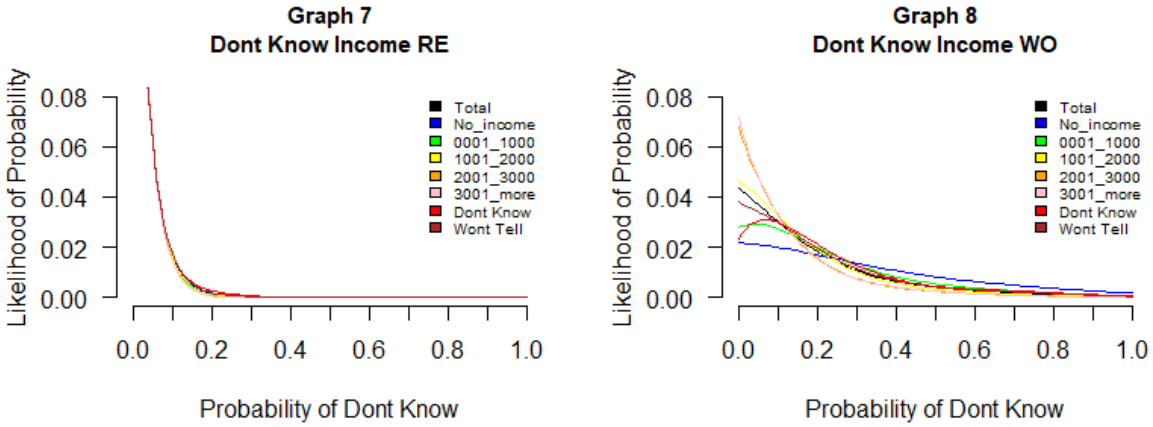

Figure 3 (continued). Consistently Relatively *Less* 'Don't Know'-Answers for Respondents without an Income (blue) and who filled out 'Don't Know' as a Background Characteristic (red), and Consistently Relatively *More* 'Don't Know'-Answers for Respondents with an Income between 2001 and 3000 EUR (orange) and above 3000 EUR (pink) for 5 Out of 7 Surveys: Assets (AS, Graph 2), Housing (HO, Graph 4), Income (IN, Graph 5), Politics and Values (PO, Graph 6), and Work and Schooling (WO, Graph 8).

Straightlining

Concerning straightlining, respondents who are 35-44 years of age, live together with both partner and children, have a higher income (3001 EUR or more), or filled out 'don't know' or 'won't tell' for the background characteristic 'income', showed more straightlining for multiple surveys. Respondents who are retired and/or 65 years or older showed less straightlining. *We did not find a consistent relation between being male or young, lower education or using one's own device and showing straightlining.*

Socially desirable responding

Concerning socially desirable responding, respondents who only followed primary education, whose main occupation was housekeeping or 'other' (respondents who are exempted from work seeking, have a work disability, or are doing unpaid or voluntary work), or who did not receive an income, showed less of the behaviour for multiple surveys. *This is in contrast to our expectation that lower income or not having paid work would be related to giving socially desirable responses.* Respondents who own a computer showed more socially desirable responding than respondents who were provided a computer from the panel. *We did not find a consistent relation between being female or a non-Western immigrant and giving socially desirable responses.*

Acquiescence

Respondents who only followed primary education showed less acquiescence for multiple surveys than the other education groups, *which is in contrast to what we expected. We also did not find a consistent relation between being female or a non-Western immigrant, lower income or not having paid work and giving acquiescent responses.*

Primacy responding

Respondents who are young (15-24 years and school-going), who only followed primary education, who are from both the first and second generation of non-western immigrants, or who are single and live together with children, showed less primacy responding for a substantial amount of surveys. Respondents with a WO background or a higher income (2001 EUR or more) showed more primacy responding. Respondents who own a computer showed more primacy responding than respondents who were provided a computer from the panel.

Avoiding follow-up questions

Respondents aged 65 or older, who have housekeeping as their primary occupation, or who are from the first generation of non-western immigrants, showed more avoidance of follow-up questions. Respondents with an own or family business showed less avoidance of follow-up questions for multiple surveys.

In sum, the results refer to an absence of behaviour consistency across all or almost all surveys. We conclude that respondents' answer behaviour may be more influenced by the survey and its topic and items than by the characteristics of the respondent. Even when considering our expectations and the results more liberally, it is evident that many of our expectations were still not met and that several of our expectations were contrasted. We also found various outcomes across a substantial number of surveys that we did not explicitly expect. We close with a discussion in the following section.

6. CONCLUSION AND DISCUSSION

In this study, we investigated to what extent respondent characteristics are associated with a high occurrence of undesirable answer behaviour *consistently* across different surveys. The occurrence of answer behaviour is indicated by varying uncertainty, as every respondent filled out a different number of the items that were applicable to each behaviour. To take this varying uncertainty into account, we used an adaptation of the robust effect size statistic Cliff's Delta

to compare groups of respondents in the form of density distributions or *behaviour profiles*. The behaviour of respondents from a specific category (for instance 'male' for the characteristic 'gender' or 'Dutch' for the characteristic 'origin') was compared to the behaviour of respondents from the other categories of the concerned characteristic together. For our study, we included the answer behaviours 'avoiding follow-up questions', 'socially desirable responding', 'answering don't know', 'answering won't tell', 'acquiescence', 'neutral responding', 'extreme responding', 'primacy responding', 'recency responding', and 'straightlining'. We included the respondent characteristics 'gender', 'age', 'education', 'domestic situation', 'primary occupation', 'income', 'origin', and 'using a borrowed computer'.

There is no consistency present for any of the characteristic categories for any of the behaviours. However, specific forms of satisficing across surveys seem evident for certain groups of respondents in particular. Relatively more 'don't know'- and 'won't tell'-answers, and less primacy responding, is shown for respondents who are young, who go to school, and who are from the first generation non-western immigrants. On the contrary, less non-substantive answers are associated with higher educated respondents with a relatively high income. More straightlining was evident for respondents living together with both a partner and children, having a relatively high income, while less straightlining was evident for respondents who are relatively older and retired. Finally, respondents who did not know or refused to state their income as a background characteristic also showed relatively more 'won't tell'-answers and forms of strong satisficing -relatively more 'don't know'-answers and straightlining- for multiple surveys than respondents who did state their income. However, there is no category for any characteristic that showed certain answer behaviour consistently across *all or almost all* surveys.

Our results seem to go beyond the absence of behaviour consistency across surveys. As the more surveys were applicable to a behaviour, the more contrasting outcomes were found; many categories were associated with relatively *more* of a behaviour for some surveys, while relatively *less* of that behaviour for other surveys. Most contrasting results were found for giving socially desirable responses, but contrasting results appeared throughout for all categories, characteristics, surveys, and behaviours. In fact, more evidence was found for contrasting behaviour than for consistent behaviour across surveys. This evidence is not compatible to our initial theory that respondents will show consistency for at least some of the

characteristics and behaviours across most or all surveys. *Overall, we conclude that the occurrence of undesirable answer behaviour cannot unambiguously be attributed to the respondent, but may be substantially determined by the characteristics of the survey and its items instead.*

In our study, we did not focus on the answer behaviour of *identified* individual or groups of respondents. For all characteristic's categories, each respondent was considered for every applicable survey that he or she participated in. Thus, for the consistency analysis of a category, some respondents were considered for only one or two surveys, while other respondents were considered for all or almost all surveys. This means that we can neither attribute survey answer behaviour to individual or groups of identified respondents, nor compare them between surveys for the same category and behaviour. At the same time, considering respondents multiple times, for each applicable survey, was the strength of our study. Taking into account every respondent who fell into a characteristic's category for every applicable survey resulted in large groups of respondents per survey. We compared behaviour profiles of large respondent groups for a single category to behaviour profiles of large respondent groups for the remaining categories. This means that we focussed on the association between the respondent's *characteristics* and potentially consistent answer behaviour across surveys. In other words, we did not attribute answer behaviour to the concerned group of respondents, but to the characteristic's category in which they were placed.

We used these comparisons between a category and the remaining categories of a characteristic together to answer our consistency research question. For this purpose, we used an adaptation of Cliff's Delta, for which we have shown it converges quickly towards Cliff's original Delta as the number of eligible items goes up (see Appendix F). For our study, this robust effect size measure was both useful because of its many advantages regarding our data (see Appendix E) and sufficient for comparing two expected group values representing a specific category versus the remaining categories of a characteristic. In case of expected group differences, follow-up research may zoom in on these differences to reveal characteristics of subgroups showing relatively more of certain behaviour for specific surveys and their topics and items. In particular, we would be interested in single groups with higher expected values than the remaining groups for a characteristic and in the respondents who are located to the right of its distribution.

Other follow-up research may focus on the relation between *item characteristics* and answer behaviour. Just as respondent characteristics, item characteristics have their influence on data quality and may be associated to measurement error. See Bais et al., (2017), Beukenhorst et al., (2014), Campanelli et al., (2011), Gallhofer et al., (2007), and Saris & Gallhofer (2007) for overviews of item characteristics and their relation to measurement error. Items can be coded on the presence or absence of characteristics like for instance question sensitivity. Hence, items that are coded as sensitive could be compared to items that are not coded as sensitive on the occurrence of potentially undesirable answer behaviour. In this way, the presence of certain item characteristics may be connected to answer behaviour for the items of whole surveys specifically or across the items of multiple surveys more generally. Based on these associations, questionnaire profiles may be constructed that give an instant overview of the present item characteristics and their relation to answer behaviour and measurement error.

Acknowledgment

We would like to thank Joost van der Neut for his contribution to this paper.

REFERENCES

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research, 42*, 957-970. doi: http://dx.doi.org/10.1016/j.ssresearch.2013.01.002

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods & Research, 20*, 139–181.

Andersen, H., & Mayerl, J. (2019). Responding to Socially Desirable and Undesirable Topics: Different Types of Response Behaviour? *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *13*(1), 7-35. https://doi.org/10.12758/mda.2018.06

Andrews F M. & Herzog A. R. (1986). The Quality of Survey Data as Related to Age of Respondent. *Journal of the American Statistical Association, 81*(394), 403-410.

Antoni, M., Bela, D., & Vicari, B. (2019). Validating Earnings in the German National Panel Study: Determinants of Measurement Accuracy of Survey Questions on Earnings. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *13*(1), 59-90. https://doi.org/10.12758/mda.2018.08

Bachman, J. G., & O'Malley, P. M. (1984a). Black–white differences in self-esteem: Are they affected by response styles? *American Journal of Sociology, 90,* 624–639.

Bachman, J. G., & O'Malley, P. M. (1984b). Yea-Saying, Nay-Saying, and Going to : Black–White Differences in Response Styles. *Public Opinion Quarterly*, *48,* 491–509.

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M., & Vis, C. (2017). Can Survey Item Characteristics Relevant to Measurement Error be coded Reliably? A Case Study on Eleven Dutch General Population Surveys. *Sociological Methods & Research, 48*(2), 1-33. doi.org/10.1177/0049124117729692

Bais, F., Schouten, B., & Toepoel, V. (submitted). Investigating response patterns across surveys: Do respondents show consistency in undesirable answer behaviour over multiple surveys?

Bakker, B. F. M. (2012). Estimating the validity of administrative variables, *Statistica Neerlandica, 66*, 8–17.

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

Beatty, P., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse.* First Edition (pp. 71–86). New York: Wiley.

Bernardi, R. A. (2006). Associations between Hofstedes cultural constructs and social desirability response bias. *Journal of Business Ethics, 65*(1), 43-53. DOI 10.1007/s10551-005-5353-0

Beukenhorst, D., Buelens, B., Engelen, F., Van der Laan, J., Meertens, V., & Schouten, B. (2014). The impact of Survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey. CBS Discussion paper 2014-16. Statistics Netherlands, The Hague.

Billiet, J. B., & McClendon, J. M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7*, 608–628. doi: http://dx.doi.org/10.1207/S15328007SEM0704_5

Binswanger, J., Schunk, D., & Toepoel, V. (2006). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly, 77*, 783–797.

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: the pressure to answer survey questions. *Public Opinion Quarterly, 50*, 240–250.

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly, 47*, 528-546.

Bosley, J., Dashen, M., & Fox, J. E. (1999). When should we ask follow-up questions about items in lists? In Proceedings of the Survey Research Methods Section of the American Statistical Association, 749-754.

Bradburn, N., Sudman, S., Blair, E., & Stocking, C. (1978). Question Threat and Response Bias. *Public Opinion Quarterly, 42*, 221-234.

Butler R. J., & McDonald, J. B. (1987). Interdistributional income inequality. *Journal of Business and Economic Statistics,* 13-18.

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

Chen, C., Lee, S., & Stevenson, H. W. (1995), Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students. *Psychological Science*, *6*, 170–175. doi: 10.1111/j.1467-9280.1995.tb00327.x

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187-212. https://doi.org/10.1177/0022022100031002003

Chun, K. T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology*, *5*, 465-480. https://doi.org/10.1177/002202217400500407

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*, 331-350.

Cliff, N. (1996b). *Ordinal Methods for Behavioral Data Analysis*. New Jersey: Lawrence Erlbaum Associates.

DeMaio, T. J. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena.* Volume 2 (pp. 257–281). New York: Russell Sage Foundation.

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

Díaz de Rada, V., & Domínguez, J. A. (2015). The quality of responses to grid questions as used in Web questionnaires (compared with paper questionnaires). *International Journal of Social Research Methodology*, *18*, 337–348. doi: http://dx.doi.org/10.1080/13645579.2014.895289

Dolnicar, S., & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review, 31*(2), 127-143. http://ro.uow.edu.au/commpapers/251

Duan, N., Alegria, M., Canino, G., McGuire, T. G., & Takeuchi, D. (2007). Survey Conditioning in Self-reported Mental Health Service Use: Randomized Comparison of Instrument Formats. *Health Services Research, 42*, 890-907. DOI: 10.1111/j.1475-6773.2006.00618.x

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., & Presser, S. (2014). Assessing the Mechanisms of Misreporting to Filter Questions in Surveys. *Public Opinion Quarterly, 78*, 721-733. doi:10.1093/poq/nfu030

Fricker S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly, 69*, 370-392. doi: 10.1093/poq/nfi027

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-Tracking Data New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding. *Public Opinion Quarterly, 72*, 892–913.

Gallhofer, I. N., Scherpenzeel, A., & Saris, W. E. (2007). The code-book for the SQP program, available at (http://www.europeansocialsurvey.org/methodology/sqpcoding.html).

Goedhart, J. (2016). Calculation of a distribution free estimate of effect size and confidence intervals using VBA/Excel. doi: http://dx.doi.org/10.1101/073999.

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*, 328–351. http://www.jstor.org/stable/2749156

Harzing, A. W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross-Cultural Management*, *6*, 243-266. DOI: 10.1177/1470595806066332

He, J., Van de Vijver, F. J. R., Espinosa, A. D., & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14*, 306-322. DOI: 10.1177/1470595814541424

He, J., & Van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55*, 794–800. http://dx.doi.org/10.1016/j.paid.2013.06.017

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

Hess, M. R., & Kromrey, J. D. (2004). Robust confidence intervals for effect sizes: a comparative study of Cohen's d and Cliff's Delta under non-normality and heterogeneous variances. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*, 79–125.

Hox, J. J., De Leeuw, E., & Kreft, I. G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 439-461). New York: Wiley.

Hui, C. H., & Triandis, H. C. (1989), Effects of Culture and Response Format on Extreme Style. *Journal of Cross-Cultural Psychology*, *20*, 296–309.

Jann, B., Krumpal, I., & Wolter, F. (2019). Editorial: Social Desirability Bias in Surveys – Collecting and Analyzing Sensitive Data. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *13*(1), 3-6.

Jensen, P. S., Watanabe, H. K., & Richters, J. E. (1999). Who's up first? Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology, 27*, 439–445.

Johnson, T., & Van de Vijver, F. J. R. (2002). Social desirability in crosscultural research. In J. Harness, F. J. R. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 193–202). New York: Wiley.

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician, 29*, 65–78. http://www.jstor.org/stable/2987495

Kaminska, O., McCutcheon, A., & Billiet, J. (2010). Satisficing Among Reluctant Respondents in a Cross-national Context. *Public Opinion Quarterly, 74*, 880–906. doi: 10.1093/poq/nfq062

Kellogg, R. T. (2007). *Fundamentals of Cognitive Psychology*. Los Angeles: SAGE Publications.

Kessler, R. C., Wittchen, H. U., Abelson, J. M., McGonagle, K., Schwarz, N., Kendler, K. S., Knäuper, B., & Zhao, S. (1998). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *International Journal of Methods in Psychiatric Research, 7*, 33-55.

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal Style? *Quality & Quantity, 47*, 193-211. doi: 10.1007/s11135-011-9511-4

Knäuper, B. (1998). Filter questions and question interpretation: presuppositions at work. *Public Opinion Quarterly, 62*, 70–78.

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The Effects of Asking Filter Questions in Interleafed versus Grouped Format. *Sociological Methods and Research, 40*, 80-104. DOI: 10.1177/0049124110392342

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly, 72*, 847-865. doi:10.1093/poq/nfn063

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

Krosnick, J. A. (1992). The impact of cognitive sophistication and attitude importance on response order effects and question order effects. In N. Schwarz & S. Sudman (Eds), *Order effects in social and psychological research* (pp. 203–218). New York: Springer.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51*, 201–219.

Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology, 57*, 416-425.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). New York: John Wiley & Sons, Inc.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of 'no opinion' response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66*, 371–403.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in Surveys: Initial Evidence. *New Directions for Evaluation, 70*, 29–44.

Lavrakas, P. J., & the AAPOR Cell Phone Task Force. (2010). New Considerations for Survey Researchers When Planning and Conducting RDD Telephone Surveys in the U.S. with Respondents Reached via Cell Phone Numbers, available at http://www.aapor.org

Lavrakas, P. J., Tompson, T. N., & Benford, R. (2010). Investigating Data Quality in Cell Phone Surveying. Paper presented at the Annual American Association for Public Opinion Research Conference, Chicago.

Leigh, J. H., & Martin, C. R. (1987). Don't Know Item Nonresponse in a Telephone Survey: Effects of Question Form and Respondent Characteristics. *Journal of Marketing Research, 24*, 418–424.

Lucas, C. P., Fisher, P., Piacentini, J., Zhang, H., Jensen, P. S., Shaffer, D., Dulcan, M., Schwab-Stone, M., Regier, D., & Canino, G. (1999). Features of Interview Questions Associated With Attenuation of Symptom Reports. *Journal of Abnormal Child Psychology, 27*(6), 429-437.

Lynn, P., & Kaminska, O. (2012). The Impact of Mobile Phones on Survey Measurement Error. *Public Opinion Quarterly, 77*, 586–605. doi:10.1093/poq/nfs046

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme Response Style and Acquiescence Among Hispanics. *Journal of Cross-Cultural Psychology, 23*, 498–509.

Marshall, R., & Lee, C. (1998), A cross-cultural, between-gender study of extreme response Style. In B. G. Englis & A. Olofsson (Eds.), *European Advances in Consumer Research*. Volume 3 (pp. 90-95). Provo, UT: Association for Consumer Research. http://acrwebsite.org/volumes/11158/volumes/e03/E-03

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly, 67*, 205–211.

McClendon, M. J. (1991). Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods and Research, 20*, 60-103.

Medway, R., & Tourangeau, R. (2015). Response quality in telephone surveys. Do pre-paid cash incentives make a difference? *Public Opinion Quarterly, 79*, 524–543. doi:10.1093/poq/nfv011

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010

Messick, S. J. (1966). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Eds.), *Response set in personality assessment* (pp. 115–145). Chicago: Aldine.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*, 58–88.

Olson, K., Smyth, J. D., & Ganshert, A. (2018). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, *0*, 1-34. doi: 10.1093/jssam/smy006

Olson, K., & Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology, 3*, 361–396. doi: 10.1093/jssam/smv021

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). Middle alternatives, acquiescence, and the quality of questionnaire data. Retrieved, October 1, 2009, from http://harrisschool.uchicago.edu/About/publications.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum.

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

Redline, C., & Dillman, D. A. (2002). The Influence of Alternative Visual Designs on Respondent's Performance with Branching Instructions in Self-administered Questionnaires. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse*. New York: Wiley.

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. ESRC National Centre for Research Methods, NCRM Methods Review Paper 008, UK. Retrieved July 2019 from http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf

Roberts C., & Jäckle, A. (2012). Causes of Mode Effects: Separating Out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys. ISER Working Paper, 2012-27. Colchester: University of Essex.

Rousselet, G. A., Foxe, J. J., & Bolam, J. P. (2016). A few simple steps to improve the description of group results in neuroscience. *European Journal of Neuroscience, 44*, 2647–2651. doi:10.1111/ejn.13400

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience,* 1-27. doi: http://dx.doi.org/10.1101/121079

Saris, W. E., & Gallhofer, I. N. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods, 1*(1), 29-43. http://w4.ub.uni-konstanz.de/srm

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61–79. http://www.surveymethods.org

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88. doi: 10.1146/annurev.soc.29.110702.110112

Scholtus, S., Bakker, B. F. M., & Van Delden, A. (2015). Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. CBS Discussion paper 2015-17. Statistics Netherlands, The Hague.

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web survey panels over time. *Survey Research Methods, 9*, 125–137. doi:10.18148/srm/2015.v9i2.6128

Schouten, B., & Calinescu, M. (2013). Paradata as Input to Monitoring Representativeness and Measurement Profiles: A Case Study of the Dutch Labour Force Survey. In F. Kreuter (Eds.), *Improving Surveys With Paradata: Analytic Uses of Process Information* (pp. 231-258). Hoboken, NJ: Wiley.

Schräpler, J. P. (2004). Response behavior in panel studies: A case study for income nonresponse by means of the German socio-economic panel (SOEP). *Sociological methods and research, 33*, 118–156. DOI: 10.1177/0049124103262689

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context*. New York: Academic Press.

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The Impact of Administrative Mode on Response Effects in Survey Measurement. *Applied Cognitive Psychology, 5*, 193–212.

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item Nonresponse: Distinguishing Between Don't Know and Refuse. *International Journal of Public Opinion Research, 14*, 193-201.

Si, S. X., & Cullen, J. B. (1998). Response Categories and Potential Cultural Bias: Effects of an Explicit Middle Point in Cross-Cultural Surveys. *International Journal of Organizational Analysis, 6*, 218-230.

Smith, P. B. (2004). Acquiescent response bias as an aspect of crosscultural communication style. *Journal of Cross-Cultural Psychology, 35*(1), 50-61. https://doi.org/10.1177/0022022103260380

Stening, B. W., & Everett, J. E. (1984). Response Styles in a Cross-Cultural Managerial Study. *Journal of Social Psychology, 122*, 151–156. http://dx.doi.org/10.1080/00224545.1984.9713475

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods, 1*, 121-138. http://www.surveymethods.org

Stricker, L. J. (1963). Acquiescence and social desirability response styles, item characteristics, and conformity. *Psychological Reports, 12*, 319–341.

Tarnai, J., & Dillman, D. A. (1992). Questionnaire Context as a Source of Response Differences in Mail versus Telephone Surveys. In N. Schwarz & S. Sudman (Eds.), *Context Effects in Social and Psychological Research*. New York: Springer Verlag.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*, 859-883. DOI: 10.1037/0033-2909.133.5.859

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*, 346-360.

Vargha, A., & Delaney, H. D. (2000). A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational Behavioral Statistics, 25*(2), 101–132. doi: http://dx.doi.org/10.2307/1165329

Vis-Visschers, R., Arends-Tóth, J., Giesen, D., & Meertens, V. (2008). Het aanbieden van 'weet niet' en toelichtingen in een webvragenlijst. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.

Watkins, D., & Cheung, S. (1995). Culture, Gender, and Response Bias: An Analysis of Responses to the Self-Description Questionnaire. *Journal of Cross-Cultural Psychology, 26*, 490–504. https://doi.org/10.1177/0022022195265003

Yan T., & Tourangeau, R. (2008). Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology, 22*, 51–68.

Ye, C., Fulton, J., & Tourangeau, R. (2011). More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice. *Public Opinion Quarterly, 75*(2), 349–365. doi: 10.1093/poq/nfr009

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: comparisons of Japanese and American college students. *Journal of Social Psychology*, *71*, 3-10. http://dx.doi.org/10.1080/00224545.1967.9919760

Zhang, C. (2013). Satisficing in web surveys: implications for data quality and strategies for reduction (Ph.D.) Ann Arbor, MI: University of Michigan. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97990

Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: the tendency to answer very fast and its association with straightlining. *Survey Research Methods, 8*, 127–135.

APPENDIX A

Table 6. Respondent Characteristics, Their Categories, and Relevant Literature.

Respondent characteristics	*Categories of the respondent characteristics in this study*	*Relevant literature*
Gender	1. male 2. female	Bernardi (2006); Hox et al. (1991); Marshall & Lee (1998); O'Muircheartaigh et al. (2000); Pickery & Loosveldt (1998); Zhang & Conrad (2013)
Age	1. 15-24 years old 2. 25-34 years old 3. 35-44 years old 4. 45-54 years old 5. 55-64 years old 6. 65 years and older	Alwin & Krosnick (1991); Andrews & Herzog (1986); Greenleaf (1992); He et al. (2014); Hox et al. (1991); Kieruj & Moors (2013); Meisenberg & Williams (2008); O'Muircheartaigh et al. (2000); Pikcery & Loosveldt (1998); Schonlau & Toepoel (2015); Zhang & Conrad (2013)
Education	1. primary school 2. vmbo: intermediate secondary education 3. havo/vwo: higher secondary education 4. mbo: intermediate vocational education 5. hbo: higher vocational education 6. wo: university	Aichholzer (2013); Alwin & Krosnick (1991); Greenleaf (1992); He et al. (2014); Krosnick (1991); Krosnick & Alwin (1987); Krosnick et al. (2002); Marin et al. (1992); McClendon (1986, 1991); Narayan & Krosnick (1996); O'Muircheartaigh et al. (2000); Pickery & Loosveldt (1998); Schuman & Presser (1981); Zhang & Conrad (2013)
Domestic situation	1. single 2. (un)married co-habitation without children 3. (un)married co-habitation with children 4. single with children 5. other	Alwin & Krosnick (1991); Holbrook et al. (2003); Kellogg (2007); Lavrakas (2010); Lavrakas et al. (2010); Lynn & Kaminska (2012); Olson et al. (2018); Schwarz et al. (1991)
Primary occupation	1. paid employment 2. family business or self-employed 3. job seeker 4. other: exempted from job seeking, work disability, unpaid/voluntary work 5. attends school or is studying 6. takes care of the housekeeping 7. retired	Butler et al. (1987); Lynn & Kaminska (2012); McClendon (1991); Schräpler (2004)
Income	1. no income 2. 1 to 1000 EUR 3. 1001 to 2000 EUR 4. 2001 to 3000 EUR 5. 3001 or more EUR 6. 'don't know' 7. 'won't tell'	Greenleaf (1992); Lynn & Kaminska (2012); McClendon (1991); Schräpler (2004)
Origin	1. Dutch background 2. 1st generation foreign western background 3. 1st generation foreign non-western background 4. 2nd generation foreign western background 5. 2nd generation foreign non-western background	Bachman & O'Malley (1984ab); Baumgartner & Steenkamp (2001); Bernardi (2006); Chen et al. (1995); Chun et al. (1974); Cheung & Rensvold (2000); Dolnicar & Grun (2007); Harzing (2006); He & Van de Vijver (2013); Hui & Triandis (1989); Johnson & Van de Vijver (2003); Marin et al. (1992); Marshall & Lee (1998); Si & Cullen (1998); Smith (2004); Stening & Everett (1984); Van Herk et al. (2004); Watkins & Cheung (1995); Zax & Takahashi (1967)
Received a PC?	1. no 2. yes	Schonlau & Toepoel (2015); Zhang (2013); Zhang & Conrad (2013)

APPENDIX B

Table 7. The Form of Answer Behaviour, the Kind of Items Eligible for the Answer Behaviour, and the Operationalization of the Answer Behaviour.

Answer behaviour and label	*Eligible items*	*Operationalization*
Avoiding follow-up questions (FQ)	All filter items containing at least one answer category factually leading to follow-up questions and at least one answer category not leading to follow-up questions	Number of filter items for which a category *not* leading to follow-up questions was filled out divided by Number of actually filled out eligible filter items
Socially desirable responding (SD)	All items formerly coded as asking for sensitive information, containing at least one answer category coded as possibly being socially desirable and at least one category coded as not being socially desirable	Number of items for which a socially desirable answer was filled out divided by Number of actually filled out eligible sensitive items
Answering don't know (DK)	All items containing a 'don't know' answer category	Number of items for which 'don't know' was filled out divided by Number of actually filled out 'don't know' items
Answering won't tell (WT)	All items containing a 'won't tell' answer category	Number of items for which 'won't tell' was filled out divided by Number of actually filled out 'won't tell' items
Acquiescence (AC)	All more or less subjective (battery) items in the form of an ordinal agree/disagree or yes/no answer scale	Number of items for which an agreeable or affirmative answer was filled out divided by Number of actually filled out 'acquiescent' items
Neutral responding (NE)	All (battery) items with an odd and minimum number of five answer categories on an ordinal scale, containing a neutral middle answer category	Number of (battery) items for which the neutral middle answer category was filled out divided by Number of actually filled out eligible (battery) items
Extreme responding (EX)	All (battery) items with a minimum number of four answer categories on an ordinal scale, containing non-neutral first and last answer categories	Number of (battery) items for which an extreme answer category was filled out divided by Number of actually filled out eligible (battery) items
Primacy responding (PR)	All (battery) items containing at least four response options	Number of (battery) items for which the first or second answer category was filled out divided by Number of actually filled out eligible (battery) items
Recency responding (RE)	All (battery) items containing at least four response options	Number of (battery) items for which one of the last two answer categories was filled out divided by Number of actually filled out eligible (battery) items
Straightlining (ST)	The items of all batteries containing at least 3 items and at least 4 answer categories, only in case all items of the battery were actually filled out	Number of filled out battery items for which the same answer category was filled out for a complete battery divided by Number of actually filled out eligible battery items

APPENDIX C

Table 8. The Number of Items and Batteries per Survey, the Average Number of Items per Battery, and the Proportions of Items for which the Answer Behaviours are Applicable for the Surveys Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), and Labour Force Survey (LF), and in Total (TT).

	AS	FA	HE	HO	IN	PE	PO	RE	WO	LF	TT
Nr. of items	50	409	243	73	286	200	148	71	471	123	2074
Nr. of batteries	-	11	5	-	3	16	12	4	2	-	53
Ave. nr. of items/battery	-	5.5	7.6	-	5.7	11.1	6.0	5.8	12.0	-	7.8
Avoiding FU questions	.62	.33	.12	.27	.48	-	.06	.20	.21	.13	.23
Soc. Des. responding	.20	.12	.62	.01	.25	.30	.51	.42	.19	.32	.28
Answering ‘don’t know’	.52	.08	.01	.33	.47	.02	.45	.49	.11	.01	.18
Answering ‘won’t tell’	.28	-	-	.30	.31	-	.01	-	.04	.81	.12
Acquiescence	-	.03	-	-	.01	.96	.68	.24	.05	.03	.17
Neutral responding	-	.10	-	-	.05	.93	.66	-	.04	-	.17
Extreme responding	-	.13	-	-	.05	.93	.66	-	.06	-	.18
Primacy responding	-	.37	.23	-	.24	.93	.73	.55	.19	.27	.35
Recency responding	-	.37	.23	-	.24	.93	.73	.55	.19	.27	.35
Straightlining	-	.15	.16	-	.06	.89	.49	.32	.05	-	.20

APPENDIX D: RE-CODING SOCIALLY DESIRABLE RESPONDING AND ACQUIESCENCE

Socially Desirable Responding

Eligible items for socially desirable responding were items that were coded as potentially asking for sensitive information by at least one of three coders (see Bais et al., 2017). For these items, all answer categories were initially coded by a student assistant relatively liberally, on which categories may *possibly* refer to as evoking social desirability. This resulted in relatively many socially desirable answer options and relatively high percentages of socially desirable responding for most surveys and respondents. As a consequence, relatively little variability between respondents across surveys was present. Therefore, we re-coded this behaviour *more conservatively* for the current research, meaning that we coded less answer categories as a socially desirable response for a number of surveys and items. In this way, respondents who are clearly sensitive to responding socially desirable may be better distinguished from respondents who are not across surveys. Let us consider the following example:

Example of an item that was re-coded for 'socially desirable responding' from the survey 'Income':

'Can you indicate, on a scale from 0 to 10, how hard or how easy it is for you to live off your income?

0 means that it is very hard to live off your income, 10 means that it is very easy.

very hard very easy

0 1 2 3 4 5 6 7 8 9 10'

In Bais et al., (2017), the answer options from 5 through 10 were coded as socially desirable options. The idea was that it is socially desirable to state that it is relatively easy to live off one's income (options 6 through 10). The neutral middle option 5 could be used as a socially desirable option in case respondents actually found it hard to live off their income but were reluctant to admit. For the current study, we only considered the answer options 8 through 10 as socially desirable options.

Acquiescence:

For acquiescence, the answer categories of all items were initially evaluated by a student assistant on whether they expressed an extent of agreeableness or affirmativeness (see Medway & Tourangeau, 2015). Both battery and non-battery items were considered and also subjective

variants of the typical answer option 'agree', like 'satisfied', 'applicable', and 'yes', were considered for acquiescence. All categories were coded relatively conservatively; a category was coded as agreeable only in case the category was obviously agreeable. This resulted in relatively few strong acquiescent answer options and relatively low percentages of acquiescence for most surveys and respondents. As a consequence, respondents who consistently acquiesce but not too an extreme extent, may have remained undetected. Therefore, we re-coded this behaviour *more liberally* for the current research, meaning that we coded more answer categories as an acquiescent response for a number of surveys and items. In this way, respondents who acquiesce to only a certain extent may also be detected and distinguished from respondents who do not acquiesce across surveys. Let us consider the following example:

Example of an item that was re-coded for 'acquiescence' from the survey 'Personality':
'I really enjoy responding to questionnaires through the mail or Internet.
totally disagree totally agree
1 2 3 4 5 6 7'

In Bais et al., (2017), only the answer option 7 was coded as a clear acquiescent option. For the current study, we considered the answer options 5 through 7 as acquiescent options.

APPENDIX E: AN ADAPTATION OF CLIFF'S DELTA

The Original Cliff's Delta for Data Observations

Cliff's Delta calculates the probability that a random data observation X_a from the one group A is larger than a random data observation X_b from the other group B, minus the reverse probability (Hess & Kromrey, 2004; Rousselet et al., 2016; Rousselet et al., 2017):

$$\delta = P(X_a > X_b) - P(X_a < X_b). \tag{5}$$

To estimate Cliff's Delta, each observation from group one is compared to each observation from group two using the sign function:

$$\text{sign}(X_a - X_b), \tag{6}$$

which gives 1 if $X_a > X_b$, 0 if $X_a = X_b$, and -1 if $X_a < X_b$, and where the total number of comparisons is the product of both group sample sizes. After assigning +1 when the observation from group one is larger than the observation from group two, -1 when the observation from group one is smaller than the observation from group two, and 0 when both observations are equally large, the -1's, 0's, and 1's are summed and divided by the product of both group sizes:

$$\delta = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\delta_{ab}}{AB} = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\text{sign}(X_a - X_b)}{AB}, \tag{7}$$

where A and B are the sizes of group A and group B respectively. Calculating Cliff's Delta may be considered a dominance analysis, referring to the extent to which the one data distribution overlaps the other (Hess & Kromrey, 2004). The smaller the overlap between two data distributions, the larger the dominance and the more difference between the two groups. A Cliff's Delta of -1 or 1 indicates absence of overlap between two groups and a Cliff's Delta of 0 refers to group equivalence (Hess & Kromrey, 2004).

Adapting Cliff's Delta for Density Distributions

Consider Cliff's original Delta for which each specific observation from sample one is compared to each specific observation from sample two. This means that when an observation with a certain value from the first sample occurs three times, this observation value is compared to all observations from the second sample three times as well. Therefore, we may regard both

observations for each such comparison on its own as having a 'frequency' or 'weight' of 1. Surely, the two data observations of all possible pairs of observations from two samples are compared exactly once. Implementing these frequencies into (7) gives

$$\delta = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\text{sign}(X_a - X_b)(W_a W_b)}{\sum_{a=1}^{A}\sum_{b=1}^{B}(W_a W_b)}, \tag{8}$$

where W_a and W_b are the frequencies of the data observations from group A and group B respectively. In the case of Cliff's original Delta, these frequencies are all 1 by definition, making this formula identical to (7).

When extrapolating this method of analysis for data observations to the analysis of our likelihoods, we may consider the probabilities from 0 to 1 with a step size interval of 0.01 our observations and the likelihoods for each probability their accompanying 'frequencies' or 'weights'. For instance, when a probability of 0.50 has a likelihood of 0.09 and a probability of 0.40 has a likelihood of 0.03, the probability of 0.50 versus 0.40 is three times more likely to occur. Just as in case of the value of an observation being evaluated three times when this observation value occurs three times instead of only once, a probability's likelihood three times larger than another probability's likelihood must be evaluated with a relative weight of three instead of just one. Therefore, the likelihood of the probability may be considered the weight of this probability. Implementing the step size probabilities and the likelihoods into (8) gives Cliff's adapted Delta for density distributions:

$$\delta = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\text{sign}(P_a - P_b)\bar{\lambda}(P_a)\bar{\lambda}(P_b)}{\sum_{a=1}^{A}\sum_{b=1}^{B}\bar{\lambda}(P_a)\bar{\lambda}(P_b)}, \tag{9}$$

where P_a and P_b are the probabilities from 0 to 1 from group A and group B respectively, $\bar{\lambda}(P_a)$ and $\bar{\lambda}(P_b)$ are the average likelihoods of the probabilities P_a and P_b respectively, and A and B are the same number of step size intervals for both groups. In the case of our step size 0.01, we have 101 of these intervals for each distribution (the first and the 101st interval have step size 0.005 and run from 0.000 to 0.005 and from 0.995 to 1.000 respectively). Here, the midpoints of these intervals may be considered our 'observations' that all have their own accompanying weight in the form of a likelihood.* In this adapted way, we use Cliff's Delta to compare the

likelihood distributions and thus the expected values of two categories of respondents for the same characteristic.

Cliff's Delta has many advantages with respect to answering our research question. Cliff's Delta makes no assumption about the shape of the underlying distribution (Cliff, 1993, 1996ab; Goedhart, 2016; Vargha & Delaney, 2000) and is robust in case of outliers or skewed or otherwise non-normal distributions (Goedhart, 2016). Cliff's Delta is easy to calculate, straightforward to interpret, and standardized, meaning different effect size categories can be distinguished (Goedhart, 2016; see section 5.3 for these categories). For our adapted Cliff's Delta, relatively small or unequal sample sizes are no issue.

Taking into Account Location and Shape

To see how Cliff's Delta works regarding the *location* of the distributions, let us consider Figure 4. In Graph 1 of Figure 4, we have the behaviour profiles for two fictitious groups of respondents, Group 1 and Group 2, with their closely resembling expected values of 0.40 and 0.42 respectively. For reasons of clarity, we show the actual 101 points of both distributions for this example. Here, every data point refers to a probability occurrence with its accompanying likelihood and is considered an observation. As can be seen in Graph 1 of Figure 4, the two behaviour profiles *largely* overlap, meaning that the expected values for both groups are alike. Roughly, the likelihoods accompanying all pairs of observations more or less cancel each other out in (9), resulting in a relatively small absolute Cliff's Delta.

In Graph 2 of Figure 4, the same Group 1 is compared to another fictitious group of respondents, Group 3. In this case, the two behaviour profiles *hardly* overlap, meaning that the expected values for both groups are different, 0.40 and 0.60 respectively. Here, the pairs of observations for which the probability occurrence from Group 3 is larger, are *frequently* accompanied by two relatively larger likelihoods. This is the case when the probability occurrences for Group 1 and Group 3 are for instance 0.35 and 0.65 respectively. However, the pairs of observations for which the probability occurrence from Group 1 is larger, are *rarely* accompanied by two relatively larger likelihoods. This is only the case for the observations surrounding the

* Note that whenever we mention a behaviour occurrence or 'observation' and its likelihood, we refer to the likelihood for the accompanying *interval* for that observation. For instance, an observation of 0.50 refers to the likelihood for the whole interval of behaviour occurrences from 0.495 to 0.505 regarding our chosen step size 0.01. For convenience, we simply mention the midpoints of the intervals.

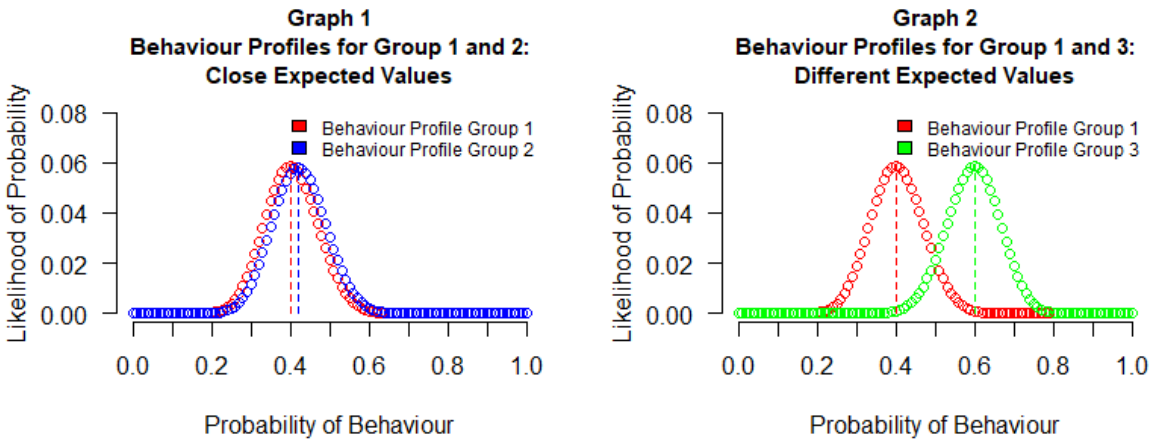

Figure 4. Examples for Two Behaviour Profiles that are Located Close to Each Other (Graph 1) versus Apart from Each Other (Graph 2).

overlapping area, roughly between the expected group values of 0.40 and 0.60, and only to a restricted extent. Hence, in the case of different expected group values and little overlap between the two behaviour profiles, the outcome is a relatively large absolute Cliff's Delta.

To see how Cliff's Delta works regarding the *shape* of the distributions, let us consider Figure 5. Here, we show the behaviour profiles line-shaped. In Graph 1 of Figure 5, we have Respondent 1 and 2 who have expected values of 0.40 and 0.60 respectively. Both respondents filled out *30* items for which they could show certain answer behaviour. Quite some uncertainty exists around the expected values for these respondents, which is marked by the *stretched* spread of the profiles across the probability range from 0 to 1. In general, this comes along with a relatively *large* area of overlap between profiles, which is roughly between 0.30 and 0.70 for these respondents. Now, let us consider Graph 2 of Figure 5 containing the behaviour profiles for Respondent 3 and 4. Their expected values are identical to those of Respondent 1 and 2 respectively. However, Respondent 3 and 4 filled out *80* items for which the concerned behaviour was applicable. As can be seen in Graph 5, quite less uncertainty exists around the expected values for these respondents, marked by the *squeezed* spread of the profiles across the probability range. In general, this comes along with a relatively *small* area of overlap between profiles, which is roughly between 0.45 and 0.55 for these respondents. Hence, a larger Cliff's Delta is expected for comparing the profiles for Respondent 3 and 4 than for the profiles for Respondent 1 and 2. In sum, the further the profiles are apart from each other and the more squeezed they are, the larger Cliff's Delta will be.

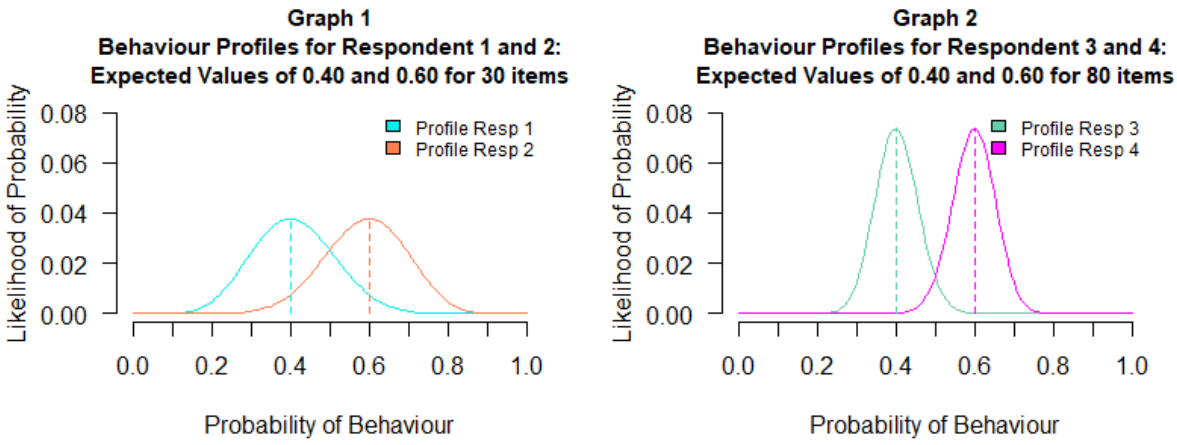

Figure 5. Examples for Two Behaviour Profiles that are Stretched (Graph 1) versus Squeezed (Graph 2).

APPENDIX F: SIMULATING UNCERTAINTY REGARDING THE NUMBER OF ITEMS

Here, we give evidence that Cliff's adapted Delta can be used for our density distributions by illustrating that Cliff's adapted Delta approaches Cliff's original Delta as the number of eligible items filled out by respondents becomes larger. As a first step, we randomly generate ten data observations for a first group A and ten data observations for a second group B from the series {0.0, 0.2, 0.4, 0.6, 0.8, 1.0}. By confining the observations to this series of probabilities, it is easy to vary in the number of eligible items, as long as this number gives an integer when divided by five. Now, let us assume that each data observation is one respondent's fixed probability for a specific answering behaviour. This means there is 100% certainty that a probability is always a respondent's exact and true behaviour occurrence, regardless of the eligible number of items filled out by that respondent. These fixed probabilities are used to calculate Cliff's original Delta by means of (7).

As a second step, suppose that there is *no* 100% certainty for these same fixed probabilities, as each respondent filled out a restricted and variable number of items. In this case, we can estimate the occurrence of behaviour for each respondent and his or her accompanying probability for a pre-specified number of eligible items. This can be done by multiplying each probability by the number of eligible items and using formula 1 to construct the accompanying density distribution with its likelihoods. By averaging the resulting ten density distributions from group A and from group B using formula 2, we can now use (9) to calculate Cliff's adapted Delta for two density distributions. In this way, we can compare Cliff's original and adapted Delta to each other, as we made sure the density distributions were exactly based on the initial data observations.

See Table 9 for the resulting differences between Cliff's original and adapted Delta in various scenario's. Note that we varied the number of times we calculated the difference before taking the average ('Loop'), we varied the magnitude of the step size ('Step'), and we varied the number of eligible items ('ei'), starting by 5 and going towards 100. From Table 9, it is clear that both Loop and Step do not have any influence on the difference between Cliff's original and adapted Delta; choosing a larger number of simulations or a smaller magnitude of step size does not have any implications for the differences. The component that does have an influence on the differences is the number of eligible items. When choosing 5 eligible items, the difference lies around 0.13, regardless of simulation number and step size magnitude.

Table 9. Differences Between Original and Adapted Cliff's Delta for Varying Numbers of Simulations (Loop), Varying Numbers of Step Sizes (Step), and Varying Numbers of Eligible Items (ei).

Loop	Step	5 ei	10 ei	15 ei	20 ei	25 ei	50 ei	100 ei
100	0.1	0.13	0.07	0.04	0.03	0.02	0.005	0.001
	0.05	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.02	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.01	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.005	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.002	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
200	0.1	0.13	0.07	0.04	0.03	0.02	0.005	0.001
	0.05	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.02	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.01	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.005	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.002	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
500	0.1	0.13	0.07	0.04	0.03	0.02	0.005	0.001
	0.05	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.02	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.01	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.005	0.13	0.07	0.04	0.03	0.02	0.005	0.0005
	0.002	0.13	0.07	0.04	0.03	0.02	0.005	0.0005

Compared to Cliff's original Delta, Cliff's adapted Delta is a bit underestimated, which may be explained by the uncertainty that comes along with only 5 eligible items per respondent. When choosing 10 eligible items, this uncertainty is already quite a bit lower, with a difference of only 0.07 between the two Delta's. The uncertainty continues to decrease towards zero as the number of items becomes larger. This refers to the importance of the *shape* of the distributions; the larger the number of eligible items, the smaller the uncertainty, the more squeezed the distributions, the smaller the overlap between both distributions, and the smaller the difference between Cliff's original and adapted Delta.

Explanation of figures

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2018–2019	2018 to 2019 inclusive
2018/2019	Average for 2018 to 2019 inclusive
2018/19	Crop year, financial year, school year, etc., beginning in 2018 and ending in 2019
2016/17–2018/19	Crop year, financial year, etc., 2016/17 to 2018/19 inclusive

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

Colopfon

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

Prepress: Statistics Netherlands
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information