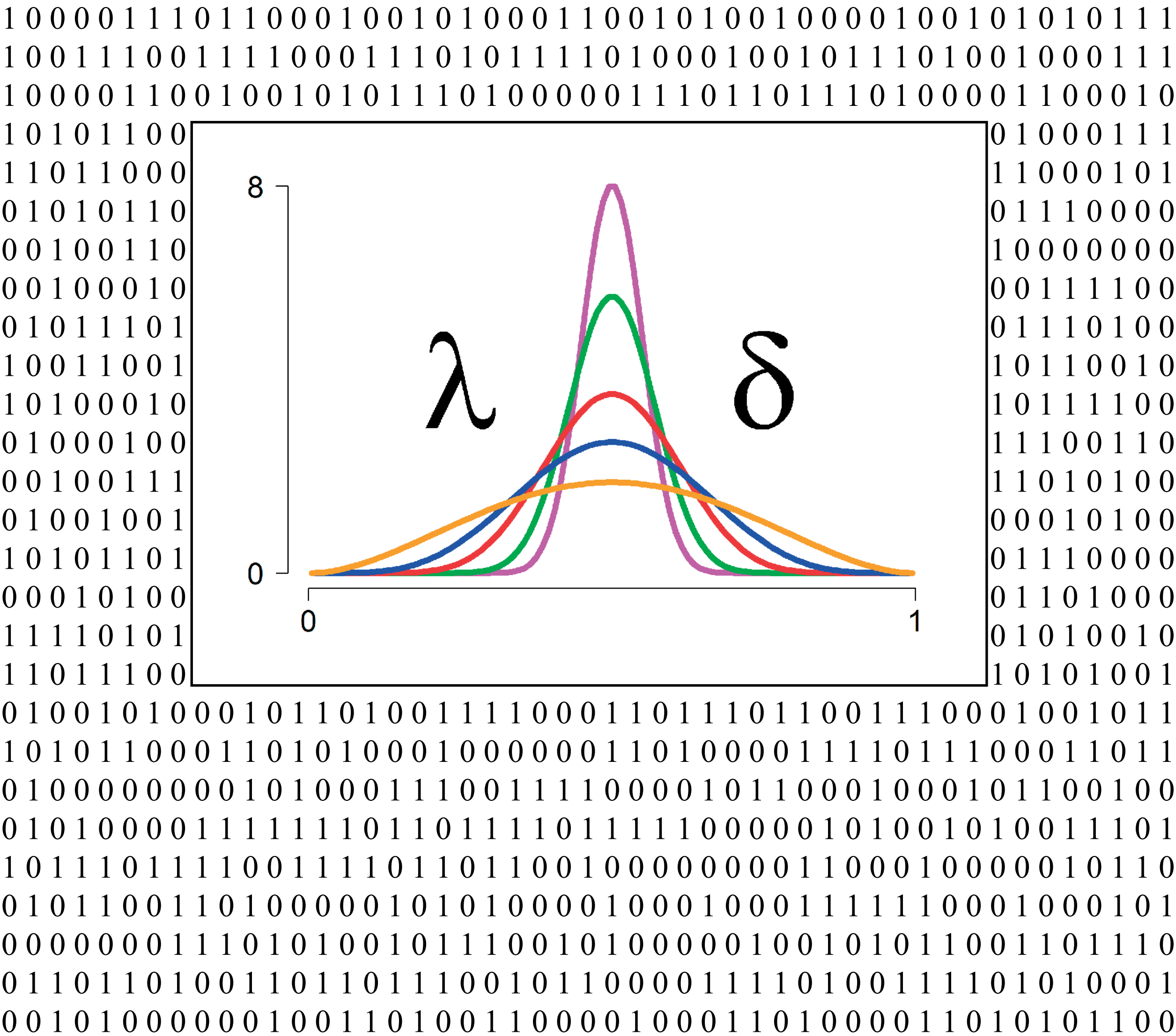
Constructing Behaviour Profiles for Answer Behaviour Across Surveys



Frank Bais

Constructing Behaviour Profiles for Answer Behaviour Across Surveys

Frank Bais

Constructing Behaviour Profiles for Answer Behaviour Across Surveys
Proefschrift Universiteit Utrecht
DOI: https://doi.org/10.33540/538

ISBN 978-94-6419-160-8
Printed by: Gildeprint
Cover design by Frank Bais

Constructing Behaviour Profiles for Answer Behaviour Across Surveys

Het Construeren van Gedragsprofielen voor Antwoordgedrag over Surveys
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 31 maart 2021 des middags te 2.30 uur

door

Frank Bais

geboren op 22 december 1979
te Den Helder

Promotor:
Prof. dr. J.G. Schouten

Copromotor:
Dr. V. Toepoel

Leden van de beoordelingscommissie:

Prof. Dr. B. F. M. Bakker	(Vrije Universiteit Amsterdam)
Prof. Dr. H. van Herk	(Vrije Universiteit Amsterdam)
Prof. Dr. E. D. de Leeuw	(Utrecht University)
Prof. Dr. K. M. Olson	(University of Nebraska-Lincoln)
Dr. B. Struminskaya	(Utrecht University)

To Jennie

and

to Joost

Deep in the chaotic regime, slight changes in structure almost always cause vast changes in behavior. Complex controllable behavior seems precluded.

(Stuart Kauffman)

List of Contents

List of Tables

List of Figures

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

Chapter 1

General Introduction

According to biologist Midas Dekkers, we should be careful in interpreting answers that are obtained from respondents to survey questions: 'Psychologists listen or they send you a survey form to fill out. This method on its own is questionable, considering the fact that language is particularly invented to enable lying.' (Dekkers, 2017). Although somewhat cynical, Dekkers has a point in that survey responses may differ in their degree of truthfulness. For instance, when respondents are asked on which political party they voted, a substantial number of them may report a party that they did not factually vote for (Van der Ploeg, 2008; Van der Ploeg, Van de Pol, & Kampen, 2008, 2009).

We can explain survey responses that do not correspond with reality by means of the cognitive response model. This model consists of four subsequent steps: 1) Comprehension of the question; 2) retrieval of relevant information from memory; 3) integration of the information into judgments, and; 4) reporting of the final answer (Jenkins & Dillman, 1997; Tourangeau, 1984, 1987; Tourangeau & Rasinski, 1988; Tourangeau, Rips, & Rasinski, 2000; see Cannell, Marquis, & Laurent, 1977; Cannell, Miller, & Oksenberg, 1981; Oksenberg & Cannell, 1977 for initial versions of the model). See Figure 1.1. During each of these four steps in the sequence, errors may occur that lead to an inaccurate response. The extent to which such a response deviates from the true value that a survey question was intended to measure is referred to as *measurement error* (De Leeuw, Hox, & Dillman, 2008). Measurement error is the core survey concept that is central to this thesis. In particular, *the focus of this thesis lies on the relation of respondent characteristics and survey item characteristics to measurement error across multiple surveys*.

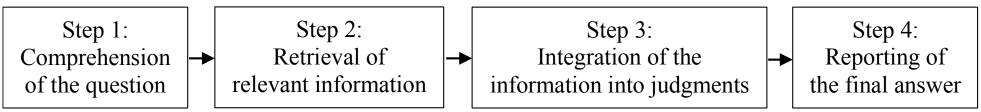


Figure 1.1. Cognitive Response Model.

1.1. Satisficing and sensitivity-based behaviour

There can be many reasons for respondents to give a response that is not or only partly corresponding with reality. A first reason is that a survey question asks for information that the respondent is not *able* to give. More concrete, a respondent may not be able to answer survey items because of a lower level of cognitive ability. Specifically, respondents who are not adept at performing complex cognitive operations may have difficulty in retrieving information (step 2 in Figure 1.1) and making judgments (step 3 in Figure 1.1) while answering survey items (Krosnick, 1991). Respondents may also be unable to answer when they have little practice at thinking about the topic or when they do not already have a well-formed attitude on the issue asked about in the particular questions (Krosnick, 1991). Two easy-to-measure respondent characteristics that are presumed to be directly related to cognitive ability are age and educational level. Both characteristics have been shown to be related to the quality of responses to survey items (Krosnick 1991, 1999; Krosnick, Narayan, & Smith, 1996). Older and lower educated respondents show less accurate answers than younger respondents (Andrews & Herzog, 1986) and higher educated respondents (Antoni, Bela, & Vicari, 2019). They also show a less stable attitude reliability measurement than younger and higher educated respondents (Alwin & Krosnick, 1991). Hence, a lower level of cognitive ability, as indicated by higher age or a lower educational level, is associated with less accurate response quality.

A second reason why responses may not fully correspond with reality is that respondents may not be *motivated* to give an accurate response to a survey question. When respondents are not motivated to answer, a suboptimal execution for any of the four steps of the cognitive response model may occur (Krosnick, 1991). Respondents may be less motivated to answer survey questions when they have a low need for cognition (see Cacioppo & Petty, 1982). Such respondents do not enjoy thinking, dislike spending cognitive effort, and avoid burdensome cognitive tasks (Krosnick, 1991). Other cases in which respondents may not be motivated to answer survey questions according to Krosnick (1991) are when the topic of the question is personally not important to the respondent, when respondents think that the survey in which they participate is not important or not useful to society, or when respondents do not consider themselves accountable for the responses they give. Finally, interviewer behaviour and the length of the interview may influence respondent motivation. In general, the longer the interview continues, the lower the level of respondent motivation, and the less optimal response data quality (Galesic & Bosnjak, 2009; Gummer & Roßmann, 2014; Holbrook, Green, & Krosnick, 2003; Krosnick, 1991). Thus, a lower degree of respondent motivation is related to less accurate responses.

A third reason why responses may be inaccurate is because of *item difficulty* (see chapter 2 for specific kinds and examples of item difficulty). Also a high degree of item difficulty can have its influence on any of the four steps of the cognitive response model. It may lead to difficulty in comprehending a question, to a burdensome retrieval of relevant information, to a demanding integration of the information into judgments, and to difficulty in selecting the most accurate response (Krosnick, 1991). According to Krosnick (1991), items may become more difficult to fill out when the interviewer maintains a high pace of interviewing or when distraction is present within the setting of the interview. Hence, a high level of item difficulty may lead to less accurate response quality. In summary, a lower level of cognitive ability, a lower degree of motivation, and a higher level of item difficulty may lead to the occurrence of inaccurate responses and thus measurement error.

The interaction of the specific degrees of cognitive ability, motivation, and item difficulty determines the extent of *satisficing*, a concept introduced by Simon (1957). When respondents satisfice while answering a survey item, they execute at least one of the four steps of the cognitive response model less optimally. They take short-cuts in the question-answering process and settle for an answer that is merely satisfactory to them (Krosnick, 1991). Krosnick (1991) presents the relations between cognitive ability, motivation, and item difficulty, and hence the likelihood that respondents will satisfice in answering a specific question as

$$p(\text{Satisficing}) = \frac{\alpha_1(\text{ITEM DIFFICULTY})}{\alpha_2(\text{COGNITIVE ABILITY})\ \alpha_3(\text{MOTIVATION})}, \tag{1.1}$$

into which measures of cognitive ability, motivation, and item difficulty can be entered, and where α_1, α_2, and α_3 can be estimated by means of regression procedures (see Krosnick, 1991). Here, a larger item difficulty is relatively more likely to induce satisficing and is therefore put in the numerator. A larger cognitive ability and motivation are relatively less likely to induce satisficing, and are therefore put in the denominator.

A fourth reason why respondents may give inaccurate answers is that an item asks for information that they are not *willing* to give. Items that ask for sensitive information may involve a threat of information disclosure to third parties (Lensvelt-Mulders, 2008; Tourangeau & Yan, 2007), can be experienced as intrusive because the issue is considered private or taboo (Tourangeau et al., 2000; Tourangeau & Yan, 2007), or may evoke a socially desirable answer (Johnson & Van de Vijver, 2003; Kreuter, Presser, & Tourangeau, 2008; Tourangeau et al., 2000; Tourangeau & Yan, 2007). The result may be a lack of willingness from the respondents to give a ‘true’ answer (Bradburn, Sudman, Blair, & Stocking, 1978; Shoemaker, Eichholz, & Skewes, 2002; Tourangeau et al., 2000).

In summary, the interaction of cognitive ability, motivation, and item difficulty determines the degree of what we call *satisficing answer behaviour*. The extent to which items ask for sensitive information determines the degree of what we call *sensitivity-based answer behaviour*. See Figure 1.2 for a graphic summary. As can be seen in Figure 1.2, when satisficing or sensitivity-based behaviour is triggered, respondents may show several specific answer behaviours. We need to emphasize that these specific answer behaviours may or may not be a consequence of satisficing or sensitivity-based behaviour. For instance, respondents may fill out a ‘don’t know’-answer because they truthfully do not know, but it may also be the result of satisficing behaviour. In other words, these specific answer behaviours may or may not refer to applicable or truthful answers.

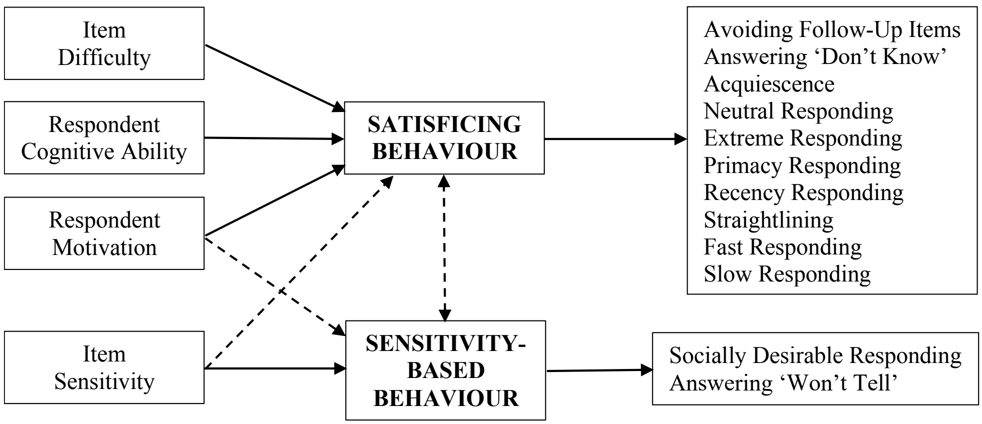


Figure 1.2. Overarching Literature-Based Theoretical Framework.

Before elaborating on the specific answer behaviours, we need to note that we do not emphasize the role of the mode in which the survey is administered for this thesis. From the literature, it is well known that responses may differ depending on the applicable survey mode (see for instance Buelens & Van den Brakel, 2011; De Leeuw, 2005; Jäckle, Roberts, & Lynn, 2010; Roberts, 2007; Schouten, Van den Brakel, Buelens, Van der Laan, & Klausch, 2013). These mode differences may also differ in magnitude or direction between specific answer behaviours (see for instance Heerwegh & Loosveldt, 2011; Holbrook et al., 2003; Kwak & Radler, 2002). As this thesis is based on survey data from a computer-assisted mode, the findings may not be generalizable to results from studies using other survey modes. It is for future research to investigate the relation of respondent and item characteristics to measurement error across multiple surveys considering mode differences. From here, we elaborate on the specific answer behaviours in Figure 1.2. We distinguish the case in which they refer to truthful answers from the case in which they may be the result of satisficing or sensitivity-based behaviour.

1.2. The specific answer behaviours

A first specific kind of satisficing behaviour is *the omission of follow-up questions* (see for instance Eckman et al., 2014; Knäuper, 1998; Kreuter, McCulloch, Presser, & Tourangeau, 2011). This omission is justified when the chosen answering option that does not lead to a follow-up question is the most applicable or truthful option for a specific respondent. But respondents may also *presume* some answering options of an item to lead to a follow-up question. Consequently, they may avoid choosing the most applicable answering option, as they expect this option to lead to follow-up questions. Instead, they choose a less applicable or truthful option in order to restrict the cognitive burden of filling out more items than necessary or the processing of items at all.

A second kind of satisficing behaviour is *answering 'don't know'*, a category of non-substantive response that is frequently offered as an answering option (see for instance Beatty & Herrmann, 2002; Binswanger, Schunk, & Toepoel, 2013; Leigh & Martin, 1987;). Respondents may choose the option 'don't know' because they actually do not know the requested content. But they may also choose the 'don't know'-option while they *do* know the applicable answer. They may have filled out a substantive answer if they exerted more cognitive effort or motivation into their response process, or when the item was less difficult to answer.

Another kind of satisficing behaviour is *straightlining* (see for instance Krosnick, 1991; Krosnick & Alwin, 1989; Schonlau & Toepoel, 2015). When respondents straightline, they show no variation in their answers in filling out a set of multiple items containing the same range of fixed answering options. For instance, respondents may fill out the option 'disagree' for all attitudinal statements that form such a set of items. Respondents may truly disagree with each statement, but they may also consistently choose this option because they are not processing each statement thoroughly. *Acquiescence* refers to satisficing behaviour by which respondents tend to answer agreeably or affirmatively, regardless of the content of the question or statement (see for instance Messick, 1966; O'Muircheartaigh, Krosnick, & Helic, 2000; Stricker, 1963). Respondents may truly agree with many statements, but they may also frequently show agreeableness as this means less effort in their cognitive processing.

Respondents may also give *neutral or extreme answers* as a result of satisficing behaviour. For attitudinal statements, respondents may refrain from expressing a judgment by filling out the middle or neutral response of the answering options (see for instance Kalton, Roberts, & Holt, 1980; Krosnick & Fabrigar, 1997; Tarnai & Dillman, 1992). Or they may choose the most extreme option that expresses their judgment (see for instance Aichholzer, 2013; Díaz de Rada & Domínguez, 2015; Ye, Fulton, & Tourangeau, 2011). In both cases, the answer may refer to their true opinion, but may also be the result of satisficing. Choosing a neutral response may be considered more or less equal to answering 'no opinion'. This could mean that respondents take a short-cut in processing information to avoid deciding which side they are on (for instance 'agree' versus 'disagree'). When respondents can make this decision, they may frequently choose an extreme response (for instance 'totally agree' or 'totally disagree'). This could mean that once respondents roughly know the direction of their opinion, they may stop processing information and choose for the most extreme answer that conforms with that opinion.

Primacy and recency responding are two other kinds of satisficing behaviour. These two behaviours mean that respondents choose one of the first few or one of the last few answering options from all ordered possibilities (see for instance Krosnick, 1991; Krosnick, 1992; Krosnick & Alwin 1987; McClendon, 1991). Primacy responses can be expected in survey modes where the items are presented visually (in online or postal interviews), while recency responses can be expected in modes where items are presented aurally by an interviewer (in face-to-face and telephone interviews). The primacy or recency answer may be truly applicable, but the answer may also be the result of little effort in processing all response options.

Two final types of specific behaviours that may follow from satisficing are answering items at a high or a low pace; '*fast responding'* (see for instance Malhotra, 2008; Yan & Tourangeau, 2008; Zhang & Conrad, 2014) or '*slow responding'* respectively (see for instance Couper & Kreuter, 2013; Yan & Tourangeau, 2008). Respondents who report their responses quickly after items are presented may process information less thoroughly than respondents who go through the interview at a more customary pace. This more superficial processing is more likely as the pace at which items are answered during the interview increases. In contrast, responding slowly may refer to difficulty in the processing of items and hence lead to lower data quality. The more these specific types of answer behaviour are the result of satisficing, the more the responses deviate from the content that was intended to be obtained. Thus, a higher level of satisficing is related to a larger degree of measurement error and hence lower response data quality.

A first kind of specific answer behaviour that may result from sensitivity-based behaviour is *socially desirable responding* (see for instance Johnson & Van de Vijver, 2003; Kreuter et al., 2008; Tourangeau & Yan, 2007). Respondents who provide a socially desirable answer tend to minimize showing behaviour that violates a social norm (DeMaio, 1984; Krosnick, 1999; Paulhus, 2002). For instance, respondents may state that they do never use drugs. This answer may be truly applicable to the respondent. But as using drugs may be considered a violation of the social norm that one does not use drugs, the answer may also be a consequence of being socially desirable.

A second specific kind of sensitivity-based behaviour is *refusing to answer* (see for instance Bradburn et al., 1978; Shoemaker et al., 2002; Tourangeau et al., 2000). Respondents can be asked to reveal sensitive information that is considered intrusive or private. For instance, they may be asked to share their monthly or yearly income. Especially when their income is very low or high, they may refuse to answer and fill out the non-substantive option *'won't tell'* (see Juster & Smith, 1997). Thus, item sensitivity may lead to socially desirable and 'won't tell'-answers. Also the respondents' motivation to reveal sensitive information may have its influence on the occurrence of these specific behaviours (see Figure 1.2).

1.3. The potential undesirability and consistency of answer behaviour

Above, we elaborated on how each specific answer behaviour may or may not be applicable or truthful. One could argue that specific answer behaviour that is not applicable, but instead the result of satisficing or sensitivity-based behaviour, can be considered '*undesirable*'. As non-applicable or undesirable answer behaviour deviates from what was intended to be asked, it may be considered an indication of measurement error. However, we need to emphasize that we cannot know when answer behaviour is applicable versus undesirable without at least some validation method (see for instance Bakker, 2012; Scholtus, Bakker, & Van Delden, 2015). Therefore, the specific answer behaviours are not undesirable by definition, but they are *potentially* undesirable. From here, we elaborate on how the potential character of undesirability becomes more likely as more survey occasions play a role.

A single survey is usually filled out in the more or less same state of mind, mood, and circumstance. For instance, when respondents give many 'don't know'-answers across the items of a single survey, it is difficult to say whether they satisfice versus factually do not know the answers. However, it is reasonable to assume that this state of mind, mood, and circumstance differs from occasion to occasion, from survey to survey. This means that when respondents show many 'don't know'-answers on multiple occasions, it is unrealistic to attribute this to the idea that the respondents do not know many answers for many surveys. More likely, this means that such a respondent may have a personal or natural *tendency* to give 'don't know'-answers. Hence, the potential non-applicability or undesirability of answer behaviour becomes more likely, as the behaviour is shown to a substantial extent on several occasions. In other words, the potential undesirability of answer behaviour becomes more likely as the behaviour is shown *consistently*, across multiple surveys. Moreover, the undesirability not only becomes more likely, but also becomes more problematic as a specific behaviour is shown across the more surveys. We use this consistency in this thesis to indicate the likelihood of the undesirability of answer behaviour.

1.4. Relating respondent and item characteristics to undesirable answer behaviour

Now that we defined undesirable answer behaviour and the importance of its consistency across surveys, we turn to the relation of such behaviour to respondent and item characteristics. See the four variables that may lead to satisficing or sensitivity-based behaviour in Figure 1.2: Cognitive ability, motivation, item difficulty, and item sensitivity. Cognitive ability and motivation can be considered characteristics of respondents, while item difficulty and sensitivity can be considered characteristics of items. Let us start by elaborating on respondent characteristics. While respondent motivation is likely to vary by mood or occasion and therefore not suitable to connect to stable answer behaviour, cognitive ability may be considered to be more or less stable through time. Two easy-to-measure variables or proxies of cognitive ability are age and educational level. Suppose that we want to investigate the extent to which the relation between educational level and a high occurrence of 'don't know'-answers for respondents is structural and consistent. This would not be reasonable to do on the basis of a single survey, since we do not know the behaviour occurrence for other surveys. When a high occurrence of the behaviour would appear consistently across multiple surveys, it may be justified to structurally relate educational level to frequently showing ‘don’t know’-answers. Thus, in order to structurally relate respondent characteristics to undesirable answer behaviour, the characteristic needs to be stable through time and the behaviour needs to be shown consistently across surveys.

The same idea applies to item difficulty and item sensitivity. These characteristics may only be structurally related to answer behaviour when the difficulty or sensitivity of survey items can be determined unambiguously (see chapter 2) and the behaviour is shown consistently across surveys. Although it can be valuable to investigate the relation between item characteristics and answer behaviour for a single survey, difficult and sensitive items are designed differently and focus on a different topic for each separate survey. Therefore, our aim is to look for a consistent relation between item characteristics and answer behaviour across surveys. In case of consistent relations of respondent and item characteristics to undesirable answer behaviour, we may summarize these relations in detail in useful overviews or schemes. We may call such overviews respondent schemes and questionnaire schemes. These schemes would then consist of the characteristics of respondents and survey items respectively that are consistently associated with specific answer behaviour across surveys. Such schemes may be used as informative overviews of these associations. The schemes could be taken into account by survey designers and administrators in order to constrain respondents’ tendency to show specific undesirable answer behaviour (see for instance Calinescu & Schouten, 2016; Schouten, Calinescu, & Luiten, 2013; Schouten, Peytchev, & Wagner, 2017). Thus, by accounting for the characteristics of respondents and items, surveys could be constructed in a way that minimizes measurement error and optimizes survey data quality.

As a final note, the literature suggests that specific answer behaviour may be referred to as a ‘response style’ (see for instance Baumgartner & Steenkamp, 2001; He & Van de Vijver, 2013; He, Van de Vijver, Espinosa, & Mui, 2014; Schouten & Calinescu, 2013; Van Herk, Poortinga, & Verhallen, 2004; Van Rosmalen, Van Herk, & Groenen, 2010). The concept of ‘style’ implies that showing a specific behaviour would be a more or less stable respondent characteristic. As far as we know, specific answer behaviours have not been investigated over a large number of surveys for the same respondents. This means that there is no evidence so far for response styles in terms of stable and consistently high occurrences of specific answer behaviour across surveys. Therefore, we omit the concept of ‘response style’ throughout this thesis and simply speak of answer behaviour.

1.5. Methodological approach

Investigating data quality and modeling answer behaviour can be done by using item response models (Bolt & Johnson, 2009; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Tijmstra, Bolsinova, & Jeon, 2018; Zickar, Gibby, & Robie, 2004), latent class models (Van Rosmalen et al., 2010), multitrait multimethod (MTMM) data (Andrews, 1984; Campbell & Fiske, 1959; Hox, 1995; Saris & Gallhofer, 2000, 2007ab; Saris, Revilla, Krosnick, & Shaeffer, 2010; Saris, Satorra, & Coenders, 2004; Scherpenzeel, 1995; Scherpenzeel & Saris, 1997), multilevel models (Beretvas, 2011; Fielding & Goldstein, 2006; Gummer & Roßmann, 2014; Leckie, 2019; Olson & Smyth, 2015; Olson, Smyth, & Ganshert, 2019; Raudenbush & Bryk, 2002), or a combination of the two latter (Maas, Lensvelt-Mulders, & Hox, 2009). We need to note that we do not use a customary statistical method for the main part of this thesis. Although we use cross-classified multilevel models to explore individual answer behaviour across surveys (see chapter 3), we propose a novel method to model answer behaviour and use this method to investigate its relation to respondent characteristics in later chapters (see chapters 4 and 5 respectively). We chose for this new method as; 1) we want to visualize complete summaries of the occurrence of answer behaviour for groups of respondents along with the uncertainty that surrounds this occurrence, and; 2) we do not want to investigate answer behaviour of identified individual respondents, but behaviour of groups of respondents sharing the same characteristic. We consider investigating the relation between answer behaviour and respondent characteristics a first step in exploring our method for survey data in practice.

1.6. Thesis contribution and overview

Central in this thesis is the relation between respondent characteristics and answer behaviour (see chapter 5). In order to investigate this relation, an exploration is executed into the degree of consistency of specific answer behaviours (see chapter 3), and a novel methodology (see chapters 4 and 5) and an adaptive statistic (see chapter 5) are proposed. To anticipate on investigating the relation between item characteristics and answer behaviour in future research, we explored the extent to which characteristics of items can be determined reliably (see chapter 2).

The thesis is based on data from ten extensive Dutch population surveys from CentERdata and Statistics Netherlands. The surveys are administered in computer-assisted format by CentERdata in the Longitudinal Internet studies for the Social Sciences (LISS) Panel. The LISS Panel consists of about 7000 respondents, and over 2000 items from all surveys together vary broadly in their design, topic, and characteristics. This is a solid basis to investigate respondent answer behaviour on its consistency across surveys and its potential undesirability as an indicator of measurement error. Emerging relations between characteristics and answer behaviour that appear to be consistent can be helpful in designing and administering adapted surveys in the future.

In chapter 2, we investigate to what degree a broad variety of item characteristics can be determined unambiguously. This chapter was inspired by overviews of relations between item characteristics and measurement error, as presented by Gallhofer, Scherpenzeel, and Saris (2007), Campanelli et al. (2011), and Beukenhorst et al. (2014), a discussion paper of Statistics Netherlands. Specific kinds of item difficulty and item sensitivity are coded by two or three coders in order to obtain intercoder reliabilities. An intercoder reliability close to 1 would mean that the presence or nature of a specific characteristic for all applicable items could be determined reliably. This means that this characteristic could be analyzed on its relation to answer behaviour. An intercoder reliability substantially lower than 1 would mean that it was not possible to determine the presence or nature of a specific characteristic reliably. Potential solutions are provided for the case of lower intercoder reliability. A step for future research is to investigate the relation of reliable item characteristics to consistent answer behaviour.

In chapter 3, we investigate the occurrence of answer behaviour and to what extent various answer behaviours are shown consistently across surveys. We use multilevel models to disentangle the variances for respondents, surveys, and the interaction between respondents and surveys. By means of these variances, we obtain indications for the extent to which respondents show a high level of specific behaviour across multiple surveys. Especially answer behaviours that show a high respondent variance and a low respondent-survey interaction variance would indicate answer behaviour consistency across surveys. Most answer behaviours are further investigated in the following chapters of this thesis.

In chapter 4, we propose a new methodology to visualize and compare answer behaviour for groups of respondents or items: Behaviour profiles. The behaviour profiles visualize the occurrence of answer behaviour for a group of respondents or items. They take into account the uncertainty regarding the size of the group of respondents (or items) and the number of items (or respondents) that they are based on. Behaviour profiles can be used for individual or groups of respondents, for single surveys or multiple surveys at once, and for any survey mode.

Behaviour profiles for any two groups of respondents or items can be compared and these groups may be based on any combination of respondent and item characteristics. In this chapter, we elaborate on the statistical properties and practical utility of behaviour profiles.

In chapter 5, we construct behaviour profiles that are based on categories of the respondent characteristics age and educational level, which are used as proxies for cognitive ability. Additionally, categories of several other respondent characteristics are selected to construct behaviour profiles for. An adaptation of the robust effect size measure Cliff’s Delta is used to compare the behaviour profiles for these subgroups. In this chapter, we address our main research question about the relation of respondent characteristics to undesirable answer behaviour across multiple surveys. Finally, in chapter 6, we present a recapitulation of the chapters, address the limitations of this thesis, and make suggestions for future research.

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

Chapter 2

Can survey item characteristics relevant to measurement error be coded reliably? A case study on eleven Dutch general population surveys

This chapter is published in Sociological Methods & Research (SMR) as:
Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M., & Vis, C. (2019). 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), 263-295. https://doi.org/10.1177/0049124117729692

Author contributions: FB, BS, JA, SD, NK, MM, and CV designed the study. CentERdata and Statistics Netherlands provided the surveys. FB, BS, PL, JA, SD, NK, MM, and CV coded the data. FB prepared the data, performed the statistical analyses, and wrote the paper. BS and VT critically reviewed the paper.

Abstract

Item characteristics can have a significant effect on survey data quality and may be associated with measurement error. Literature on data quality and measurement error is often inconclusive. This could be because item characteristics used for detecting measurement error are not coded unambiguously. In our study, we use a systematic coding procedure with multiple coders to investigate the extent to which the coding of item characteristics could be done reliably. For this purpose, we constructed a list of item characteristics that is based on existing typologies of item characteristics. High intercoder reliability indicates a clear relation between item characteristic, item content, and measurement error. Our results show that intercoder reliability is often low, especially for item characteristics that are hard to code due to subjectivity. Low intercoder reliability complicates comparisons between studies about item characteristics and measurement error. We give suggestions for coping with low intercoder reliability.

2.1. INTRODUCTION

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

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

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

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

To our knowledge, no research so far has reported a systematic procedure to code many items of multiple surveys on their characteristics by two or more coders to evaluate intercoder reliability. By their experiment, Beukenhorst et al. (2014) made a first attempt to characterize a whole survey questionnaire to investigate mode-specific measurement error by using an item coding typology that was partly based on the SQP typology of Saris and Gallhofer (2007b) and Gallhofer et al. (2007), and on the typology of Campanelli et al. (2011). On the basis of these typologies, we constructed a list of questionnaire characteristics consisting of both question and answer characteristics. By coding 11 questionnaires of Statistics Netherlands and CentERdata, we can investigate the intercoder reliability on these characteristics for 2470 items that range over various general population topics such as income, education, work, leisure, and personality. In case the intercoder reliability is *high* on certain characteristics, a questionnaire scheme based on these characteristics may be constructed relatively easily. In case the intercoder reliability is *low* on certain characteristics, we need to explain this low reliability and how to cope with it. In this study, we 1) investigate the intercoder reliability for each item characteristic over the items of all surveys together; 2) try to explain potential low intercoder reliability, and; 3) give suggestions about how to cope with such low reliability.

From here, we first motivate the chosen item characteristics and accompanying literature background in section 2.2. In section 2.3, we present all surveys for which these characteristics are coded, and elaborate on the actual coding procedure and the statistics that were calculated. In section 2.4, we present all statistical results of the actual coding procedure. In section 2.5, we suggest ways of coping with low intercoder reliability. In section 2.6, we conclude with a discussion of these results.

2.2. THE ITEM CHARACTERISTICS

In this section, we present the list of 16 item characteristics as used in the current study and elaborate on the literature background of these characteristics. Thirteen other item characteristics were considered to be codable on their true category unambiguously. For instance, one such characteristic is the amount of words that the item contains up till the first answering category. Therefore, these 13 characteristics were coded by a single coder and are not taken into consideration for this paper. See Table A.1 in Appendix A for an overview of these item characteristics. Table 2.1 below presents an overview of the 16 item characteristics and their references that are involved in the current study.

An important note is that we came to our conclusive list of item characteristics based on a pilot study. The pilot study was set up to investigate the actual occurrence of each item characteristic and to check for potential difficulties during the coding process. See Appendix B to read about the execution of the pilot study and about what changes the pilot study resulted in before coming to the conclusive list of item characteristics. From here, we give a motivation for the inclusion of the item characteristics, considering their influence on data quality in general and on measu-

Table 2.1. Definitions of the Item Characteristics, Their Coding Numbers and Categories, and References.

Item characteristic	*Definition of the item characteristic as used in the current study*	*Coding number and categories*	*References*
Content of the question *	What kind of topic or aspect is the item about?	1 factual behaviour 2 otherwise factual 3 opinion 4 satisfaction 5 otherwise subjective	Campanelli et al. 2011; Gallhofer et al. 2007; Lozar Manfreda & Vehovar 2002; Saris & Gallhofer 2007b; Schonlau et al. 2004
Emotional charge *	Does the item contain potentially emotional words or a potentially emotional charge?	0 not applicable 1 applicable	Lensvelt-Mulders 2008
Sensitive information *	Does the item contain sensitive information of some societal, menial, or personal kind?	0 not applicable 1 applicable	Campanelli et al. 2011; Gallhofer et al. 2007; Kreuter et al. 2008; Lensvelt-Mulders 2008; Saris & Gallhofer 2007b; Tourangeau & Yan 2007
Presumption of filter question *	Might the respondent be able to presume the item to be a filter question?	0 not applicable 1 applicable	Bosley et al. 1999; Eckman et al. 2014; Kreuter et al. 2011
Centrality *	Does the item go beyond the interest, knowledge, or experience of the respondent?	0 not applicable 1 applicable	Gallhofer et al. 2007; Saris & Gallhofer 2007b; Van der Zouwen 2000
Question complexity 1: Difficult language usage *	Does the question contain unknown or difficult words, or complex sentences?	0 not applicable 1 applicable	Beukenhorst et al. 2014; Van der Zouwen 2000
Question complexity 2: Conditions **	Does the item contain conditions?	0 not applicable 1 applicable	Beukenhorst et al. 2014; Van der Zouwen 2000
Question complexity 3: Memory **	Does answering require some kind of memory?	0 no memory 1 non-specific memory 2 memory < one month ago 3 memory > one month ago	Van der Vaart et al. 1995; Van der Zouwen 2000
Question complexity 4: Hypothetical situation **	Does the item refer to a concrete, specific hypothetical situation in the future?	0 not applicable 1 applicable	Van der Zouwen 2000; Van der Zouwen & Dijkstra 1996
Question complexity 5: Calculations **	Does answering require the performance of some kind of calculation?	0 not applicable 1 applicable	Beukenhorst et al. 2014; Van der Zouwen 2000
Question complexity 6: Ambiguity **	Does the item contain multiple subquestions or is the item otherwise potentially confusing?	0 not applicable 1 applicable	Campanelli et al. 2011; Foddy 1993; Fowler & Mangione 1990; Van der Zouwen 2000

* The 'hard' item characteristics; ** The 'easy' item characteristics; see section 2.3

Table 2.1 (continued). Definitions of the Item Characteristics, Their Coding Numbers and Categories, and References.

Item characteristic	*Definition of the item characteristic as used in the current study*	*Coding number and categories*	*References*
Response complexity *	Do the answering options contain unknown or difficult words, or complex sentences, or do they require the execution of some kind of performance?	0 not applicable 1 applicable	Campanelli et al. 2011; Gallhofer et al. 2007; Saris & Gallhofer 2007b
Time reference **	What time period does the item refer to?	1 past 2 present 3 future	Gallhofer et al. 2007; Saris & Gallhofer 2007b
Mismatch **	Do the question and its answering options match?	0 not applicable 1 applicable	Beukenhorst et al. 2014; Smyth & Olson 2019; Van der Zouwen 2000
Formulation **	Is the item formulated as a statement?	0 not applicable 1 applicable	Fowler 1995; Gallhofer et al. 2007; Saris & Gallhofer 2007b; Saris et al. 2010; Ye et al. 2011
Clarification **	Does the item contain some kind of clarification?	0 not applicable 1 applicable	Gallhofer et al. 2007; Saris & Gallhofer 2007b; Van der Zouwen 2000

* The 'hard' item characteristics; ** The 'easy' item characteristics; see section 2.3

rement error. According to the literature, some item characteristics may be sensitive to mode-specific measurement error in particular. Therefore, we finish this section by briefly elaborating on how these characteristics may be associated with mode-specific measurement error.

2.2.1. Motivating the item characteristics

2.2.1.1. Question complexity

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

The characteristic *difficult language usage* refers to the use of unknown or difficult words, or complex sentences within the item (Beukenhorst et al., 2014), possibly having a negative influence on response quality (Van der Zouwen, 2000). The characteristic *conditions* refers to specifically including and/or excluding certain aspects in/from the answer and the characteristic *calculations* refers to the performance of some kind of mathematical computation (Beukenhorst et al., 2014; Van der Zouwen, 2000). Both characteristics may relate to a relatively high cognitive burden on the respondent while answering a question (Lenzner, Kaczmirek, & Lenzner, 2009; Tourangeau et al., 2000; Van der Zouwen, 2000).

The characteristic *hypothetical situation* refers to imagining a fictitious or hypothetical situation (Van der Zouwen & Dijkstra, 1996). Respondents may have difficulty in accepting the reality of a hypothetical situation or with imagining a situation in the far future (Van der Zouwen, 2000). The characteristic *memory* refers to retrieving information from the past. Questions requiring information retrieval from the past are retrospective questions that may have a negative effect on response quality (Van der Vaart, Van der Zouwen, & Dijkstra, 1995; Van der Zouwen, 2000), especially when no recall aiding devices are used (Van der Vaart, 1996). The characteristic *ambiguity* refers to questions that are double-barrelled (Bassili & Scott, 1996; Campanelli et al., 2011; Foddy, 1993; Fowler & Mangione, 1990) or otherwise have an unclear meaning of wording (Van der Zouwen, 2000).

2.2.1.2. Response complexity

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

2.2.1.3. Centrality

Centrality is particularly about the concept or content of the question. When the item is about a topic that extends beyond the knowledge, experience, or interest of the respondent, the content or topic is not ‘central’ in the mind of the respondent (Gallhofer et al., 2007; Saris & Gallhofer, 2007b). This is for instance the case when an item deals with a political or religious topic, which is not ‘central’ in the life of relatively many respondents. The respondent might be either reluctant or incapable to answer items that are non-central or hardly accessible (Van der Zouwen, 2000) to them.

2.2.1.4. Content of the question

Concerning content of the question, an item is about factual behaviour, otherwise factual, opinions, satisfaction, or otherwise subjective (Campanelli et al., 2011; Gallhofer et al., 2007; Saris & Gallhofer, 2007b). Here, otherwise factual refers to items asking for factual data other than factual behaviour, and otherwise subjective refers to items asking for thoughts, feelings, or emotions other than opinions or satisfaction of the respondent. We defined factual behaviour and otherwise factual as objective categories that are observable and measurable, as opposed to opinions, satisfaction, and otherwise subjective, which are considered subjective categories. The goal is to distinguish objective versus subjective categories, with the latter categories being more sensitive to the predispositions of the respondent.

2.2.1.5. Sensitive information

Some items ask for sensitive information that may be perceived as being more or less threatening by respondents (Lensvelt-Mulders, 2008). Sensitive questions are about private, stressful, or sacred issues. Answering sensitive questions may evoke emotional responses or the potential fear of stigmatization on the part of the respondent or his social group (Lensvelt-Mulders, 2008). Tourangeau et al. (2000) define a sensitive question as being experienced as intrusive, involving a threat of disclosure, or to some extent eliciting an answer that is socially undesirable. In effect, a question is sensitive when it asks respondents to admit that they have violated a social norm (Tourangeau & Yan, 2007). This may for instance be the case when items ask for information about former or current drug or alcohol use. As a result, respondents might be reluctant to answer the question and may tend to avoid or distort their answer.

2.2.1.6. Emotional charge

This item characteristic is related to the characteristic sensitive information, but is more narrow and specific. In some cases, emotional charge may be considered an intrinsic subcategory of the characteristic sensitive information, potentially evoking strong personal negative emotions (Lensvelt-Mulders, 2008). An item contains a potentially emotional charge when it is about for instance a former traumatic experience or another event that the respondent fell victim to. Emotionally charged items and items asking for sensitive information may be distinguished by the idea that the former, in contrast to the latter, will probably be answered candidly. Nevertheless, when a question contains an emotional charge or word, respondents might be either reluctant or very eager to answer it (Beukenhorst et al., 2014).

2.2.1.7. Presumption of a filter question

In some surveys more than in others, certain questions may lead to follow-up items. These questions are so-called filter questions. Dependent on the content of a question and on the format of asking filter questions, respondents may presume a question to be a filter question (Eckman et al., 2014; Kreuter et al., 2011). When presuming a question to be a filter question, respondents might be motivated to give an answer that avoids them from having to answer follow-up questions (Bosley, Dashen, & Fox, 1999). The item characteristic presumption of a filter question was considered a separate characteristic by the involved researchers as a result of a pilot study (see Appendix B). The coders experienced difficulty in distinguishing an item as a true filter question versus as a question of which the respondent could presume to be a filter question, regardless of whether the question *is* a true filter question. Some respondents could avoid a filter question in case they presume a question to be one.

The remaining item characteristics that may have their influence on data quality are *time reference*, which refers to whether the item concerns the past, the present, or the future (Gallhofer et al., 2007; Saris & Gallhofer, 2007b); *mismatch*, which refers to whether the question matches its accompanying answering options (Beukenhorst et al., 2014; Smyth & Olson, 2019; Van der Zouwen, 2000); *formulation*, which refers to whether the item is formulated as a statement (Gallhofer et al., 2007; Saris & Gallhofer, 2007b), and; *clarification*, which refers to whether the item contains instruction or clarification for the respondent (Gallhofer et al., 2007; Saris & Gallhofer, 2007b; Van der Zouwen, 2000).

2.2.2. The item characteristics and mode-specific measurement error

The characteristics can all have their influence on data quality and may be associated with measurement error. This may however differ for each survey mode, possibly leading to mode-specific measurement error. Considering *question complexity*, differences in interviewer-administered versus self-administered survey modes may be expected. In interviewer-administered modes, the respondent can be assisted in answering a particular question containing some form of complexity. In self-administered modes however, the respondent does not have this assistance. Respondents can take as much time as they need to understand and answer the particular question (Beukenhorst et al., 2014), but the probability of some form of satisficing may be relatively high in self-administered modes (Krosnick, 1991). Concerning *centrality*, the respondent may be assisted or stimulated by the interviewer in interviewer-administered modes concerning topics that are not 'central' to the respondent, while this assistance or stimulation is less evident in self-administered modes.

Regarding *content of the question*, especially subjective questions are sensitive to the presence of an interviewer and may be more prone to mode-specific measurement error than factual questions (Campanelli et al., 2011; Lozar Manfreda & Vehovar, 2002; Schonlau et al., 2004). Considering *sensitive information*, interviewer-administered modes may strongly facilitate the avoidance or distortion of the respondent's answers, while this effect will be much less strong in case of self-administered modes. Therefore, this characteristic in particular is sensitive to mode-specific measurement error and may well evoke socially desirable answering (Campanelli et al., 2011; Kreuter et al., 2008; Tourangeau & Yan, 2007). In interviewer-administered modes, the interviewer may mitigate the effect of *emotional charge* by stimulating the respondent to answer in any case. In self-administered modes however, there is no interviewer present to regulate potential emotions of the respondent.

Concerning *presumption of a filter question*, respondents may be able to scroll through the survey to check for follow-up questions in mail and web mode. Filter questions that are repeated later in the survey may also be recognized more easily. In personal and telephone mode however, respondents do not have the option to scroll through the survey, making filter questions relatively more difficult to detect. It is important to note however, that we used the characteristic presumption of a filter question without considering the mode in which surveys were administered. This means that we did not account for possible mode differences concerning visual aspects or scroll-through options during the coding process. The benefit of a mode-free coding process is that items are purely judged on their content, meaning that coding results can be used regardless of the mode in which a survey is executed.

2.3. METHOD

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

2.3.1. Surveys

This coding research is based on 11 Dutch general population surveys. These are the first wave of the Dutch Labour Force Survey (LFS) administered by Statistics Netherlands and the most recent waves of the ten core studies from the Longitudinal Internet studies for the Social Sciences (LISS) of CentERdata. See Table 2.2 for an overview of these surveys with a brief description of the topics of their content and the total number of items they contain. In total, the surveys together contain 2470 items of a broad range of topics that covers virtually the whole area of general population statistics. All items of these surveys were coded by a group of survey researchers on all 16 item characteristics. In the following, we describe the steps of the coding procedure.

Table 2.2. Overview of All Surveys and a Description of Their Content.

Survey (Wave: Number of items)	*Topics of the content*
Labour Force Survey (LFS) (LFS-A: N = 123)	Education; employment and labour
Economic Situation Assets (Wave 3: N = 50)	Income, property, and investment
Economic Situation Housing (Wave 6: N = 73)	Housing and household; income, property, and investment
Economic Situation Income (Wave 6: N = 286)	Employment, labour, and retirement; income, property, and investment; social security and welfare
Family and Household (Wave 6: N = 409)	Housing and household; social behaviour
Health (Wave 6: N = 243)	Health and well-being
Personality (Wave 6: N = 200)	Psychology
Politics and Values (Wave 6: N = 148)	Politics; social attitudes and values
Religion and Ethnicity (Wave 6: N = 71)	Religion; social stratification and groupings
Social Integration and Leisure (Wave 6: N = 396)	Communication, language, and media; leisure, recreation, and culture; social behaviour; travel and transport
Work and Schooling (Wave 6: N = 471)	Education; employment, labour, and retirement

The URL to the LISS Panel Survey Data: https://www.dataarchive.lissdata.nl/study_units/view/1

2.3.2. The allocation of coders

The coding procedure consisted of three steps. First, as described in section 2.2, we set up the list of candidate item characteristics based on existing literature. Second, this tentative list was coded on a small but diverse subset of items for executing the pilot study. Based on these coding results, the list was refined and revised. Third, all items of all selected surveys were coded by either two or three coders, depending on the anticipated complexity of the coding task. Throughout these steps, the same group of survey researchers was involved. Altogether, eight researchers from Utrecht University, CentERdata, and Statistics Netherlands with knowledge of and experience with survey research were involved in coding the 11 surveys on the final 16 selected item characteristics. All coders were allocated randomly to the surveys, but each coder was assigned a different amount of surveys and survey items to code.

To each survey, two main coders were randomly allocated to code all item characteristics. A third coder was randomly allocated to code only seven specific item characteristics that appeared to be hard to code during the pilot study. Therefore, we have called these characteristics the ‘hard’ item characteristics. The hard item characteristics are content of the question, difficult language usage, emotional charge, presumption of a filter question, sensitive information, centrality, and response complexity. For reasons of clarity, we have called the remaining item characteristics that will be coded by only two coders the ‘easy’ item characteristics. The easy item characteristics are time reference, conditions, memory, hypothetical situation, calculations, ambiguity, mismatch, formulation, and clarification. All coders were instructed to abide by the agreed definitions and coding categories as strictly as possible during the coding process.

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

2.3.3. Statistics

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

The intercoder agreement for the easy item characteristics was calculated on the basis of two coders and for the hard item characteristics on the basis of three coders. Therefore, these two different kinds of intercoder agreement are not directly comparable. Here, it seems logical to calculate Fleiss' kappa, which is an indicator of the interrater agreement between multiple coders. Fleiss' kappa incorporates a correction for the degree of agreement that may be expected by chance alone (Fleiss, 1971). However, we do not believe that the coding of items by coders involves an element of chance. The coders were instructed on the coding procedure precisely and are assumed to have coded conscientiously and consistently. This means that differences between coders are real differences in the sense that the coders considered the item characteristics differently for certain items, based on their own perspective. Therefore, we did not use Fleiss' kappa, but instead calculated the fixed probability λ that a coder correctly indicates the true category for an item characteristic. This probability was calculated on the basis of the accompanying intercoder agreement for the concerned item characteristic. Then, the probability λ for an item characteristic is the number of correctly coded items divided by the total number of all items. For this calculation, we assumed that each coder acted independently and that this probability is the same for each coder. See Table 2.5 in section 2.4.2 for the probability λ and its accompanying intercoder agreement for each item characteristic. See Appendix C for an elaboration on the probability λ and see Table C.1 in Appendix C for an overview of specific values for the probability λ and its accompanying intercoder agreement for two or three coders.

2.4. RESULTS

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

2.4.1. Relative frequencies

Three coders were assigned to each survey, meaning that 33 sets of coding data for 11 surveys were collected. For each survey, this consisted of two sets of coding data for all item characteristics and one set of coding data for only the seven so-called hard item characteristics. For each coding category, we calculated the relative frequencies for all item characteristics. The calculations were done over all surveys, giving an overview of these frequencies for the broad range of all 11 surveys together in proportions. See Table 2.3 for the overall relative frequencies for the item characteristics with more than two coding categories. Over all surveys, all categories were coded to at least some extent. Factual questions (content of the question), questions for which no memory was needed (memory), and questions about the present (time reference) were coded most frequently. Questions that ask for a degree of satisfaction (content of the question), questions about events from the past one month (memory), and questions about the future (time reference) were coded relatively infrequently.

Table 2.3. The Relative Frequencies of the Coding Categories for the Item Characteristics Content of the Question, Memory, and Time Reference over All Surveys (2470 Items).

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

See Table 2.4 for the item characteristics with only two coding categories. Over all surveys, the category indicating that the characteristic is applicable was coded to at least some extent for each characteristic. The applicability of an item being formulated as a statement and an item containing some form of clarification were coded most frequently. Complexity of the answering options, questions about a hypothetical situation, ambiguous questions, and questions being a mismatch were coded relatively infrequently. The lowest proportion of 0.02 for questions being a mismatch indicates an applicability of still roughly 40 items of all survey items per coder on average. Because of this substantial amount of items, we decided to include all item characteristics and their coding categories in further analyses.

Table 2.4. The Relative Frequencies for the Item Characteristics with Two Coding Categories over All Surveys.

Item characteristic	*Applicability characteristic*	*Item characteristic*	*Applicability characteristic*
Conditions	0.14	Difficult language usage	0.19
Hypothetical situation	0.03	Emotional charge	0.12
Calculations	0.20	Presumption of filter question	0.26
Ambiguity	0.02	Sensitive information	0.25
Mismatch	0.02	Centrality	0.21
Formulation	0.31	Response complexity	0.04
Clarification	0.36		

2.4.2. Intercoder reliabilities

Following this overview of the relative frequencies of the item characteristics over all surveys together, we now deal with our first research question and present to what extent coding of these item characteristics is actually reliable. As a rule of thumb and for reasons of convenience, we consider proportions of 0.80 and higher as reasonably high intercoder reliability and proportions of 0.79 and lower as low intercoder reliability. Therefore, we focus on proportions below 0.80 when we try to explain potential low intercoder reliability. For clarity reasons, we present the intercoder reliabilities for the hard and easy item characteristics separately. See Table 2.5 for the intercoder reliabilities for the easy item characteristics on the left side and the hard item characteristics on the right side of the table. Regarding the hard item characteristics, see Table 2.6 for the intercoder reliabilities for the three pairs of coders.

2.4.2.1. Intercoder reliabilities for the easy item characteristics

As can be seen in the left part of Table 2.5, the intercoder reliabilities for most easy item characteristics were reasonably high, indicating that coding of these item characteristics can be done relatively reliably. For the item characteristics formulation and clarification however, low intercoder reliabilities were evident. Although formulation and clarification were defined as easy item characteristics and thus coded by only two coders, coding of these two item characteristics could not be done reliably. This means that coders did often not agree on whether

Table 2.5. Intercoder Reliabilities for the Easy and Hard Item Characteristics (and Their Fixed Coder Probability λ).

Easy item characteristics	*Intercoder reliability*	*Hard item characteristics*	*Intercoder reliability*
Time reference	0.85 (.92)	Content of the question (5 cat.)	0.56 (.82)
Conditions	0.89 (.94)	Content of the question (2 cat.)	0.90 (.97)
Memory	0.85 (.92)	Difficult language usage	0.61 (.85)
Hypothetical situation	0.98 (.99)	Emotional charge	0.75 (.91)
Calculations	0.94 (.97)	Presumption of filter question	0.62 (.85)
Ambiguity	0.96 (.98)	Sensitive information	0.53 (.81)
Mismatch	0.98 (.99)	Centrality	0.59 (.84)
Formulation	0.57 (.68)	Response complexity	0.91 (.97)
Clarification	0.71 (.82)		

Table 2.6. The Intercoder Reliabilities for the Three Pairs of Coders for the Hard Item Characteristics.

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

the concerned item was formulated as a question or a statement, and whether it contained a clearly present clarification or not.

2.4.2.2. Intercoder reliabilities for the hard item characteristics

For the item characteristic content of the question, a second kind of intercoder reliability was calculated to investigate to what extent this characteristic could be coded reliably with only an objective and a subjective category. For this specific intercoder reliability, the categories ‘factual behaviour’ and ‘otherwise factual’ were merged into one overall *objective* category, and the categories ‘opinion’, ‘satisfaction’, and ‘otherwise subjective’ were merged into one overall *subjective* category. As can be seen in the right part of Table 2.5, for the initial item characteristic content of the question, the intercoder reliability was relatively low. For content of the question with merely the objective and subjective category however, the intercoder reliability was reasonably high. This indicates that this item characteristic could not be coded reliably with five subcategories, but could be coded reliably when only one objective and one subjective category were used. For the items for which no consensus was found, this means that coders usually agreed on whether an item contained either objective or subjective content, but did often not agree on the category within the objective or subjective content.

As can be seen in the right part of Table 2.5, the intercoder reliabilities for most other hard item characteristics were also relatively low, indicating that coding of these item characteristics cannot be done reliably. For relatively many items, this means that coders did often not agree on when an item contained unknown or difficult words, or complex sentences (difficult language usage), when an item was about a topic or contained words that could evoke an emotional reaction (emotional charge), when an item could make respondents presume that follow-up questions might result depending on the answer they would give (presumption of a filter question), when an item asked for some kind of sensitive information so that it may evoke socially desirable answer behaviour (sensitive information), or when an item was difficult to answer as it goes beyond the interest, knowledge, or experience of the respondent (centrality). In the following section, we try to explain low intercoder reliability for the concerned item characteristics.

2.4.3. Explaining low intercoder reliability

Following this overview of the intercoder reliability statistics, we now deal with our second research question and try to explain the low intercoder reliabilities that we found. Overall, the interaction of two related key factors is probably associated with the obtained low intercoder reliabilities. First, we briefly discuss these key factors to indicate the difficulty in obtaining reasonably high intercoder reliabilities. Second, with the two key factors in mind, we discuss the item characteristics that had a fixed coder probability λ below the value of 0.90 (see section 2.3.3 and Table 2.5 in section 2.4.2). We do not believe that coders had the same coding probabilities nor that the correct probabilities are equal for each category, but the criterion allows for a more objective and intuitive decision. See Appendix C and Table C.1 for a brief explanation. Regarding the hard item characteristics, we also discuss those characteristics that had an intercoder reliability below the value of 0.80 for at least one of the three pairs of coders (see Table 2.6).

2.4.4. Key factors associated with low intercoder reliability

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

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

2.4.5. Explaining low intercoder reliability

2.4.5.1. Formulation and clarification

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

2.4.5.2. Content of the question

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

2.4.5.3. Difficult language usage

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

2.4.5.4. Emotional charge

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

2.4.5.5. Presumption of a filter question

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

2.4.5.6. Sensitive information

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

2.4.5.7. Centrality

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

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

2.5. COPING WITH LOW INTERCODER RELIABILITY

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

2.5.1. Option 1: Excluding survey items

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

2.5.2. Option 2: Redefining and refining item characteristics

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

2.5.3. Option 3: Computerizing the definition and demarcation of item characteristics

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

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

2.5.4. Option 4: Using item characteristic scales with multiple applicability categories

For a way to avoid redefining and re-demarcating the item characteristics or programming strict rules for the coding procedure, a fourth option for coping with low intercoder reliability is to construct scales with multiple applicability categories for the item characteristics with two categories that are coded by three coders. Let us consider the item characteristic presumption of a filter question here. This characteristic was coded by three coders, meaning that either no, one, two, or three coders indicated its applicability for a certain item. Based on all items for which either no, one, two, or three coders indicated the characteristic's applicability, a questionnaire scheme consisting of four respective categories could be constructed. Then, for the items of a survey, the characteristic presumption of a filter question is expressed on a gradual scale with four applicability categories, rather than on a dichotomous scale with only the categories applicable and not applicable. This scheme can be used to investigate to what extent it explains variation in the influence of this item characteristic on evoking measurement error. For instance, consider items that were coded as presumed to be a filter question by three coders versus two coders. Here, the influence on evoking measurement error may appear relatively larger for items for which all three coders versus for items for which only two coders presumed them as filter questions. Exactly the same may be true for two coders versus one coder and for one coder versus no coders. In this way, the relative influences of each of these four categories can be compared directly to check for their potential different relations to the occurrence of measurement error.

To be able to investigate and compare the categories of such an applicability scale, each category should contain enough items to base its scheme on. For the current study, we calculated the relative frequencies of each category for all item characteristics with two coding categories that were coded by three coders. As can be seen in Table 2.7, the applicability of the item characteristics is coded by all three coders for only relatively few items. Hence, it may not be feasible to construct a scale for all four category schemes, as relatively few items may not contain enough power to expose potential measurement error. Here, an alternative option might be to pool the two categories with two and three coders into a single third category. Then, this third category may contain enough items and will consist of all items that were coded as applicable to the concerned item characteristic by at least two coders.

Table 2.7. Relative Frequencies of the Applicability of the Hard Item Characteristics with Two Coding Categories for the Number of Coders over All Surveys.

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

2.6. DISCUSSION

In this study, we used a systematic coding procedure to code all 2470 items of 11 Dutch surveys on 16 item characteristics that we expected to be relevant in evoking measurement error according to the literature. We have investigated to what extent the coding of these item characteristics could be done by multiple coders *reliably*. In case of reasonably high intercoder reliability, this would be indicative for an unambiguous relation between item characteristic, item content, and measurement error. Hence, so-called questionnaire schemes may be constructed, which summarize the characteristics of the items of a survey. If questionnaire schemes could be identified and would appear to be related to varying answer behaviour of the part of the respondent, they might be helpful in controlling for measurement error. In case of relatively low intercoder reliability however, questionnaire schemes cannot be constructed without difficulty. Low intercoder reliability would then need to be explained and suggestions should be made for coping with low intercoder reliability.

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

At the start of our study, we distinguished item characteristics that were coded by either two or three coders. In principle, we wanted the characteristics to be coded by two coders, but we assigned a third coder to characteristics that appeared to be hard to code during the pilot study. Considering the study results, the intercoder reliability for characteristics coded by three coders was generally lower than for characteristics coded by two coders. However, it is difficult to say to what extent this can be explained by the different degree of difficulty of coding the characteristics versus to what extent this can be attributed to the different number of coders; the characteristics coded by three coders may actually have been relatively more difficult to code, but it is also obvious that consensus decreases as more coders are involved. First, the fixed intercoder probabilities for most characteristics coded by three coders were clearly *below* the value of 0.90 that we set as a minimum as a reasonable intercoder probability, while the fixed intercoder probabilities for most characteristics coded by only two coders were clearly *above* this value (see Table 2.5 in section 2.4.2 and Table C.1 in Appendix C). Second, for most characteristics coded by three coders, the intercoder reliabilities for all three pairs of coders showed that either one, two, or all three pairs of coders had an intercoder reliability *below* the value of 0.80 that we set as a minimum for reasonable intercoder reliability (see Table 2.6 in section 2.4.2). Based on both the intercoder probabilities that are assumed to be fixed and equal for each coder, and the intercoder reliabilities for the pairs of coders, this means that characteristics coded by three coders were indeed relatively more difficult to code.

It must be noted that, according to the coders, the occurrence of some characteristics was relatively rare (see Table 2.4 in section 2.4.1). The rareness of a characteristic is logically related to the intercoder reliability of a characteristic. For instance, let us consider the characteristic mismatch with an intercoder reliability of 0.98 and a relative frequency of 0.02. This means that, for almost all items, both coders did not indicate its applicability, explaining the high intercoder reliability of 0.98. Thus, for the remaining two percent of all items, one of the two coders indicated the applicability of the characteristic mismatch and the other coder did not. In fact, there were *no* items at all for this characteristic for which both coders indicated the applicability. This means that the high intercoder reliability for this characteristic is solely based on the majority of items for which both coders did not indicate the applicability. In short, when a characteristic appears to be rare, a high intercoder reliability is a logical result and may mask a low consensus for those items on the boundary of having the characteristic.

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

ρ
$\rho \quad \sigma$
$\rho \quad \sigma \quad \tau$
$\rho \quad \sigma \quad \tau \quad \varphi$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta \quad \rho$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta \quad \rho \quad \sigma$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta \quad \rho \quad \sigma \quad \tau$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta \quad \rho \quad \sigma \quad \tau \quad \varphi$
$\rho \quad \sigma \quad \tau \quad \varphi \quad \theta \quad \rho \quad \sigma \quad \tau \quad \varphi \quad \theta$

Chapter 3

Investigating response patterns across surveys: Do respondents show consistency in undesirable answer behaviour over multiple surveys?

This chapter (additional section 3.6 excluded) is published in Bulletin de Méthodologie Sociologique (BMS) as:
Bais, F., Schouten, B., & Toepoel, V. (2020). Investigating response patterns across surveys: Do respondents show consistency in undesirable answer behaviour over multiple surveys? *Bulletin de Méthodologie Sociologique, 147-148*(1-2), 150-168. https://doi.org/10.1177/0759106320939891

Author contributions: FB, BS, and VT designed the study. CentERdata and Statistics Netherlands provided the survey data. FB prepared the data, performed the statistical analyses, and wrote the paper. BS and VT critically reviewed the paper.

Abstract

The relation between answer behaviour and measurement error has been studied extensively. Answer behaviour may be considered undesirable, like answering ‘don’t know’ or ‘won’t tell’. It is not clear to what degree undesirable answer behaviour from the same respondents is present across different surveys. In this study, we investigated to what extent respondents show undesirable answer behaviours consistently over multiple surveys. First, we investigated to what extent the answer behaviours occurred in ten large general population surveys of CentERdata and Statistics Netherlands. Second, we explored the respondent variances and respondent-survey interaction variances to obtain an indication for respondent consistency for each answer behaviour. The results showed that respondents only occasionally give ‘don’t know’- and ‘won’t tell’-answers. An indication for respondent consistency was found for fast responding, slow responding, and ‘won’t tell’-answers in particular. We recommend follow-up research to investigate the relation between respondent characteristics and consistent answer behaviour.

3.1. INTRODUCTION

The relation between survey answer behaviour and measurement error has been studied extensively. Various forms of answer behaviour are considered undesirable, like answering 'don't know' (Beatty & Herrmann, 2002; Shoemaker et al., 2002). The extent to which undesirable answer behaviour of the same respondents is present across different surveys is unclear. When a respondent only incidentally shows a behaviour, for example for one specific survey, it is not to say whether that behaviour is typical for that respondent. The behaviour may then be the result of taking shortcuts in the question-answering process, also known as satisficing (Krosnick, 1991, 1999; Krosnick et al., 1996), but may just as well be truly attributed to the respondent. When a respondent shows the behaviour across different and multiple surveys, it becomes more likely that the behaviour is typical for that specific respondent. In other words, the respondent may have a stable personal tendency to show specific undesirable behaviour, regardless of survey topic or design. *In this study, we investigate to what extent respondents show consistency in expressing potentially undesirable answer behaviour across multiple surveys.* The term 'consistency' refers to a pattern of answer behaviour that is shown over several moments in time, across multiple surveys. To answer our research question, we use a large number of panel respondents and a large number of different surveys.

We need to note that we explore consistent answer behaviour across surveys without considering characteristics of respondents, surveys, or survey items: We are neither trying to identify individual or subgroups of respondents, nor taking into account survey topic or design. First, by including many different surveys, variation will be present in survey topic and design. As a result of this variation, we assume that each survey has its own specific effect on the answer behaviours. In our study, we want to distinguish respondent behaviour that is survey-specific from behaviour that occurs consistently across surveys. In order for respondent consistency to appear, it needs to predominate a single survey topic or design effect by definition. In other words, we need the full presence of topic and design variability to investigate behaviour consistency across various surveys. Thus, for the purpose of our study, it is sufficient to presume variability in survey topic and design without taking these into further account.

Second, in case of little or no respondent consistency, the presence of specific types of respondents showing specific behaviour across surveys is unlikely. It may then not be recommendable to further investigate the role of respondent characteristics in relation to consistent undesirable answer behaviour and measurement error. Only in case of a substantial degree of respondent consistency in showing specific behaviour, it may be worthwhile for further research to investigate to what extent subgroups of respondents share the same stable characteristics. Therefore, we want to execute an exploration of respondent consistency without differentiating between types of respondents first. Subsequently, future research may focus on the relation between stable respondent characteristics and undesirable answer behaviour across surveys. This relation could then be used by survey constructors and administrators to control for measurement error.

It is important to emphasize that when we speak of undesirable answer behaviour, we refer to *potentially* undesirable answer behaviour. Our idea is that the more consistently the behaviour is present over the more surveys, the more likely it is that the behaviour is typical. In this study, we only include answer behaviours that are relatively straightforward to consider in checking the answering options from survey response data. For instance, answering don't know and primacy responding can be coded by respectively selecting the 'don't know'-answers and the first two options of an answering scale. To empower finding potential respondent consistency, we use ten large national population computer-assisted surveys administered by CentERdata in the LISS Panel. These surveys vary broadly in topic and contain many different kinds of items. With this broad range of surveys, we aim to present a complete overview of the prevalence of several typical answer behaviours in web surveys in a panel.

First, we investigate the frequency of potentially undesirable answer behaviours in the different surveys. Second, we investigate to what extent respondents are consistent over multiple surveys in showing these behaviours. We analyze various forms of answer behaviours, as motivated in section 3.2. In section 3.3, we present the included surveys of Statistics Netherlands and the LISS Panel, and elaborate on the programming and operationalization procedure, the consistency methodology, and the statistics. In section 3.4, we show all statistical results and give answers to our research questions. In section 3.5, we conclude with a discussion of these results and make suggestions on how to proceed.

3.2. ANSWER BEHAVIOURS: SELECTION AND MOTIVATION

In this section, we elaborate on seven relevant answer behaviours, selected from the literature: Answering 'don't know', answering 'won't tell', primacy responding, recency responding, straightlining, slow responding, and fast responding. Below, we give a short overview of the literature for each answer behaviour. We motivate their inclusion by elaborating on why they may be referred to as undesirable and how they may be related to measurement error.

3.2.1. Answer behaviours

3.2.1.1. Answering 'don't know' and 'won't tell'

The answering options 'don't know' or 'won't tell' are often added to the 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 and that questions requiring more cognitive effort are likely to receive more don't knows (Shoemaker et al., 2002). On the one hand, respondents may not give a substantive answer in case; they are relatively inexperienced as a respondent (Binswanger et al., 2013); they are reluctant or lacking motivation to answer (Beatty & Herrmann, 2002; Krosnick et al., 2002; Roßmann, Gummer, & Silber, 2018); items ask for sensitive information (Bradburn et al., 1978; Tourangeau et al., 2000). On the other hand, respondents may give an actual answer without knowing the answer or having an opinion (Beatty & Herrmann, 2002; Bishop, Tuchfarber, & Oldendick, 1986). This implies that a non-response option should only be included when deemed a realistic plausible option (Vis-Visschers, Arends-Tóth, Giesen, & Meertens, 2008). In these situations, measurement error can be the result, which may be mode-dependent (Fricker, Galesic, Tourangeau, & Yan, 2005; Roberts, 2007).

3.2.1.2. Primacy and recency responding

Depending on the order in which answering options are offered, response order effects may occur; an increased probability that an option will be chosen at the beginning or end of a list, respectively called a primacy and a recency effect (Krosnick & Alwin, 1987). Response order effects occur when respondents do not give equal consideration to all the response alternatives (Krosnick & Alwin, 1987; McClendon, 1991). Primacy effects may be expected in self-administered surveys, case items are presented visually (Galesic, Tourangeau, Couper, & Conrad, 2008; Krosnick, 1991; Krosnick, 1992; Krosnick & Alwin, 1987). Recency effects may be expected in interviewer-administered surveys, in case items are presented orally (Krosnick, 1991; Krosnick, 1992; Krosnick & Alwin, 1987). Both situations may lead to measurement error (see Galesic et al., 2008). Order effects may affect respondents on different levels of their characteristics in similar ways, but in different degree (Stern, Dillman, & Smyth, 2007).

3.2.1.3. Straightlining

Questions followed by a common answering scale are often clustered together (Krosnick, 1991). This may lead 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 (Díaz de Rada & Domínguez, 2015; Schonlau & Toepoel, 2015). Straightlining seems more common towards the end than the beginning of a questionnaire (Krosnick, 1991). It tends to increase for respondents who give answers very quickly or 'speed' (Zhang, 2013; Zhang & Conrad, 2014) or had relatively longer panel experience (Schonlau & Toepoel, 2015). The behaviour may partly be dependent on the type of survey topic or question (Schonlau & Toepoel, 2015), on whether the questions are constructed in grid versus item-by-item design (Roßmann et al., 2018), and on survey mode (Fricker et al., 2005).

3.2.1.4. Fast and slow responding

Response times might be indicative for problems in the response process and linked to measurement error (Olson & Smyth, 2015; Yan & Olson, 2013). Spending relatively much or little time in filling out a survey may be of concern. Filling out survey items very rapidly, known as 'speeding' (Greszki, Meyer, & Schoen, 2015; Zhang & Conrad, 2014), may refer to poor processing of the items (Malhotra, 2008) and is associated with lower data quality (Revilla & Ochoa, 2014). Filling out survey items very slowly may mean that respondents have potential difficulty in processing the items (Couper & Kreuter, 2013; Yan & Tourangeau, 2008). The interview duration in web surveys is dependent on several factors on both the survey level and the respondent level (Gummer & Roßmann, 2014). Variation in interview duration can be explained by internet experience (Yan & Tourangeau, 2008) and longer online survey experience (Toepoel, Das, & Van Soest, 2008; Yan & Tourangeau, 2008).

3.2.2. Hypotheses

As far as we know, respondent consistency has not been investigated by means of a panel consisting of many respondents across the broad range of ten national population surveys. For the purpose of this study, we assume that showing undesirable answer behaviour may be attributed substantially to part of the respondent. This means that we presume that respondents either do not show or do show a specific behaviour consistently across surveys. Therefore, in case an answer behaviour occurs in the applicable surveys to some extent at all, we expect the behaviour to occur consistently across the surveys to a substantial degree.

3.3. METHOD

3.3.1. Surveys and LISS Panel

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 3.1.

Table 3.1. Overview of All Surveys, a Description of Their Content, and Their Response Rates (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)

The URL to the LISS Panel Survey Data: https://www.dataarchive.lissdata.nl/study_units/view/1

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 first administration period for each survey was approximately a month. In case of initial non-response, the respondent was sent one or two reminders within this period. To increase the response rate, a second administration period of about a month including one or two reminders was executed for each survey. The respondents were compensated for each survey that they completed. 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 almost eight. Altogether, the surveys contain 2074 items that were used to cover the possible answer behaviours as presented in section 3.2.

3.3.2. Coding the answer behaviours

To come to the factual occurrence of the behaviours, all items of the ten surveys were investigated on whether they were eligible for the selected answer behaviours first. For instance, this means that only items that actually present the don't know-option are eligible for the behaviour 'answering don't know'. For the items that appeared eligible, the accompanying answering categories were coded for each specific answer behaviour. In case a category was filled out for which the behaviour occurred, the response was coded as 1; in case a category was filled out for which the behaviour did not occur, the response was coded as 0. For all behaviours, the coding was relatively straightforward. For instance, for answering 'don't know' and 'won't tell', the don't know- and won't tell-options were coded as 1, while all other options were coded as 0. For primacy and recency responding, the first two and the last two options respectively were coded as 1, while all other options were coded as 0. The coding method was based on Medway and Tourangeau (2015).

After coding all eligible items for all behaviours, the relative frequencies for the behaviours were calculated. A relative frequency refers to the proportion of behaviour occurrence. For instance, consider a respondent who fills out 100 eligible survey items for the behaviour 'answering don't know'. And suppose that the respondent fills out 'don't know' for 20 of these 100 items and a non-'don't know'-answer for the other 80 items. Then the relative frequency is 20/100 or 0.20. See Table 3.2 for an overview of the answer behaviours, the kind of eligible items, and the operationalization for each behaviour in calculating the relative frequencies. See Table D.1 in Appendix D for the numbers and proportions of items for which each answer behaviour is applicable for each survey and in total. From here, we discuss the coding process of the answer behaviours that need more elaboration.

3.3.2.1. Straightlining: Choosing the same answering category for all items in a battery

All batteries for all surveys were investigated on straightlining. Our idea was to consider straightlining for a battery only when the very same answering options were filled out *for all*

Table 3.2. The Kind of Answer Behaviour, the Kind of Items Eligible for the Answer Behaviour, and the Operationalization of the Answer Behaviour.

Answer behaviour	*Eligible items*	*Operationalization*
Answering 'don't know'	All items containing a 'don't know' answering 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'	All items containing a 'won't tell' answering category	Number of items for which 'won't tell' was filled out divided by Number of actually filled out 'won't tell' items
Primacy responding	All (battery) items containing at least four response options	Number of (battery) items for which the first or second answering category was filled out divided by Number of actually filled out eligible (battery) items
Recency responding	All (battery) items containing at least four response options	Number of (battery) items for which one of the last two answering categories was filled out divided by Number of actually filled out eligible (battery) items
Straightlining	The items of all batteries containing at least 3 items and at least 4 answering categories, only in case all items of the battery were actually filled out	Number of filled out battery items for which the same answering category was filled out for a complete battery divided by Number of actually filled out eligible battery items
Fast responding	All items	0-1 indicator per survey per respondent for average duration per item being smaller than the 10% quantile over all items answered by all respondents
Slow responding	All items	0-1 indicator per survey per respondent for average duration per item being larger than the 90% quantile over all items answered by all respondents

its items (see Schonlau & Toepoel, 2015). When this is the case, the number of times that a '1' is coded is equal to the number of items that the battery consists of. For instance, the occurrence of straightlining for a battery of five items was assigned the code '1' five times. This means that we took into account the length of the battery for this behaviour. Note that we considered the concept of straightlining as it is defined by Krosnick (1991), as the full non-differentiation between the categories of a set of battery items. We did not consider diagonal answering patterns in batteries, as we argue that responding in a straight diagonal line involves relatively more effort for the respondent and may not occur very often. We also did not consider seemingly random answering patterns in batteries, as such patterns are difficult to distinguish from non-random patterns.

3.3.2.2. Responding faster or slower than average

We were able to determine the duration of filling out an item of a survey for respondents on average, but we did not have the duration of filling out each specific item separately. This means that we did not consider response time of particular items, but only of each survey as a whole. We constructed an absolute response time threshold for fast and slow responding that was the same for each survey. We chose these thresholds so that the proportion of 'fast' and 'slow' respondents differed per survey and was roughly 0.10 on overall average. We chose to use overall quantiles instead of survey-specific quantiles to avoid a survey variance of zero in the analyses.

3.3.3. Model for variance decomposition and consistency indication

First, the number of eligible items differs per behaviour and per survey. Second, most surveys contain routing of questions. This means that the number of responses differed per respondent and per survey. These responses, consisting of 0's and 1's, were used for the analyses per answer behaviour. Therefore, we chose for logistic data analyses with answer behaviour as the dichotomous dependent variable. We considered both 'respondent' and 'survey' as two higher order levels that are not unambiguously hierarchical or nested, but that are cross-classified (Hox, 2010). This means that we used two-way cross-classified logistic multilevel models to estimate the variance components of the random intercepts for respondent, survey, and respondent-survey interaction (simply called 'interaction' from here).

The larger the respondent variance and survey variance, the more respondents and surveys respectively differ in showing specific answer behaviour. The interaction variance refers to the extent to which the occurrence of behaviour for respondents is dependent on the survey level. Hence, a relatively high interaction variance means that the occurrence of that behaviour differs for respondents and, depending on the respondent, also differs for surveys. To discover potential respondent consistency, the behaviour must be shown; 1) to a certain degree at all, and; 2) for relatively many to all surveys. In the scenario of ultimate consistency, respondents either rarely or frequently show the behaviour over all (or many) surveys. Hence, in order to find such consistency, answer behaviour should *not* be dependent on the survey level. This means that interaction variance needs to be absent or low relative to respondent variance. *In summary, an indication for respondent consistency in answer behaviour over surveys would be relatively large respondent variance and relatively small interaction variance*. Note that we are not interested in behaviour that is consistently rare or absent, but in behaviour that is clearly present across surveys. Only undesirable answer behaviour that actually and consistently occurs is of concern considering potential measurement error.

As we want to obtain these variances and give an indication about respondent consistency across surveys, note that we assume that our surveys are a representative sample of all possible surveys. We justify this assumption by emphasizing that we have the substantial number of ten large surveys that broadly vary in topic. As we stated in the introduction, we deliberately do not control for survey topic or design effect. By our multilevel model, we obtain pure variances for respondents, surveys, and their interactions for an indication of behaviour consistency across surveys that vary in topic and design. Again, also note that we deliberately do not incorporate respondent characteristics into our multilevel model. This study is to explore the presence of respondent consistency without differentiating between types of respondents.

In our study, consistency refers to showing about the same relatively high occurrence for an answer behaviour over multiple surveys and within each survey over multiple items. One cannot speak of consistent answer behaviour over surveys by including only one survey or only a few items. Therefore, for each separate answer behaviour and analysis, we chose to include respondents who filled out at least two surveys, for which they filled out at least five items each. For instance, the behaviour 'answering won't tell' can occur only in the surveys Assets, Housing, Income, Work and Schooling, and the Labour Force Survey. Only respondents who filled out at least five eligible items from at least two of these surveys are included for the analysis of this behaviour.

3.3.4. Statistical analyses

We calculated the relative frequencies of each answer behaviour in proportions per survey and in total over all surveys together. In calculating the relative frequencies for the total over all surveys together, we considered all items of all surveys as if they were all part of one large survey. For response time, we did not calculate average response times because of extreme outliers that refer to respondents who distributed filling out a survey over multiple days. While including all outliers, we calculated the median response time for each survey and in total. For fast and slow responding, we calculated the proportions of respondents answering faster or slower than the fixed absolute response time threshold per survey and in total over all surveys together. For the total over all surveys, all items of all surveys were considered together as if they were all part of one large survey. As each survey differed in the amount of time that was needed to answer an average item, the fixed threshold ensures survey variation in the proportions of fast- and slow-answering respondents.

By means of the multilevel models, we calculated the respondent, survey, and interaction random intercepts, and their profile likelihood-based confidence intervals (see Venzon & Moolgavkar, 1988). Profile likelihood-based confidence intervals examine the likelihood of a single parameter separately, holding all other model parameters constant. The intervals refer to a robust method for estimating non-symmetric likelihood functions about the variances (Joshi, 2015). Finally, we calculated the intraclass correlation coefficients (ICCs) to present the proportions of explained variance. This is done for each variance component and answer behaviour separately over all surveys.

3.4. RESULTS

Before answering our research questions, we address the potential influence of the administration period during which respondents filled out the surveys on the results. The idea is that respondents who filled out all or most of their surveys during the second administration period may deviate in their answer behaviour from respondents who did so during the first administration period. Concretely, it may be that respondents who repeatedly need multiple invitations/reminders to fill out a survey may show other or more undesirable response patterns than respondents who do not repeatedly need multiple invitations/reminders. Therefore, we briefly explored the percentages of surveys that were filled out during the second administration period for each respondent. Here, a relatively large group of respondents who filled out all or most of their surveys during the second administration period may be of concern. The exploration shows that only 0.1%, 0.2%, and 1.3% of all respondents filled out all, more than 75%, and more than 50% of their surveys respectively in the second administration period. We consider these percentages too small to make a significant potential difference with respect to our study results. Therefore, we decided not to further investigate the influence of administration period on the results.

From here, we elaborate on the occurrence of answer behaviours by evaluating the observed relative frequencies of the behaviours resulting from Table 3.2. See Table D.1 in Appendix D for the maximum proportions of items for which each answer behaviour can factually occur. See Table 3.3 for the relative frequencies of all answer behaviours per survey and in total.

Table 3.3. Relative Frequencies for the Answer Behaviours and Median Response Times (in Seconds) for All Surveys* and in Total (TT).

Answer behaviour	*AS*	*FA*	*HE*	*HO*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*	*TT*
Answering 'don't know'	.37**	.06	-	.12	.20	-	.09	.00	.09	-	.08
Answering 'won't tell'	.15	-	-	.04	.07	-	-	-	.41	.03	.05
Primacy responding	-	.40	.65	-	.39	.25	.19	.69	.32	.47	.33
Recency responding	-	.30	.14	-	.27	.37	.19	.25	.29	.42	.28
Straightlining	-	.15	.43	-	.29	.03	.10	.23	.01	-	.12
Median response time; sec./item	6.6	6.6	5.0	6.9	8.5	6.2	6.3	4.2	6.1	8.9	7.4
Fast responding ***	.11	.05	.23	.11	.08	.08	.08	.34	.13	.01	.13
Slow Responding ***	.12	.08	.07	.14	.24	.12	.09	.03	.12	.15	.11

* Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).

** Consider this proportion of .37 as a clarifying example. This proportion means that for 37% of all items of the survey Assets for which a 'don't know'-answer was an option, this option was actually chosen.

*** Fast and slow responding concern the respondent proportions with the respective response times of < 3.5 seconds per item and > 15 seconds per item.

Note. Regarding the answer behaviours except for fast and slow responding, the relative frequencies are based on respondents who filled out at least five items for a specific behaviour and survey, and who did so for at least two surveys for a specific behaviour, to remain consistent with the final analyses. This means that the relative frequencies of respondents who did not fill out at least five items for a specific behaviour for at least two surveys were omitted. Consequently, some relative frequencies may be biased compared to the factual overall relative frequencies, although to a restricted extent.

3.4.1. Prevalence of answer behaviour

See Table 3.3. Overall, some answer behaviours were filled out more frequently than others and the relative frequencies clearly differ per survey. Answering don't know and won't tell did not occur frequently. In general, respondents had the tendency to answer don't know in the surveys Assets (AS) and Income (IN). These surveys may be relatively difficult to fill out. Further, it is striking that the relative frequency of answering don't know is almost zero in the survey Religion and Ethnicity (RE). It is not clear how this may be explained. The presence of answering won't tell was clear in the surveys Assets (AS) and especially Work and Schooling (WO). This seems plausible considering the sensitive topics in the former, but is less clear concerning the latter. Primacy and recency responding appeared relatively frequently overall. Straightlining occurred relatively frequently in the surveys Health (HE), Income (IN), and Religion and Ethnicity (RE).

See Table 3.3 for the median response time in seconds per item and for the proportions of respondents who were faster than 3.5 seconds and slower than 15 seconds per item in filling out a survey respectively. These thresholds roughly conform to the 10% and 90% quantiles of response times over all surveys and respondents. The median response time per item was relatively high for the survey Income (IN) and the Dutch Labour Force Survey (LF). This may be explained by topic complexity for the former and by relatively many difficult terms in both questions and answering categories for the latter. The survey Income (IN) stands out with a high respondent proportion of .24 that answered slower than 15 seconds per item. A low respondent proportion of .01 answered faster than 3.5 seconds per item for the Dutch Labour Force Survey (LF).

The median response time per item was relatively low for the surveys Health (HE) and Religion and Ethnicity (RE). This may be explained by a relatively good health and the non-religious nature of many respondents respectively. These may result in rapid negative responses about illness and religion respectively for many batteries of items in these surveys. High respondent proportions of .23 and .34 answered faster than 3.5 seconds per item for the surveys Health (HE) and Religion and Ethnicity (RE) respectively. A low respondent proportion of .03 answered slower than 15 seconds per item for the survey Religion and Ethnicity (RE).

3.4.2. Variance components and behaviour consistency

In this section, we elaborate on the model variance components and respondent consistency for each behaviour. First, we compared various multilevel models to test whether variance components should be included into the model. See Appendix E for an elaboration on these models. See Table E.1 in Appendix E for the results of the likelihood ratio tests (LRTs) for the model comparisons. See Table E.2 in Appendix E for the random intercepts and their profile likelihood-based confidence intervals of the final models. See the figures in Appendix F for the estimation and distribution of the random intercepts for the final models. Second, we calculated the intraclass correlation coefficients (ICCs) for the final models to present their proportions of explained variances. See Table 3.4. Note that the ICC for a single variance component coincides with the variance partition coefficient (VPC) -the proportion of explained variance- in multilevel models with only random intercepts (see Leckie, 2013). See Appendix E for an elaboration on calculating the ICCs. Also note that a separate multilevel model is used for the analysis of each answer behaviour. This means that for each behaviour, only those surveys that contain sufficient eligible items for that specific behaviour are included in the analysis.

Table 3.4. Proportions of Explained Variances (ICCs) for the Answer Behaviours for the Final Models over All Applicable Surveys Together.

Behaviour (nr. of respondents)	*Explained respondent variance (ICC)*	*Explained survey variance (ICC)*	*Explained interaction variance (ICC)*
Answering ‘don’t know’ (6145)	.12*	.38	.13
Answering ‘won’t tell’ (3682)	.37	.12	.15
Primacy responding (6319)	.01	.15	.03
Recency responding (6319)	.01	.06	.04
Straightlining (6234)	.06	.46	.22
Fast responding (6697)	.48	.18	NA**
Slow responding (6697)	.25	.10	NA**

* Consider this value of .12 as a clarifying example. This proportion means that 12% of the total variance can be explained by variance between respondents in giving ‘don’t know’-answers.
** These cells are empty, as the model without the interaction variance was the final model for these answer behaviours. The interaction variances for the behaviours fast and slow responding do not exist.

3.4.3. Consistency of answer behaviour

See Table 3.4. For the behaviours answering don’t know and answering won’t tell, the respondent ICCs were relatively large. Especially for answering won’t tell, the respondent ICC was large, also clearly when compared to the relatively small interaction ICC. This means that respondents varied in their frequency of answering won’t tell and that this variance was not particularly dependent on survey level. Therefore, it is likely that part of the respondents who answered won’t tell relatively frequently did so over the same and multiple surveys. This indicates the presence of consistency in showing this behaviour for part of the respondents. For answering don’t know, this interpretation seems less convincing, as the interaction ICC was about as large as the respondent ICC.

For the behaviours primacy responding and recency responding, the respondent ICCs were close to zero. Therefore, it is not likely that a substantial part of the respondents showed consistency in expressing these behaviours over the surveys. For these behaviours, the combination of relatively high frequencies, a relatively small respondent variance, and a relatively large survey ICC implies that respondents roughly followed the same overall behaviour pattern per survey. This means that respondents overall tended to show the behaviour; less frequently for surveys that show a relatively lower occurrence of the behaviour, and; more frequently for surveys that show a relatively higher occurrence of the behaviour. Additional evidence for this were the relatively small interaction ICCs. These imply that the survey differences in showing the behaviour were not dependent on the respondent level.

For the behaviour straightlining, the respondent ICC was a bit smaller than the survey and interaction ICC. In case of a relatively low respondent ICC in combination with an interaction ICC about equally high or higher, relatively much of the small respondent variance can be attributed to the survey level. Considering the low respondent ICC and the relatively modest occurrence, respondents who showed these behaviours did so for only a restricted number of surveys and above all for different surveys. This means that the occurrence of straightlining may be related to survey topic or difficulty, and that consistency in showing the behaviour was not likely.

Concerning fast and slow responding, the interaction variances and accompanying ICCs were lacking, as we assigned each respondent only one 0 or 1 per survey. This resulted in an unidentifiable within-respondent variance. This means that the model without interaction variance was the final model for these behaviours. Here, the relatively large respondent ICCs indicate that respondents were 'fast' or 'slow' for a variable number of surveys. This means that relatively many respondents may have a stable tendency to be either fast or slow in answering survey items overall.

In summary, respondent consistency only seems evident for answering 'won't tell'. And to a lesser extent, respondent consistency may be indicated for fast and slow responding. For these behaviours, our expectations regarding the presence of respondent consistency across surveys were confirmed to some degree. For the behaviours answering 'don't know', primacy responding, recency responding, and straightlining, these expectations were not met.

3.5. DISCUSSION

In this study, we investigated seven typical survey answer behaviours on their occurrence and on the extent to which respondents seem to show them consistently across different and multiple surveys. For this purpose, we used data from 2074 items of ten Dutch general population surveys that broadly vary in topic. Overall, the answer behaviours 'answering don't know' and 'answering won't tell' occurred relatively rarely with respective frequencies of .08 and .05. The behaviours 'primacy responding', 'recency responding', and 'straightlining' had moderate frequencies of .33, .28, and .12 respectively. Considering the behaviours 'fast responding' and 'slow responding' overall, the surveys Health, and Religion and Ethnicity were filled out relatively rapidly. The surveys Income and the Labour Force Survey were filled out relatively slowly.

For our study, we used two-way cross-classified logistic multilevel models. By these models, we were able to investigate the amount of variance between respondents, between surveys, and in their interaction with respect to the occurrence of answer behaviour. The respondent and survey variance referred to the differences in answer behaviour between respondents and surveys respectively. The variance in their interaction referred to the degree to which the behaviour occurrence for respondents is dependent on the survey level. To detect consistent presence of answer behaviour across multiple surveys, we stated two conditions. First, the behaviour must be shown to a substantial extent at all. This means that a relatively large respondent variance would indicate that the behaviour is shown by at least part of the respondents. And second, when the behaviour is shown at all, this is done for many to all surveys. This would indicate that the behaviour is absent for part of the respondents, while consistently present across surveys for another part of the respondents. This means that we would expect an absent or small interaction variance, as the behaviour is not dependent on the survey level in this case.

Hence, we used a multilevel model to investigate the variation between respondents and the respondent-survey interaction variation in showing each answer behaviour across surveys. Our idea was to detect relatively much respondent variation in combination with relatively little interaction variation. This refers to specific respondents filling out behaviour more frequently than other respondents, doing so over the same and multiple surveys. This would indicate respondent behaviour consistency across surveys. This indicative consistency appeared for the behaviours fast responding, slow responding, and especially answering won't tell. There was more respondent than survey variance present for fast and slow responding. For these three answer behaviours in particular, it may be useful to investigate what kind of respondent characteristics, such as age or education, could be attributed to respondents who are consistent in showing the behaviour. In case groups of consistent respondents can be linked to the same characteristics, the resulting relations could be used to control for measurement error.

It is crucial to state that the results of our study are not based on certainty but on likelihood. First, the nature of the respondent consistency is above all indicative. Relatively large respondent variation and relatively little interaction variation refer to respondents varying in their frequency of filling out the behaviour and doing so over the same and multiple surveys. It is however not straightforward for how many or what type of respondents, or for how many or which surveys this is the case. Yet, in spite of the indicative nature of the consistency and the necessity to identify types of respondents, our results disentangle the different respondent, survey, and interaction variations for each behaviour (see Table 3.4 in section 3.4).

Second, the undesirable answer behaviours are above all potential, as there is no straightforward relation between answer behaviour and measurement error. For instance, in case a respondent filled out don't know-answers relatively frequently for a specific survey, it is unknown to what extent this behaviour may be attributed to a stable respondent tendency. The respondent either may indeed not know much about a certain topic, or was perhaps struggling with a temporary lack of motivation to answer with respect to the specific survey content. However, the potentially undesirable nature of behaviour becomes likely as the behaviour is clearly shown over multiple surveys. Hence, the same behaviour may be associated with a stable respondent tendency. This tendency to express a specific behaviour across surveys might then predominate the occurrence of different behaviours for different survey topics and designs. In summary, our research presented an informative overview on the relative frequencies for the selected surveys and resulted in insightful variations for these behaviours. This gives us suggestions about which behaviours may need further investigation. See Table 3.5 for a brief overview of future recommendations in response to this discussion section.

Besides the single answer behaviours used in this study, it seems obvious to also include clusters of behaviours in further research that logically belong together. One of the most plausible clusters would be satisficing. Satisficing may be divided into weak versus strong satisficing (Krosnick, 1991) and could be clustered by means of the behaviours used in the current study (see Roßmann, 2017 for an example of such a satisficing cluster). Consider for instance the behaviours answering don't know and straightlining, both indicators of strong satisficing (Krosnick, 1991), together as a simple example of a satisficing cluster. Some respondents might have a personal tendency either to answer don't know or to straightline without necessarily being 'satisficers', but it is also plausible that other respondents satisfice more globally. Such 'global satisficers' may be likely to either answer don't know or to straightline for any item for

Table 3.5. Future Recommendations for the Answer Behaviours.

Answer behaviour	*Recommendation for further research*
Answering 'don't know'	Further investigate and increase respondent variance using clusters of surveys that are similar in difficulty and/or topic
Answering 'won't tell'	Identify groups of consistent respondents and relate these to characteristics and socio-demographics for the construction of respondent schemes
Primacy responding	Further investigate and increase respondent variance using items containing at least seven answering categories and/or using more conservative coding by only considering the first category (see Medway & Tourangeau, 2015)
Recency responding	Further investigate and increase respondent variance using items containing at least seven answering categories and/or using more conservative coding by only considering the last category (see Medway & Tourangeau, 2015)
Straightlining	Further investigate and increase respondent variance using clusters of surveys that are similar in difficulty and/or topic
Fast responding	Identify groups of consistent respondents and relate these to characteristics and socio-demographics for the construction of respondent schemes
Slow responding	Identify groups of consistent respondents and relate these to characteristics and socio-demographics for the construction of respondent schemes

which one of these behaviours is possible. In this way, not only respondents who either tend to answer don't know or straightline are distinguished, but also respondents who satisfice more generally could be detected. In other words, where respondents may not be considered a don't know answerer or straightliner, they could be identified as a 'strong satisficer' taking these two behaviours together.

Future research has the challenge to investigate such potential clusters of behaviours and the behaviours for which relatively much respondent variation and relatively little interaction variation was found in the current study. It is worthwhile to try to identify respondents who show consistency for these behaviours and to attribute characteristics to potential groups of respondents showing the same behaviour pattern over surveys. Such characteristics could refer to customary variables like gender, age, and education, but also to for instance current employment status, total income, or household composition. When characteristics could be identified for groups of respondents being consistent in showing specific behaviour, respondent schemes could be constructed. These respondent schemes would consist of the specific types of respondents, the identified characteristics, and the accompanying answer behaviours that are likely to relate to measurement error because of their consistent nature. The respondent schemes may be used to immediately look into the presumable undesirability of particular answer behaviour for specific types of respondents. This idea could be extended to survey items and questionnaires in follow-up research as well. The item characteristics of a questionnaire may be connected to consistent undesirable answer behaviour. A scheme for the questionnaire could then be constructed for an instant overview on the relation between its items and undesirable answer behaviour across surveys.

3.6. ADDITIONAL SECTION

We also investigated the occurrences and variances for five other relevant answer behaviours: 'Avoiding follow-up questions', 'socially desirable responding', 'acquiescence', 'neutral responding', and 'extreme responding'. The background and results for these behaviours were present in early versions of the manuscript, but not in the published article. In this additional section, we elaborate on the background, coding method, and results for these additional answer behaviours. We close with a brief discussion and some recommendations for future research.

3.6.1. Additional answer behaviours

3.6.1.1. Avoiding follow-up questions

Filter questions are questions containing answering options that may lead to follow-up questions. 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 situations may possibly lead to measurement error. Respondents may also presume a question to be a filter question, depending on the content of the question and on the format of asking the question (Eckman et al., 2014; Kreuter et al., 2011). This may motivate respondents to give an answer that avoids follow-up questions (Bosley et al., 1999). Adding a filter may result in more 'don't knows' and affect conclusions about public opinions on more abstract or less familiar topic content (Bishop, Oldendick, & Tuchfarber, 1983). For filter questions in grouped or ensemble format, respondents report higher disorder prevalence rates (Kessler et al., 1998), more mental health service use (Duan, Alegria, Canino, McGuire, & Takeuchi, 2007), and more times 'yes' triggering follow-up questions (Jensen, Watanabe, & Richters, 1999; Lucas et al., 1999) than for filter questions in interleafed format.

3.6.1.2. Socially desirable responding

Socially desirable responding refers to the tendency to minimize showing socially *undesirable* behaviour (DeMaio, 1984; Krosnick, 1999; Paulhus, 2002). Socially desirable responding can refer to both automatic and deliberate answer behaviour (Andersen & Mayerl, 2019). Questions asking for sensitive information (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 a socially desirable answer. Socially desirable responding depends on the data collection mode and is expected particularly in personal and telephone interviews (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, Krumpal, and Wolter (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 particularly dependent on the situation (see Paulhus, 2002) or survey (Johnson & Van de Vijver, 2003). This means that respondents may show this behaviour consistently over multiple surveys.

3.6.1.3. 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 (Díaz de Rada & Domínguez, 2015) and potentially lead to 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). When regarding acquiescence as a human tendency rather than survey- or item-dependent, it is likely that respondents who show this behaviour may do so consistently over multiple surveys.

3.6.1.4. Neutral and extreme responding

Choosing a neutral or 'mid-point' answer option may indicate satisficing (Krosnick & Fabrigar, 1997). Offering a middle option increases the probability that respondents show the accompanying neutral response (Kalton et al., 1980). 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 the amount of 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 with more extreme answers in interviewer-administered than 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 (Díaz de Rada & Domínguez, 2015). An increase in extreme responding may refer to stable response behaviour. For instance, answering the extreme option ‘very satisfied’ may be a true answer from a respondent, but it may also be a preference of the respondent for extreme answers (De Leeuw, 1992).

3.6.2. Coding the additional answer behaviours

In this section, we describe the coding process of the additional behaviours that need more elaboration: Avoiding follow-up questions, socially desirable responding, and acquiescence. The coding of neutral responding and extreme responding was straightforward. For neutral responding, the neutral middle response option was coded as ‘1’ for the eligible items. For extreme responding, only the most extreme responses on both end points of the answering scale were coded as ‘1’ for the eligible items. All other options were coded as ‘0’ for these behaviours. See Table 3.6 for an overview of these answer behaviours, the kind of eligible items, and the operationalization for each behaviour in calculating the relative frequencies.

3.6.2.1. An answer that avoids follow-up questions

Each survey contains a specific number of filter questions. Each filter question contains one or more answers that either do or do not continue towards one or more follow-up questions. First, we determined which items were filter questions. Second, we distinguished between answering options that make the item sequence stop immediately versus those that branch out into at least one follow-up question. Our idea is that respondents cannot estimate the shortest follow-up route between two answering options that both branch out into follow-up questions, but that respondents may be able to distinguish between answering options that lead to follow-up questions versus those that do not. Therefore, we only differentiated between answering categories branching out into one or more follow-up questions versus those that do not.

3.6.2.2. Socially desirable responding

About 50% of all items of the involved surveys together were coded as potentially asking for sensitive information by at least one of three coders (see Bais et al., 2019). The answering categories of these items were coded on whether they could refer to a socially desirable answer. All categories were coded liberally. This means that a category was coded as socially desirable when the answer may or may not refer to socially desirable behaviour. In other words, we con-

Table 3.6. The Kind of Additional Answer Behaviour, the Kind of Items Eligible for the Answer Behaviour, and the Operationalization of the Answer Behaviour.

Answer behaviour	*Eligible items*	*Operationalization*
Avoiding follow-up questions	All filter items containing at least one answering category factually leading to follow-up questions and at least one answering 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	All items coded as asking for sensitive information, containing at least one answering 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
Acquiescence	All more or less subjective (battery) items in the form of an ordinal agree/disagree or yes/no answering 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	All (battery) items with an odd and minimum number of five answering categories on an ordinal scale, containing a neutral middle answering category	Number of (battery) items for which the neutral middle answering category was filled out divided by Number of actually filled out eligible (battery) items
Extreme responding	All (battery) items with a minimum number of four answering categories on an ordinal scale, containing non-neutral first and last answering categories	Number of (battery) items for which an extreme answering category was filled out divided by Number of actually filled out eligible (battery) items

sidered an answer socially desirable when this could even only potentially be the case. As a result, relatively many answering options were coded as being socially desirable. Consider the following two item coding examples:

Item example 1 for 'socially desirable responding' from the survey Politics and Values:
'*There are too many people of foreign origin or descent in the Netherlands.*
1 fully disagree / 2 disagree / 3 neither agree nor disagree / 4 agree / 5 fully agree'

Here, the categories 'fully disagree', 'disagree', and 'neither agree nor disagree' were coded as a potential socially desirable response. Choosing one of these categories avoids going against the dominant norm that one should not express a form of hostility against people of foreign origin.

Item example 2 for 'socially desirable responding' from the survey Work and Schooling:
'*I perform paid work (even if is it just for one or several hours per week or for a brief period).*
0 no / 1 yes'

Here, the option 'yes' was coded as a potential socially desirable response. When one is not retired or disabled, the dominant norm is that one performs some kind of paid work to make a living. In case respondents do not have paid work, they may still fill out 'yes' to meet this dominant norm.

3.6.2.3. Acquiescence: Responding agreeably/affirmatively to a question
The answering categories of all items were evaluated on whether they showed 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 answering option 'agree', like 'satisfied', 'applicable', and 'yes', were considered for acquiescence. All categories were coded conservatively. This means that a category was coded as agreeable only in case the category was clearly agreeable. As a result, relatively few answering options were coded as being acquiescent. Consider the following two item coding examples:

Item example 1 for 'acquiescence' from the survey Family and Household:
'I find that my housekeeping work gives me stress.
1 disagree entirely / 2 disagree / 3 neither agree nor disagree / 4 agree / 5 agree entirely'

Here, only the category 'agree entirely' was coded as acquiescent.

Item example 2 for 'acquiescence' from the survey Personality:
'I am sometimes irritated by people who ask favours of me.
1 False / 2 True'

Here, the option 'True' was coded as acquiescent.

We need to note that the coding of the behaviours socially desirable responding and acquiescence is more or less arbitrary; the coding of both behaviours may have been executed either more liberally or more conservatively. On the one hand, this means that a response option that was coded as socially desirable or acquiescent may be a socially desirable or acquiescent response for some respondents, but the intended response for others. On the other hand, a response option that was *not* coded as socially desirable or acquiescent may indeed be the intended response for some respondents, but should have been coded as socially desirable or acquiescent for others. However, in order to investigate socially desirable responding and acquiescence at all, a coding threshold that distinguishes the occurrence from the non-occurrence of these behaviours simply needs to be placed at some point. Our idea is that enough variability between respondents across surveys should be present to distinguish respondents who consistently respond socially desirable or acquiesce from those who do not. We argue that this variability is sufficiently present by the way we coded these behaviours.

3.6.3. Results for the additional answer behaviours

For the additional behaviours, see Table D.1 in Appendix D for the maximum proportions of items for which the behaviours can factually occur. See Table E.1 in Appendix E for the results of the likelihood ratio tests (LRTs) for the model comparisons. See Table E.2 in Appendix E for the random intercepts and their profile likelihood-based confidence intervals of the final models. See Appendix F for the estimation and distribution of the random intercepts for the final models.

3.6.3.1. Prevalence of the additional behaviours

See Table 3.7. The potential avoidance of follow-up questions showed up frequently. This may partly be explained by the fact that relatively many single and blocks of follow-up items were simply not applicable to the respondent. This effect seems less evident for the Dutch Labour Force Survey (LF) that consists almost solely of filter questions. Most answering options of these LF filter questions branch out into follow-up questions that are applicable to most respondents. This means that these questions cannot be omitted easily. Concerning neutral responding, it stands out that the behaviour was chosen relatively often in the survey Family and Household (FA) and relatively rarely in the survey Work and Schooling (WO). Interestingly, also the behaviour extreme responding was chosen relatively often in the survey Family and Household (FA), mostly concerning items containing five or seven answering categories. This indicates that both neutral and distinctive attitudes concerning social household behaviour (see Table 3.1) occur frequently. The presence of potential socially desirable responding was striking for most surveys. This may partly be related to the liberal coding of the answering categories; they were coded as socially desirable when this could even only possibly be the case. It may also be that many respondents are relatively sensitive to social desirability in general. The potential behaviour acquiescence had a relatively modest frequency that did not differ a lot between the surveys in which it occurred.

Table 3.7. Relative Frequencies for the Additional Answer Behaviours for All Surveys* and in Total (TT).

Answer behaviour	*AS*	*FA*	*HE*	*HO*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*	*TT*
Avoiding follow-up questions	.84	.75	.72	.59	.86	-	.79	.66	.78	.35	.78
Socially desirable responding	.92	.52	.68	-	.90	.96	.66	.28	.75	.46	.71
Acquiescence	-	.13	-	-	-	.14	.07	.21	.08	-	.12
Neutral responding	-	.32	-	-	.25	.22	.21	-	.05	-	.22
Extreme responding	-	.30	-	-	.11	.22	.12	-	.14	-	.18

* Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).
Note. The relative frequencies are based on respondents who filled out at least five items for a specific behaviour and survey, and who did so for at least two surveys for a specific behaviour, to remain consistent with the final analyses. This means that the relative frequencies of respondents who did not fill out at least five items for a specific behaviour for at least two surveys were omitted. Consequently, some relative frequencies may be biased compared to the factual overall relative frequencies, although to a restricted extent.

3.6.3.2. Consistency of the additional behaviours

See Table 3.8. For the behaviour avoiding follow-up questions, there was no respondent variance and also no interaction variance. For this behaviour, only survey variance was present. This means that respondents more or less followed the same specific pattern of going through the filter question process for each survey. For the behaviour socially desirable responding, the respondent ICC was close to zero. Thus, it is unlikely that many respondents showed consistency in showing these behaviours over the surveys. The interpretations for primacy and recency responding may be extended to avoiding follow-up questions and socially desirable responding; high frequencies, small respondent variance, and larger survey ICC imply that respondents roughly followed the same overall behaviour pattern per survey. The relatively small interaction ICC for socially desirable responding implies that survey differences in showing the behaviour were not dependent on the respondent level. Hence, respondents overall tended to show the behaviour according to the specific occurrences in the individual surveys.

For the behaviours acquiescence and neutral responding, the respondent ICCs were a bit smaller than the survey and interaction ICCs. As for straightlining, relatively much of the small respon-

Table 3.8. Proportions of Explained Variances (ICCs) for the Additional Answer Behaviours for the Final Models over All Surveys Together.

Behaviour (nr. of respondents)	*Explained respondent variance (ICC)*	*Explained survey variance (ICC)*	*Explained interaction variance (ICC)*
Avoiding follow-up questions (6333)	.00	.13	NA*
Socially desirable responding (6394)	.01	.34	.03
Acquiescence (6133)	.04	.05	.08
Neutral responding (5974)	.04	.15	.05
Extreme responding (5985)	.11	.07	.13

* This cell is empty, as the model without the interaction variance was the final model for this answer behaviour. The interaction variance for the behaviour avoiding follow-up questions was zero.

dent variance can be attributed to the survey level. Respondents who showed these behaviours did so for few and above all different surveys. This means that their occurrence may point to survey topic or difficulty, rather than to consistency. As for answering 'don't know', the behaviour extreme responding refers to a relatively larger respondent ICC. However, the interaction ICC was about as large as the respondent ICC; the variance in extreme respondent answers is partly dependent on the survey level. This means that consistency in extreme responding for respondents is not convincing. In summary, none of the additional answer behaviours seems to be evidently related to consistent respondent behaviour across surveys.

3.6.4. Brief discussion of the additional answer behaviours

While coding avoiding follow-up questions, neutral responding, extreme responding was straightforward, coding the answering options of the eligible items for socially desirable responding and acquiescence was more or less arbitrary. As a result, an answering option may refer to a socially desirable or acquiescent response, but may also actually be the intended response. A neutral or extreme answer, or an answer that avoids follow-up questions, may be an intended response as well and does not necessarily point to undesirable behaviour. Still, our idea is that the more these behaviours are shown consistently across surveys, the more likely it becomes that they refer to a stable respondent tendency that may be undesirable. Thus, also considering these behaviours, it may be useful to distinguish respondents for whom the behaviours do and do not consistently occur.

Almost all surveys contained many filter questions and showed high overall relative frequencies for the behaviour avoiding follow-up questions. This may be explained by the fact that relatively many follow-up questions were not applicable to many respondents. For instance, respondents did not own specific kinds of income and property, did not meet criteria concerning illness, and were not part of a religious family. These factors may explain the respective high frequencies for the surveys Assets, Health, and Religion and Ethnicity. As a consequence of these high frequencies overall, there was not much respondent and interaction variation for this behaviour. Therefore, potential consistency is unlikely to be detected.

This is also applicable to socially desirable responding that showed relatively high frequencies for almost all surveys as well. An explanation for these high frequencies may be the liberal coding of the answering options of the applicable items; all options that could ‘possibly refer to a socially desirable answer’ were ticked as potentially ‘socially desirable’. Especially for ordinal rating scales with relatively many answering options, this means that the probability of choosing a potential socially desirable answer is relatively large. Another explanation may be the fact that every item that was formerly coded as containing sensitive information was considered for this behaviour. As a result, many respondents may simply have answered many items within the range of a generally acceptable answer. A more strict selection of items that consists of clearly sensitive items or topics could increase both respondent and interaction variance.

Considering acquiescence, the relative frequencies appeared to be modest overall. This may be explained by the conservative coding of this behaviour; only the most evident agreeable answering option was ticked as potentially ‘acquiescent’ for all ordinal rating scales. It may be useful to enlarge respondent variation and the likelihood that consistency may be detected for these behaviours. Further research might consider; including surveys with more filter questions for which options that lead to follow-up items are likely to be chosen; coding socially desirable responding more conservatively or choosing a more strict selection of sensitive items that contain socially desirable responses; and coding acquiescence more liberally. See Table 3.9 for a brief overview of future recommendations in response to this discussion.

Table 3.9. Future Recommendations for the Additional Answer Behaviours.

Answer behaviour	*Recommendations for further research*
Avoiding follow-up questions	Further investigate and increase respondent variance using surveys in which it is less likely that follow-up questions are being avoided
Socially desirable responding	Further investigate and increase respondent variance using more conservative coding of answering categories, and/or using a more strict coding of what constitutes an item containing socially desirable responses
Acquiescence	Further investigate and increase respondent variance using clusters of surveys that are similar in difficulty and/or topic, and/or using more liberal coding of answering categories
Neutral responding	Further investigate and increase respondent variance using clusters of surveys that are similar in difficulty and/or topic
Extreme responding	Further investigate and increase respondent variance using clusters of surveys that are similar in difficulty and/or topic

λ

$\lambda \ \lambda$

$\lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda$

$\lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda \ \lambda$

CHAPTER 4

Translating survey answer behaviour into meaningful summaries: The concept and estimation of survey behaviour profiles

This chapter is internally published at Statistics Netherlands as a discussion paper:
Bais, F., Schouten, B., & Toepoel, V. (2021). Translating survey answer behaviour into meaningful summaries: The concept and estimation of survey behaviour profiles. CBS Discussion paper. Statistics Netherlands, The Hague.

Author contributions: FB and BS designed the study. CentERdata and Statistics Netherlands provided the survey data. FB prepared the data, performed the statistical analyses, and wrote the paper. BS and VT critically reviewed the paper.

Abstract

Survey answer behaviour may be considered undesirable and related to measurement error, like answering ‘don’t know’ or ‘won’t tell’. In this paper, we propose a new method to translate answer behaviour in surveys into meaningful statistical summaries – *behaviour profiles*. Behaviour profiles are density distributions that visualize the occurrence of answer behaviour and the uncertainty that comes along with this occurrence. They also visualize the locations where types of respondents or items may show a high occurrence of answer behaviour. We introduce our pseudo-Bayesian-based method by defining the concept of behaviour profiles and by elaborating on how the profiles can be estimated for respondents and items. We illustrate the statistical properties of behaviour profiles by means of simulation and by estimating profiles from surveys in the probability-based LISS Panel administered by CentERdata. Behaviour profiles can be compared for single surveys or across multiple surveys to expose relations of respondent and item characteristics to undesirable answer behaviour. Hence, survey administrators may use behaviour profiles to explore the presence of such relations in their survey data. Survey constructors may adapt their survey designs in accordance with the impact of these relations on measurement error.

4.1. THEORETICAL BACKGROUND

The relation between survey answer behaviour and measurement error has been studied extensively. The occurrence and size of measurement error and hence data quality can be influenced by both respondent characteristics (see for instance Olson & Smyth, 2015; Pickery & Loosveldt, 1998; Tourangeau et al., 2000) and item characteristics (see for instance Campanelli et al., 2011; Saris & Gallhofer, 2007b; Tourangeau et al., 2000). Respondent characteristics can for instance refer to age, educational level, or income level. An example of an item characteristic is non-centrality. A large degree of non-centrality means that the item may go beyond the interest, knowledge, or experience of the respondent (Gallhofer et al., 2007; Saris & Gallhofer, 2007b; Van der Zouwen, 2000). Another example of an item characteristic is sensitive information. An item may contain sensitive information of some societal, menial, or personal kind (Campanelli et al., 2011; Gallhofer et al., 2007; Kreuter et al., 2008; Lensvelt-Mulders, 2008; Saris & Gallhofer, 2007b; Tourangeau & Yan, 2007).

Both item and respondent characteristics may lead to undesirable answer behaviour, like satisficing (Holbrook et al., 2003; Kaminska, McCutcheon, & Billiet, 2010). When respondents satisfice, they take short-cuts in the question-answering process. Satisficing may be considered the outcome of the interaction of cognitive ability, motivation, and item difficulty (Krosnick, 1991, 1999; Krosnick et al., 1996). Both age and educational level, both proxies for cognitive ability, have been shown to be related to undesirable answer behaviours and hence survey data quality (Krosnick 1991, 1999; Krosnick et al., 1996). Older and lower educated respondents show less accurate answer behaviour than younger respondents (Andrews & Herzog, 1986) and higher educated respondents (Antoni et al., 2019), and a less stable attitude reliability measurement than younger and higher educated respondents (Alwin & Krosnick, 1991). Also a high degree of item difficulty has a negative effect on the quality of the response to that item (Van der Zouwen, 2000). According to the cognitive response model (Jenkins & Dillman, 1997; Tourangeau et al., 2000), the presence of difficulty-related characteristics in items may impose a problem for the respondent in understanding the question, or in retrieving or judging relatively complex information. A relatively lower cognitive ability and/or motivation from the respondent may then lead to satisficing behaviour (Krosnick, 1991, 1999; Krosnick et al., 1996).

Answer behaviour can also be the result of sensitive survey or item content. Sensitive items may involve a threat of disclosure (Lensvelt-Mulders, 2008) or can be experienced as intrusive (Tourangeau et al., 2000; Tourangeau & Yan, 2007). Item sensitivity is related to answer behaviour that may be the outcome of a lack of willingness from the respondent to give a 'true' answer (Bradburn et al., 1978; Shoemaker et al., 2002; Tourangeau et al., 2000). As a result of sensitive content, respondents may give a socially desirable response or refuse to answer. In summary, lower cognitive ability and motivation, and a higher degree of item difficulty and item sensitivity may all result in measurement error and suboptimal data quality. See Figure 4.1 for an overview of the relation of item and respondent characteristics to answer behaviour. See Table 4.1 for the definition of the specific answer behaviours that are shown in Figure 4.1.

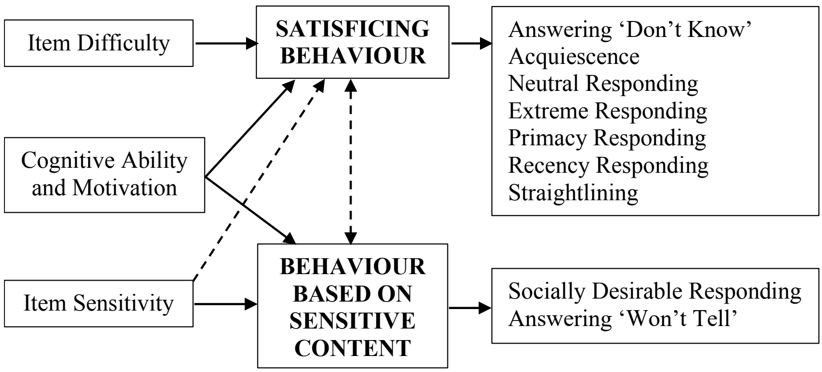


Figure 4.1. Literature-Based Theoretical Framework.

A lot is known from the literature about the relation of respondent and item characteristics to answer behaviour for single surveys. However, panel studies about the degree of consistency of answer behaviour for specific respondent and item characteristics across multiple surveys are lacking. It is important to fill this gap. When a respondent shows specific answer behaviour only for specific types of items within a single survey, it is not to say whether this behaviour is typical for this specific respondent or these types of items. The answers may refer to a form of satisficing (Krosnick, 1991, 1999), but may just as well be truly attributed to the type of respondent or item. In order to investigate the relation of answer behaviour to the characteristics of items and respondents, the behaviour should be stable and typical for types of items and respondents. That is, the behaviour for a specific type of item or respondent with specific characteristics must be shown *consistently* in order to be typical for that type of item or respondent. Here, the term 'consistent' refers to a pattern of answer behaviour that is shown over several points in time, across multiple types of surveys. In case of consistent answer beha-

Table 4.1. Answer Behaviours, Their Definitions, and Selected Relevant Literature.

Answer behaviour	*Definition*	*Selected relevant literature*
Socially desirable responding	The tendency to minimize showing socially undesirable behaviour	Andersen & Mayerl 2019; Campanelli et al. 2011; DeMaio 1984; Heerwegh & Loosveldt 2011; Holbrook et al. 2003; Jann et al. 2019; Johnson & Van de Vijver 2003; Kreuter et al. 2008; Krosnick 1999; Paulhus 2002; Roberts 2007; Roberts & Jäckle 2012; Tourangeau et al. 2000; Tourangeau & Yan 2007
Answering ‘don’t know’ and ‘won’t tell’	The tendency to give a ‘don’t know’- or a ‘won’t tell’-answer to a question	Beatty & Herrmann 2002; Binswanger et al. 2013; Bishop et al. 1986; Bradburn et al. 1978; Fricker et al. 2005; Krosnick et al. 2002; Leigh & Martin 1987; Roberts 2007; Roßmann et al. 2018; Schuman & Presser 1981; Shoemaker et al. 2002; Tourangeau et al. 2000; Vis-Visschers et al. 2008
Acquiescence	The tendency to answer affirmatively, regardless of the content of the question	Billiet & McClendon 2000; De Leeuw 1992; Díaz De Rada & Domínguez 2015; Heerwegh & Loosveldt 2011; McClendon 1991; Messick 1966; O’Muircheartaigh et al. 2000; Saris et al. 2010; Schaeffer & Presser 2003; Stricker 1963
Neutral responding	The tendency to choose the neutral midpoint category from a bipolar answering scale	He & Van de Vijver 2013; Kalton et al. 1980; Krosnick & Fabrigar 1997; O’Muircheartaigh et al. 2000; Si & Cullen 1998; Stern et al. 2007; Tarnai & Dillman 1992
Extreme responding	The tendency to choose an extreme category from the answering scale	Aichholzer 2013; De Leeuw 1992; Díaz De Rada & Domínguez 2015; Ye et al. 2011
Primacy responding	The tendency to choose an option at the beginning of an answering list	Galesic et al. 2008; Krosnick 1991; Krosnick 1992; Krosnick & Alwin 1987; McClendon 1991; Stern et al. 2007
Straightlining	The tendency to give the same answers to a series of questions arranged in a grid format	Díaz De Rada & Domínguez 2015; Fricker et al. 2005; Krosnick 1991; Krosnick & Alwin 1989; Roßmann et al. 2018; Schonlau & Toepoel 2015; Zhang 2013; Zhang & Conrad 2014

viour across surveys, we may connect the behaviour to stable characteristics of the concerned types of respondents and items.

In order to compare types of items or respondents on consistent answer behaviour, a few aspects need to be taken into account regarding the answer behaviour. First, the number of respondents that participates in a survey and the number of items that is applicable to a specific answer behaviour per survey can be relatively small. This means that an extent of uncertainty exists around the actual occurrence of behaviour, since it is based on a by definition delimited number of respondents or items. In other words, the actual occurrence of behaviour for respondents may be more biased as they fill out a smaller number of items. And a smaller number of items also reduces the probability that the specific set of items contains those items that are more likely to induce a specific behaviour. At the same time, the actual occurrence of behaviour for items may be more biased as they are filled out by a smaller number of respondents. Second, when a survey contains filter questions that may or may not branch out into follow-up questions, each respondent is likely to fill out a different number of items for that survey. Therefore, the actual occurrence of behaviour is indicated with varying uncertainty across different respondents and items within a survey. These uncertainties must be taken into account in estimating behaviour occurrences before comparing types of respondents or items.

Techniques that account for item and respondent characteristics in modelling answer behaviour are cross-classified multilevel models (see for instance Olson & Smyth, 2015; Olson et al., 2019) and mixed item response theory models (see for instance Tijmstra et al., 2018; Zickar et al., 2004). We have good reasons to provide an alternative method to model answer behaviour. *In this paper, we present a method to translate survey answer behaviour into meaningful summaries - behaviour profiles - that consider the aforementioned uncertainties.* Behaviour profiles are density distributions that show the likelihood of specific answer behaviour for respondents or items for any probability occurrence (see section 4.2). Behaviour profiles give an immediate visual overview of the distribution of specific answer behaviour. This means that the uncertainty that comes along with behaviour occurrences is visible instantly. The behaviour profile also indicates where subgroups of respondents or items that show undesirable behaviour to a high degree may be distinguished from the rest of a specified group of respondents or items. These visual benefits are illustrated in sections 4.2 through 4.4. Behaviour profiles can be constructed for both individual and groups of respondents, for both single surveys and across multiple surveys, for any combination of respondent and item characteristics, and for any survey mode.

In this paper, we first show how behaviour profiles can be derived across respondents, items, and surveys. Second, we present how the statistical properties of behaviour profiles can be estimated from simulated examples and survey data. In section 4.2, we elaborate on the concept of behaviour profiles and provide definitions. In section 4.3, we show how the profiles can be estimated for respondents and items separately, we discuss the extent of uncertainty around the profiles by the use of bootstrapping, and we elaborate on how peaked profiles can be smoothed. In section 4.4, we give examples of respondent profiles and item profiles by means of data from surveys administered to the probability-based LISS panel owned by CentERdata. In section 4.5, we discuss the implications of using behaviour profiles in comparing groups of respondents and items, and give suggestions for future research.

4.2. DEFINITION OF ANSWER BEHAVIOUR PROFILES

In this section, we elaborate on the concept of the answer behaviour profile and provide its definition and accompanying notation. We distinguish two behaviour profiles: The respondent profile in which we focus on respondent behaviour and characteristics, and the item profile in which we focus on survey item behaviour and characteristics. We elaborate on the respondent profile and the item profile separately. For convenience, we choose a single answer behaviour and omit this reference in the notation.

4.2.1. The respondent profile

Let us assume that the answer behaviour of interest can occur in S surveys, labelled $s = 1, 2, \dots, S$. Let survey s have I_s items for which the behaviour is possible, labelled $i = 1, 2, \dots, I_s$. The survey items can be answered by R_S respondents, labelled r = 1, 2, … , R_S, which composition may vary per survey. Let $I_{s,r}$ be the number of items from survey s and I_r the total number of items over all surveys that was actually presented to respondent r. Let $p_{s,i,r}$ be the probability that respondent r shows the behaviour for item i of survey s. Let p_r be the probability that respondent r shows the behaviour for a random item from a random survey. Although the real probabilities p_r are unknown, the mean of $p_{s,i,r}$ over many surveys and items converges to p_r. We assume that all p_r are coming independently from the same probability distribution F_R with support [0,1]. We define the corresponding density function f_R as the respondent answer behaviour profile or simply the *respondent profile*.

The density function f_R is an unknown distribution. Given that we have a finite sample of respondents and that we can thus obtain a smooth estimate of this distribution at best, we use a pseudo-Bayesian analysis: We set prior distributions for the unknown answer behaviour profiles. We consider all values of the probabilities and any density function with support [0,1] as equally likely. This means that the p_r have uniform priors. The respondent profiles are then updated to posteriors by the respondent data.

Respondents usually complete a survey within a relatively short time span, which means that individual behaviour will show some degree of interdependence across items in practice.

Theoretically, a respondent could show a behaviour for all or none items, referring to full interdependence. We assume that full interdependence is unlikely to occur in practice. For this paper, we obviously use examples that omit the case of full interdependence. Although we acknowledge this interdependence across items in practice, we assume that individual behaviour is independent across items. Our method is a first step in representing answer behaviour visually. For taking into account interdependence across items, a few follow-up steps are suggested in section 4.5. Finally, in modelling answer behaviour, we assume that the behaviour of different respondents is independent.

We can now model the actual occurrence of the answer behaviour given the probabilities. When respondent r is presented I items for which the behaviour is possible, the number of times G that the behaviour is shown has a binomial distribution $\mathrm{Bin}(I, p_r)$. The complete model for respondents becomes

$$p_r \sim \mathrm{Unif}(0,1) \text{ and } (G_r|p_r) \sim \mathrm{Bin}(I_r|p_r). \tag{4.1}$$

The respondent profile f_R can be further specified by vectors of respondent characteristics x_R and item characteristics x_I. Then $f_R(.|x_R)$ is the respondent profile when restricting answer behaviour to respondents with characteristics x_R. And $f_R(.|x_I)$ is the respondent profile when restricting answer behaviour to drawing only items with characteristics x_I. And $f_R(.|x_R, x_I)$ is the respondent profile when restricting answer behaviour to both respondents with characteristics x_R and drawing only items with characteristics x_I. The overall respondent profile f_R follows from integration over the respondent and item characteristics.

As an example of respondent profiles, see the fictitious profiles for men $f_R(.|x_{R_{men}})$ and for women $f_R(.|x_{R_{women}})$ in Graph 1 of Figure 4.2. As can be seen from the graph, the expected occurrences for a specific behaviour for men and women are around 0.70 and 0.40 respectively. The profiles also differ in shape; the profile for men is relatively more steep and peaked, while the profile for women is relatively more broad and smooth. This means that the expected behaviour occurrence is less precisely estimated for women than for men in this example.

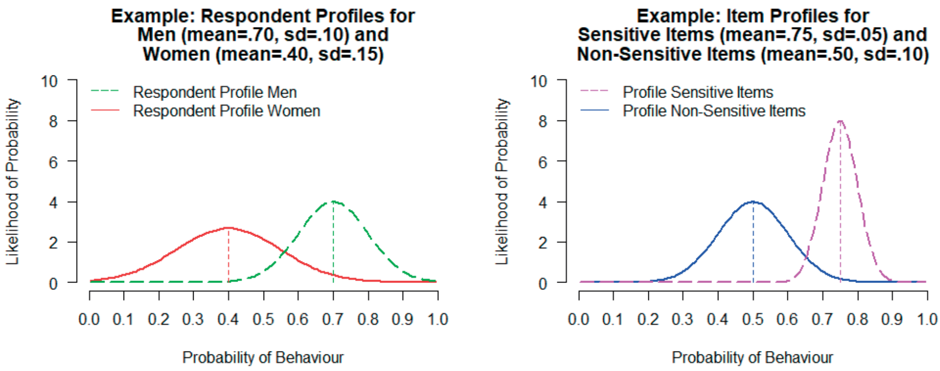


Figure 4.2. Simple Examples of Respondent and Item Behaviour Profiles.

4.2.2. The item profile

Let $R_{s,i}$ be the number of respondents who were presented item i from survey s. Let $p_{s,i}$ be the probability that a random respondent shows the behaviour for item i from survey s. For large survey sample sizes, the mean of $p_{s,i,r}$ over respondents converges to $p_{s,i}$. We assume that the $p_{s,i}$ are coming independently from the same probability distribution F_I with support [0,1]. We define the corresponding density function f_I as the item answer behaviour profile or simply the *item profile*.

As the situation and our assumptions for the density function f_I are the same as for f_R, we use the same pseudo-Bayesian analysis to model the actual occurrence of answer behaviour: When item i from survey s is presented to R respondents, the number of times L that the answer behaviour is shown has a binomial distribution $\text{Bin}(R, p_{s,i})$. The complete model for items becomes

$$p_{s,i} \sim \text{Unif}(0,1) \text{ and } \left(L_{s,i} \middle| p_{s,i}\right) \sim \text{Bin}\left(R_{s,i} \middle| p_{s,i}\right). \tag{4.2}$$

As for the respondent profile, the item profile f_I can be further specified by vectors of item characteristics x_I and respondent characteristics x_R into $f_I\,(.\,|x_I)$, $f_I\,(.\,|x_R)$, and $f_I\,(.\,|x_I\,, x_R)$. The overall item profile f_I follows from integration over the item and respondent characteristics.

As an example of item profiles, see the fictitious profiles for sensitive items $f_I\ (.\,|x_{I_{sense}})$ and for non-sensitive items $f_I\ (.\,|x_{I_{nonse}})$ in Graph 2 in Figure 4.2. Just as the respondent profiles, the item profiles have their own expected behaviour occurrence and show a difference in shape in this example. Here, the expected behaviour occurrence is more precisely estimated for sensitive items than for non-sensitive items.

4.3. ESTIMATION OF ANSWER BEHAVIOUR PROFILES

In this section, we first elaborate on the method to estimate behaviour profiles. Second, we discuss the accuracy of the estimated profiles by means of simulation.

4.3.1. Profile values

We estimate the density functions or behaviour profiles underlying the respondent and item probabilities. We start by estimating the respondent profile. The observed data for respondent r consist of $(I_{1,r}, G_{1,r};\ \ldots;\ I_{S,r}, G_{S,r})$. Let $I_r = \sum_{s=1}^{S} I_{s,r}$ and $G_r = \sum_{s=1}^{S} G_{s,r}$. The following posterior for p_r is the individual profile for respondent *r*:

$$p_{post}(p|G_r, I_r) = \binom{I_r}{G_r} p^{G_r}(1-p)^{I_r - G_r}. \tag{4.3}$$

We consider each respondent to have a value of the same weight and normalize each individual respondent profile to obtain a distribution area of 1. Then, we average the posteriors of all concerned respondents *R* to obtain an estimate $\hat{f}_R$ of the respondent profile f_R:

$$\hat{f}_R(p) = \frac{1}{R}\sum_{r=1}^{R} p_{post}\ (p|G_r, I_r) = \frac{1}{R}\sum_{r=1}^{R} \binom{I_r}{G_r} p^{G_r}(1-p)^{I_r - G_r}. \tag{4.4}$$

We continue by estimating the item profile. The observed data for item *i* from survey *s* consist of $(R_{s,i}, L_{s,i})$. The following posterior for $p_{s,i}$ is the individual profile for item *i* from survey *s*:

$$p_{post}\left(p\middle|L_{s,i}, R_{s,i}\right) = \binom{R_{s,i}}{L_{s,i}} p^{L_{s,i}}(1-p)^{R_{s,i} - L_{s,i}}. \tag{4.5}$$

We consider each item to have a value of the same weight and normalize each individual item profile to obtain a distribution area of 1. Then, we average the posteriors of all concerned items *I* for all surveys *S* to obtain an estimate $\hat{f}_I$ of the item profile f_I:

$$\hat{f}_I(p) = \frac{1}{I}\sum_{s=1}^{S}\sum_{i=1}^{I} p_{post}\left(p \middle| L_{s,i}, R_{s,i}\right) = \frac{1}{I}\sum_{s=1}^{S}\sum_{i=1}^{I}\binom{R_{s,i}}{L_{s,i}} p^{L_{s,i}}(1-p)^{R_{s,i}-L_{s,i}}. \quad (4.6)$$

Estimators (4.4) and (4.6) can be applied across surveys or to single surveys to evaluate the clustering effect of surveys, as the items within a survey are usually clustered on a related topic. They can also be applied to respondent and items with specified characteristics (see section 4.2) to evaluate the influence of types of respondents and items.

The estimated respondent or item profile may become very peaked. This happens when the number of items is large relative to the number of respondents in respondent profiles or when the number of respondents is large relative to the number of items in item profiles. The reason for this is that behaviour profiles are estimated using a finite sample of either respondents or items. In these cases, it may be desirable to use kernel density estimation methods to smooth the estimated profiles. From here, we elaborate on standard kernel density estimation with regard to our method of using posterior distributions for the respondent profile perspective.

Let K_h be a kernel function with bandwidth h: $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, where K is a specific kernel like uniform or Gaussian. Using the kernel function to construct a smoothed version of an averaged respondent group profile gives

$$\hat{f}_{R,h}(p) = \int_0^1 K_h\,(p-x)\hat{f}_R(x)dx. \quad (4.7)$$

Since the behaviour profiles are computed at discrete probability points, we follow the discretization in estimating the integral. Let the discrete step size be Δ, for which we use 0.01. With $m = n = 1,2, \dots, 1/\Delta$, and $M = N = 1/\Delta$, the estimated profile becomes

$$\hat{f}_{R,h,\Delta}(n\Delta) = \Delta\sum_{m=1}^{M} K_h\,(n\Delta - m\Delta)\hat{f}_R(m\Delta). \quad (4.8)$$

As a final step, we normalize the estimated profile in order to obtain an area of 1. This is necessary as a small part of the distribution may have been smoothed towards the left of 0 or towards the right of 1. See section 4.3.2.3 for a few smoothing examples.

4.3.2. Profile confidence intervals

There are two factors that affect the shape and uncertainty of the behaviour profile; the number of items per respondent and the number of respondents per item. We illustrate the effect of increasing numbers for the respondent and item profile perspective separately. We use bootstrapping to map the profile uncertainty in the form of profile confidence intervals. In general, the confidence intervals are determined by group size and group variance; the larger the group and the smaller the variance within the group of respondents or items, the smaller the uncertainty surrounding the profile. Considering the focus of this paper on item and respondent group size, we do not take into account the influence of group variance on the confidence intervals. Therefore, we use examples of profiles with approximately the same group variance to keep them comparable in this respect.

4.3.2.1. Respondent perspective without smoothing

The respondent profile is based on a group of respondents who filled out a specific number of items. To calculate the 99% confidence intervals for a profile, we bootstrap 10000 profiles. For each bootstrap profile, respondents are randomly sampled with replacement and their individual profiles are averaged conform formula (4.4). The number of sampled respondents for each bootstrap profile is equal to the group size. In calculating behaviour profiles, we consider the behaviour probabilities from 0 to 1 with a step size interval of 0.01. This means that we base the calculation of the confidence intervals on the probabilities $\{0.005, 0.015, 0.025, \dots, 0.995\}$ for 100 equal intervals. For each specific probability, we calculate the standard deviation of the accompanying 10000 likelihoods obtained from the bootstraps. These standard deviations are used for the calculation of the 99% confidence intervals.

For the *respondent perspective*, we discuss the impact of increasements in the *number of items per respondent* and in *respondent group size.* We do so by elaborating on several fictitious examples by means of simulation. The examples are based on random sampling from two fixed normal distributions with pre-specified means and standard deviations. Only the numbers of respondents and items are varied across the examples; we use the numbers of 20 and 500 respectively to refer to small and large numbers of respondents or items. We choose for the large number of 500, as we want to be able to combine the responses for items of multiple surveys. Considering that the number of items per respondent is not infinitely large, the estimate of a respondent profile is smooth by definition. This implies that a group of respondents that is sampled from a fixed distribution will not contain the degree of detailed information of that distribution. By sampling enormous numbers of respondents from a fixed distribution, we can obtain the best possible estimates of respondent profiles for varying number of items.

See Figure 4.3 for a comparison of such estimated profiles for the case of 20 and 500 items per respondent respectively. Let us begin by describing what can actually be seen in the graphs. The fixed distributions for men (green) and women (red) are shown in both graphs. Both distributions vary around their expected behaviour occurrences. The occurrence can be read from the x-axis, where the probability of a behaviour runs from 0 to 1. The y-axis shows how likely a specific probability for a behaviour is to occur. We can see that a behaviour probability of around 0.70 is most likely to occur for men and a probability of around 0.40 is most likely to occur for women. Note that the shape of the distribution is more concentrated around the value of 0.70 for men than around the value of 0.40 for women. Both Graph 1 and 2 show that respondent profiles for men (magenta) and women (blue) are 'underestimated' relative to the fixed distributions; the lower the number of items per respondent, the more the sampled profile will deviate from the fixed normal distribution. In case of only 20 items per respondent (Graph 1), this difference is large, while in case of 500 items per respondent (Graph 2), the difference is hardly visible, although still present. This means that the lower the number of items per respondent, the less power the sampled profile contains to accurately approach the fixed distribution.

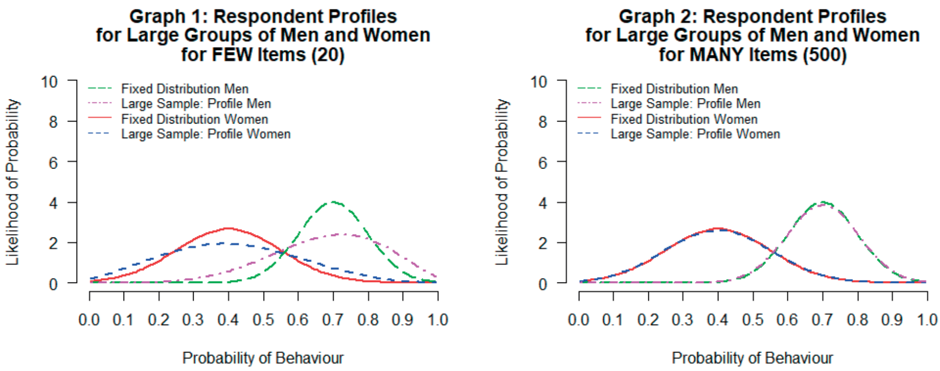


Figure 4.3. Fictitious Examples of Sampled Behaviour Profiles for Large Groups of Respondents.

For our fictitious examples, we use the best possible estimates of profiles for large groups of respondents for 20 and 500 items as our reference profiles. Note that we sample examples of respondent profiles to elaborate on their statistical properties in order to show their usefulness for survey data. Also note that the sampled profile examples for men and women are accompanied by their 99% confidence intervals (the thin dotted lines) in the graphs. In the first example, we consider the probabilities for a specific behaviour for two groups of sampled respondents. One group consists of *few* men, the other group consists of *few* women, who all filled out *few* items. See Graph 1 in Figure 4.4. As the respondents filled out only few items, the profiles are smooth and stretched. And because there are only few respondents in the groups, the confidence intervals are relatively broad and may be formed erratically along the behaviour profiles. This indicates that the expected behaviour occurrences and respondent profiles are relatively uncertain. Note that whenever a behaviour profile is accompanied by its 99% confidence interval in a graph, the interval has the same colour as the behaviour profile.

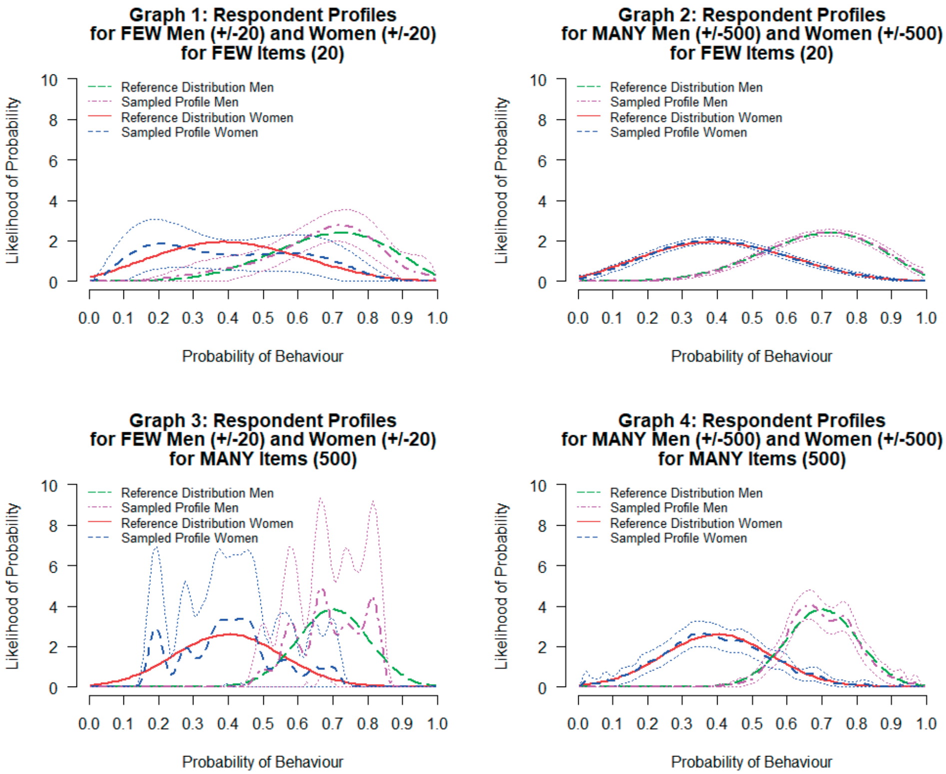


Figure 4.4. Fictitious Examples of Sampled Behaviour Profiles Accompanied by Their 99% Confidence Intervals.

Graph 2 in Figure 4.4 shows sampled groups of *many* men and women who filled out *few* items. Compared to Graph 1, the sampled profiles look about the same, but have more narrow confidence intervals. Although the expected behaviour occurrences and sampled profiles are relatively certain due to the large number of respondents, they may not be very precise due to the small number of items. The result is a broad and smooth profile, hence a possible ‘underestimation’ of a population profile. Therefore, we need to emphasize that a certain profile cannot be used unconditionally; the profile must first be checked on its precision.

Graph 3 in Figure 4.4 shows an example in which *few* men and women filled out *many* items. Now that the respondents filled out more items, their individual behaviour probabilities are relatively more precise, resulting in a more peaked distribution. In case of randomly many items, the individual respondent profiles become a collection of peaks that ultimately converge to a discrete distribution. Note however that the confidence intervals for the sampled respondent group profiles are even broader than in the first example. This is explained by the combination of a small group size and large individual behaviour precision. This results in relatively uncertain expected behaviour occurrences and respondent profiles.

Graph 4 in Figure 4.4 shows the fourth example in which *many* men and women filled out *many* items. Now that larger groups of respondents filled out many items, the individual peaks partly cancel each other out, resulting in smoother averaged respondent profiles compared to Graph 3. These profiles resemble the profiles in Graph 1, with the important difference that the profiles are less stretched and the confidence intervals are more narrow. This means that the profiles have become relatively certain due to the larger groups of respondents. In general, profiles that contain both the precision of individuals and the certainty of large groups are preferred. As the profiles in Graph 4 are still modestly peaked, this implies that the optimal situation refers to marginally less items than respondents.

4.3.2.2. Item perspective without smoothing

The bootstrap procedure and calculation of the confidence intervals for the item perspective equal those of the respondent perspective. The only difference is that now items are sampled, with the item profile based on a group of items that were filled out by a specific number of respondents. For the *item perspective*, the impact of increasements in the *number of respondents per item* and in *item group size* are discussed. The graphs of the respondent profiles above can also be considered for the item perspective, with the concepts of respondent and item exchanged. Parallel to the respondent perspective, the probabilities for a specific behaviour for the same four cases are considered. The results for the item perspective are identical to those of the respondent perspective. See Table 4.2 for a summary of the features for the different examples.

Table 4.2. Summary of the Features for the Different Examples for Few/Many Numbers of Respondents and Items for the Respondent Profile (and for the Item Profile).

	Group perspective:	
	Few respondents (items)	*Many* respondents (items)
Few items per respondent (respondents per item)	Smooth and stretched group profile Imprecise individual profiles Uncertain group profile • Broad confidence intervals Uncertain group behaviour occurrence	Smooth and stretched group profile Imprecise individual profiles Certain but 'underestimated' group profile • Narrow confidence intervals Certain group behaviour occurrence
Many items per respondent (respondents per item)	Peaked and less stretched group profile Precise individual profiles Uncertain group profile • Broad confidence intervals Uncertain group behaviour occurrence	Smooth and less stretched group profile Precise individual profiles Certain group profile • Narrow confidence intervals Certain group behaviour occurrence

4.3.2.3. Simulation examples with smoothing

The behaviour profiles in Graphs 3 and 4 of Figure 4.4 in section 4.3.2.1 show large and modest peaks respectively. As an illustration, see Figure 4.5 for a few smoothing results using a Gaussian kernel. Based on the minimization of the mean integrated squared error (see Dekking, Kraaikamp, Lopuhaä, & Meester, 2005; Silverman, 1986), respective bandwidths of 0.07 and 0.04 are used. Note that we selected these bandwidths on the basis of standard heuristics, as our main purpose here is to simply illustrate the idea of profile smoothing. More research is needed into the selection of the optimal bandwidth for the use of smoothing behaviour profiles.

Graphs 1 and 2 in Figure 4.5 show the sampled profiles from Graphs 3 and 4 in Figure 4.4 respectively, overlaid with their smoothed profiles. Concerning Graph 1, note that although the profiles have changed firmly and lost much of their detail, the confidence intervals are still quite broad. Relative to the profiles in Graph 1, the shape of the sampled profiles in Graph 2 is largely preserved by the smoothed profiles. Also the confidence intervals of the smoothed profiles in Graph 2 have become more narrow in comparison to those of the sampled profiles. See section 4.4.3 for a smoothing example based on LISS Panel survey data.

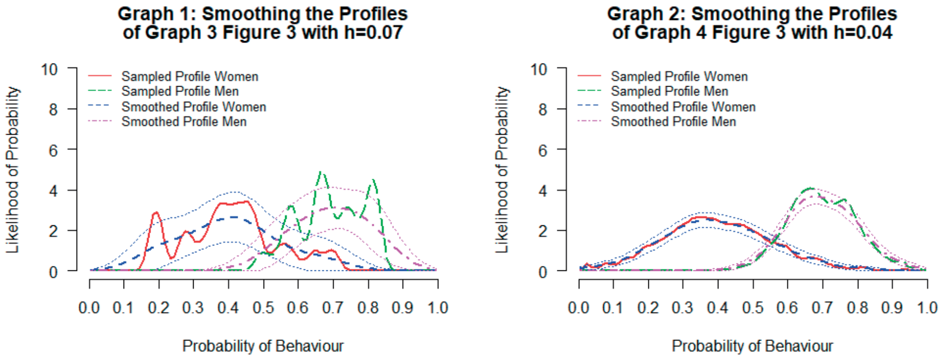


Figure 4.5. Fictitious Examples of Smoothed Behaviour Profiles Accompanied by Their 99% Confidence Intervals.

4.4. SURVEY DATA PROFILE EXAMPLES

To illustrate the estimation of behaviour profiles for real data, we give examples in this section based on survey data from the online LISS Panel. The examples are selected randomly, to illustrate the broad variety of options in constructing profiles. For both the item and respondent perspective, the profiles are estimated following the methods described in section 4.3.1 and their uncertainty is shown following section 4.3.2. See Table 4.1 in section 4.1 for the definitions and relevant references, and see Bais, Schouten, and Toepoel (2020) for the coding of the answer behaviours that are used in the examples. See Bais et al. (2019) for the coding of the item characteristics sensitive information and non-centrality as used in the examples. See Table 3.1 in section 3.3 for the surveys and their topics that were used. As our main purpose here is to simply illustrate behaviour profiles from survey data, we do not further elaborate on the design and context of the LISS Panel and surveys.

4.4.1. Examples of respondent profiles

We start by providing two survey data examples of respondent profiles. The first example shows the profile for respondents who stated that they do not know their income. Their probability for the behaviour neutral responding is shown for the eligible items of the survey Personality, which consists of items about one's personal characteristics. See Graph 1 in Figure 4.6. The group consists of relatively *few* respondents who filled out relatively *many* items, which is clear from the graph; several bumpy peaks are present that are accompanied by *broad* confidence intervals. Although the peaks indicate precise individual profiles, this means that the group profile is relatively *uncertain*. Now see the second example in Graph 2 in Figure 4.6. This is the profile for respondents who are 65 years or older. Their probability for the behaviour socially desirable responding is shown for all eligible items across all applicable surveys. The group consists of relatively *many* respondents who filled out relatively *many* items. As can be seen in Graph 2, peaks are absent; the profile is smooth and accompanied by *narrow* confidence intervals. Besides the precise individual profiles, this means that the group profile is also relatively *certain*.

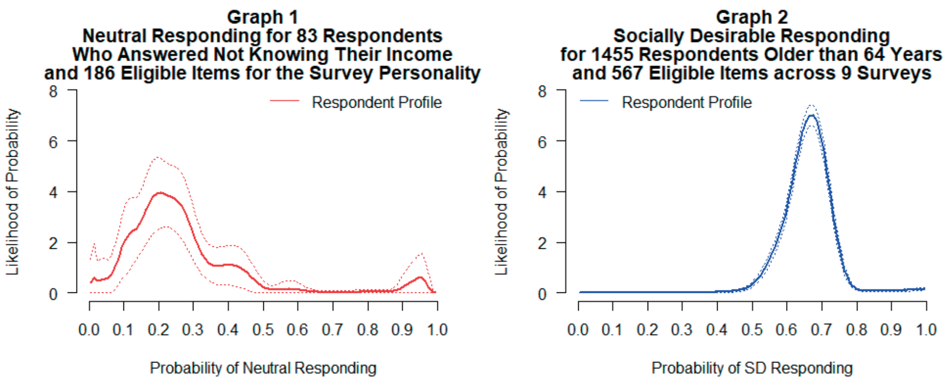


Figure 4.6. Respondent Profiles Accompanied by Their 99% Confidence Intervals, Based on Surveys in the LISS Panel.

4.4.2. Examples of item profiles

Now we provide two survey data examples of item profiles. The first example shows the profile for the eligible items of the survey Politics, consisting of items about political attitudes, that were coded as sensitive. See Graph 1 in Figure 4.7 for their probability for the behaviour extreme responding. The group consists of relatively *few* sensitive items that were filled out by *many* respondents, which is clear from the graph; several peaks are present. The peaks are accompanied by *broad* confidence intervals along the profile. Although individual profiles are precise, this means that the group profile is relatively *uncertain*. See Graph 2 in Figure 4.7 for the second example. This is the profile for all items that were coded as non-central. Their probability for the behaviour answering 'don't know' is shown for all eligible non-central items across all applicable surveys. The group consists of relatively *many* non-central items that were filled out by *many* respondents. As can be seen in Graph 2, peaks are small or absent; the profile is relatively smooth and accompanied by *narrow* confidence intervals. This means that the group profile consists of precise individual profiles and is relatively *certain*.

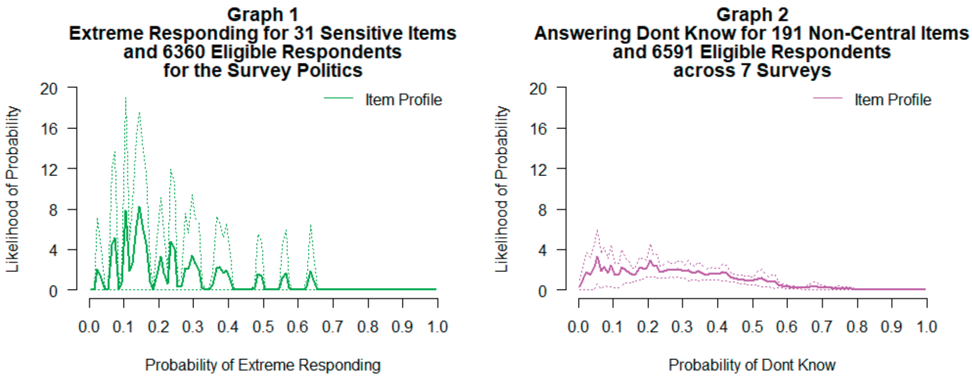


Figure 4.7. Item Profiles Accompanied by Their 99% Confidence Intervals, Based on Surveys in the LISS Panel.

4.4.3. Examples of smoothing an item profile

We use the peaked LISS Panel item profile of Graph 1 in Figure 4.7 consisting of few items and many respondents to discuss smoothing results in practice. See Figure 4.8. In Graph 1, the item profile is smoothed by a Gaussian kernel with a bandwidth of 0.08, based on the minimization of the mean integrated squared error (Dekking et al., 2005; Silverman, 1986). Compared to the original profile, the smoothed profile in Graph 1 does not contain peaks anymore. The profile has become more certain considering the more narrow confidence intervals, but may also be a bit oversmoothed. Using the somewhat lower bandwidth of 0.03, the smoothed profile in Graph 2 contains relatively more of the original shape, but has also become a bit more uncertain compared to the smoothed profile in Graph 1. This refers to a trade-off between the informativity versus the certainty of the profile. By smoothing a peaked profile only modestly, the profile preserves part of its shape and detailed information, but is still not very certain in this situation. By smoothing the profile further, it becomes more certain, but loses more of its shape and information content.

4.5. CONCLUSION AND DISCUSSION

In this paper, we presented the concept of behaviour profiles as a method to summarize survey answer behaviour. We elaborated on defining and estimating the profiles for both groups of res-

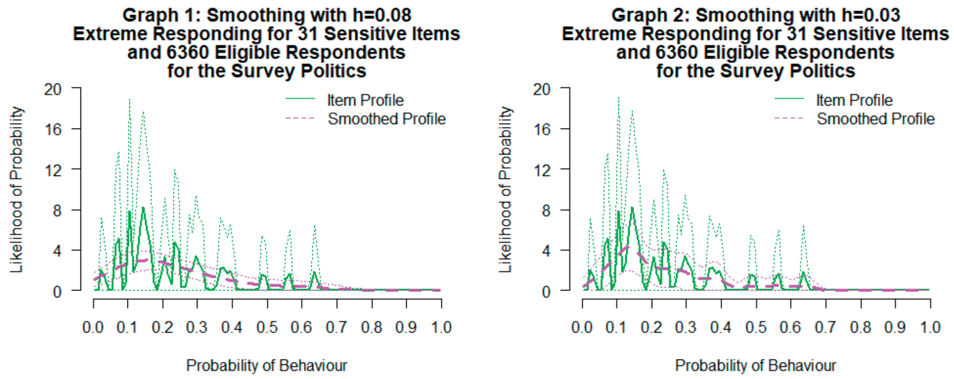


Figure 4.8. Smoothed Item Profiles Accompanied by Their 99% Confidence Intervals, Based on Survey Data in the LISS Panel.

pondents and groups of survey items. The profiles for these groups are respectively called respondent profiles and item profiles. For a group of respondents, the individual respondent profiles are based on the number of applicable items per individual respondent for a specific behaviour and then averaged. For a group of items, the individual item profiles are based on the number of applicable respondents per individual item and averaged. In this way, the respondent and item profiles take into account the varying number of items per respondent and respondents per item respectively.

Behaviour profiles may also be restricted to specific respondent and item characteristics. By means of LISS Panel survey data, we presented real data example profiles that were restricted to specified respondent or item characteristics. For the respondent perspective, we illustrated that the individual respondent profiles become more precise, but the averaged respondent profile becomes more uncertain as *few* respondents fill out more items. As *many* respondents fill out more items, the averaged profile becomes more smooth and certain. The same applies to the item perspective; individual item profiles become more precise, but the averaged item profile becomes more uncertain as *few* items are filled out by more respondents. As *many* items are filled out by more respondents, the averaged profile remains precise and becomes more smooth and certain.

It is not recommended to use behaviour profiles in case of few respondents or few items. Considering the respondent perspective, when group size is large but the occurrence of behaviour for an individual respondent is based on only few items, the respondent profile cannot be estimated precisely. The individual profiles and averaged respondent profile will be relatively flat and broad. This means that the value of the expected behaviour occurrence is estimated only roughly, although the profile appears to be certain. This means that the profile may be either accurate or inaccurate due to the imprecision of the individuals. As our method does not distinguish between accurate and inaccurate in this case, it needs to be determined what numbers of respondents and items can be considered reasonable in follow-up research.

When the individual behaviour occurrence is based on many items but the number of respondents as a group is small, the group behaviour occurrence and respondent profile cannot be estimated well, in spite of the fact that the individual profiles are precise. This means that group size must be large and hence have enough strength in order to empower the high precision of the individuals. Thus, to use respondent profiles, a large group size along with a reasonable number of items per respondent is a prerequisite. The same applies to item profiles, with the concepts of respondent and item exchanged. Determining optimal numbers of respondents and items should be investigated in future research.

Respondent profiles consisting of few respondents who filled out many items (or item profiles consisting of few items that are filled out by many respondents) can be very peaked. In this case, the profiles are in fact ‘too precise’ or ‘too informative’ relative to the group size and can be smoothed by kernel density estimation. Then the profiles lose their peaks and obtain less broad confidence intervals, although they also lose part of their information content. By the use of smoothing in this situation, the profile loses much of its informativity in order to obtain reasonably certain profiles. Moreover, regardless of the degree of smoothing and certainty gain, the actual group size remains small and has too little power to smoothly map the high precision of the individual group members. However, in case of equally large numbers of respondents as numbers of items, the peaks are modest at most. In this situation, one could choose to smoothen the profile only slightly. Then the result clearly shows a more certain profile, while losing only a small amount of informativity. Aside the option of smoothing, we conclude that having marginally less items than respondents in the respondent profile (or marginally less respondents than items in the item profile) refers to an optimal ratio of respondent to item numbers in order to construct accurate profiles that are smooth, precise, and certain. However, this optimal ratio of respondent to item numbers needs further investigation.

Follow-up research may also focus on the model assumption of independence across items, which may not be considered realistic in practice. For a single survey, there is often some degree of interdependence across items for individual answer behaviour. Respondents usually fill out a survey in more or less the same state of mind, mood, and circumstance. Moreover, within a single survey, items sharing the same true values, characteristics, or subtopic are often clustered together. Within such a cluster, items may have their subsequent influence on answer behaviour. One option to investigate the degree of such interdependence is to analyze the extent of clustering that would mean a change in behaviour profiles and their statistical properties. When the effect of clustering on behaviour profiles appears to be limited, this may justify assuming independence across items. But when behaviour profiles would change substantially due to clustering, a second option could be to refine the models by statistically implementing such a cluster effect. The model itself then takes into account interdependence across items. As a result though, behaviour profiles may lose some of their informativity and become less certain.

In summary, behaviour profiles can be constructed and applied best in the situation of relatively many respondents and many items. Then the individual profiles are relatively precise, and the group profile and group behaviour occurrence are relatively certain. In the case of two precise and certain profiles, the profiles may be compared. For instance, men and women can be compared on their probability of answering don’t know, and sensitive and non-sensitive items can be compared on their probability of answering won’t tell. While taking into account the varying number of respondents per item or items per respondent, the relations of respondent and item characteristics to undesirable answer behaviour may be investigated by behaviour profiles. This can be done for single or various surveys, in order to gain insight into the potential consistency of these relations for different survey topics and across multiple surveys.

Survey administrators and practitioners can use behaviour profiles in focussing on these relations. The behaviour profile method visualizes the distribution of survey answer behaviour for any probability occurrence, the uncertainty that comes along with the behaviour occurrences, and the locations where types of respondents or items may show a high degree of a behaviour. This means that a single behaviour profile is relevant and informative on its own, as subgroups that share a characteristic that is not shared by the remainder of the group may be detected in the profile. Behaviour profiles can also be constructed for all combinations of multiple respondent and item characteristics. For instance, respondent profiles could be constructed for younger men and older women who fill out a subset of items that contain both sensitive information and difficult language. Hence, any specific subgroup may be compared to any other specific subgroup in order to clarify relations of respondent and item characteristics to answer behaviour. Additionally, our behaviour profile method can be used for any survey mode and device.

Survey constructors and researchers may use relations that are exposed in practice to adapt survey and questionnaire design features to respondent and item characteristics (see Calinescu & Schouten, 2016; Schouten et al., 2013a). This may be considered per survey and topic, but also across surveys when specific relations appear to be structural and consistent, regardless of survey or topic. In this case, overviews or schemes of such consistent relations may be constructed that give an instant and detailed summary of the relations. These relations can be used to adapt surveys to subgroups of respondents or items that are sensitive to showing specific answer behaviour. Thus, schemes could function as a guide for survey constructors and administrators in putting together and setting out a survey.

δ

$\delta\ \delta$

$\delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta$

$\delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta\ \delta$

Chapter 5

Is undesirable answer behaviour consistent across surveys? An investigation into respondent characteristics

This chapter (additional section 3.6 excluded) is currently under review for Survey Methodology (SMJ). An earlier version of this chapter was denoted a revision by SMJ.

Author contributions: FB, BS, and VT designed the study. CentERdata and Statistics Netherlands provided the survey data. FB prepared the data, performed the statistical analyses, and wrote the paper. BS and VT critically reviewed the paper.

Abstract

In this study, we investigate to what extent the respondent characteristics age and educational level may be associated with undesirable answer behaviour consistently across surveys. We use data from panel respondents who participated in ten general population surveys of CentERdata and Statistics Netherlands. A new method to visually present answer behaviour and an inventive adaptation of a non-parametric effect size measure are used. The occurrence of undesirable answer behaviour of respondents with specific characteristics is summarized in density distributions that we refer to as respondent profiles. An adaptation of the robust effect size Cliff's Delta is used to compare respondent profiles on the potentially consistent occurrence of undesirable answer behaviour across surveys. The results show more undesirable answer behaviour for specific age and educational subgroups than others by considering all surveys taken together. These groups vary dependent on the specific undesirable answer behaviour. The results do *not* show consistent undesirable answer behaviour across individual surveys: Age and educational level are associated with a relatively *higher* occurrence of undesirable answer behaviour for some surveys, but a relatively *lower* occurrence for other surveys. We conclude that the occurrence of undesirable answer behaviour across surveys may be more dependent on the survey and its items than on respondent's cognitive ability.

5.1. INTRODUCTION

The relation between answer behaviour in surveys and measurement error has been studied extensively. Measurement error refers to the extent to which a response deviates from the true value that a survey question was intended to measure (De Leeuw, Hox, & Dillman, 2008). The occurrence and size of measurement error and hence data quality can be influenced by respondent characteristics (Olson & Smyth, 2015; Tourangeau et al., 2000). Respondent characteristics can be thought of as fixed tendencies of a respondent that may lead to undesirable answer behaviour, like satisficing (Holbrook et al., 2003; Kaminska et al., 2010). When respondents satisfice, they take short-cuts in the question-answering process. Satisficing can be seen as the outcome of the interaction of question difficulty, motivation, and cognitive ability (Krosnick, 1991, 1999; Krosnick et al., 1996). Cognitive ability may be considered a characteristic of the respondent that is relatively constant over time. A straightforward proxy for cognitive ability like age or educational level may be used as a background variable to investigate its relation to answer behaviour. Background variables may not be free of measurement errors themselves, but these errors are assumed not to relate to answer behaviour and to be relatively stable over 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 specific answer behaviour, it is hard to say whether this is typical for that specific respondent. For instance, a respondent could fill out a single battery or set of five 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 study, we investigate the relation between cognitive ability and consistent undesirable answer behaviour.* For this purpose, we use the respondent background variables age and educational level as proxies for cognitive ability.

Investigating the relation between cognitive ability and undesirable answer behaviour is not new. However, this relation has not previously been investigated for a large sample of panel respondents across many surveys. To empower finding potential consistency for types of respondents in showing specific undesirable answer behaviour, we use data from ten large population surveys administered by CentERdata in the LISS Panel. These surveys vary broadly in topic and contain many different kinds of items. By including many different surveys, variation will be present in survey topic and design. As a result of this variation, we assume that each survey has its own specific effect on the answer behaviours. In our study, we want to distinguish respondent behaviour that is survey-specific from behaviour that occurs consistently across surveys. In order for respondent consistency to appear, it needs to predominate a single survey topic or survey design effect by definition. In other words, we need the full presence of topic and design variability to investigate behaviour consistency across various surveys.

Note that answer behaviour may or may not be undesirable (see Bais et al., 2020 for an elaboration). Considering the aforementioned example, filling out the first answering option for all five items of a battery may refer to satisficing or to truthful responses. In the case of satisficing, we could say that this answer behaviour is undesirable. In the case of truthful responses, the behaviour is not undesirable. Our idea is that answer behaviour may refer to being undesirable as it is consistently shown for the more surveys. The more consistent the behaviour, the more likely it becomes that the respondent is showing a personal pattern or style, and the more undesirable the behaviour may be considered. Therefore, the term 'undesirable' is inherently potential when used throughout this study. In summary, by using ten large different surveys to detect potential behaviour consistency and to indicate the extent to which behaviour may be undesirable, the foundation to answer our research question is solid and powerful.

This study reads as follows: In section 5.2 of this study, we briefly elaborate on the theoretical framework on which our main research question is based. In section 5.3, we describe the data, methods, and statistics that were used to compare the different age and educational categories for each undesirable answer behaviour across surveys. As a method to detection of consistent undesirable answer behaviour, we use so-called 'respondent profiles', as suggested and explored by Bais, Schouten, and Toepoel (2021). In section 5.4, we show all statistical results and give answers to our main research question. In section 5.5, we conclude with a discussion of these results and make suggestions on how to proceed.

5.2. THEORETICAL FRAMEWORK

Cognitive ability may be considered a stable personal characteristic that has its influence on answer behaviour (Krosnick 1991, 1999; Krosnick et al., 1996). For our study, we consider the respondent characteristics age and educational level as proxies for cognitive ability to investigate its relation to specific undesirable answer behaviour. Both age and educational level have been shown to be related to undesirable answer behaviour and hence survey data quality (Krosnick 1991, 1999; Krosnick et al., 1996). Older and lower educated respondents show less accurate answer behaviour than younger respondents (Andrews & Herzog, 1986) and higher educated respondents (Antoni et al., 2019), and a less stable attitude reliability measurement than younger and higher educated respondents (Alwin & Krosnick, 1991). See Table 5.1 for an overview of the age and educational categories as used in this study, and relevant literature.

Table 5.1. Respondent Characteristics, Their Categories, and Selected Relevant Literature.

Respondent characteristic	*Categories of the respondent characteristics in this study*	*Selected relevant literature*
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; Pickery & Loosveldt 1998; Schonlau & Toepoel 2015; Zhang & Conrad 2014
Education	1. Primary school 2. VMBO: intermediate secondary 3. HAVO/VWO: higher secondary 4. MBO: intermediate vocational 5. HBO: higher vocational 6. WO: university	Aichholzer 2013; Alwin & Krosnick 1991; Greenleaf 1992; He et al. 2014; Krosnick 1991; Krosnick & Alwin 1987; Krosnick et al. 2002; Marín et al. 1992; McClendon 1986, 1991; Narayan & Krosnick 1996; O'Muircheartaigh et al. 2000; Pickery & Loosveldt 1998; Schuman & Presser 1981; Zhang & Conrad 2014

5.2.1. Satisficing and sensitivity-based answer behaviour

In this study, we include two overarching kinds of undesirable answer behaviour: Satisficing behaviour, and behaviour that is based on sensitive content. Satisficing behaviour refers to taking short-cuts in the question-answering process. The behaviour is positively related to item difficulty and can be the outcome of low cognitive ability (Heerwegh & Loosveldt, 2011; Krosnick 1991, 1999; Krosnick et al., 1996). As a result of satisficing, respondents may show one of the following six specific undesirable answer behaviours: Answering 'don't know', acquiescence, neutral responding, extreme responding, primacy responding, and straightlining. See Table 4.1 in section 4.1 for the meaning of these behaviours and their relevant literature.

Undesirable answer behaviour can also be the result of sensitive survey content. Such behaviour is positively related to item sensitivity and may be the outcome of a lack of willingness from the respondent to give a true answer (Bradburn et al., 1978; Shoemaker et al., 2002; Tourangeau et al., 2000). Sensitive items may involve a threat of disclosure (Lensvelt-Mulders, 2008) or can be experienced as intrusive (Tourangeau et al., 2000; Tourangeau & Yan, 2007). As a result of sensitive content, respondents may give one of the following two specific undesirable answer behaviours: Socially desirable responding and answering ‘won’t tell’. Note that in spite of its somewhat misleading name, the behaviour ‘socially desirable responding’ is in fact undesirable because of its relation to measurement error (see for instance DeMaio, 1984; Heerwegh & Loosveldt, 2011). See Table 4.1 in section 4.1 for the meaning of the behaviours and relevant literature. See Figure 5.1 for the complete theoretical framework.

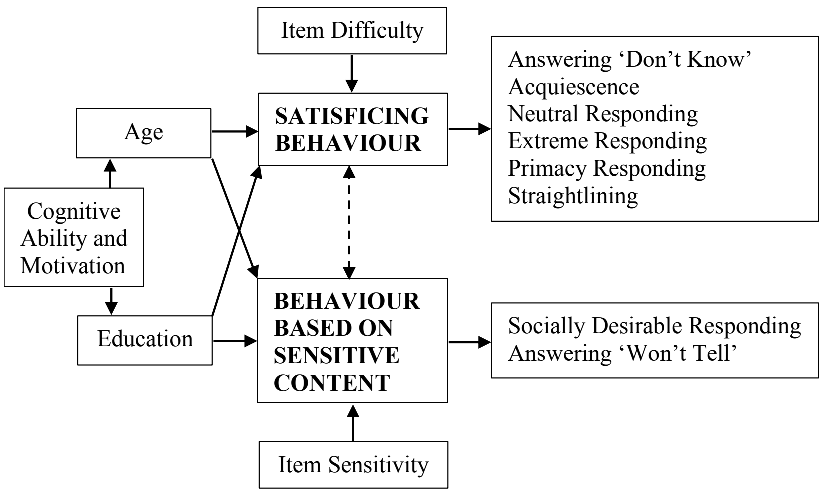


Figure 5.1. Literature-Based Theoretical Framework Including Age and Education.

We need to emphasize that most of the specific undesirable answer behaviours in this study are referred to in some literature as ‘response styles’ (see for instance He & Van de Vijver, 2013; He et al., 2014; Van Herk et al., 2004; Van Rosmalen et al., 2010). We deliberately do not use the concept of response style throughout this paper. The goal of this study is to investigate whether respondents express a stable and consistent pattern or style of specific undesirable answer behaviour across surveys. It would be misleading and confusing to conclude that ‘a specific response style did not appear to be a personal style’. In other words, using the concept of response style would mean the use of two different kinds of styles throughout the study. Therefore, we distinguished between undesirable answer behaviour and the pattern or style of behaviour across surveys that we are actually expecting to find.

5.2.2. Literature overview: Age and education

5.2.2.1. Age

Age and education seem to be related to non-substantive answer behaviour, giving neutral, extreme, and acquiescent answers, and straightlining. Some 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, De Leeuw, & Kreft, 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). Schonlau and Toepoel (2015) found more straightlining for younger than for older respondents, while Zhang and Conrad (2014) did not find a relation between age and straightlining for respondents who speed. 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.

5.2.2.2. Education

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. Some 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, 1984b; 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; Marín, Gamba, & Marín, 1992 - but see Bachman & O’Malley, 1984b for different findings), while mixed results exist concerning choosing middle or neutral options; see Narayan and Krosnick (1996) versus He et al. (2014). Among respondents who speed, more straightlining was found for lower than for higher educated respondents (Zhang & Conrad, 2014). Although response order effects are stronger among less educated respondents

(Krosnick, 1991; Krosnick & Alwin, 1987), evidence for the relation between education and primacy responding was mixed; see Krosnick and Alwin (1987) versus McClendon (1991).

5.2.3. Propositions

Based on this brief literature overview, we expect that age and education are related to satisficing behaviour across surveys. As summarized above, the literature shows that the relation between age or education and satisficing is not unambiguous. Based on relations from the literature that were not rejected or contrasted by other findings, we construct the following propositions:

1) *Older respondents show more extreme and more 'don't know'-answers across surveys than younger respondents.*

2) *Younger respondents show more neutral answers and straightlining across surveys than older respondents*.

3) *Lower educated respondents show more extreme, acquiescent, 'don't know'-answers, and straightlining across surveys than higher educated respondents.*

We do not construct propositions considering behaviour that is based on sensitive content, as the literature does not show a clear relation to age or education. We explore to what extent giving socially desirable and 'won't tell'-answers for different age and educational groups of respondents may be evident across surveys.

5.3. METHOD

5.3.1. LISS Panel and surveys

We selected ten Dutch general population surveys that were administered by CentERdata to 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. The data for the background variables as presented in section 5.2 were also provided by CentERdata. All surveys were administered in computer-assisted format. The ten surveys cover a broad range of topics in the field of general population statistics. See Table 3.1 in section 3.3. Also note the relatively high response rates for all surveys in this table, ensuring a comparable sample representativeness across the surveys. Considering these high and comparable response rates, we do not expect them to have a substantial relation to the occurrence of undesirable answer behaviour within the context of this study.

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 first administration period for each survey was approximately a month. In case of initial non-response, the respondent was sent one or two reminders within this period. To increase the response rate, a second administration period of about a month including one or two reminders was executed for each survey. The respondents were compensated for each survey that they completed. This whole procedure was standardized for each survey, ensuring the comparability of the response rates for the surveys. 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 almost eight. Altogether, the surveys contain 2074 items that were used to cover the undesirable answer behaviours as presented in section 5.2.

5.3.2. Coding the undesirable answer behaviours

All items of all surveys were investigated on whether they were eligible for the selected undesirable answer behaviours separately. The answering categories of the eligible items were coded for each undesirable answer behaviour. In case a category was filled out for which the behaviour occurred, the response was coded as 1; in case a category was filled out for which the behaviour did not occur, the response was coded as 0. For all behaviours, the coding was relatively straightforward. For neutral responding and answering ‘don’t know’ and ‘won’t tell’, the neutral, don’t know- and won’t tell-options respectively were coded as 1, while all other options were coded as 0. For extreme responding, the most negative and most positive option were coded as 1, while all other options were coded as 0. For primacy responding, the first two options were coded as 1, while all other options were coded as 0. This coding method was based on Medway and Tourangeau (2015) for the undesirable answer behaviours that matched our research. See Table 3.2 in section 3.3 and Table 3.6 in section 3.6 for an overview of the undesirable answer behaviours, their eligible kind of items, and the operationalization for each behaviour. See Table D.1 in Appendix D for the proportions of items for which the undesirable answer behaviours are applicable per survey and in total. From here, we discuss the coding process of the undesirable answer behaviours that need more elaboration: Socially desirable responding, acquiescence, and straightlining.

5.3.2.1. Socially desirable responding

About 50% of all items of the involved surveys together were coded as potentially asking for sensitive information by at least one of three coders (see Bais et al., 2019). Next, the answering categories of these items were coded by one other independent coder on whether they may refer to a socially desirable answer. Let us consider the following example:

'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'

The idea is that it is socially desirable to state that it is relatively easy to live off one's income. For our study, we only considered the answering options 8 through 10 as socially desirable options. In this way, we hoped to distinguish respondents who are clearly sensitive to responding socially desirable across surveys from those who are not.

5.3.2.2. Acquiescence: Responding agreeably/affirmatively to a question

The answering categories of all items were evaluated on whether they showed an extent of agreeableness or affirmativeness (see Medway & Tourangeau, 2015). Both positively and negatively worded items were present throughout the surveys to measure acquiescence. Both battery and non-battery items were considered and also subjective variants of the typical answering option 'agree', like 'satisfied', 'applicable', and 'yes', were considered for acquiescence. We chose to include those variants as acquiescent options to capture a broad range of possible acquiescent behaviour across many items. Such a broad range may result in more variation between respondents in showing acquiescence, so that we may better distinguish acquiescent from non-acquiescent respondents. Let us consider the following example:

'I really enjoy responding to questionnaires through the mail or Internet.

totally disagree							totally agree
	1	2	3	4	5	6	7'

For our study, we considered the answering options 5 through 7 as acquiescent options. We decided to consider the option 'somewhat agree' (option 5 in the example) as an acquiescent response as well, as we hoped to distinguish respondents who acquiesce clearly or to only a specific extent from respondents who do not acquiesce.

We need to note that the coding of socially desirable responding and acquiescence is more or less arbitrary; the coding of both behaviours may have been executed either more or less strictly. On the one hand, this means that a response option that was coded as socially desirable or acquiescent may be a socially desirable or acquiescent response for some respondents, but the intended response for others. On the other hand, a response option that was *not* coded as socially desirable or acquiescent may indeed be the intended response for some respondents, but should have been coded as socially desirable or acquiescent for others. However, in order to investigate socially desirable responding and acquiescence at all, a coding threshold that distinguishes the occurrence from the non-occurrence of these behaviours simply needs to be placed at some point. By the current way of coding these behaviours, enough variability between respondents is present in order to distinguish age and educational subgroups that may differ in the occurrence of behaviour. Note that we coded socially desirable responding more conservatively and acquiescence more liberally for this study compared to the coding of these behaviours in Bais et al. (2020) in accordance with their follow-up recommendations (see Table 3.9 in section 3.6.4).

5.3.2.3. Straightlining: Choosing the same answering category for all items in a battery

Our idea is to consider straightlining for a battery only when the very same answering options were filled out *for all its items* (see Schonlau & Toepoel, 2015). When this is the case, the number of times that a '1' is coded is equal to the number of items that the battery consists of. For instance, the occurrence of straightlining for a battery of five items was assigned the code '1' five times. This means that we took into account the length of the battery for this behaviour. In other words, the more items a battery consists of, the stronger the behaviour refers to straightlining in case a respondent filled out the same option for each item. See the following section for an elaboration on how the coding at the item level for all behaviours is transformed into meaningful respondent behaviour summaries.

5.3.3. Respondent profiles

In order to compare respondents on consistent undesirable answer behaviour across surveys, a few aspects need to be taken into account regarding the behaviour. First, the number of items that is applicable to a specific behaviour per survey can be relatively small. This means that an extent of uncertainty exists around the actual occurrence of behaviour, since it is based on a by definition delimited number of items per respondent. To give an example, suppose a respondent A fills out ten items and gives a 'don't know'-answer five times, while another respondent B fills out 100 items and gives a 'don't know'-answer 50 times. Although both respondents can be attributed a probability of 0.50 for answering 'don't know', this probability is relatively more certain for respondent B since it is based on more response data. In other words, the actual occurrence of behaviour for respondents may be more biased as respondents fill out a smaller number of items.

Second, when a survey contains filter questions that may or may not branch out into follow-up questions, each respondent is likely to fill out a different number of items for that survey. Therefore, the actual occurrence of behaviour is indicated with varying uncertainty across different respondents within a survey. Hence, to compare respondents sharing the same characteristic on their undesirable answer behaviour across surveys, simply using individual behaviour proportions is insufficient: A method must be used that takes into account these uncertainties. For this purpose, we introduce the method of using respondent profiles. See Bais et al. (2021) for an extensive statistical elaboration on this method.

5.3.3.1. The respondent profile

The respondent profile is a summary of answer behaviour for a group of respondents. It represents the relative proportions of a specified population group (for instance lower educated respondents) in showing a specified behaviour (for instance answering 'don't know') at all possible probabilities from 0 to 1. In constructing a respondent profile, we make use of the binomial distribution to take into account the abovementioned uncertainties.

Consider an individual respondent *r* who fills out a survey consisting of 50 items of which each offers the answering option 'don't know'. Suppose that the respondent chooses for the 'don't know'-option 10 times out of the 50 possible occasions. Then these numbers are used to construct a binomial distribution. This binomial distribution shows the occurrence of answering 'don't know' for respondent *r*. The likelihood of the behaviour occurrence is calculated for each probability along the probability range from 0 to 1. For practical calculation, we chose for a probability step size interval of 0.01 in order to construct the binomial distribution on the basis of 100 probabilities. We call the resulting binomial distribution for respondent *r* an individual respondent profile. An individual respondent profile is the likelihood curve for the behaviour occurrence and is calculated for each probability from 0 to 1. Hence, to construct the individual profile for respondent r, the likelihood of the behaviour occurrence is calculated on the basis of 10 actual 'don't know'-answers out of 50 possible occasions for all 100 probabilities:

$$\lambda_r(p) = \binom{I_r}{G_r} p^{G_r}(1-p)^{I_r-G_r}, \tag{5.1}$$

where λ_r is the likelihood curve or individual profile for respondent r, p is the probability between 0 and 1 with step size 0.01, I_r is the number of items for which choosing the answer behaviour is possible for respondent r, and G_r is the number of items for which the behaviour is actually shown by respondent r. In order to make individual respondent profiles comparable, we normalize the resulting distribution to obtain an area below the curve of 1 regardless of step size. This is done by dividing each of the likelihoods that the profile consists of by the sum of all likelihoods:

$$\tilde{\lambda}_r(p) = \frac{\lambda_r(p)}{\int_{p=0}^{1} \lambda_r(p)dp}, \tag{5.2}$$

where $\tilde{\lambda}_r$ is the normalized individual profile for respondent r. For a single respondent r, the average or expected value E_r for the behaviour occurrence can be estimated on the basis of the respondent's profile and the integral over p. This means that each probability from 0 to 1 is multiplied by its accompanying likelihood:

$$E_r = \int_{p=0}^{1} p\tilde{\lambda}_r(p)dp. \tag{5.3}$$

The likelihood curve resulting from formula (5.1) and (5.2) is an individual respondent profile. The profile delineates the expected behaviour occurrence across the full potential probability range from 0 to 1 and gives consideration to the amount of occurrence uncertainty. To illustrate the uncertainty on the individual level, consider two respondents who may both have an expected behaviour value of 0.50, but who filled out a different number of items for which the behaviour was possible. For instance, respondents A and B showed a behaviour for 10 out of 20 items and for 30 out of 60 items respectively. See Graph 1 in Figure 5.2. Here, our method takes into account that the expected value of 0.50 is more precisely estimated for respondent B than for respondent A. This is visible by the relatively more narrow and peaked profile for respondent B, indicating that this respondent's behaviour occurrence is relatively more certain.

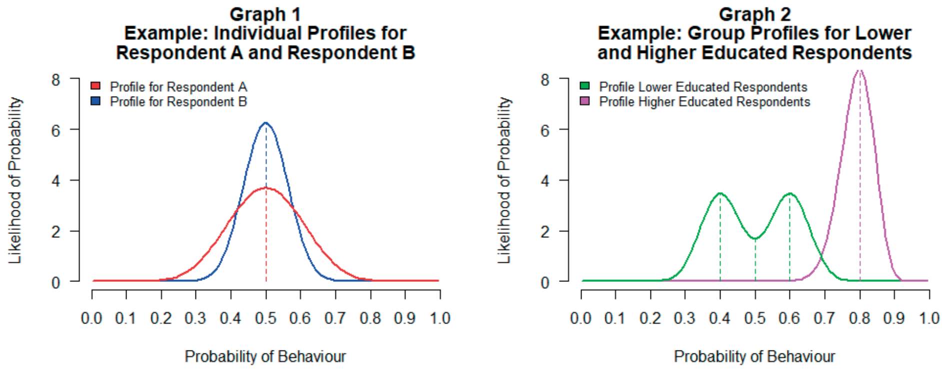


Figure 5.2. Examples of Respondent Profiles with Similar Expected Values (Graph 1) and Different Expected Values (Graph 2).

By considering all respondents who meet the condition of a specific category for a characteristic (for instance lower educated respondents for educational level), the average respondent group profile can be calculated by simply summing their comparable individual profiles and dividing the outcome by the number of respondents:

$$\bar{\lambda}(p) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\lambda}_r(p), \qquad (5.4)$$

where $\bar{\lambda}$ is the respondent profile of the group behaviour occurrence averaged over all respondents, and R is the total number of respondents in the group. By means of this average respondent profile, the averaged expected value $\bar{E}$ for the behaviour occurrence for this group of respondents can be calculated as follows:

$$\bar{E} = \int_{p=0}^{1} p\bar{\lambda}(p)dp. \qquad (5.5)$$

The likelihood curve resulting from formula (5.4) is a group respondent profile. To illustrate the uncertainty on the group level, consider the two groups of lower and higher educated respondents showing a specific behaviour. See Graph 2 in Figure 5.2. The expected values for the groups are 0.50 and almost 0.80 respectively. Our method shows that the expected behaviour occurrence is more precisely estimated for higher than for lower educated respondents. It is also visible that for lower educated respondents, the behaviour occurrence is not centred around the expected group value of 0.50, but around the values of 0.40 and 0.60.

5.3.3.2. Detection of subgroups

Although formula (5.4) refers to a profile for a group of respondents, it does give an indication of individual behaviour. Consider the respondent profile in Figure 5.3 containing individuals on all educational levels. The majority of individuals does not show a specific answer behaviour very often considering the large bump left of the centre. On the right, a small peak is visible that refers to a subgroup of individuals showing the behaviour very often. These respondents may be either lower or higher educated respondents, or they may share another characteristic that is associated with a high behaviour occurrence. The point here is that the respondent profile takes into account the individual behaviour and that subgroups of individuals showing a specific occurrence of behaviour may be identified in the profile.

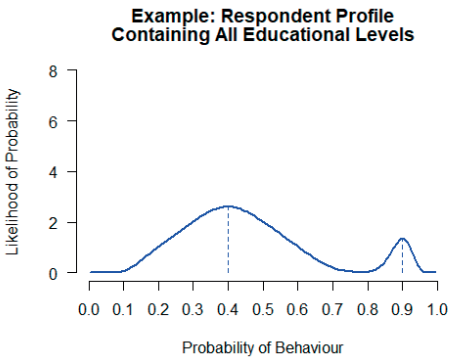


Figure 5.3. Example of a Respondent Profile Containing All Educational Levels.

Note that by using this method of constructing respondent profiles, we assume that individual undesirable answer behaviour is independent across items. This assumption may be partly unjustified, as there may be interdependence across items to some extent in practice. Elaborating on taking into account interdependence across items is beyond the scope of this study. We refer to Bais et al. (2021) for suggestions on how to cope with interdependence across items in future research using respondent profiles.

In summary, 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 specific undesirable answer behaviour (for instance answering don't know) between two groups (for instance lower and higher educated respondents). The next step is to use a solid analysis to compare the behaviour occurrences of two groups.

5.3.4. Cliff's Delta for comparing groups of respondents

To compare two groups or categories of respondents meeting a specific characteristic, an adaptation of the effect size Cliff's Delta (Cliff, 1993, 1996ab) is used. Cliff's Delta δ can be used as a robust alternative to using two independent group means. Using Cliff's Delta for our research asks for an adaptive version of the statistic, as we are not considering data observations but density distributions.

5.3.4.1. The original Cliff's Delta for data observations

Cliff's Delta δ is a robust effect size that indicates to what extent two groups are different. It calculates the probability that a random data observation X_a from a group A is larger than a random data observation X_b from another group B, minus the reverse probability (Hess & Kromrey, 2004; Rousselet, Foxe, & Bolam, 2016; Rousselet, Pernet, & Wilcox, 2017). In practice, this means that each data observation in group A is compared to each data observation in group B. Then a value is assigned to each such comparison. If an observation from group A is larger than an observation in group B, this value is 1. If an observation in group A is smaller than an observation in group B, this value is -1. If the observations in group A and B are equal, this value is 0. Then the total sum of all these values is divided by the total number of comparisons, giving Cliff's Delta. The smaller the overlap between the distributions of two groups, 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). The sample estimate of Cliff's Delta $\hat{\delta}$ is

$$\hat{\delta} = \frac{\sum_{a=1}^{A} \sum_{b=1}^{B} \operatorname{sgn}(X_a - X_b)}{AB}, \tag{5.6}$$

where $(X_a - X_b)$ results in a positive or negative number, or 0, the sign function 'sgn' transforms each positive number into 1 and each negative number into -1, and preserves each 0, and A and B are the sizes of group A and group B respectively.

5.3.4.2. Adapting Cliff's Delta for density distributions

We need to adapt the original Cliff's Delta for our respondent profiles that consist of likelihood distributions. Consider Cliff's Delta for which each specific observation from sample A is compared to each specific observation from sample B exactly once. This means that when an observation with a specific value from sample A occurs three times, this observation value is compared to all observations from sample B 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 (see also Appendix G). When we transpose this idea to respondent profiles, we may consider the behaviour probabilities from 0 to 1 (with a specific step size interval) our 'observations' and the likelihoods for each probability their 'frequencies' or 'weights':

$$\hat{\delta} = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\operatorname{sgn}(P_a - P_b)\bar{\lambda}(P_a)\bar{\lambda}(P_b)}{\sum_{a=1}^{A}\sum_{b=1}^{B}\bar{\lambda}(P_a)\bar{\lambda}(P_b)}, \tag{5.7}$$

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 averaged 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.

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 5.4 for these categories). For our adapted Cliff's Delta, relatively small or unequal sample sizes are no issue.

5.3.5. Confidence intervals for Cliff's Delta and statistics

For each Cliff's Delta, we use 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 bootstrap 10000 category profiles and 10000 overall profiles. We use the so-called empirical bootstrap method, as we cannot make assumptions about the profiles that are non-parametric by definition (see for instance Dekking et al., 2005 for more on this bootstrap method). For each profile, respondents are randomly sampled with replacement and their individual profiles are averaged using formula (5.4). The number of sampled respondents is equal to the number of respondents in the category or overall group respectively. By means of these averaged bootstrap profiles, we calculate 10000 Cliff's Deltas and rank them from low to high. Because of the large number of Cliff's Deltas in our study, we choose to use 99% confidence intervals. This means that we use the 51st and the 9950th Cliff's Delta in the ranking to construct each confidence interval. In the results section, we show Cliff's Delta outcomes for the respondent characteristics and their categories for all undesirable answer behaviours. Each Cliff's Delta is accompanied by its 99% confidence interval. All calculations were done in the programming language R.

5.4. RESULTS

In this section, we first show the Cliff's Deltas for all surveys together as if they were one large survey. Second, we consider the Cliff's Deltas per survey to give an indication about behaviour consistency across surveys to answer our research question. All Cliff's Deltas are obtained by comparing each category profile to the combined profile of the remaining categories. For instance, this means that the profile for respondents aged 15-24 are compared to the profile for the respondents from all other age categories. We chose for this type of comparison, as we are interested in whether a specific subgroup deviates from the complete sample of respondents, considered representative regarding age and education, minus that subgroup.

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 Health, Income, and Personality, this respondent is included in the data analyses for all these surveys. Second, respondents are classified in one category for both age and education. This means that a respondent can be older than 64 years and highly educated, and is included in the data analyses for both characteristics. Hence, respondents are included in each survey and characteristic analysis that is applicable to them. From this, it should be clear that we do not analyze individual respondents in this study, but that we focus on *groups* of respondents sharing the same characteristic. The reason is that we want to relate behaviour to characteristics that are known from the literature to affect undesirable answer behaviour, rather than to isolate individuals and explore potentially related characteristics.

We consider an individual respondent profile based on less than five items non-informative and too imprecise to take into account. Therefore, for each respondent group profile, we only include respondents who filled out at least five items. This means that part of the respondents may be excluded from several subgroups for the analyses. As a result, the occurrence of undesirable answer behaviour for a subgroup after excluding respondents may differ from the initial occurrence of undesirable answer behaviour for that subgroup. Thus, after excluding respondents from a subgroup, the remainder of the subgroup may not be representative for the original subgroup anymore in terms of the initial behaviour occurrence. Therefore, we used two inclusion criteria to guarantee the representativeness of each original subgroup for the analyses: 1) Each subgroup should consist of more than 30% of the number of respondents in the original group, and; 2) the behaviour occurrence in each subgroup should not differ more than 0.02 from the original group's behaviour occurrence. In effect, the second criterion excludes those subgroups for which respondents who fill out less than five items show a different behaviour occurrence than the other respondents in the specific subgroup. See the brief notes about the consequences for group size after applying these criteria for the overall and consistency results in sections 5.4.2 and 5.4.3 respectively.

5.4.1. Exploring survey participation and respondents aged 65 or older

Before elaborating on the main results, we give the outcomes of a few explorations. First, we investigated to what extent frequency of survey participation may have differed between the various age and educational subgroups. See Table 5.2. The average number of surveys that was filled out per respondent overall is 7.6. The average number of surveys per educational subgroup appeared to be relatively high and not to differ much between subgroups. For the age subgroups however, it is evident that younger respondents filled out a lower number and older respondents a higher number of surveys on average.

Second, we used respondent profiles and Cliff's Delta to explore whether the degree of participation made a difference in the occurrence of the specific undesirable answer behaviours taking all surveys together. See section 5.4.2 for the effect size rules for Cliff's Delta. We split up the complete sample of panel respondents into a group who filled out at most eight surveys

Table 5.2. Overall Survey Participation in Total and per Subgroup in Average Number of Surveys (and Absolute Number of Respondents).

	TOTAL	*15_24*	*25_34*	*35_44*	*45_54*	*55_64*	*> 64*
Age	7.6 (6700)	6.0 (838)	6.8 (803)	7.3 (1083)	7.7 (1223)	8.3 (1289)	8.5 (1464)
	TOTAL	*Primary*	*VMBO*	*HAVWO*	*MBO*	*HBO*	*WO*
Education	7.6 (6688)	7.3 (601)	7.7 (1634)	7.3 (791)	7.6 (1549)	7.7 (1504)	7.6 (609)

and a group who filled out at least nine surveys. See Table 5.3. It is clear that participation rate did not affect the occurrence of most behaviours. Not surprisingly, respondents who participated in relatively few surveys showed relatively more 'won't tell'-answers. A second effect was relatively more straightlining in case of a lower participation rate.

Table 5.3. Cliff's Delta δ for Low (Filled Out at Most Eight Surveys) versus High (Filled Out at Least Nine Surveys) Survey Participation per Answer Behaviour*.

	δ SD	*δ PR*	*δ DK*	*δ ST*	*δ WT*	*δ AC*	*δ NE*	*δ EX*
At most eight vs. at least nine surveys	-.09	.07	.08	**.14 ~**	**.29 ***	-.06	.02	-.10

~ → small effect; * → medium effect; # → large effect
* Socially Desirable Responding (SD), Primacy Responding (PR), Answering 'Don't Know' (DK), Straightlining (ST), Answering 'Won't Tell' (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX).

Lastly, respondents aged 75 or older may be even more vulnerable to difficulty in cognitive processing and hence showing undesirable answer behaviour than respondents aged 65-74. Therefore, we compared respondents aged 65-74 to respondents aged 75 or older on their group behaviour proportion. See Table H.1 in Appendix H. Age subgroups of respondents aged 65-74 and aged 75 or older did not or hardly differ for most behaviours and surveys. Only regarding straightlining there were a few striking differences, but interestingly, these showed that respondents aged 75 or older expressed *less* straightlining than respondents aged 65-74. This means that we do not have a reason to split up the age subgroup of 65 years or older into two smaller subgroups.

5.4.2. Overall outcomes for Cliff's Delta

The overall results for Cliff's Delta concern the global picture for specific subgroups for all surveys taken together. We use the rules that $|\delta| < .11$ indicates no effect, $.11 \leq |\delta| < .28$ a small effect, $.28 \leq |\delta| < .43$ a medium effect, and $|\delta| \geq .43$ a large effect, as investigated by

Vargha and Delaney (2000), see also Goedhart (2016). For the analyses, we want to investigate the degree to which each specific subgroup deviates from the complete group. Therefore, a subgroup is always compared to the aggregated total of all remaining applicable subgroups regarding the specific characteristic. See Table 5.4 for the Cliff's Deltas for all surveys taken together. We note that the vast majority of subgroups are both large and representative for the original subgroups after excluding respondents who filled out less than five items (see Table H.2 in Appendix H). See Appendix I for the expected values and Appendix K for the overall profiles.

Table 5.4. Overall Cliff's Delta δ (and its 99% Confidence Interval) Taken over All Surveys, for All Age Categories and All Educational Categories for All Answer Behaviours**.

Characteristic and its category	*Satisficing behaviour*						*Sensitivity-based behaviour*	
	δ DK	*δ AC*	*δ NE*	*δ EX*	*δ PR*	*δ ST*	*δ SD*	*δ WT*
Age 15_24 yrs	**.30** * (.25, .35)	-.06 (-.12,-.00)	-.02 (-.08, .04)	**-.15** ~ (-.21,-.10)	**-.24** ~ (-.30,-.18)	.00 (-.06, .07)	-.04 (-.09, .01)	**.25** ~ (.20, .31)
Age 25_34 yrs	**.11** ~ (.05, .16)	.05 (-.00, .11)	-.06 (-.12,-.00)	-.08 (-.14,-.02)	.08 (.03, .14)	**.12** ~ (.06, .17)	.02 (-.03, .08)	.09 (.04, .14)
Age 35_44 yrs	.08 (.04, .13)	-.01 (-.06, .04)	.03 (-.02, .07)	.01 (-.04, .06)	**.13** ~ (.08, .17)	**.19** ~ (.15, .24)	-.02 (-.07, .02)	.08 (.03, .12)
Age 45_54 yrs	.02 (-.02, .07)	-.04 (-.09, .00)	.01 (-.04, .05)	.04 (-.01, .08)	**.13** ~ (.08, .17)	**.11** ~ (.07, .16)	-.01 (-.05, .03)	.02 (-.02, .06)
Age 55_64 yrs	**-.15** ~ (-.19,-.11)	.03 (-.01, .07)	-.02 (-.06, .02)	.06 (.01, .10)	.06 (.03, .10)	**-.12** ~ (-.16,-.08)	.02 (-.02, .06)	-.06 (-.10,-.02)
Age 65_older	**-.20** ~ (-.24,-.16)	.02 (-.02, .06)	.04 (.00, .08)	.05 (.01, .09)	**-.17** ~ (-.20,-.13)	**-.22** ~ (-.26,-.18)	.02 (-.02, .06)	**-.17** ~ (-.20,-.14)
Education Primary	**.20** ~ (.14, .26)	**-.13** ~ (-.19,-.06)	**.14** ~ (.08, .20)	.03 (-.04, .10)	**-.21** ~ (-.27,-.15)	**-.14** ~ (-.20,-.07)	**-.13** ~ (-.20,-.08)	.08 (.02, .14)
Education VMBO	.10 (.06, .14)	**-.18** ~ (-.22,-.14)	**.14** ~ (.10, .18)	.04 (-.00, .08)	**-.13** ~ (-.17,-.09)	-.04 (-.08, .00)	-.08 (-.12,-.04)	.07 (.04, .11)
Education HAVWO	.00 (-.05, .06)	.01 (-.04, .06)	-.10 (-.16,-.05)	-.02 (-.08, .03)	.00 (-.05, .06)	-.04 (-.09, .02)	-.06 (-.10,-.01)	.02 (-.03, .07)
Education MBO	.07 (.03, .11)	-.04 (-.08, .00)	.05 (.01, .09)	-.02 (-.07, .02)	.02 (-.02, .06)	.05 (.00, .09)	-.02 (-.06, .02)	.08 (.04, .11)
Education HBO	**-.17** ~ (-.21,-.13)	**.18** ~ (.14, .22)	**-.12** ~ (-.16,-.08)	-.03 (-.07, .01)	**.12** ~ (.09, .16)	.02 (-.02, .06)	**.13** ~ (.10, .17)	**-.13** ~ (-.16 -.09)
Education WO	**-.21** ~ (-.27,-.16)	**.22** ~ (.16, .27)	**-.18** ~ (-.23,-.12)	.02 (-.04, .08)	**.19** ~ (.14, .24)	**.12** ~ (.07, .18)	**.14** ~ (.08, .19)	**-.13** ~ (-.18,-.08)

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

** Answering 'Don't Know' (DK), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Primacy Responding (PR), Straightlining (ST), Socially Desirable Responding (SD), Answering 'Won't Tell' (WT).

From Table 5.4, it is clear that subgroups for age and education differ in various forms of specific satisficing behaviours overall. Younger and lower educated respondents showed more 'don't know'-answers than older and higher educated respondents. Higher educated respondents showed more acquiescent, but less neutral responses than lower educated respondents. Younger respondents showed less extreme responses than respondents from other age categories. Respondents from the middle age categories showed more primacy responses than both younger and older respondents (see Graph 1 in Figure 5.4), while higher educated respondents showed more primacy responses than lower educated respondents. Respondents from the middle age categories showed more straightlining than older respondents, while higher educated respondents showed more straightlining than lower educated respondents. From Table 5.4, it is also evident that some subgroups for age and education differ for sensitivity-based answer behaviour overall. Younger respondents showed more 'won't tell'-answers than older respondents. Higher educated respondents showed more socially desirable responses (see Graph 2 in Figure 5.4), but less 'won't tell'-answers than lower educated respondents. In summary, overall satisficing and sensitivity-based behaviours are clearly present, in most cases particularly for the youngest, oldest, lowest educated, or highest educated respondent groups.

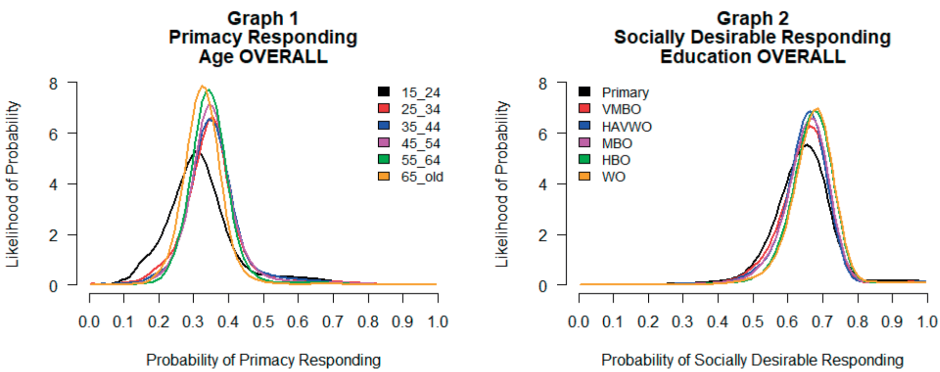


Figure 5.4. Less Primacy Responding for Respondents Aged 15-24 (black) and 65 or Older (orange), and More Primacy Responding for Respondents Aged 35-44 (blue) and 45-54 (magenta) in Graph 1; Less Socially Desirable Responding for Respondents Who Finished only Primary School (black), and More Socially Desirable Responding for Respondents Who Finished HBO (green) or WO (orange) in Graph 2.

An overall effect size for a specific age or educational category does not by definition mean that various surveys show effect sizes; 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 effect 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 and answer our main research question.

5.4.3. Consistency outcomes for Cliff's Delta

These results for Cliff's Delta concern the consistency of subgroups across surveys. This means that we examined effect sizes for all surveys separately. To reveal consistency, we considered the number of surveys for which at least a small effect ($|\delta| \geq .11$) was the result. Such an effect for many surveys that points to the same direction would indicate behaviour consistency across surveys. First, we considered consistency conservatively, as an at least small effect for a specific behaviour and category for *all or almost all* applicable surveys. In this situation, we would draw the conclusion that there is no consistency to be found: *There is no consistent satisficing or sensitivity-based behaviour evident across surveys*. Strictly, this means that the results do not meet any of our propositions. See Table 5.5 containing all results for the behaviours and categories for which more than half of the applicable surveys showed either positive or negative effect sizes: Not one category for a behaviour shows an effect for all or almost all surveys. We note that most subgroups are both large and representative for the original subgroups for each survey after excluding respondents who filled out less than five items (see Table H.3 in Appendix H). See Appendix I for the expected values, Appendix J for the Cliff's Deltas and their confidence intervals, and Appendix K for the profiles for all separate surveys.

Therefore, for each behaviour and category, we considered the number of surveys for which at least a small either positive or negative effect was found. See Table 5.6. It is striking that relatively many cells or category-behaviour pairs showed both positive and negative effects (marked by a '2' in Table 5.6). This means that a category may show *more* of a specific undesirable answer 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) in Table 5.6. Here, this age category showed more 'won't tell'-answers than the other categories combined for one survey, while less 'won't tell'-answers for another survey. For a more liberal perspective on consistency, we elaborate on the cases for which more than half of the applicable surveys showed either positive or negative effect sizes (see Table 5.5). Strikingly, this is applicable to only seven out of the 96 possible cases (as we have results for eight behaviours and twelve categories) and at a maximum of only 75% of the applicable surveys.

Table 5.5. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviours Answering Don't Know, Primacy Responding, and Neutral Responding, for the Applicable Age Categories and Educational Categories for the Applicable Surveys**.

	δ FA	*δ HE*	*δ HO*	*δ IN*	*δ PE*	*δ PO*	*δ RE*	*δ WO*
Answering 'don't know'								
Age 15_24	.09 (.05, .12)			**.46 #** (.41, .51)		**.28 *** (.22, .34)	.05 (.03, .07)	**.24 ~** (.19, .30)
Age 65_old			**-.13 ~** (-.17,-.09)	**-.20 ~** (-.24,-.16)		**-.14 ~** (-.17,-.10)	-.02 (-.03,-.01)	
Education Primary	**.15 ~** (.08, .23)		.08 (-.00, .15)	**.16 ~** (.10 ,.23)		**.17 ~** (.11, .24)	.02 (.00, .05)	**.23 ~** (.15, .31)
Primacy responding								
Age 15_24	**-.36 *** (-.40,-.32)	-.10 (-.13,-.06)		**-.31 *** (-.37,-.26)	**-.18 ~** (-.24,-.12)	**-.11 ~** (-.17,-.06)	-.09 (-.14,-.04)	-.05 (-.09,-.01)
Education Primary	.03 (-.03, .09)	**-.11 ~** (-.16,-.06)		**-.23 ~** (-.29,-.17)	**-.15 ~** (-.22,-.08)	-.08 (-.15,-.01)	**-.14 ~** (-.20,-.09)	-.09 (-.15,-.04)
Education WO	-.10 (-.14,-.05)	.06 (.02, .10)		**.18 ~** (.12, .24)	**.18 ~** (.12, .24)	.03 (-.02, .09)	**.16 ~** (.11, .21)	**.24 ~** (.19, .28)
Neutral responding								
Education WO	.05 (.01, .10)			**-.14 ~** (-.20,-.09)	**-.16 ~** (-.23,-.09)	**-.18 ~** (-.23,-.13)		-.04 (-.09,-.00)

~ → small effect; * → medium effect; # → large effect
** Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO).

For the behaviour answering 'don't know', Table 5.5 shows that respondents 15-24 years of age gave more 'don't know'-answers and respondents of 65 years or older gave less 'don't know'-answers than other respondents for multiple surveys (see Graphs 1 through 4 in Figure 5.5). Both outcomes are in contrast to what we expected. In accordance with what we expected, respondents who finished only primary education gave more 'don't know'-answers than other respondents for various surveys. For primacy responding, we found that respondents 15-24 years of age or who finished only primary education chose less early response options than other respondents for multiple surveys. Respondents who finished the highest educational level chose more early response options and less neutral responses than other respondents for various surveys. These results for primacy and neutral responding were not expected. Our propositions regarding the behaviours acquiescence, straightlining, neutral responding, and extreme responding were not met.

In summary, the results refer to an absence of undesirable answer behaviour consistency across all or almost all surveys: Both satisficing and sensitivity-based answer behaviours did not emerge consistently across surveys. We conclude that respondents' answer behaviour across

Table 5.6. The Categories for Age and Education with either at Least Two Positive *or* Two Negative Effect Sizes Being Assigned a '1' (Unidirectional Results) and the Categories with at Least One Positive *and* One Negative Effect Size Being Assigned a '2' (Contrasting Results) for All Behaviours*.

	Number of surveys	*3*	*4/5*	*4/5*	*4/5*	*4/5/6*	*7*	*7*	*8/9*
	Answer behaviour	*WT*	*NE*	*EX*	*AC*	*DK*	*ST*	*PR*	*SD*
Age	15-24 years	2 **				1	2	1	2
	25-34 years						2	2	2
	35-44 years						1	2	2
	45-54 years						1	1	
	55-64 years					1			2
	65 years or older					1	1	2	2
Education	Primary education		1 ***		1	1		1	2
	VMBO				1			2	2
	HAVWO								2
	MBO								
	HBO				1	1		1	1
	WO		1	2		1	1	1	2

* Answering Won't Tell (WT), Neutral Responding (NE), Extreme Responding (EX), Acquiescence (AC), Answering Don't Know (DK), Straightlining (ST), Primacy Responding (PR), Socially Desirable Responding (SD).

** Consider this value of 2, indicating a contrasting result. This value means that respondents aged 15-24 provided 'won't tell'-answers *more* frequently for *some* surveys, while *less* frequently for *other* surveys, relative to the other age subgroups combined.

*** Consider this value of 1, indicating a unidirectional result. This value means that respondents who only followed primary education showed a frequency of neutral answers consistently in the same direction (either positive or negative) across all applicable surveys showing an effect, relative to the other educational subgroups combined.

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

surveys may be more influenced by the survey and its topic and items than by the age or educational level of the respondent. Even when considering our propositions and the results more liberally, it is evident that many of our propositions were still not met and that several of our propositions were contrasted. From the liberal perspective, 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.

5.5. CONCLUSION AND DISCUSSION

In this study, we investigated to what extent cognitive ability is associated with a high occurrence of undesirable answer behaviour *consistently* across different surveys. For cognitive ability, we used the respondent characteristics age and educational level. The occurrence of undesirable answer behaviour is indicated by varying uncertainty, as every respondent filled

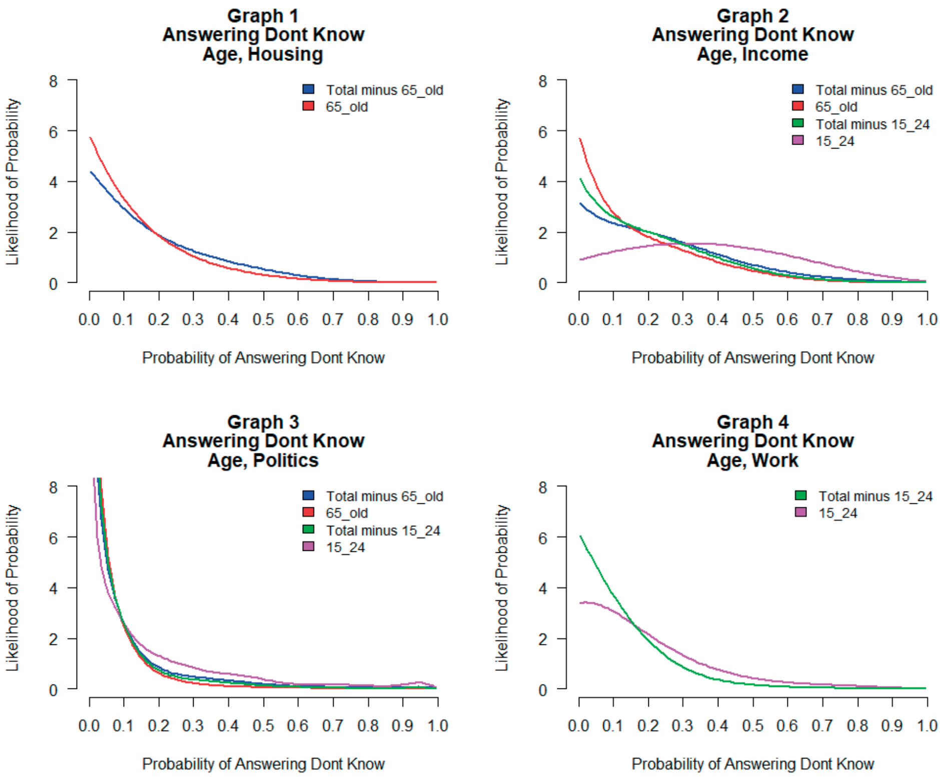


Figure 5.5. Consistently *More* 'Don't Know'-Answers for Respondents Aged 15-24 (magenta) for the Surveys Income, Politics, and Work (see Graphs 2, 3, and 4 Respectively); Consistently *Less* 'Don't Know'-Answers for Respondents Aged 65 or Older (red) for the Surveys Housing, Income, and Politics (see Graphs 1, 2, and 3 Respectively).

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 *respondent profiles*. The behaviour of respondents from a specific category (for instance '15-24 years' for the characteristic 'age') was compared to the behaviour of respondents from the other categories of the characteristic together. For our study, we included the specific satisficing behaviours 'answering don't know', 'acquiescence', 'neutral responding', 'extreme responding', 'primacy responding', and 'straightlining'; the specific sensitivity-based behaviours 'socially desirable responding' and 'answering won't tell'; and the respondent characteristics 'age' and 'education'.

Considering all surveys together overall, specific satisficing and sensitivity-based behaviours are evident for specific age and educational groups. However, *there is no consistency across surveys present for the age and educational categories for any of the behaviours*. This study used response data from a panel consisting of the same respondents. In general, if undesirable answer behaviour consistency was to be expected at all, this should particularly be found in such a panel. If respondents would have any predisposition to show a behaviour style or pattern, this should especially occur while getting familiar with filling out multiple panel surveys within a specific time span. The fact that we did not find such patterns means that cognitive ability is most likely not a predictor of consistent undesirable answer behaviour across surveys.

Considering consistency from a more liberal perspective, specific forms of satisficing across surveys seem evident for specific respondents in particular. Young and lower educated respondents gave relatively more 'don't know'-answers; higher educated respondents chose relatively more answering options early in the list; young and lower educated respondents chose relatively less answering options early in the list; and higher educated respondents showed relatively less neutral responses for multiple surveys. However, there is no category for age or education that showed specific undesirable answer behaviour consistently across *all or almost all* surveys. Obviously, survey items are not randomly clustered, as each survey refers to a specific central topic and contains different types of items. This means that variation between surveys will arise in spite of the fact that respondents have the same disposition to show behaviour from one survey to the other. Still, when considering undesirable answer behaviour consistent, relatively similar patterns should have occurred.

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. More evidence was found for contrasting behaviour than for consistent behaviour across surveys. This evidence is not compatible to our initial theory that specific groups will show consistency for at least some of the specific undesirable answer behaviours across most or all surveys. *Overall, we conclude that the occurrence of undesirable answer behaviour cannot unambiguously be attributed to the respondent's cognitive ability, but may be substantially determined by the characteristics of the survey and its items instead.*

In our study, we did not focus on undesirable answer behaviour of *identified* individual or groups of respondents. For all age and educational categories, each respondent was considered for every applicable survey that the respondent 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. Our purpose was neither to attribute undesirable answer behaviour to individual or groups of identified respondents, nor to compare them between surveys for the same category and behaviour. Considering respondents multiple times, for each applicable survey, was the strength of our study. Taking into account every respondent who fell into a category for every applicable survey resulted in large groups per survey. We compared respondent profiles of large groups for a single category to respondent profiles of large groups for the remaining categories. This means that we focussed on the association between the respondent's *characteristics* and potentially consistent undesirable answer behaviour across surveys. In other words, we did not attribute undesirable answer behaviour to the particular group, but to the specific category in which they were placed. Considered from this approach, we note that we deliberately did not use a more classic method like cross-classified multilevel analysis (see for instance Olson & Smyth, 2015; Olson et al., 2019) that takes into account repeated measurements of individual respondents. The focus of our study was placed on visualizing summaries of undesirable answer behaviour and comparing subgroups who share the same characteristic.

We used the comparisons between a category and the remaining categories together for age and education to answer our consistency research question. For this purpose, we used an adaptation of Cliff's Delta; a robust effect size measure that was both useful because of its many advantages regarding our data, and sufficient for comparing two groups representing a specific category versus the remaining categories. In case of differences in expected group value or group shape, follow-up research may zoom in on these differences to reveal characteristics of subgroups showing relatively more of an undesirable answer behaviour for specific surveys and their topics and items. Other relevant characteristics like respondent gender and origin may also be investigated. In particular, we would be interested in single groups with higher expected values than the other groups for a characteristic and in the respondents who are located to the right of the respondent profile.

Other follow-up research using the profile method may focus on the relation between *item characteristics* and undesirable answer behaviour. Just as respondent characteristics, item characteristics have their influence on data quality and may be associated with measurement error. See Bais et al. (2019), Beukenhorst et al. (2014), Campanelli et al. (2011), Gallhofer et al. (2007), and Saris and Gallhofer (2007b) 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 undesirable answer behaviour. In this way, the presence of item characteristics may be connected to undesirable answer behaviour for the items of whole surveys specifically or across the items of multiple surveys more generally. Based on these associations, questionnaire schemes may be constructed that give an instant overview of the present item characteristics and their relation to undesirable answer behaviour and measurement error.

5.6. ADDITIONAL SECTION

We also investigated the relation of undesirable answer behaviour to a few additional respondent characteristics across surveys. A literature overview and the results for these respondent characteristics were present in early versions of the manuscript, but not in the manuscript that is currently under revision. From here, we elaborate on the selected additional respondent characteristics: ‘Gender’, ‘origin’, ‘income’, ‘primary occupation’, ‘domestic situation’, and ‘being provided a computer’. See Table 5.7 for an overview of the categories of these respondent characteristics as used in this study, and relevant literature. Following this elaboration, we present the results of the analyses for which we used respondent profiles and the adapted Cliff’s Delta.

5.6.1. Additional respondent characteristics

5.6.1.1. 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 specific 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 (Cole, McCormick, & Gonyea, 2012; Zhang & Conrad, 2014). Thus, men and women may differ in various kinds of undesirable answer behaviour across surveys.

Table 5.7. Additional Respondent Characteristics, Their Categories, and Relevant Literature.

Respondent charac-teristics	*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 2014
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 2013; Olson et al. 2019; 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 2013; 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 2013; McClendon 1991; Schräpler 2004
Origin	1. Dutch 2. 1st generation foreign western 3. 1st generation foreign non-western 4. 2nd generation foreign western 5. 2nd generation foreign non-western	Bachman & O'Malley 1984ab; Baumgartner & Steenkamp 2001; Bernardi 2006; Chen et al. 1995; Chun et al. 1974; Cheung & Rensvold 2000; Dolnicar & Grün 2007; Harzing 2006; He & Van de Vijver 2013; Hui & Triandis 1989; Johnson & Van de Vijver 2003; Marín 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
Provided a PC?	1. no 2. yes	Schonlau & Toepoel 2015; Zhang 2013; Zhang & Conrad 2014

5.6.1.2. Origin

Answer behaviour may be influenced by cultural factors and have substantive culture-specific meaning (Cheung & Rensvold, 2000; Smith, 2004). Cultural differences in answer behaviour 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, Lee, &

Stevenson, 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; Marín 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, Campbell, & Yoo, 1974; Dolnicar & Grün, 2007; Watkins & Cheung, 1995; Zax & Takahashi, 1967). But see Marshall and Lee (1998), and Stening and 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 & Grün, 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 (Marín 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 Mexican and Afro-Americans, Hispanics, and Asians, while less for US-born and European Americans, non-Hispanic whites, and Mexicans (see Johnson & Van de Vijver, 2003). Hence, subgroups of different origin may differ in showing undesirable answer behaviour across surveys.

5.6.1.3. Income and primary occupation

Income and primary occupation have been shown to be related to answer behaviour and measurement error (see Butler & McDonald, 1987; Greenleaf, 1992; Lynn & Kaminska, 2013; 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 and McDonald (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. Schräpler 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. Thus, various answer behaviours are to be expected for the respondent characteristics income and primary occupation.

5.6.1.4. 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, Tompson, & Benford, 2010; Lynn & Kaminska, 2013; Olson et al., 2019; Schwarz, Strack, Hippler, & Bishop, 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., 2019), 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 regard this characteristic as an indicator for risk of distraction or multitasking (see Olson et al., 2019; Schwarz et al., 1991).

5.6.1.5. Being provided a computer

Respondents who do not own a personal computer were provided a computer from the panel administrators to complete the LISS core surveys that we used for this thesis (see also Schonlau & Toepoel, 2015). Studies showed a lower prevalence of speeders among respondents who were provided a computer from the panel administrators than the respondents who were not (Zhang, 2013; Zhang & Conrad, 2014). 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, 2014). Both explanations refer to the plausibility of more accurate answer behaviour and hence better survey data quality for respondents being provided a computer.

5.6.2. Results for the additional respondent characteristics

From here, we follow the outline of sections 5.4.2 and 5.4.3. This means that we briefly elaborate on the overall outcomes, the consistency outcomes, and the extent to which both positive and negative effects are found for categories and behaviours.

5.6.2.1. Overall Cliff's Delta outcomes for the additional respondent characteristics

Also for the additional respondent characteristics, it is clear that subgroups differ in specific forms of satisficing behaviours over all surveys taken together. From here, we briefly discuss some of the main overall results from Table 5.8. See Appendix I for the expected values and Appendix K for the overall profiles. Overall, female respondents showed more 'don't know'-answers than male respondents. School-going respondents gave more 'don't know'- and 'won't tell'-answers, while retired respondents gave less 'don't know'- and 'won't tell'-answers than other occupational groups. Respondents without an income, with a lower income (1-1000 EUR), or who filled out 'don't know' or 'won't tell' for the background characteristic 'income', showed more 'don't know'-answers, while respondents with a higher income (2001 EUR or more) showed less 'don't know'-answers than other income groups. Respondents who are from the first generation of non-western immigrants gave more 'don't know'-answers than Dutch and western groups of origin.

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 than most other occupational groups. Respondents without an income or who are from the first generation of non-western immigrants showed less socially desirable responses, but more 'won't tell'-answers than other income groups or groups of origin respectively. The opposite was true for respondents with a higher income (2001 EUR or more), who gave more socially desirable responses, but less 'won't tell'-answers than the other applicable groups. Finally, respondents who filled out 'don't know' or 'won't tell' for the background characteristic 'income' showed more 'won't tell'-answers than the substantive income groups.

Table 5.8. Overall Cliff's Delta δ (and its 99% Confidence Interval) Taken over All Surveys, for All Categories for the Characteristics Gender, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC), for each Answer Behaviour**.

Characteristic and its category	*Satisficing behaviour*						*Sensitivity-based behaviour*	
	δ DK	*δ AC*	*δ NE*	*δ EX*	*δ PR*	*δ ST*	*δ SD*	*δ WT*
Gender ma vs fe	**-.27** ~ (-.30, -.23)	.06 (.02, .09)	-.05 (-.09, -.02)	-.09 (-.13, -.05)	.05 (.01, .08)	.07 (.03, .10)	-.01 (-.04, .02)	-.08 (-.11, -.04)
Domestic Single	-.05 (-.09, -.01)	-.03 (-.07, .02)	-.01 (-.06, .03)	-.01 (-.05, .04)	-.10 (-.14, -.05)	**-.13** ~ (-.18, -.08)	-.03 (-.07, .01)	**-.12** ~ (-.16, -.08)
Domestic Cohab	**-.16** ~ (-.20, -.13)	.06 (.02, .09)	-.02 (-.05, .02)	.06 (.02, .09)	.01 (-.03, .04)	**-.14** ~ (-.17, -.10)	.06 (.03, .09)	-.08 (-.12, -.05)
Domestic Cohab+ch	**.15** ~ (.11, .18)	-.02 (-.05, .02)	.01 (-.02, .05)	-.05 (-.09, -.01)	.09 (.05, .12)	**.21** ~ (.17, .24)	-.02 (-.05, .02)	**.16** ~ (.12, .19)
Domestic Single+ch	**.18** ~ (.10, .25)	**-.13** ~ (-.21, -.05)	.07 (-.02, .15)	.01 (-.07, .09)	**-.12** ~ (-.19, -.03)	.06 (-.02, .15)	**-.11** ~ (-.18, -.04)	.09 (.01, .17)
Domestic Other	.08 (-.06, .21)	.10 (-.04, .23)	-.06 (-.22, .09)	-.05 (-.20, .11)	**-.14** ~ (-.28, .01)	-.09 (-.25, .06)	-.02 (-.15, .11)	.01 (-.13, .15)
Occupation Paid	-.02 (-.06, .01)	.07 (.04, .11)	-.06 (-.09, -.03)	-.04 (-.08, .00)	**.20** ~ (.17, .23)	**.18** ~ (.15, .22)	**.15** ~ (.12, .19)	.03 (.00, .07)
Occupation Famself	.05 (-.02, .12)	.01 (-.06, .08)	-.02 (-.09, .06)	.03 (-.04, .11)	**.13** ~ (.06, .20)	**.14** ~ (.06, .20)	.02 (-.04, .09)	.06 (-.01, .13)
Occupation Seek	.04 (-.06, .14)	-.07 (-.17, .04)	-.05 (-.15, .05)	.06 (-.05, .16)	.02 (-.08, .12)	-.03 (-.15, .07)	**-.20** ~ (-.29, -.11)	.02 (-.07, .12)
Occupation Other	.06 (-.01, .13)	**-.13** ~ (-.19, -.06)	.07 (.01, .14)	**.15** ~ (.08, .22)	**-.12** ~ (-.18, -.06)	**-.19** ~ (-.26, -.12)	**-.40** * (-.46, -.33)	-.02 (-.08, .04)
Occupation School	**.29** * (.23, .35)	-.04 (-.10, .02)	-.04 (-.11, .02)	**-.16** ~ (-.22, -.10)	**-.26** ~ (-.32, -.20)	-.04 (-.10, .03)	-.05 (-.11, .00)	**.22** ~ (.15, .28)
Occupation Housekeep	**.14** ~ (.07, .20)	**-.12** ~ (-.18, -.06)	**.13** ~ (.07, .19)	**.12** ~ (.06, .19)	-.03 (-.09, .03)	-.07 (-.13, .00)	-.08 (-.13, -.02)	.09 (.03, .14)
Occupation Retired	**-.25** ~ (-.29, -.21)	.03 (-.01, .08)	.04 (.00, .08)	.01 (-.04, .05)	**-.15** ~ (-.19, -.11)	**-.20** ~ (-.24, -.16)	.03 (-.01, .07)	**-.17** ~ (-.21, -.14)
Income No income	**.24** ~ (.19, .29)	-.10 (-.16, -.05)	.06 (.00, .11)	-.01 (-.06, .05)	**-.17** ~ (-.23, -.12)	-.04 (-.10, .02)	**-.12** ~ (-.17, -.07)	**.21** ~ (.16, .27)
Income 0001_1000	**.17** ~ (.13, .21)	-.10 (-.14, -.06)	.06 (.02, .10)	.03 (-.02, .07)	-.08 (-.12, -.04)	-.07 (-.12, -.03)	-.10 (-.14, -.06)	.05 (.01, .08)
Income 1001_2000	-.04 (-.08, -.01)	.04 (.00, .07)	.00 (-.03, .04)	-.01 (-.04, .03)	.00 (-.03, .04)	-.03 (-.06, .01)	.01 (-.03, .04)	-.07 (-.10, -.04)
Income 2001_3000	**-.31** * (-.35, -.27)	**.15** ~ (.10, .19)	**-.11** ~ (-.15, -.06)	-.01 (-.06, .04)	**.15** ~ (.11, .20)	.04 (-.01, .08)	**.15** ~ (.11, .20)	**-.17** ~ (-.21, -.13)
Income 3001_more	**-.34** * (-.41, -.27)	**.20** ~ (.12, .27)	**-.18** ~ (-.26, -.11)	-.03 (-.11, .05)	**.23** ~ (.16, .30)	**.17** ~ (.10, .24)	**.16** ~ (.09, .24)	**-.18** ~ (-.24, -.11)
Income DK	**.33** * (.21, .44)	**-.18** ~ (-.31, -.06)	**.15** ~ (.00, .29)	-.02 (-.16, .12)	-.10 (-.24, .05)	.10 (-.05, .23)	-.05 (-.18, .08)	**.28** * (.16, .40)
Income WT	**.26** ~ (.18, .34)	**-.20** ~ (-.30, -.11)	**.11** ~ (.01, .20)	.01 (-.09, .11)	-.01 (-.11, .09)	**.22** ~ (.12, .31)	.01 (-.08, .09)	**.58** # (.50, .64)

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

** Answering 'Don't Know' (DK), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Primacy Responding (PR), Straightlining (ST), Socially Desirable Responding (SD), Answering 'Won't Tell' (WT).

Table 5.8 (continued). Overall Cliff's Delta δ (and its 99% Confidence Interval) Taken over All Surveys, for All Categories for the Characteristics Gender, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC), for each Answer Behaviour**.

Characteristic and its category	*Satisficing behaviour*						*Sensitivity-based behaviour*	
	δ DK	*δ AC*	*δ NE*	*δ EX*	*δ PR*	*δ ST*	*δ SD*	*δ WT*
Origin Dutch	**-.13** ~ (-.18, -.07)	.01 (-.05, .07)	-.06 (-.11, .00)	-.07 (-.12, -.01)	.08 (.02, .14)	-.03 (-.09, .03)	.10 (.05, .15)	-.07 (-.12, -.02)
Origin 1st West	.09 (-.02, .20)	-.02 (-.12, .09)	.05 (-.05, .16)	.06 (-.06, .17)	-.10 (-.20, .01)	.02 (-.08, .13)	-.07 (-.17, .02)	.03 (-.06, .12)
Origin 1st Nonw	**.25** ~ (.15, .35)	-.01 (-.11, .10)	.08 (-.03, .20)	.08 (-.03, .19)	**-.19** ~ (-.31, -.08)	.09 (-.02, .19)	**-.20** ~ (-.29, -.10)	**.19** ~ (.10, .29)
Origin 2nd West	.03 (-.06, .11)	-.02 (-.10, .07)	.02 (-.06, .11)	.03 (-.06, .12)	.04 (-.05, .12)	-.04 (-.13, .04)	-.02 (-.10, .06)	-.01 (-.08, .06)
Origin 2nd Nonw	**.12** ~ (-.01, .27)	.02 (-.14, .18)	.06 (-.10, .21)	**.11** ~ (-.05, .26)	-.08 (-.25, .08)	**.11** ~ (-.04, .26)	-.09 (-.23, .04)	.10 (-.06, .26)
Sim PC no vs yes	-.07 (-.14, .00)	.05 (-.03, .13)	-.08 (-.15, .00)	**-.12** ~ (-.20, -.03)	**.19** ~ (.12, .26)	**.23** ~ (.16, .31)	**.18** ~ (.10, .25)	.04 (-.02, .10)

~ → small effect; * → medium effect; # → large effect
** Answering 'Don't Know' (DK), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Primacy Responding (PR), Straightlining (ST), Socially Desirable Responding (SD), Answering 'Won't Tell' (WT).

5.6.2.2. Consistency Cliff's Delta outcomes for the additional respondent characteristics

Considering the consistency results for the additional respondent characteristics conservatively, we draw the very same conclusion as for age and educational level in section 5.4: *There is no consistent satisficing or sensitivity-based behaviour evident across surveys.* See Table 5.9 that contains all outcomes for the additional behaviours and categories for which more than half of the applicable surveys showed either positive or negative effect sizes: As for age and educational level, not one category for an additional behaviour shows an effect for all surveys. See Appendix I for the expected values, Appendix J for the Cliff's Deltas and their confidence intervals, and Appendix K for the profiles for all separate surveys.

For each additional behaviour and its categories, we considered the number of surveys for which at least a small either positive or negative effect was found. See Table 5.10. The results are completely in line with the results from section 5.4.3: Many cells or category-behaviour pairs showed both positive and negative effects (marked by a '2' in Table 5.10). For a more liberal perspective on consistency, we elaborate on the cases for which more than half of the applicable surveys showed either positive or negative effects sizes (see Table 5.9). Strikingly, this is applicable to only 16 out of the more than 200 possible cases (as we have results for eight behaviours and many categories).

Table 5.9. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviours Answering Don't Know, Primacy Responding, Answering Won't Tell, Neutral Responding, and Straightlining, for the Applicable Categories of the Characteristics Gender, Primary Occupation, Income, and Origin, for the Applicable Surveys**.

	δ FA	*δ HE*	*δ HO*	*δ IN*	*δ PE*	*δ PO*	*δ RE*	*δ WO*	*δ LF*
Answering 'don't know'									
Gender mal vs fem	.05 (.02, .07)		**-.17** ~ (-.21,-.13)	**-.18** ~ (-.21,-.14)		**-.26** ~ (-.30,-.23)	-.01 (-.01, .00)	**-.13** ~ (-.17,-.10)	
Occupation Retired			**-.15** ~ (-.19,-.11)	**-.23** ~ (-.28,-.19)		**-.18** ~ (-.22,-.14)	-.02 (-.03,-.01)		
Income No income	.04 (.00, .08)		**.19** ~ (.09, .29)	**.28** * (.23, .34)		**.19** ~ (.13, .25)	.03 (.02, .05)	**.34** * (.27, .41)	
Income 2001_3000	-.06 (-.09,-.03)		**-.15** ~ (-.20,-.10)	**-.29** * (-.33,-.24)		**-.24** ~ (-.28,-.21)	-.02 (-.03,-.01)	**-.23** ~ (-.27,-.20)	
Income 3001_more	-.04 (-.10, .03)		**-.18** ~ (-.25,-.10)	**-.29** * (-.36,-.22)		**-.30** * (-.36,-.24)	-.02 (-.03,-.01)	**-.21** ~ (-.28,-.15)	
Income DK	05 (-.05, .16)		**.23** ~ (.01, .45)	**.35** * (.21, .48)		**.22** ~ (.09, .36)	.05 (.00, .12)	**.19** ~ (.06, .32)	
Income WT	**.12** ~ (.05, .19)		**.17** ~ (.05, .29)	**.38** * (.30, .45)		**.16** ~ (.06, .26)	.03 (.00, .07)	.10 (.01, .20)	
Primacy responding									
Income 2001_3000	-.02 (-.06, .02)	**.15** ~ (.12, .18)		**.16** ~ (.12, .20)	**.14** ~ (.09, .19)	.04 (-.01, .08)	**.15** ~ (.11, .20)	**.13** ~ (.09, .17)	
Income 3001_more	-.01 (-.07, .06)	**.11** ~ (.06, .16)		**.25** ~ (.18, .32)	**.18** ~ (.09, .25)	.02 (-.06, .09)	**.18** ~ (.11, .24)	**.20** ~ (.14, .26)	
Origin 1st Nonw	**-.22** ~ (-.30,-.13)	**-.14** ~ (-.23,-.06)		-.08 (-.19, .04)	**-.23** ~ (-.34,-.11)	-.06 (-.17, .06)	**-.23** ~ (-.31,-.14)	.08 (.00, .16)	
Origin 2nd Nonw	**-.26** ~ (-.37,-.13)	**-.11** ~ (-.24, .02)		**-.22** ~ (-.39,-.04)	**-.17** ~ (-.36, .01)	-.05 (-.19, .10)	**-.16** ~ (-.28,-.04)	-.01 (-.12, .09)	
Answering 'won't tell'									
Income DK			**.12** ~ (-.04, .29)	**.19** ~ (.06, .33)					**.11** ~ (-.01, .25)
Income WT			**.43** # (.31, .54)	**.54** # (.44, .63)					**.14** ~ (.06, .23)
Origin 1st Nonw			.06 (-.03, .16)	**.11** ~ (.02, .20)					**.14** ~ (.05, .24)
Neutral responding									
Income 3001_more	.01 (-.05, .07)			**-.16** ~ (-.23,-.10)	**-.17** ~ (-.26,-.07)	**-.16** ~ (-.23,-.08)		**-.12** ~ (-.17,-.07)	
Straightlining									
Income WT	.09 (-.01, .19)	**.12** ~ (.02, .22)		**.16** ~ (.05, .27)	**.18** ~ (.08, .27)	**.20** ~ (.10, .30)	.04 (-.05, .14)	.01 (-.02, .06)	

~ → small effect; * → medium effect; # → large effect
** Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).

See Table 5.9 for the more liberal consistency results. Respondents who are female, have no income, or filled out 'don't know' or 'won't tell' for the background characteristic 'income', showed more 'don't know'-answers for a substantial number of surveys. Respondents who are

Table 5.10. The Categories for Gender, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) with either at Least Two Positive *or* Two Negative Effect Sizes Being Assigned a '1' (Unidirectional Results) and the Categories with at Least One Positive *and* One Negative Effect Size Being Assigned a '2' (Contrasting Results) for All Behaviours*.

	Number of surveys	*3/4*	*3/4/5*	*4/5*	*4/5*	*4/5/6*	*6/7*	*7/8*	*8/9*
	Answer behaviour	*WT*	*NE*	*EX*	*AC*	*DK*	*ST*	*PR*	*SD*
Gender	Mal vs fem					1	2	2	2
Domestic situation	Single		1 ***					1	2
	Cohab					1	1	2	2
	Cohab+ch					1	1	2	2
	Single+ch				1	1		1	2
	Other							1	2
Primary occupation	Paid		2	1			2	2	1
	Famself						1		
	Seek								2
	Other						1	1	2
	School	2 **	2			1	2	2	2
	Housekeep			1		1		2	2
	Retired					1	1	2	2
Income	No income				1	1		1	2
	0001_1000				1	1		1	1
	1001_2000								
	2001_3000				1	1	1	1	2
	3001_more		1		1	1	1	1	2
	DK	1			1	1	1		
	WT	1				1	1		
Origin	Dutch				2	1		1	1
	1st West					1		1	2
	1st Nonw	1	1		2	1	1	1	2
	2nd West								
	2nd Nonw		2		2	2		1	2
Sim PC	No vs yes						1	2	2

* Answering Won't Tell (WT), Neutral Responding (NE), Extreme Responding (EX), Acquiescence (AC), Answering Don't Know (DK), Straightlining (ST), Primacy Responding (PR), Socially Desirable Responding (SD).

** Consider this value of 2, indicating a contrasting result. This value means that school-going respondents provided 'won't tell'-answers *more* frequently for *some* surveys, while *less* frequently for *other* surveys, relative to the other occupational subgroups combined.

*** Consider this value of 1, indicating a unidirectional result. This value means that respondents who are single showed a frequency of neutral answers consistently in the same direction (either positive or negative) across all applicable surveys showing an effect, relative to the other domestic subgroups combined.

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

male, retired, or have a higher income (2001 EUR or more), showed less ‘don’t know’-answers. Respondents with a higher income (2001 EUR or more) more frequently chose an early answering option for multiple surveys. Respondents who are from the first or second generation of non-western immigrants less frequently chose an early answering option. Respondents who 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 neutral responding for multiple surveys. Respondents who filled out ‘won’t tell’ for the background characteristic ‘income’ showed more straightlining for multiple surveys.

In summary, also for the additional respondent characteristics, we did not find consistency in undesirable answer behaviour across all or almost all surveys: Both satisficing and sensitivity-based answer behaviours did not emerge consistently across surveys. By these additional results, we empower our conclusion from section 5.4 that answer behaviour across surveys may be more influenced by the survey and its topic and items than by respondent characteristics.

Δ
Δ Δ
Δ Δ Δ
Δ Δ Δ Δ
Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ Δ Δ Δ Δ

Chapter 6

General Conclusion

6.1. Thesis foundation and overall outcome

The starting point of this thesis was to investigate the relation of respondent and item characteristics to survey answer behaviour. Specifically, we aimed to examine whether this relation is consistent across multiple surveys in order to explore the characteristics' potentially structural influence on measurement error and response data quality. To empower this exploration, the thesis is based on ten large Dutch general population surveys that were administered by CentERdata in computer-assisted format to respondents of the Longitudinal Internet studies for the Social Sciences (LISS) Panel. The LISS Panel consists of about 7000 respondents and is based on a probability sample of about 4500 households that is drawn from the population registry by Statistics Netherlands. The surveys are the first wave of the Dutch Labour Force Survey from Statistics Netherlands and nine of the different LISS Panel core studies from CentERdata. These surveys together consist of more than 2000 different items and cover a broad range in topic and design in the field of general population statistics. As far as we know, the relation of respondent or item characteristics to answer behaviour across such a large number of different surveys and items has not been investigated before.

In this thesis, we essentially introduced the new concept of answer behaviour profiles, explored their statistical properties, and used these profiles to investigate the relation between respondent characteristics and potentially undesirable answer behaviour across surveys. As a prelude, we mapped the occurrence of a large number of specific answer behaviours for all surveys and we used cross-classified logistic multilevel models to examine the variances in behaviour between respondents, surveys, and their interaction. Based on these varying occurrences and variances for the different surveys and behaviours, we decided to switch to a method that is new in analyzing survey answer behaviour: The use of 'behaviour profiles'. By constructing such profiles, the answer behaviour is summarized and visualized for the complete behaviour probability range from 0 to 1. This can be done for subgroups of respondents whose profiles are based on a number of items that they filled out (the respondent profile) or for subgroups of items for which profiles are based on a number of respondents who filled out the items (the item profile). The profiles become more certain as group size increases and more precise as they are based on more respondents or items per group member. The combination of large group size certainty and high group member precision results in an accurate behaviour profile.

To answer our main research question about the relation between respondent characteristics and potentially undesirable answer behaviour across surveys, we constructed respondent profiles to examine the following specific answer behaviours: Answering ‘don’t know’, answering ‘won’t tell’, straightlining, socially desirable responding, acquiescence, primacy responding, neutral responding, and extreme responding. These behaviours were investigated on their relation to the following respondent characteristics: Age, educational level, gender, domestic situation, primary occupation, income, origin, and borrowing a sim PC. The categories that the characteristics were divided into were all examined on whether they differed from the remaining categories of the applicable characteristic combined. These comparisons were executed by using an adaptation of the robust and non-parametric effect size measure Cliff’s Delta. This was done for all surveys separately and overall, which means considering all items of all surveys taken together as if they formed one large survey. Regarding most categories of all characteristics, the outcomes showed many significant effects that differed depending on the behaviour and characteristic. This means that a category could show more of one specific behaviour, while less of another specific behaviour, either overall or in considering surveys separately. We concluded that there is no consistent relation between respondent characteristics and undesirable answer behaviour across surveys. Thus, although respondent characteristics are clearly related to undesirable answer behaviour, the occurrence and direction of this relation is substantially dependent on the specific characteristic, survey, and answer behaviour.

6.2. Considerations for follow-up research

One purpose of this thesis was to anticipate on constructing respondent schemes that could present an instant overview of revealed relations between respondent characteristics and measurement error that are consistent across surveys. In case of consistent relations, adaptive survey design could prevent types of respondents to show specific undesirable answer behaviour. As a result, measurement error could be tackled and hence response data quality could be improved. We did however not find consistent relations for any characteristic or behaviour, which means that we cannot construct such overviews. Therefore, we do not recommend adaptive survey design for types of respondents from the perspective of consistent behaviour across surveys. Nevertheless, separate surveys about specific topics could be adapted for types of respondents, although two aspects should be considered. First, a specific survey about for instance health from one administrator is unlikely to be comparable one-on-one to a specific health survey from another administrator in terms of item design. This means that surveys about the same topic from different administrators should be explored on design differences before generalizing survey adaptations towards other surveys about the same topic. Second, many effects that were found for various surveys and behaviours in this thesis were relatively small. This finding raises the question whether it is useful and cost-effective at all to adapt separate surveys to respondents who share the same characteristic.

Following our overall conclusion, we deem it likely that item characteristics have a more substantial and consistent effect on answer behaviour than respondent characteristics have. As a prelude to examining this association in follow-up research, we explored intercoder reliability in coding many different item characteristics for all items of the included surveys. The outcomes showed that intercoder reliability was relatively low for several important item characteristics. Mostly, this meant that the presence versus absence of specific characteristics for items could not be determined unambiguously. Nevertheless, item profiles could be used to investigate the relation between item characteristics and answer behaviour. The characteristics for which intercoder reliability was low mainly concerned those that were coded by three independent coders. One option could be to construct and compare four item profiles based on items for which either zero, one, two, or three coders respectively judged a specific characteristic as present. A variant on this option is to aggregate items for which zero or one coder, and items for which two or three coders judged a characteristic as present. A more conservative option would be to consider only the items for which there was full intercoder reliability. In other words, item profiles based on items for which zero versus three coders judged a characteristic as present could be constructed and compared. For all these options, the profile for items containing for instance difficult language could be directly compared to the profile for items not containing difficult language. When items containing a specific characteristic would be related to a relatively high occurrence of answer behaviour consistently across surveys, item schemes could be constructed. Such schemes would consist of consistent relations between item characteristics and measurement error. Adaptive survey designs could be based on these schemes and minimize specific undesirable answer behaviour with respect to specific item characteristics.

We consider two aspects important in constructing and comparing item profiles. First, groups need to consist of substantial numbers of items. In case a group consists of few items, the profile will be relatively uncertain. When two item profiles are compared of which at least one is uncertain, any statistical comparison measure will be accompanied by a relatively large standard deviation and hence large uncertainty. Therefore, item profiles should ideally be based on large groups of items. Second, one should realize that several options exist in how to analyze item profiles. One option is to compare item profiles as complete groups, for which the adaptation of the robust and assumption-free Cliff's Delta is suitable. Another option is to detect subgroups within a complete group of items, which can be done without necessarily comparing two item profiles. Smaller or larger peaks or bumps in the profile recommend a further zoom-in to consider other item (and/or respondent) characteristics. A subgroup may for instance show a particularly high occurrence of some behaviour compared to the remainder of the group. The members of such a subgroup may share more than just one characteristic. The subgroup could be isolated by using behaviour probability thresholds and hence examined on its additional shared characteristics. How to isolate potential subgroups of items, choose such thresholds, and investigate additional characteristics is for future research.

A final consideration refers to survey mode and device. The results of this thesis are fully based on survey data from computer-assisted self-administered survey mode. This means that we did not take into account mode-specific measurement error. The literature shows that response behaviour may differ between various survey modes like telephone, face-to-face, paper and pencil, and web-based interviews (see for instance Klausch, Hox, & Schouten, 2013; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010). In more recent years, mobile devices like tablets and smartphones are on the rise for respondents to fill out surveys on. Response behaviour from these mobile devices may also differ from behaviour from more traditional devices like personal computers or laptops (see for instance Couper, Antoun, & Mavletova, 2017; Daikeler, Bach, Silber, & Eckman, 2020; Lugtig & Toepoel, 2016). Because of differences in questionnaire and item design between survey modes and devices, we recommend to also include mode and device in comparing answer behaviour for item characteristics in follow-up research. This could be done by constructing two item profiles for the same set of survey items that contain for instance sensitive information, but that are obtained from two different modes or devices. Also revealed mode and device differences could then be added to the aforementioned item schemes and function as guides in adaptive survey design for specific modes and devices.

6.3. Method limitations and strengths

A first limitation of our methodology is that it does not take into account potential interdependence across items within single surveys. As respondents normally fill out a single survey in the same state of mind, mood, and circumstance, and as clustered items may share the same design and subtopic, some extent of item interdependence may occur. As we only made a first effort in shaping and implementing the behaviour profile method, we deemed it beyond the scope of this thesis to take into account item interdependence. Additionally, we have two reasons to believe that item interdependence may not have a substantial influence on the comparison of different subgroups. First, there is no a priori argument to expect why some subgroups would show more interdependence than others. Therefore, we do not presume different subgroups to be affected in their answer behaviour by a clearly varying magnitude. Second, regardless of the degree of item interdependence, we expect patterns of styles of answer behaviour to emerge if these would be actually present in the data. It is reasonable to assume that respondents who have the personal tendency to show specific behaviour do so more frequently for both independent items and clusters of items than respondents who lack this tendency. We recommend potential item interdependence to be investigated in follow-up research. One option is to apply sensitivity analysis to explore the degree of interdependence and its effect on potential change in research outcomes. A more rigid option is to refine the profile method by statistically implementing cluster effects to take into account interdependence across items.

A second limitation of our methodology is that we chose not to use a more traditional model like multilevel analysis to analyze our data. This means that we did not follow identified individual respondents across surveys. We deliberately chose to analyze subgroups of respondents sharing the same characteristic and consider this investigation a first explorative step towards employing our behaviour profile method for survey data. We are enthusiastic about the solid results and the advantages of this new method. First, we succeeded in fully summarizing and graphically visualizing survey answer behaviour for subgroups of respondents. Second, constructing respondent profiles takes into account the uncertainty that comes along with the delimited and varying number of respondents and/or items; the more respondents in the group, the more stable and certain the respondent profile. And the more items that respondents fill out, the more informative and precise the respondent profile can be estimated. And third, by means of full respondent profiles, relatively small subgroups that deviate from the main body of a larger group may be detected. In general, behaviour profiles can be constructed for both individual and groups of respondents or items, for both single surveys and across multiple surveys, for any combination of respondent and item characteristics, and for any survey mode.

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

Appendices

APPENDIX A

Table A.1. Definitions of the Item Characteristics Coded by One Coder and Their Coding Numbers and Categories.

Item characteristic	*Definition*	*Coding numbers and categories*
Number of words	How many words does the text of an item (clarifications included) contain up to the answering categories?	0 <= 25 words 1 > 25 words
Factual filter question	Is the question a factual filter question?	0 no 1 yes
Measurement level	What is the measurement level of the answering categories?	0 closed nominal 1 closed ordinal 2 open numeric 3 open non-numeric
Number of answering options	How many answering options does the item contain?	0 not applicable 1 one or two categories 2 three to five categories 3 six or more categories
Answer as a mark	Does the question need to be answered as a mark from 0 or 1 to 10?	0 not applicable 1 no 2 yes
Polarity of the scale	Is the polarity of the answering scale unipolar or bipolar?	0 not applicable 1 unipolar 2 bipolar
Balance of the scale	Is the answering scale balanced or unbalanced?	0 not applicable 1 balanced answering scale 2 unbalanced answering scale
Neutrally formulated middle category	Does the answering scale contain a neutrally formulated middle category?	0 not applicable 1 with middle category 2 without middle category
Direction of the scale	What is the direction of the answering scale?	0 not applicable 1 from positive to negative 2 from negative to positive
Labels of the scale	Do the categories of the answering scale contain labels?	0 not applicable 1 no labels 2 partly labelled 3 fully labelled
‘Don’t know’ explicitly present	Is ‘don’t know’ an explicit answering option?	0 no 1 yes
Item part of a battery	Is the item part of an item battery?	0 no 1 yes
Relative position of item in battery	What is the relative position of the concerned item in the item battery?	

APPENDIX B: THE PILOT STUDY

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

In the pilot study, a selection of 31 items of the Dutch Labour Force Survey (LFS) and 50 items of the three LISS core studies ‘Personality’, ‘Politics & Values’, and ‘Religion & Ethnicity’ was coded on its item characteristics. This was done by the main author of this study, four researchers from Statistics Netherlands, and three researchers from CentERdata. Six out of these seven researchers were also co-author of this study. We chose these 81 items in such a way to capture as many of the item characteristics as possible. The four chosen surveys differed substantially in content, so that a relatively broad range of topics was covered. After the pilot study, its evaluation, and several follow-up meetings, the list of item characteristics was finalized for the actual coding study with the adjustments from the pilot study. In total, 29 item characteristics were selected for the actual coding study, as the characteristic presumption of a filter question was added to the final list (see section 2.2).

After the pilot study, a few important changes were made. Based on the results of the pilot study, the item characteristic filter question was split up into two separate item characteristics; one item characteristic that asks whether an item is a true filter question (see Table A.1); and a second item characteristic that asks whether an item could make the respondent presume that it would be a filter question, regardless of whether the item *is* a true filter question (see Table 2.1). Also, for the item characteristics sensitive information and centrality (see Table 2.1), the middle option was removed so that only two coding categories remained for these characteristics: Characteristic not applicable and characteristic applicable. This was done because it appeared to be difficult for the coders to choose between two gradual categories of applicability of these characteristics. Finally, a few item characteristics were defined more strictly, as it was not clear for some items what coding category had to be chosen based on their definitions.

APPENDIX C: USING THE FIXED PROBABILITY λ TO CALCULATE INTERCODER RELIABILITY

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

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

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

Table C.1. The Intercoder Reliability Based on the Fixed True Coding Probability λ and the Number of Coders m.

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

APPENDIX D

Table D.1. 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 All Surveys* and in Total (TOT).

	AS	*FA*	*HE*	*HO*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*	*TOT*
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
Average nr. of items/battery	-	5.5	7.6	-	5.7	11.1	6.0	5.8	12.0	-	7.8
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
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
Avoiding follow-up questions	.62	.33	.12	.27	.48	-	.06	.20	.21	.13	.23
Socially desirable responding	.20	.12	.62	.01	.25	.30	.51	.42	.19	.32	.28
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

* Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).

** Consider this proportion of .52 as a clarifying example. This proportion means that 52% of all items of the survey Assets contains a ‘don’t know’-answer as a possible response option.

Note. The behaviours fast and slow responding are not included in this table, as these behaviours are calculated on the basis of all applicable items per survey.

APPENDIX E: MULTILEVEL MODELS

First, we build up the different multilevel models and elaborate on their notation. For the purpose of model comparison, we started with an empty model, Model 0, not consisting of any random intercept at all:

$$\digamma_{i}^{a} = logit\{P[Y_{i}^{a} = 1]\} = \gamma_{00} + e_{i}, \tag{E.1}$$

where latent variable $\digamma_{i}^{a}$ is the propensity that behaviour *a* is shown on item *i*, Y_{i}^{a} is the outcome variable for behaviour *a* to occur on item *i*, γ_{00} is the overall intercept with which answer behaviour *a* is modelled, and e_i is the residual error term for item *i*. This residual error term has the fixed variance θ^2, which is by definition $\frac{\pi^2}{3}$ for the standard logistic distribution (Hox, 2010).

From the empty model, we started building two intercept-only models for obtaining overall respondent and survey variance respectively, resulting in Model 1A:

$$\digamma_{ir}^{a} = logit\{P[Y_{ir}^{a} = 1]\} = \gamma_{00} + u_{0r} + e_{ir}, \tag{E.2}$$

where latent variable $\digamma_{ir}^{a}$ is the propensity that respondent *r* is showing behaviour *a* on item *i*, Y_{ir}^{a} is the outcome variable for respondent *r* to show behaviour *a* on item *i*, u_{0r} is the random intercept for respondent *r* with variance σ^2, and e_{ir} is the residual error term for item *i* and respondent *r*, and Model 1B:

$$\digamma_{is}^{a} = logit\{P[Y_{is}^{a} = 1]\} = \gamma_{00} + v_{0s} + e_{is}, \tag{E.3}$$

where latent variable $\digamma_{is}^{a}$ is the propensity that survey *s* is showing behaviour *a* on item *i*, Y_{is}^{a} is the outcome variable for survey *s* to show behaviour *a* on item *i*, v_{0s} is the random intercept for survey *s* with variance τ^2, and e_{is} is the residual error term for item *i* and survey *s*.

The subsequent model was an intercept-only model for obtaining both respondent and survey variance, Model 2:

$$\digamma_{i(rs)}^{a} = logit\left\{P\left[Y_{i(rs)}^{a} = 1\right]\right\} = \gamma_{00} + u_{0r} + v_{0s} + e_{i(rs)}, \tag{E.4}$$

where latent variable $ƒ^a_{i(rs)}$ is the propensity that respondent *r* is showing behaviour a on item *i* in survey s, $Y^a_{i(rs)}$ is the outcome variable for respondent *r* to show behaviour a on item i in survey s, and $e_{i(rs)}$ is the residual error term for item *i* in the cross-classification of respondent *r* and survey s (see Hox, 2010).

The complete model was an intercept-only model for obtaining respondent, survey, and respondent-survey interaction variance, Model 3:

$$ƒ^a_{i(rs)} = logit\{P[Y^a_{i(rs)} = 1]\} = \gamma_{00} + u_{0r} + v_{0s} + w_{0r,0s} + e_{i(rs)}, \tag{E.5}$$

where $w_{0r,0s}$ is the random intercept for interaction with variance φ^2.

See Table E.1 for the results of the likelihood ratio tests for comparing the different models. As can be seen from Table E.1, the model without the random intercept for interaction refers to the best fit for the behaviours fast and slow responding, and avoiding follow-up questions. The reason is that the interaction variances for the behaviours fast and slow responding do not exist. The interaction variance for the behaviour avoiding follow-up questions was zero. For all other behaviours, the complete model including the random intercept for interaction has the best fit. See Table E.2 for the variances for the final models for all answer behaviours.

Finally, we elaborate on the calculation of the intraclass correlation coefficients (ICCs). The ICCs refer to how much variance is explained at the respondent, survey, and interaction level (see Hox, 2010). We used partially explained variances to calculate the ICCs for respondent, survey, and interaction respectively; the explained respondent variance ρ^a_{resp}, the explained survey variance ρ^a_{surv}, and the explained interaction variance ρ^a_{int} for answer behaviour a. This is done by taking the ratio of these explained variances to the explained total variance (see Hox, 2010):

$$\rho^a_{resp} = \frac{\sigma^2}{\sigma^2+\tau^2 + \varphi^2 + \theta^2}, \tag{E.6}$$

$$\rho^a_{surv} = \frac{\tau^2}{\sigma^2+\tau^2 + \varphi^2 + \theta^2}, \tag{E.7}$$

Table E.1. Chi-Square Statistics (and Accompanying *p* Values) from the Likelihood Ratio Tests for Comparing the Different Models with *df* = 1.

Behaviour (nr. of respondents)	*Model 0 * vs Model 1A **	*Model 0 vs Model 1B **	*Model 1A vs Model 2 **	*Model 1B vs Model 2*	*Model 2 vs Model 3 **
Answering ‘don’t know’ (6145)	48373**(.000)	37118 (.000)	42310 (.000)	53564 (.000)	12312 (.000)
Answering ‘won’t tell’ (3682)	6409 (.000)	1276 (.000)	596 (.000)	5729 (.000)	639 (.000)
Primacy responding (6319)	19277 (.000)	317251 (.000)	314511 (.000)	16537 (.000)	16242 (.000)
Recency responding (6319)	29100 (.000)	93969 (.000)	90183 (.000)	25313 (.000)	18752 (.000)
Straightlining (6234)	120902 (.000)	265810 (.000)	287510 (.000)	142602 (.000)	130004 (.000)
Fast responding (6697)	4921 (.000)	3594 (.000)	4606 (.000)	5932 (.000)	NA***
Slow responding (6697)	1239 (.000)	1503 (.000)	1527 (.000)	1263 (.000)	NA***
Avoiding follow-up questions (6333)	792 (.000)	32977 (.000)	32465 (.000)	280 (.000)	0 (1)
Socially desirable responding (6394)	10631 (.000)	339935 (.000)	336742 (.000)	7438 (.000)	11678 (.000)
Acquiescence (6133)	50217 (.000)	24929 (.000)	24772 (.000)	50059 (.000)	12371 (.000)
Neutral responding (5974)	59710 (.000)	12402 (.000)	12640 (.000)	59948 (.000)	14163 (.000)
Extreme responding (5985)	131632 (.000)	34378 (.000)	37076 (.000)	134330 (.000)	35505 (.000)

* Model 0 No random intercepts; Model 1A Random intercept for respondent; Model 1B Random intercept for survey; Model 2 Random intercepts for respondent and survey; Model 3 Random intercepts for respondent, survey, and their interaction.
** Consider these values as a clarifying example. The large Chi-Square statistic of 48373 is accompanied by a p-value of .000. This means that Model 1A has a better fit regarding the data for the behaviour answering ‘don’t know’ than Model 0. Thus, the random intercept for respondent and its respondent variance should be included in the model.
*** These cells are empty, as the model without the interaction variance was the final model for these answer behaviours. The interaction variances for the behaviours fast and slow responding do not exist.

$$\rho_{int}^{a} = \frac{\varphi^2}{\sigma^2+\tau^2+\varphi^2+\theta^2}. \tag{E.8}$$

Table E.2. Random Intercepts and Their Profile Likelihood-Based Confidence Intervals for the Answer Behaviours for the Final Models over All Surveys Together.

Behaviour (nr. of respondents)	*Random intercept respondent*	*Random intercept survey*	*Random intercept interaction*
Answering 'don't know' (6145)	1.05* (0.98, 1.11)	3.28 (1.34, 11.66)	1.14 (1.09, 1.19)
Answering 'won't tell' (3682)	3.40 (3.09, 3.79)	1.07 (0.35, 5.25)	1.41 (1.22, 1.62)
Primacy responding (6319)	0.03 (0.03, 0.03)	0.60 (0.26, 1.94)	0.12 (0.11, 0.12)
Recency responding (6319)	0.02 (0.02, 0.02)	0.20 (0.09, 0.65)	0.14 (0.14, 0.15)
Straightlining (6234)	0.68 (0.63, 0.73)	5.60 (2.29, 19.88)	2.68 (2.61, 2.75)
Fast responding (6697)	4.58 (4.30, 4.90)	1.73 (0.79, 4.86)	NA**
Slow responding (6697)	1.23 (1.12, 1.34)	0.49 (0.23, 1.38)	NA**
Avoiding follow-up questions (6333)	0.01 (0.01, 0.01)	0.49 (0.22, 1.45)	NA**
Socially desirable responding (6394)	0.03 (0.02, 0.03)	1.76 (0.79, 5.24)	0.15 (0.15, 0.16)
Acquiescence (6133)	0.18 (0.16, 0.19)	0.21 (0.08, 1.02)	0.33 (0.32, 0.34)
Neutral responding (5974)	0.15 (0.14, 0.16)	0.66 (0.24, 3.15)	0.21 (0.20, 0.21)
Extreme responding (5985)	0.52 (0.49, 0.52)	0.34 (0.12, 1.61)	0.61 (0.59, 0.63)

* Consider these values as a clarifying example. The value for the random intercept for respondent considering answering 'don't know' is 1.05. Here, the value has a 95% chance of falling between the profile likelihood-based confidence interval values of 0.98 and 1.11 (see Venzon & Moolgavkar, 1988; Joshi, 2015).
** These cells are empty, as the model without the interaction variance was the final model for these answer behaviours. The interaction variances for the behaviours fast and slow responding do not exist; the interaction variance for the behaviour avoiding follow-up questions was zero.

APPENDIX F: RANDOM INTERCEPT GRAPHS

See below for the caterpillar plots with error bars and the accompanying histograms with normal curves for the random intercepts for respondent, survey, and if applicable their interaction, for each behaviour.

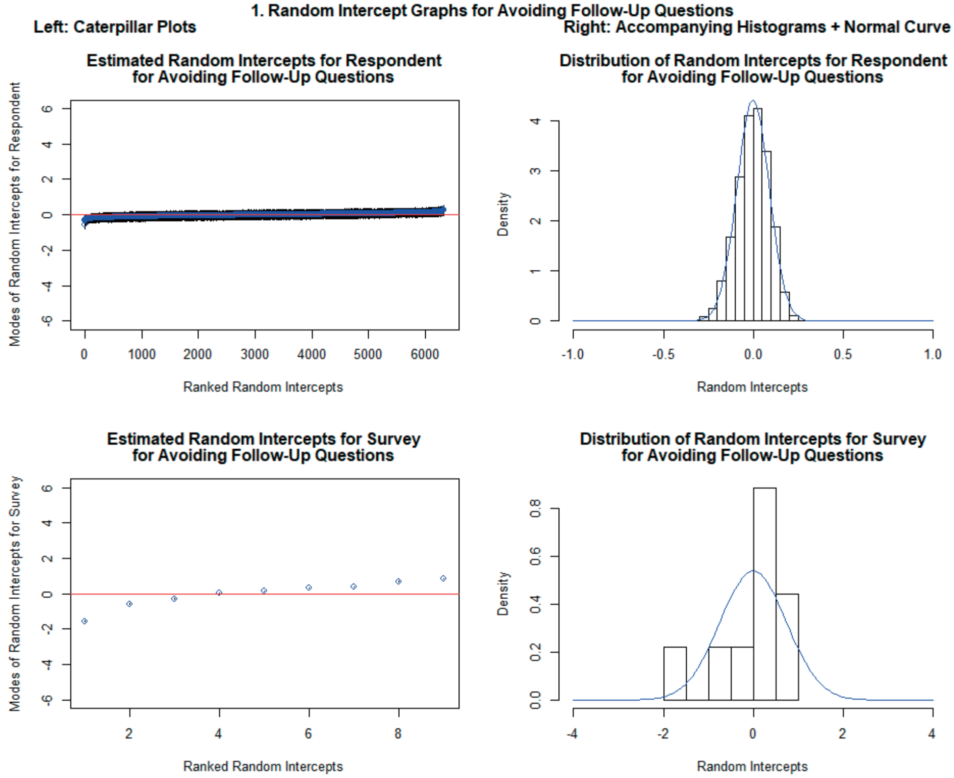


Figure F.1. Random Intercept Graphs for the Answer Behaviour ‘Avoiding Follow-Up Questions’.

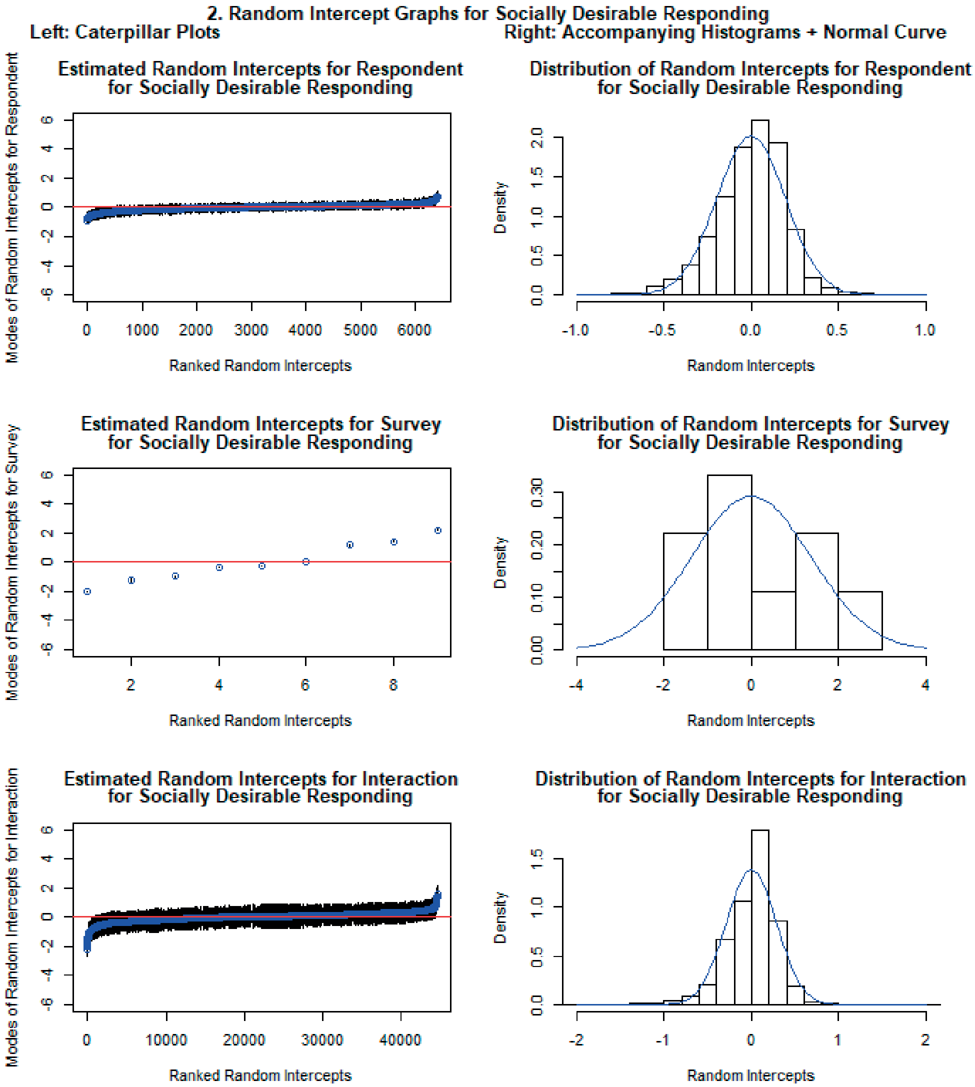


Figure F.2. Random Intercept Graphs for the Answer Behaviour ‘Socially Desirable Responding’.

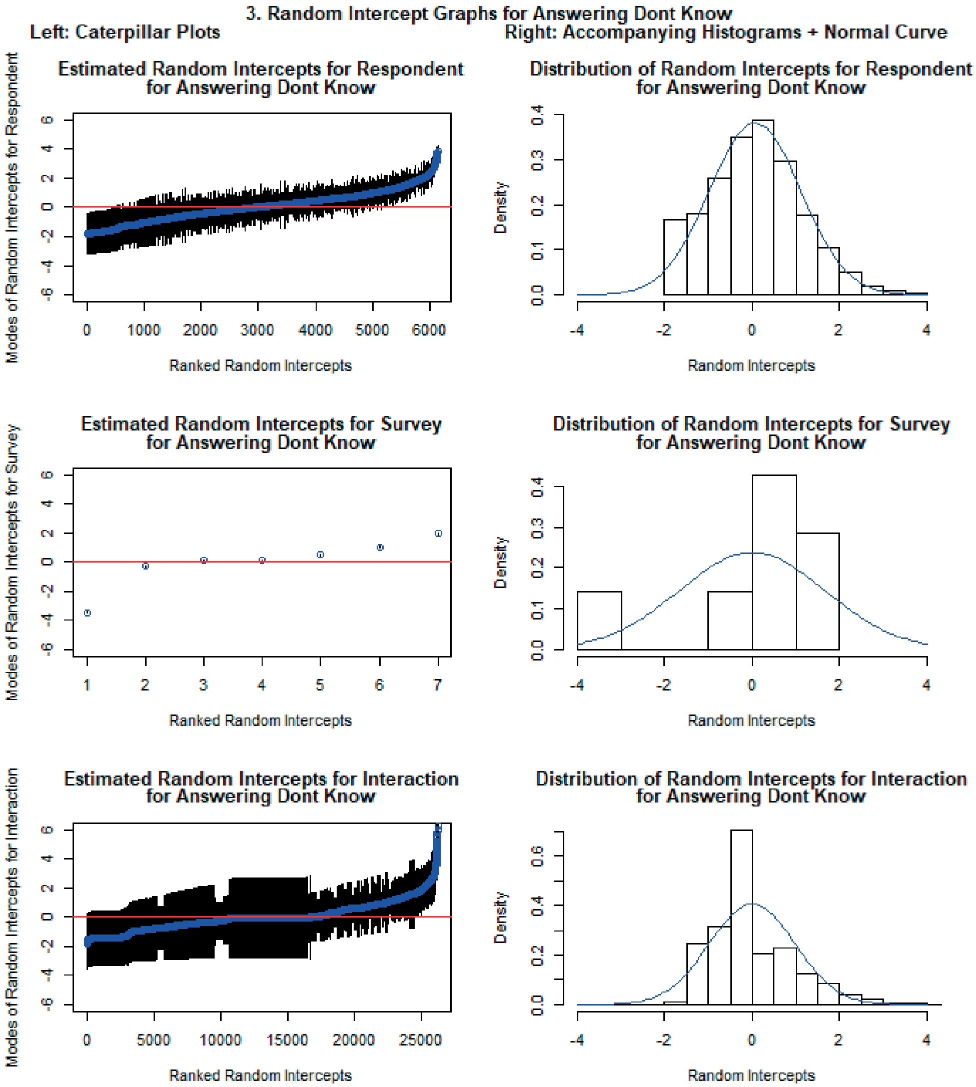


Figure F.3. Random Intercept Graphs for the Answer Behaviour ‘Answering Don’t Know’.

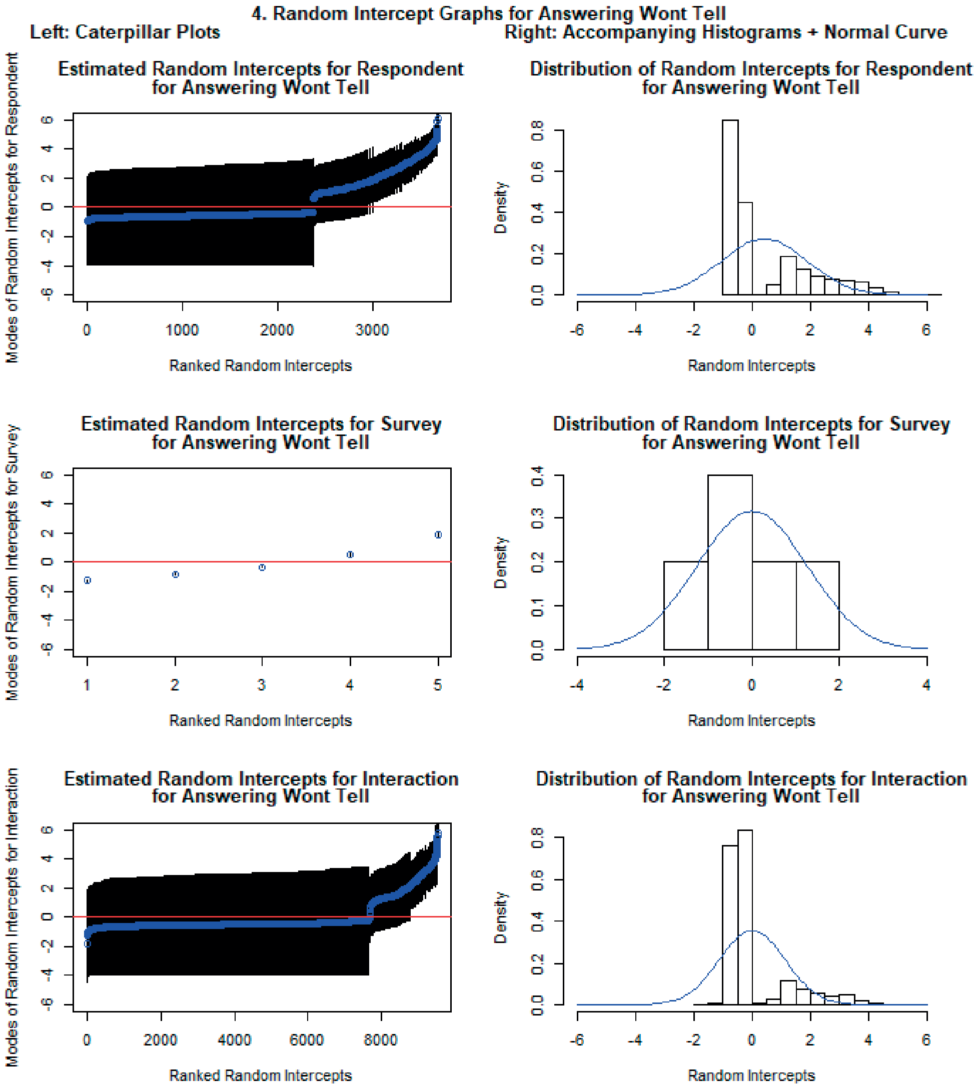


Figure F.4. Random Intercept Graphs for the Answer Behaviour ‘Answering Won’t Tell’.

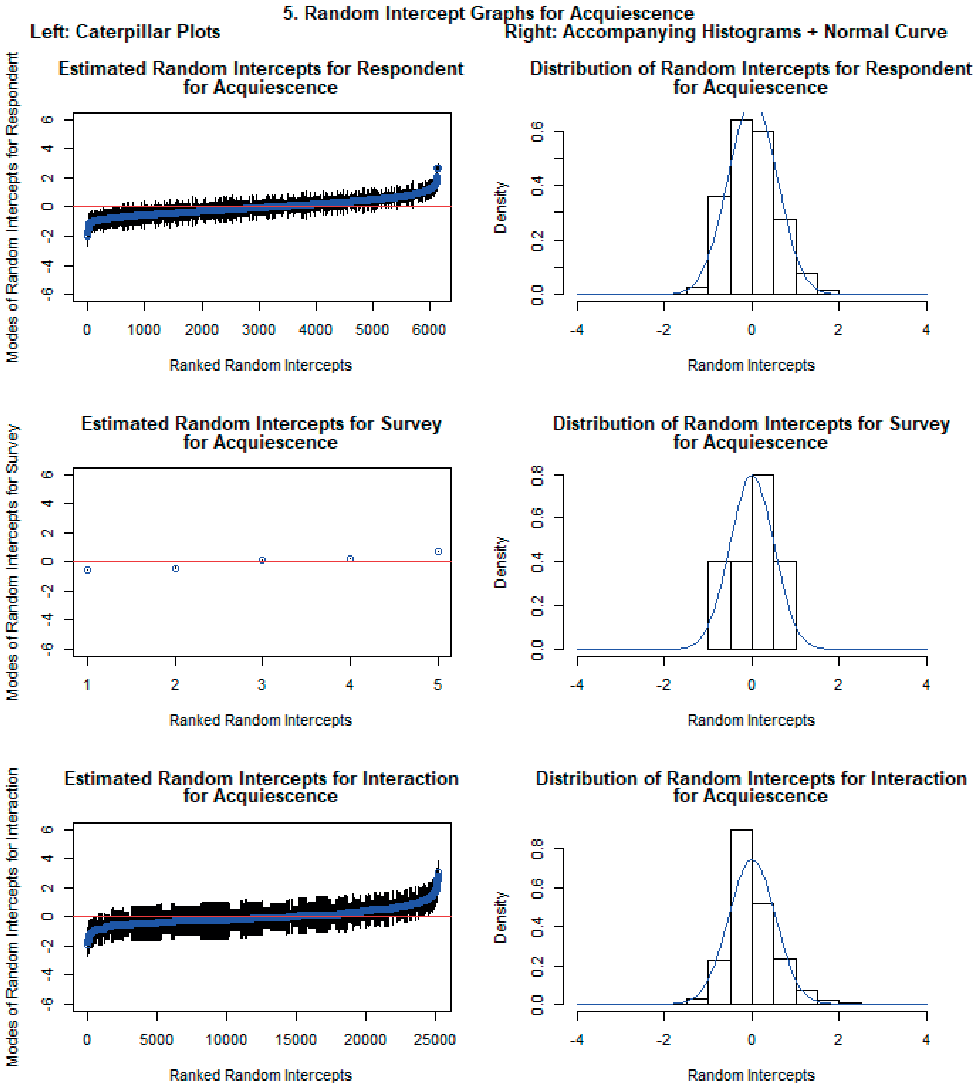


Figure F.5. Random Intercept Graphs for the Answer Behaviour ‘Acquiescence’.

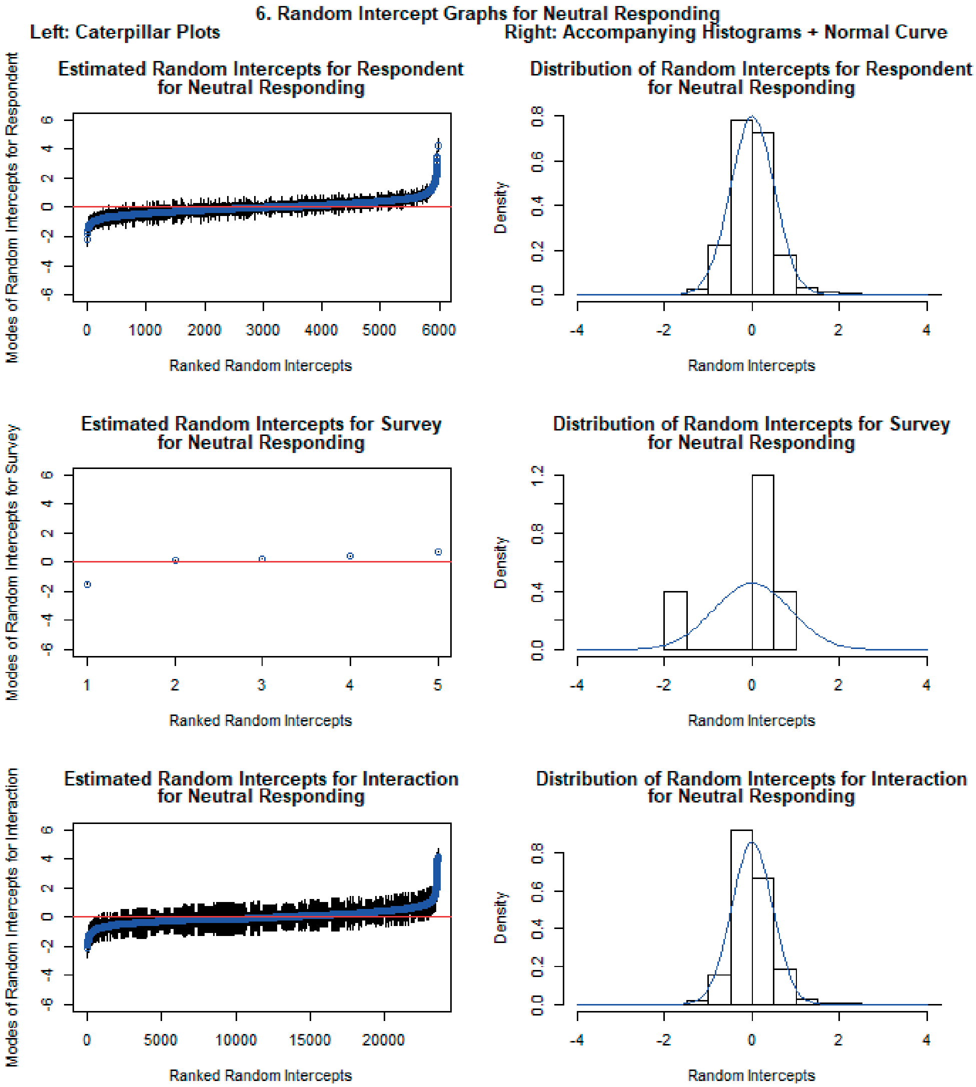


Figure F.6. Random Intercept Graphs for the Answer Behaviour ‘Neutral Responding’.

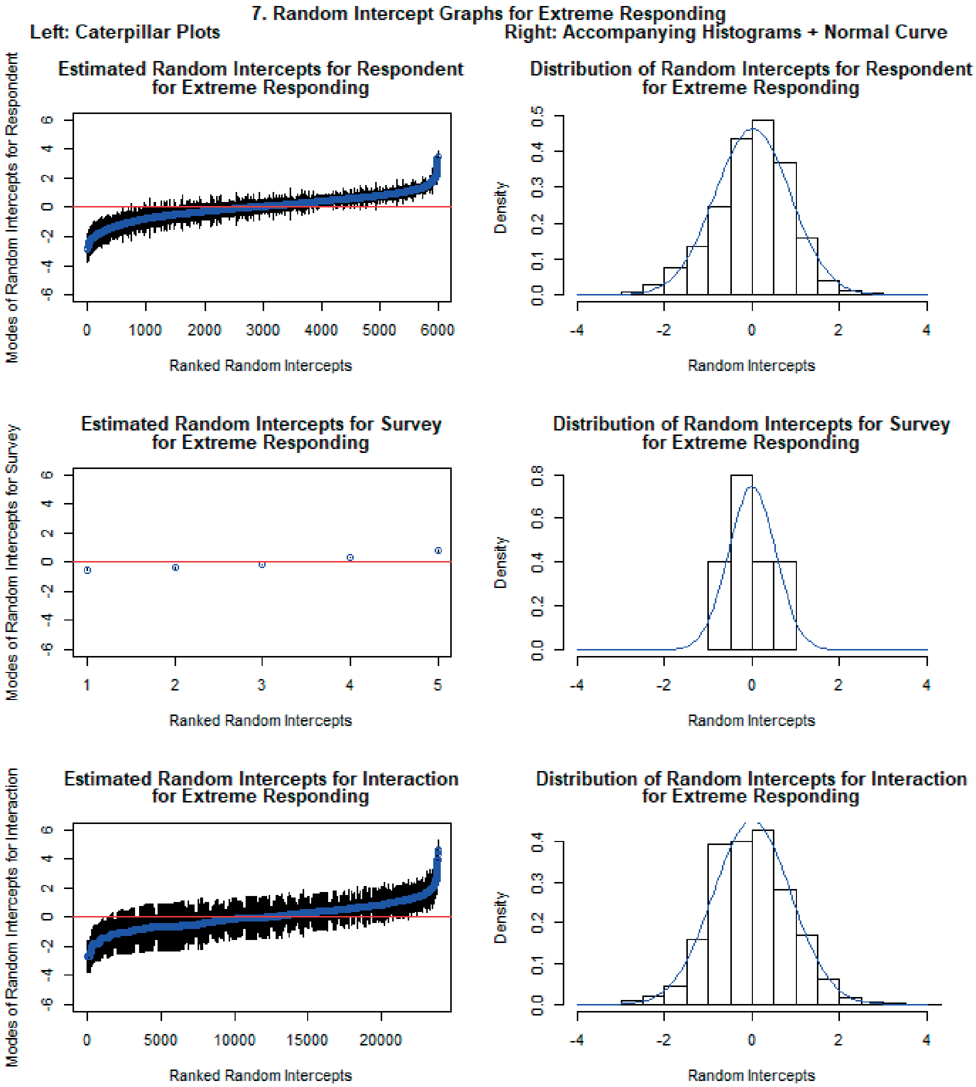


Figure F.7. Random Intercept Graphs for the Answer Behaviour ‘Extreme Responding’.

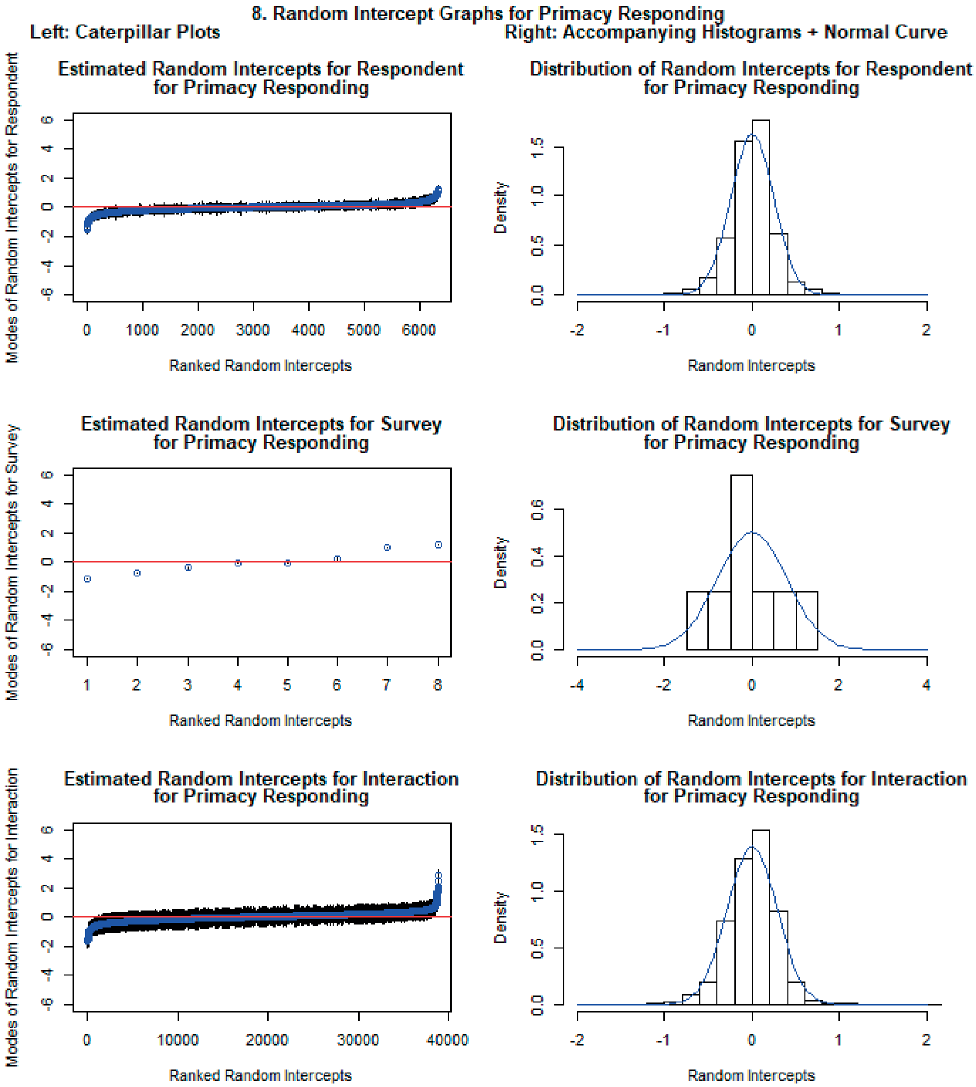


Figure F.8. Random Intercept Graphs for the Answer Behaviour ‘Primacy Responding’.

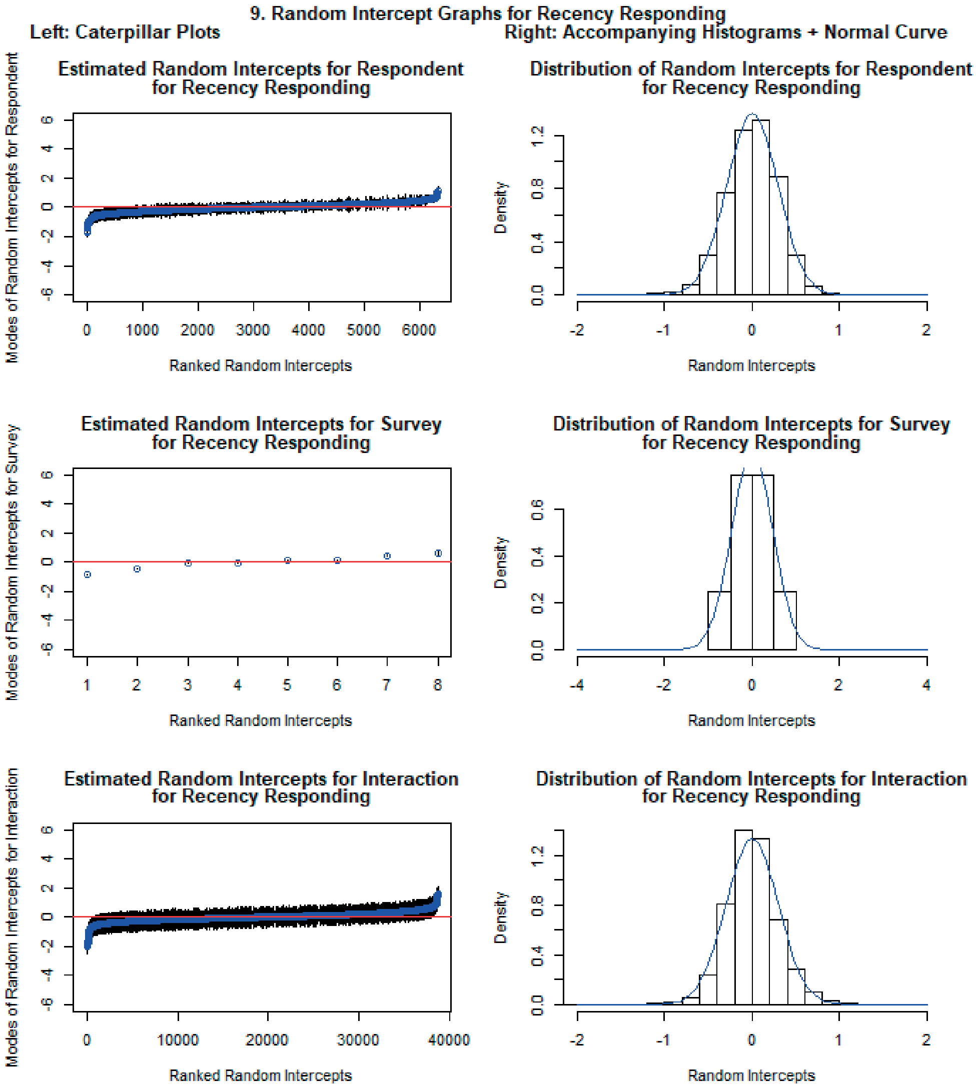


Figure F.9. Random Intercept Graphs for the Answer Behaviour ‘Recency Responding’.

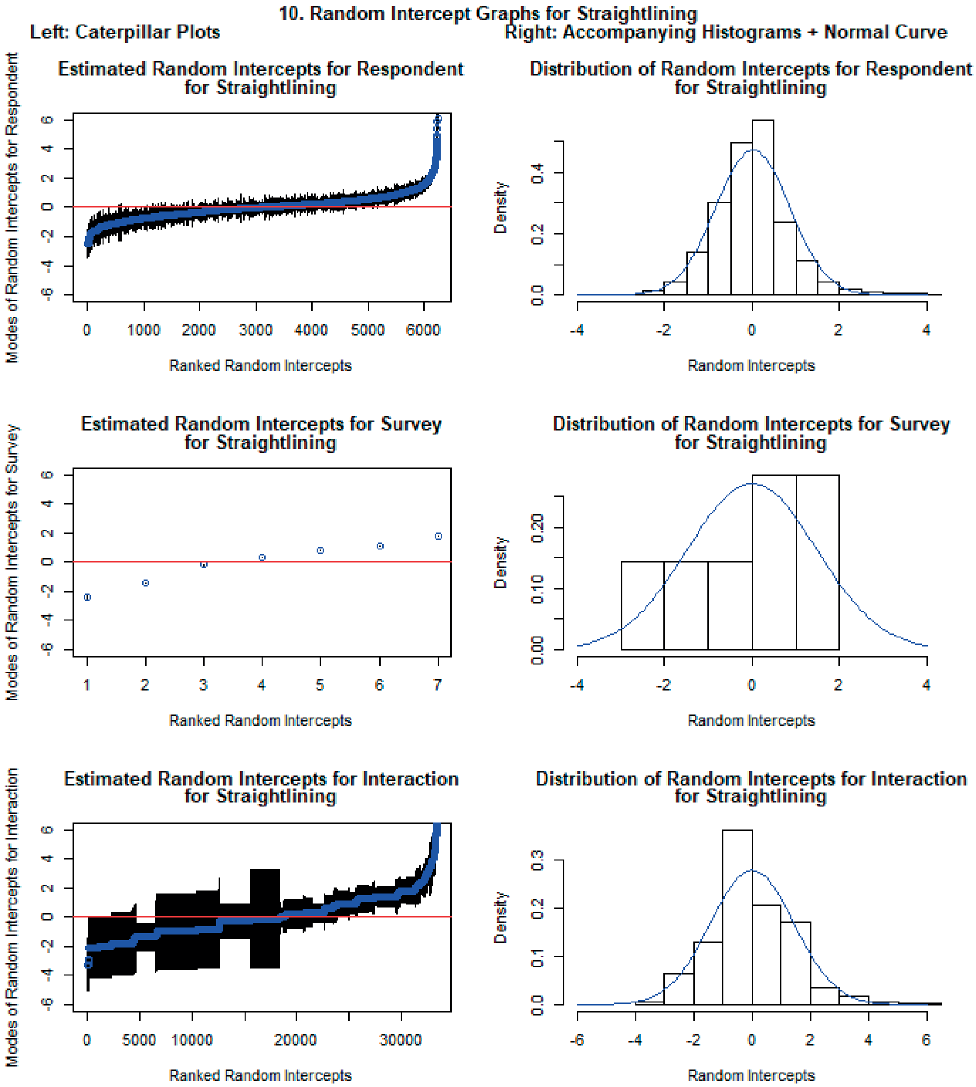


Figure F.10. Random Intercept Graphs for the Answer Behaviour ‘Straightlining’.

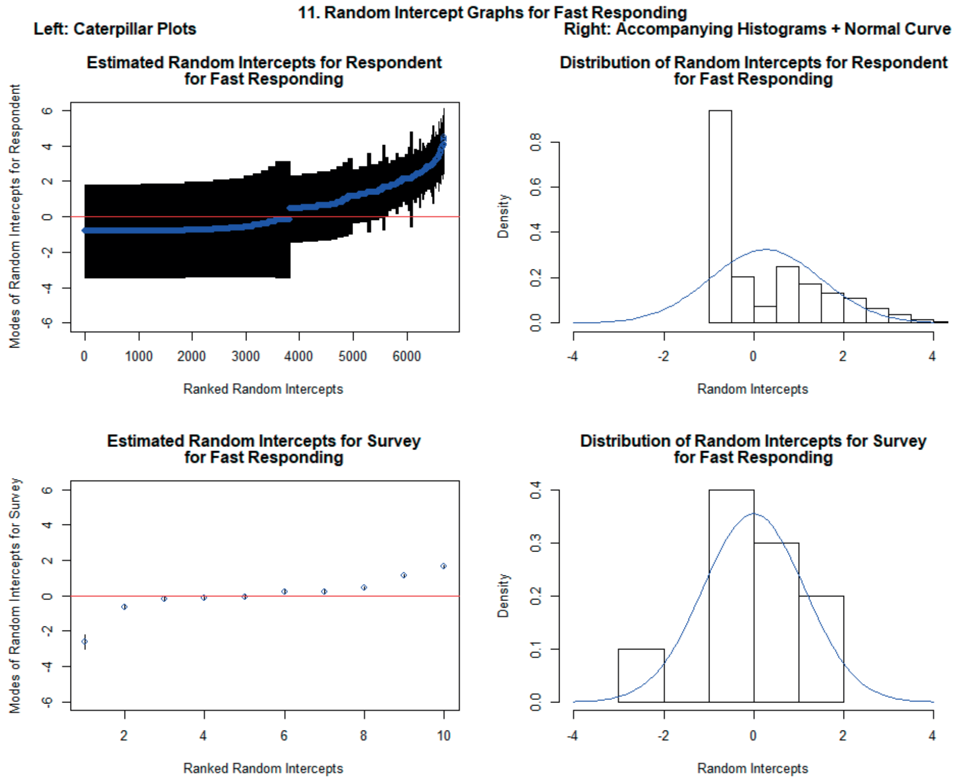


Figure F.11. Random Intercept Graphs for the Answer Behaviour ‘Fast Responding’.

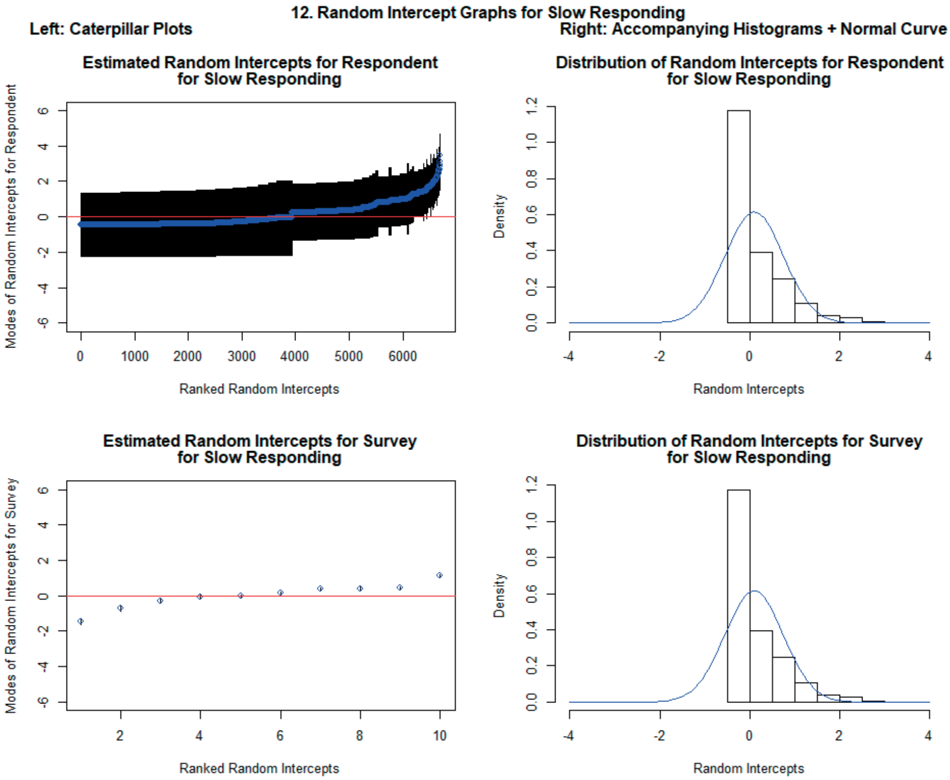


Figure F.12. Random Intercept Graphs for the Answer Behaviour ‘Slow Responding’.

APPENDIX G: THE ADAPTED CLIFF'S DELTA

As stated in section 5.3.4, for Cliff's adapted Delta, we consider the behaviour probabilities our 'observations' and the accompanying likelihoods their 'frequencies' or 'weights'. From here, we elaborate on the justification for this idea. The behaviour probabilities are the midpoints of the 100 step size intervals and hence already fixed. As these intervals run from {0.00-0.01, 0.01-0.02, 0.02-0.03, … , 0.99-1.00}, the behaviour probabilities run from {0.005, 0.015, 0.025, … , 0.995}. This means that these probabilities are by definition the observations from both groups of respondents. Now, the idea for the adapted Cliff's Delta is the same as for the original Cliff's Delta. If the observation from group A (for instance 0.015) is larger than the observation in group B (for instance 0.005), this value is 1. If the observation in group A (for instance 0.255) is smaller than the observation in group B (for instance 0.755), this value is -1. If the observations in group A and B are equal (both for instance 0.555), this value is 0.

Each observation is accompanied by its own specific likelihood that represents the observation's 'frequency'. For instance, when probabilities of 0.50 and 0.40 have likelihoods of 6 and 2 respectively, the probability of 0.50 versus 0.40 is three times more likely to occur. For comparison, where a specific observation value is evaluated three times when it occurs three times instead of only once, a probability's likelihood three times larger than another probability's likelihood is evaluated with a relative weight of three instead of just one. Therefore, the likelihood of the probability may be considered its frequency or weight. To take into account the frequency of both observations per comparison, their likelihoods are multiplied. The result of this multiplication is assigned a plus or a minus, or becomes 0, depending on the outcome value (1, -1, or 0 respectively) from comparing both observations. Comparing each observation from group A to each observation from group B gives 10000 comparisons and hence 10000 multiplication values. Implementing the probabilities and their likelihoods into formula (5.6) gives the formula for Cliff's adapted Delta $\hat{\delta}$ (see formula (5.7) in section 5.3.4).

See Graph 1 in Figure G.1 for an example of a pair of 'observations' in comparing group 1 to group 3. Note that the respondent profiles in Graph 1 present their actual 100 likelihood points to refer to the centre of each step size interval. Consider their respective probability midpoints of 0.305 and 0.655, and their respective likelihoods of about 2.1 and 4.2 (see the dotted lines). Implementing these values into the nominator of formula (5.7) gives sgn(0.305-0.655)(2.1*4.2), resulting in a negative contribution to the total nominator sum. As can be seen in Graph 1, the pairs of observations for which the probability occurrence from Group 3 is larger, are *frequently* accompanied by two relatively larger likelihoods. However, the pairs of observations for which the probability occurrence from Group 1 is larger, are *rarely* accompanied by two relatively larger likelihoods. The result will be a negative Cliff's Delta $\hat{\delta}$, indicating that the behaviour occurrence is higher for group 3 than for group 1.

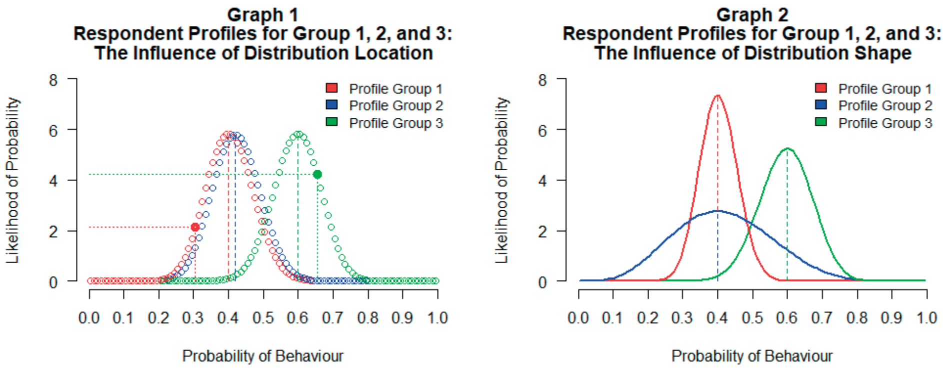


Figure G.1. Examples of Respondent Profiles with Different Locations (Graph 1) and Different Shapes (Graph 2).

In comparing respondent profiles, Cliff's adapted Delta takes into account both their location and their shape. See Graph 1 in Figure G.1. Considering the *location*, when two groups are located close to each other, they largely overlap, resulting in a smaller absolute Cliff's Delta (see group 1 and 2); when two groups are located further from each other, they hardly overlap, resulting in a larger absolute Cliff's Delta (see group 1 and 3). Considering the *shape*, see Graph 2 in Figure G.1. When a group profile is stretched, it is likely to overlap another profile, giving a smaller absolute Cliff's Delta (see group 2 and 3); when a group profile is squeezed, it is less likely to overlap another profile, giving a larger absolute Cliff's Delta (see group 1 and 3). In summary, the further the profiles are located from each other and the more squeezed they are, the larger Cliff's Delta. Hence, we can use Cliff's adapted Delta to compare the likelihood distributions of two categories of respondents for a specific characteristic.

APPENDIX H

Table H.1. The Behaviour Occurrence Proportions for Respondents Aged 65-74 (65+) and Respondents Aged 75 or Older (75+) for All Behaviours*, in Total (TT) and for All Surveys**.

	TT	*AS*	*FA*	*HE*	*HO*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*
SD 65+	.66	.95	.61	.66		.79	.77	.59	.27	.77	
SD 75+	.65	.96	.60	.64		.78	.76	.58	.30	.79	
PR 65+	.33		.49	.65		.36	.25	.18	.68	.17	
PR 75+	.31		.50	.65		.33	.24	.16	.66	.13	
DK 65+	.06				.07	.16		.06	.00		
DK 75+	.06				.07	.14		.07	.00		
ST 65+	.10		.05	.36		.32	.02	.07	.24		
ST 75+	.08		.04	.25		.29	.01	.06	.19		
WT 65+	.05				.02	.04					.03
WT 75+	.04				.01	.03					.03
AC 65+	.47		.44				.50	.45	.19		
AC 75+	.49		.42				.51	.48	.21		
NE 65+	.22		.28			.25	.21	.22			
NE 75+	.21		.28			.25	.21	.22			
EX 65+	.19		.37			.11	.23	.11			
EX 75+	.20		.40			.11	.25	.10			

* Socially Desirable Responding (SD), Primacy Responding (PR), Answering ‘Don’t Know’ (DK), Straightlining (ST), Answering ‘Won’t Tell’ (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX).
** Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).
Note. Empty cells refer either to surveys that were not applicable to the specific behaviour or to a situation in which one subgroup contained no or only a few respondents.

Table H.2. Proportions (and Absolute Numbers) of Respondents per Subgroup and Behaviour* Taken over All Surveys after Excluding Respondents Who Filled Out Less than 5 Items.

	Satisficing behaviour						*Sensitivity-based behaviour*	
	DK	*AC*	*NE*	*EX*	*PR*	*ST*	*SD*	*WT*
Age TOTAL	.98 (6424)	1 (6380)	1 (6338)	1 (6339)	.99 (6476)	1 (6451)	.99 (6636)	.84 (5353)
Age 15_24	.95 (781)	.99 (774)	.98 (765)	.98 (766)	1 (803)	.99 (786)	.99 (833)	.53 (400)
Age 25_34	.97 (757)	1 (748)	1 (740)	1 (740)	.98 (767)	1 (762)	.98 (790)	.81 (612)
Age 35_44	.98 (1044)	1 (1034)	1 (1026)	1 (1026)	.99 (1049)	1 (1046)	.99 (1071)	.85 (875)
Age 45_54	.99 (1190)	1 (1179)	1 (1165)	1 (1165)	.99 (1193)	1 (1193)	.99 (1214)	.89 (1049)
Age 55_64	.98 (1241)	1 (1238)	1 (1236)	1 (1236)	.99 (1247)	1 (1247)	.99 (1273)	.90 (1133)
Age 65_old	.98 (1411)	1 (1407)	1 (1406)	1 (1406)	.99 (1417)	1 (1417)	.99 (1455)	.91 (1284)
Education TOTAL	.98 (6413)	1 (6369)	1 (6327)	1 (6328)	.99 (6465)	1 (6440)	.99 (6625)	.84 (5342)
Education Primary	.97 (570)	1 (571)	1 (566)	1 (566)	.99 (583)	.99 (577)	.99 (596)	.73 (401)
Education VMBO	.97 (1568)	1 (1557)	.99 (1550)	.99 (1550)	.99 (1579)	1 (1570)	.99 (1621)	.84 (1312)
Education HAVWO	.97 (754)	1 (746)	.99 (743)	.99 (743)	.99 (759)	1 (756)	.99 (784)	.76 (563)
Education MBO	.98 (1487)	1 (1478)	1 (1463)	1 (1464)	.99 (1501)	1 (1496)	.99 (1536)	.86 (1284)
Education HBO	.98 (1445)	1 (1433)	1 (1427)	1 (1427)	.99 (1455)	1 (1453)	.99 (1484)	.88 (1267)
Education WO	.98 (589)	1 (584)	1 (578)	1 (578)	.99 (588)	1 (588)	.99 (604)	.88 (515)

* Answering 'Don't Know' (DK), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Primacy Responding (PR), Straightlining (ST), Socially Desirable Responding (SD), Answering 'Won't Tell' (WT).

Table H.3. The Proportions of the Original Subgroups (and the Absolute Numbers of Respondents) after Excluding Respondents Who Filled Out Less than 5 Items, for the Behaviours, Subgroups, and Surveys* that Refer to More than Half of the Applicable Surveys Showing either Positive or Negative Effects Sizes (see Table 5.5 in Section 5.4.3).

	FA	*HE*	*HO*	*IN*	*PE*	*PO*	*RE*	*WO*
Answering 'Don't Know'								
Age 15_24	.99 (647)			.91 (512)		.99 (595)	1 (661)	.74 (425)
Age 65_old			.81 (723)	.99 (1199)		1 (1377)	1 (1359)	
Primary Education	.39 (195)		.76 (186)	.94 (419)		1 (487)	1 (504)	.43 (194)
Primacy responding								
Age 15_24	1 (649)	1 (599)		.84 (472)	1 (568)	.99 (595)	1 (660)	.99 (622)
Primary Education	1 (505)	1 (487)		.93 (415)	1 (474)	1 (487)	1 (504)	.95 (475)
WO Education	1 (519)	1 (514)		.96 (423)	1 (434)	1 (517)	1 (532)	1 (506)
Neutral responding								
WO Education	1 (518)			.87 (386)	1 (434)	1 (517)		.73 (371)

* Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO).

APPENDIX I

See the expected values for all respondent characteristics, all their categories, and all answer behaviours, taken over all surveys together and for each applicable separate survey, in Tables I.1 through I.8. Note that empty cells refer to subgroups that are not considered representative for their initial subgroups after excluding respondents who did not fill out at least five items for the specific survey (see also Appendix J).

Table I.1. Expected Values per Category for the Behaviour **Socially Desirable Responding** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*AS*	*FA*	*HE*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*
Total	**.66**	**.82**	**.52**	**.68**	**.78**	**.73**	**.61**	**.28**	**.73**	**.53**
Gen Male	.65	.81	.51	.69	.78	.73	.59	.26	.73	
Gen Female	.66	.82	.53	.67	.77	.72	.63	.29	.72	
Age 15_24	.65	.84	.35	.71	.87	.66	.65	.28	.68	.49
Age 25_34	.65	.79	.46	.70	.78	.70	.63	.26	.75	
Age 35_44	.65	.80	.51	.69	.75	.71	.61	.27	.76	
Age 45_54	.65	.80	.52	.68	.76	.72	.61	.27	.74	
Age 55_64	.66	.82	.57	.66	.76	.75	.60	.28	.71	
Age 65_old	.66	.83	.59	.65	.78	.76	.59	.30	.72	
Edu Primary	.64	.83	.50	.66	.77	.69	.60	.31	.69	
Edu VMBO	.65	.83	.55	.67	.76	.72	.58	.31	.71	
Edu HAVWO	.65	.81	.47	.68	.80	.71	.62	.26	.71	
Edu MBO	.65	.82	.52	.68	.76	.72	.61	.28	.73	
Edu HBO	.67	.81	.53	.68	.79	.76	.63	.26	.76	
Edu WO	.67	.80	.52	.69	.80	.75	.64	.23	.77	
Dom Single	.65	.81	.56	.66	.75	.70	.61	.28	.72	
Dom Cohab	.66	.82	.56	.66	.78	.75	.60	.28	.73	
Dom Cohab_ch	.65	.82	.47	.70	.79	.72	.62	.27	.74	
Dom Single_ch	.64	.80	.47	.68	.78	.66	.64	.29	.71	
Dom Other	.66	.81	.49	.68	.82	.73	.61	.28	.71	
Occ Paid	.66	.81	.51	.69	.78	.73	.61	.27	.77	
Occ Famself	.66	.80	.53	.69	.76	.72	.61	.26	.78	
Occ Seek	.63	.78	.53	.68	.70	.69	.62	.30	.57	.64
Occ Other	.61	.80	.55	.60	.69	.67	.60	.28	.64	
Occ School	.65	.83	.34	.71	.86	.66	.66	.27	.66	.48
Occ Housekeep	.65	.83	.58	.66	.75	.74	.60	.30	.68	
Occ Retired	.66	.83	.59	.65	.78	.76	.59	.29	.72	
Inc No_income	.64	.84	.44	.69	.81	.68	.63	.29	.64	
Inc 0001_1000	.65	.82	.52	.67	.76	.71	.61	.30	.70	
Inc 1001_2000	.66	.81	.54	.67	.76	.73	.61	.28	.74	
Inc 2001_3000	.67	.81	.53	.69	.81	.76	.60	.25	.78	
Inc 3001_more	.67	.81	.53	.69	.83	.77	.61	.24	.79	
Inc Don't know	.65	.81	.51	.69	.76	.69	.62	.31	.72	
Inc Won't tell	.65	.84	.51	.69	.78	.70	.60	.28	.75	
Ori Dutch	.66	.82	.53	.68	.78	.73	.61	.28	.73	
Ori 1st_west	.65	.81	.56	.67	.74	.71	.61	.31	.70	
Ori 1st_nonw	.63	.78	.53	.67	.71	.66	.61	.31	.69	
Ori 2nd_west	.66	.80	.53	.67	.76	.72	.62	.28	.72	
Ori 2nd_nonw	.64	.80	.45	.70	.81	.64	.64	.28	.70	
Sim No	.66	.82	.52	.68	.78	.73	.61	.28	.73	
Sim Yes	.64	.82	.59	.64	.72	.70	.57	.31	.71	

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table I.2. Expected Values per Category for the Behaviour **Primacy Responding** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*FA*	*HE*	*IN*	*PE*	*PO*	*RE*	*WO*	*LF*
Total	**.34**	**.42**	**.65**	**.38**	**.25**	**.19**	**.68**	**.29**	**.48**
Gen Male	.35	.40	.66	.39	.25	.19	.69	.30	
Gen Female	.34	.43	.64	.38	.25	.19	.67	.28	
Age 15_24	.32	.32	.64	.30	.23	.17	.66	.29	
Age 25_34	.35	.36	.66	.38	.25	.19	.69	.35	
Age 35_44	.36	.36	.66	.42	.25	.20	.69	.34	
Age 45_54	.36	.40	.65	.41	.26	.20	.69	.34	
Age 55_64	.35	.47	.65	.40	.25	.19	.69	.30	
Age 65_old	.33	.49	.65	.36	.25	.18	.67	.19	
Edu Primary	.32	.42	.63	.33	.23	.19	.65	.27	
Edu VMBO	.33	.44	.64	.36	.24	.20	.66	.26	
Edu HAVWO	.35	.39	.64	.38	.26	.19	.69	.28	
Edu MBO	.35	.41	.65	.39	.25	.19	.68	.30	
Edu HBO	.36	.41	.66	.41	.26	.19	.70	.31	
Edu WO	.36	.39	.66	.42	.27	.19	.72	.34	
Dom Single	.34	.41	.64	.35	.25	.19	.65	.27	
Dom Cohab	.34	.47	.65	.39	.25	.19	.68	.26	
Dom Cohab_ch	.35	.37	.65	.39	.25	.19	.69	.33	
Dom Single_ch	.33	.35	.64	.35	.23	.19	.66	.32	
Dom Other	.33	.37	.65	.32	.23	.18	.66	.30	
Occ Paid	.36	.39	.66	.40	.26	.20	.69	.36	
Occ Famself	.36	.41	.66	.40	.25	.20	.69	.32	
Occ Seek	.35	.41	.64	.39	.25	.21	.65	.30	
Occ Other	.33	.44	.56	.37	.25	.21	.67	.20	
Occ School	.32	.31	.64	.30	.23	.17	.66	.27	.52
Occ Housekeep	.34	.49	.64	.38	.25	.19	.66	.24	
Occ Retired	.33	.48	.65	.36	.25	.18	.67	.20	
Inc No_income	.32	.39	.63	.34	.23	.19	.66	.27	
Inc 0001_1000	.34	.43	.63	.37	.25	.19	.66	.27	
Inc 1001_2000	.35	.42	.65	.38	.25	.19	.68	.29	
Inc 2001_3000	.36	.41	.67	.42	.27	.19	.71	.32	
Inc 3001_more	.37	.42	.67	.44	.27	.19	.72	.34	
Inc Don't know	.33	.39	.63	.35	.23	.19	.65	.30	
Inc Won't tell	.34	.40	.64	.33	.24	.21	.67	.31	
Ori Dutch	.35	.42	.65	.38	.25	.19	.68	.29	
Ori 1st_west	.34	.41	.64	.38	.23	.18	.65	.28	
Ori 1st_nonw	.32	.36	.62	.36	.22	.19	.62	.31	
Ori 2nd_west	.35	.43	.65	.38	.25	.19	.69	.28	
Ori 2nd_nonw	.34	.35	.63	.31	.22	.19	.64	.29	
Sim No	.35	.41	.65	.38	.25	.19	.68	.29	
Sim Yes	.32	.45	.63	.34	.24	.20	.65	.24	

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table I.3. Expected Values per Category for the Behaviour **Straightlining** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*FA*	*HE*	*IN*	*PE*	*PO*	*RE*	*WO*
Total	**.14**	**.16**	**.44**	**.31**	**.04**	**.11**	**.25**	**.05**
Gen Male	.14	.16	.46	.32	.04	.10	.29	.05
Gen Female	.13	.16	.41	.31	.04	.12	.22	.05
Age 15_24	.16	.17	.49	.22	.06	.14	.22	.08
Age 25_34	.16	.16	.54	.27	.05	.13	.25	.05
Age 35_44	.16	.20	.51	.32	.05	.13	.26	.05
Age 45_54	.15	.19	.46	.33	.04	.12	.27	.05
Age 55_64	.11	.14	.39	.29	.03	.09	.25	.05
Age 65_old	.10	.12	.34	.33	.03	.08	.25	.06
Edu Primary	.12	.17	.35	.30	.04	.12	.24	.05
Edu VMBO	.14	.16	.38	.31	.05	.13	.24	.06
Edu HAVWO	.13	.16	.45	.30	.04	.09	.25	.05
Edu MBO	.14	.17	.44	.32	.04	.12	.26	.06
Edu HBO	.13	.16	.47	.32	.02	.09	.25	.04
Edu WO	.14	.14	.53	.30	.02	.09	.29	.05
Dom Single	.12		.40	.29	.03	.10	.20	.06
Dom Cohab	.11	.12	.39	.30	.03	.09	.26	.05
Dom Cohab_ch	.16	.20	.49	.34	.04	.12	.27	.05
Dom Single_ch	.17	.19	.46	.30	.08	.17	.24	.06
Dom Other	.13	.12	.41	.27	.04	.09	.24	.04
Occ Paid	.15	.17	.50	.32	.04	.11	.26	.05
Occ Famself	.16	.18	.49	.32	.04	.12	.27	.07
Occ Seek	.14	.17	.45	.23	.04	.12	.23	.05
Occ Other	.11	.17	.26	.26	.04	.12	.24	.06
Occ School	.15		.48	.21	.05	.14	.22	.09
Occ Housekeep	.12	.16	.37	.30	.03	.11	.23	.04
Occ Retired	.11	.11	.35	.33	.03	.08	.25	.05
Inc No_income	.14	.16	.43	.31	.05	.13	.23	.07
Inc 0001_1000	.13	.15	.39	.29	.04	.12	.23	.06
Inc 1001_2000	.13	.16	.43	.30	.03	.10	.25	.05
Inc 2001_3000	.13	.15	.48	.33	.02	.08	.28	.04
Inc 3001_more	.15	.14	.51	.37	.02	.09	.31	.04
Inc Don't know	.16	.20	.46	.35	.06	.16	.23	.06
Inc Won't tell	.19	.21	.49	.38	.09	.18	.28	.06
Ori Dutch	.13	.16	.44	.31	.04	.10	.25	.05
Ori 1st_west	.13	.16	.42	.32	.04	.11	.22	.06
Ori 1st_nonw	.17	.22	.44	.30	.09	.15	.24	.07
Ori 2nd_west	.13	.15	.42	.27	.03	.11	.24	.05
Ori 2nd_nonw	.16	.18	.48	.27	.05	.15	.28	.06
Sim No	.14	.16	.44	.31	.04	.11	.25	.05
Sim Yes	.11	.15	.30	.28	.05	.10	.22	.05

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table I.4. Expected Values per Category for the Behaviour **Answering Don't Know** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*AS*	*FA*	*HO*	*IN*	*PO*	*RE*	*WO*
Total	**.10**	**.39**	**.16**	**.19**	**.23**	**.10**	**.04**	**.15**
Gen Male	.08		.16	.16	.20	.07	.04	.13
Gen Female	.12		.15	.21	.26	.13	.04	.16
Age 15_24	.16		.17		.39	.18	.04	.21
Age 25_34	.11		.16	.21	.26	.12	.04	.14
Age 35_44	.10		.15	.22	.24	.12	.04	.13
Age 45_54	.10		.14	.21	.23	.10	.04	.14
Age 55_64	.08			.18	.19	.07	.04	.14
Age 65_old	.07			.16	.18	.07	.04	
Edu Primary	.14		.20	.21	.29	.14	.04	.20
Edu VMBO	.11		.16	.20	.25	.12	.04	.17
Edu HAVWO	.09		.15	.17	.25	.09	.04	.17
Edu MBO	.10		.16	.21	.24	.12	.04	.14
Edu HBO	.07		.15	.17	.19	.07	.04	.12
Edu WO	.07		.15	.17	.18	.06	.04	.13
Dom Single	.09		.17	.17	.22	.08	.04	.15
Dom Cohab	.08			.18	.18	.08	.04	.14
Dom Cohab_ch	.11		.15	.22	.27	.12	.04	.15
Dom Single_ch	.13		.16	.19	.34	.15	.04	.16
Dom Other	.11		.16	.21	.33	.10	.04	.17
Occ Paid	.09		.15	.20	.22	.10	.04	.12
Occ Famself	.10		.14	.19	.26	.08	.04	.15
Occ Seek	.11		.15	.18	.25	.12	.04	.22
Occ Other	.10		.17	.19	.24	.10	.04	
Occ School	.16		.17	.19	.39	.18	.04	.22
Occ Housekeep	.12		.13	.23	.23	.13	.04	
Occ Retired	.07			.15	.17	.06	.04	
Inc No_income	.14		.17	.25	.33	.16	.04	.24
Inc 0001_1000	.12		.16	.21	.26	.13	.04	.18
Inc 1001_2000	.09		.15	.19	.21	.09	.04	.13
Inc 2001_3000	.06		.14	.15	.16	.05	.04	.10
Inc 3001_more	.05		.15	.14	.15	.04	.04	.10
Inc Don't know	.15		.17	.26	.35	.15	.04	.19
Inc Won't tell	.14		.18	.24	.34	.16	.04	.17
Ori Dutch	.09		.15	.18	.22	.09	.04	.14
Ori 1st_west	.11		.17	.23	.23	.12	.04	.18
Ori 1st_nonw	.15		.18	.22	.29	.17	.04	.21
Ori 2nd_west	.10		.15	.21	.22	.10	.04	.16
Ori 2nd_nonw	.13		.17	.15	.34	.15	.05	.16
Sim No	.10			.19	.23	.10	.04	.15
Sim Yes	.10			.19	.23	.11	.04	.17

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table I.5. Expected Values per Category for the Behaviour **Acquiescence** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*FA*	*PE*	*PO*	*RE*	*WO*
Total	**.47**	**.48**	**.49**	**.45**	**.24**	**.69**
Gen Male	.47	.45	.50	.47	.23	.69
Gen Female	.46	.50	.49	.44	.24	.69
Age 15_24	.45	.42	.49	.44	.24	.67
Age 25_34	.47	.52	.48	.46	.26	.69
Age 35_44	.46	.52	.48	.44	.24	.70
Age 45_54	.46	.49	.48	.44	.23	.70
Age 55_64	.47	.47	.50	.45	.24	.69
Age 65_old	.47	.45	.51	.46	.23	
Edu Primary	.45	.43	.49	.41	.24	.66
Edu VMBO	.44	.46	.48	.40	.24	.66
Edu HAVWO	.47	.48	.49	.47	.23	.67
Edu MBO	.46	.48	.49	.44	.24	.69
Edu HBO	.49	.49	.50	.49	.24	.72
Edu WO	.50	.50	.50	.52	.24	.70
Dom Single	.46	.48	.49	.45	.23	.69
Dom Cohab	.47	.47	.50	.46	.23	.69
Dom Cohab_ch	.46	.48	.49	.45	.25	.69
Dom Single_ch	.44	.49	.47	.41	.25	.67
Dom Other	.48	.46	.51	.48	.27	.68
Occ Paid	.47	.50	.49	.46	.23	.70
Occ Famself	.47	.49	.49	.45	.24	.68
Occ Seek	.46	.49	.50	.43	.23	.60
Occ Other	.45	.48	.49	.42	.26	
Occ School	.45	.42	.49	.45	.24	.67
Occ Housekeep	.45	.48	.49	.41	.27	
Occ Retired	.47	.45	.50	.47	.23	
Inc No_income	.45	.45	.49	.42	.26	.63
Inc 0001_1000	.45	.48	.49	.43	.25	.66
Inc 1001_2000	.47	.48	.50	.46	.23	.70
Inc 2001_3000	.49	.48	.50	.49	.24	.72
Inc 3001_more	.50	.47	.51	.52	.23	.73
Inc Don’t know	.43	.45	.46	.39	.24	.70
Inc Won’t tell	.42	.49	.47	.36	.24	.69
Ori Dutch	.47	.48	.49	.45	.23	.69
Ori 1st_west	.47	.48	.50	.43	.24	.67
Ori 1st_nonw	.46	.49	.50	.41	.42	.62
Ori 2nd_west	.47	.47	.50	.44	.23	.67
Ori 2nd_nonw	.46	.46	.52	.41	.35	.68
Sim No	.47	.48	.49	.45	.24	
Sim Yes	.46	.46	.50	.42	.26	

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table I.6. Expected Values per Category for the Behaviour **Neutral Responding** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*FA*	*IN*	*PE*	*PO*	*WO*
Total	**.22**	**.33**	**.28**	**.22**	**.21**	**.12**
Gen Male	.22	.33	.28	.22	.21	.12
Gen Female	.22	.32	.29	.22	.22	.12
Age 15_24	.23	.32		.24	.22	.15
Age 25_34	.22	.35	.27	.22	.20	.12
Age 35_44	.22	.35	.29	.23	.21	.12
Age 45_54	.22	.33	.29	.22	.21	.12
Age 55_64	.21	.32	.28	.21	.21	.12
Age 65_old	.22	.31	.28	.21	.22	
Edu Primary	.24	.32	.30	.24	.24	.14
Edu VMBO	.24	.33	.30	.24	.23	.14
Edu HAVWO	.21	.31	.27	.21	.20	.13
Edu MBO	.22	.33	.29	.23	.22	.12
Edu HBO	.20	.32	.27	.20	.20	.11
Edu WO	.19	.34	.25	.19	.18	.11
Dom Single	.22	.28	.26	.22	.22	.13
Dom Cohab	.21	.33	.29	.21	.21	.12
Dom Cohab_ch	.22	.35	.30	.22	.21	.12
Dom Single_ch	.24	.32		.26	.24	.14
Dom Other	.21	.33		.22	.21	.13
Occ Paid	.22	.35	.28	.22	.21	.12
Occ Famself	.22	.33	.29	.22	.21	.13
Occ Seek	.21	.31	.28	.21	.21	
Occ Other	.23	.32	.29	.23	.22	
Occ School	.22	.31		.23	.21	.15
Occ Housekeep	.23	.31	.30	.23	.23	
Occ Retired	.22	.31	.28	.22	.22	
Inc No_income	.23	.31		.24	.22	
Inc 0001_1000	.23	.32	.30	.23	.22	.14
Inc 1001_2000	.22	.33	.28	.22	.21	.12
Inc 2001_3000	.20	.32	.26	.20	.20	.11
Inc 3001_more	.19	.32	.24	.19	.18	.10
Inc Don't know	.25	.36	.31	.27	.24	.13
Inc Won't tell	.25	.35	.31	.26	.24	.12
Ori Dutch	.22	.33	.28	.22	.21	.12
Ori 1st_west	.22	.32	.30	.23	.23	.13
Ori 1st_nonw	.24	.35	.31	.25	.24	.15
Ori 2nd_west	.22	.32	.30	.22	.22	.14
Ori 2nd_nonw	.24	.35	.23	.22	.24	.13
Sim No	.22	.33	.28	.22	.21	
Sim Yes	.23	.32	.30	.22	.22	

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table I.7. Expected Values per Category for the Behaviour **Extreme Responding** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*FA*	*IN*	*PE*	*PO*	*WO*
Total	**.18**	**.35**	**.16**	**.22**	**.13**	**.18**
Gen Male	.17	.33	.17	.21	.13	.18
Gen Female	.19	.36	.16	.23	.12	.19
Age 15_24	.16	.36		.18	.11	.18
Age 25_34	.17	.29	.15	.20	.13	.17
Age 35_44	.18	.30	.15	.21	.14	.17
Age 45_54	.19	.32	.16	.22	.14	.17
Age 55_64	.19	.37	.17	.23	.13	.19
Age 65_old	.19	.40	.16	.24	.11	
Edu Primary	.20	.39	.16	.24	.13	.22
Edu VMBO	.19	.37	.15	.23	.13	.19
Edu HAVWO	.18	.35	.16	.21	.13	.19
Edu MBO	.18	.34	.15	.21	.13	.17
Edu HBO	.17	.33	.17	.21	.12	.17
Edu WO	.18	.31	.19	.21	.13	.19
Dom Single	.19	.38	.17	.23	.13	.17
Dom Cohab	.19	.36	.16	.23	.12	.20
Dom Cohab_ch	.18	.31	.15	.20	.13	.17
Dom Single_ch	.19	.34	.16	.21	.14	.18
Dom Other	.18	.36	.17	.21	.13	.21
Occ Paid	.18	.31	.16	.21	.13	.17
Occ Famself	.19	.35	.17	.21	.14	.22
Occ Seek	.20	.36	.18	.22	.15	.23
Occ Other	.21	.36	.17	.25	.15	
Occ School	.16	.35		.18	.11	.18
Occ Housekeep	.20	.40	.15	.25	.12	
Occ Retired	.18	.39	.16	.23	.11	
Inc No_income	.19	.38	.15	.21	.12	.22
Inc 0001_1000	.19	.36	.15	.23	.12	.20
Inc 1001_2000	.18	.34	.15	.22	.13	.17
Inc 2001_3000	.18	.33	.17	.21	.12	.17
Inc 3001_more	.17	.33	.24	.20	.13	.19
Inc Don't know	.18	.33	.17	.20	.13	.18
Inc Won't tell	.19	.35	.15	.23	.15	.19
Ori Dutch	.18	.35	.16	.22	.12	.18
Ori 1st_west	.20	.34	.16	.25	.14	.19
Ori 1st_nonw	.21	.35	.17	.25	.15	.21
Ori 2nd_west	.19	.38	.15	.23	.13	.18
Ori 2nd_nonw	.21	.35		.25	.15	.24
Sim No	.18	.35	.16	.22	.13	
Sim Yes	.21	.38	.16	.26	.13	

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table I.8. Expected Values per Category for the Behaviour **Answering Won't Tell** for Gender (Gen), Age (Age), Education (Edu), Domestic Situation (Dom), Primary Occupation (Occ), Income (Inc), Origin (Ori), and Provided a PC (Sim) in Total (TT) and for All Applicable Surveys*.

Categories	*TT*	*AS*	*HO*	*IN*	*WO*	*LF*
Total	**.13**	**.25**	**.13**	**.16**	**.44**	**.10**
Gen Male	.13		.13	.15		.10
Gen Female	.14		.14	.16		.11
Age 15_24	.20		.17	.16		.10
Age 25_34	.16		.14	.18		.10
Age 35_44	.15		.15	.17		.10
Age 45_54	.14		.14	.18		.10
Age 55_64	.12		.13	.15		.10
Age 65_old	.09		.12	.13		.11
Edu Primary	.15		.14	.16		.14
Edu VMBO	.14		.14	.16		.12
Edu HAVWO	.14		.13	.15		.10
Edu MBO	.14		.15	.17		.10
Edu HBO	.11		.12	.15		.09
Edu WO	.11		.12	.14		.09
Dom Single	.11		.14	.15		.11
Dom Cohab	.12		.13	.15		.10
Dom Cohab_ch	.16		.14	.18		.10
Dom Single_ch	.16		.15	.17		.12
Dom Other	.14		.13	.17		.10
Occ Paid	.14		.13	.17		.10
Occ Famself	.15	.29	.14	.18		.11
Occ Seek	.13		.14	.15		.09
Occ Other	.12		.14	.15		.11
Occ School	.19		.17	.17		.09
Occ Housekeep	.15		.15	.18		.12
Occ Retired	.09		.12	.13		.11
Inc No_income	.18		.16	.19		.11
Inc 0001_1000	.13		.14	.15		.11
Inc 1001_2000	.12		.13	.15		.10
Inc 2001_3000	.10		.12	.14		.09
Inc 3001_more	.10		.11	.13		.09
Inc Don't know	.21		.16	.23		.13
Inc Won't tell	.31		.27	.38		.15
Ori Dutch	.13		.13	.16		.10
Ori 1st_west	.13		.12	.17		.12
Ori 1st_nonw	.19		.15	.20		.14
Ori 2nd_west	.12		.13	.14		.10
Ori 2nd_nonw	.16		.15	.17		.10
Sim No	.13		.13	.16		.10
Sim Yes	.11		.14	.14		.13

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

APPENDIX J

See the Cliff's Deltas and their accompanying confidence intervals for all respondent characteristics, all their categories, and all answer behaviours, taken over all surveys together and for each applicable separate survey, in Tables J.1 through J.8. Note that empty cells refer to subgroups that are not considered representative for their initial subgroups after excluding respondents who did not fill out at least five items for the specific survey (see also Appendix I).

Table J.1. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Socially Desirable Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *AS*	δ *FA*	δ *HE*	δ *IN*	δ *PE*	δ *PO*	δ *RE*	δ *WO*	δ *LF*
Gender ma vs fe	-.03 (-.05,-.01)	-.08 (-.11,-.05)	**.13** ~ (.10, .16)	.03 (.00, .07)	.02 (-.02, .06)	**-.17** ~ (-.21,-.14)	**-.12** ~ (-.15,-.09)	.03 (.00, .06)	
Age 15_24 yrs	.08 (.05, .11)	**-.59 #** (-.62,-.55)	**.33 *** (.29, .36)	**.50 #** (.45, .54)	**-.25** ~ (-.31,-.19)	**.22** ~ (.16, .28)	-.03 (-.08, .02)	**-.20** ~ (-.25,-.14)	-.09 (-.32, .14)
Age 25_34 yrs	-.09 (-.14,-.05)	**-.23** ~ (-.27,-.19)	**.25** ~ (.20, .29)	.04 (-.02, .10)	-.10 (-.17,-.03)	.09 (.03, .14)	-.06 (-.11,-.01)	**.11** ~ (.05, .17)	
Age 35_44 yrs	-.04 (-.07, .00)	-.04 (-.07, .00)	**.15** ~ (.12, .19)	**-.17** ~ (-.22,-.11)	-.09 (-.14,-.04)	.02 (-.03, .07)	-.03 (-.07, .01)	**.16** ~ (.11, .21)	
Age 45_54 yrs	-.05 (-.08,-.02)	.00 (-.04, .03)	.02 (-.02, .06)	-.10 (-.15,-.05)	-.01 (-.06, .04)	.02 (-.03, .06)	-.03 (-.07, .01)	.04 (.00, .08)	
Age 55_64 yrs	.01 (-.02, .03)	**.19** ~ (.16, .22)	**-.17** ~ (-.21,-.13)	-.09 (-.13,-.04)	.10 (.05, .15)	-.05 (-.09,-.01)	.00 (-.04, .04)	-.06 (-.10,-.02)	
Age 65_older	.07 (.05, .09)	**.31 *** (.28, .34)	**-.27** ~ (-.31,-.24)	-.02 (-.06, .02)	**.16** ~ (.12, .20)	**-.14** ~ (-.18,-.10)	.09 (.06, .13)	-.04 (-.07,-.01)	
Education Primary	.06 (.03, .09)	-.04 (-.10, .02)	**-.12** ~ (-.18,-.06)	-.02 (-.09, .05)	**-.15** ~ (-.21,-.08)	-.04 (-.10, .03)	**.12** ~ (.06, .18)	**-.13** ~ (-.18,-.08)	
Education VMBO	.04 (.02, .06)	**.11** ~ (.08, .15)	-.09 (-.13,-.06)	**-.11** ~ (-.15,-.07)	-.03 (-.07, .02)	**-.15** ~ (-.20,-.11)	**.15** ~ (.11, .18)	-.09 (-.12,-.05)	
Education HAVWO	-.03 (-.07, .01)	**-.18** ~ (-.23,-.13)	.07 (.02, .12)	**.13** ~ (.06, .18)	-.07 (-.13,-.01)	.04 (-.02, .09)	-.09 (-.14,-.04)	-.09 (-.14,-.03)	
Education MBO	.01 (-.02, .03)	.00 (-.03, .03)	.05 (.02, .09)	-.10 (-.14,-.05)	-.02 (-.07, .03)	-.03 (-.07, .01)	.01 (-.03, .04)	-.02 (-.05, .02)	
Education HBO	-.03 (-.06, .00)	.02 (-.02, .05)	.02 (-.02, .05)	.09 (.05, .13)	**.13** ~ (.09, .18)	**.12** ~ (.08, .16)	-.07 (-.11,-.04)	**.13** ~ (.09, .17)	
Education WO	-.07 (-.11,-.03)	-.02 (-.07, .03)	.09 (.04, .14)	**.15** ~ (.08, .20)	.06 (.00, .13)	**.16** ~ (.10, .21)	**-.19** ~ (-.24,-.14)	**.19** ~ (.13, .24)	
Domestic Single	-.03 (-.06,-.01)	**.16** ~ (.12, .20)	**-.11** ~ (-.15,-.07)	**-.15** ~ (-.19,-.10)	**-.11** ~ (-.16,-.06)	.03 (-.02, .07)	.03 (-.01, .08)	-.06 (-.10,-.02)	
Domestic Cohab	.02 (.00, .04)	**.20** ~ (.17, .23)	**-.18** ~ (-.21,-.15)	.00 (-.04, .03)	**.15** ~ (.12, .19)	**-.11** ~ (-.14,-.07)	.03 (-.01, .06)	-.01 (-.04, .02)	
Domestic Cohab+ch	.01 (-.01, .03)	**-.26** ~ (-.29,-.23)	**.24** ~ (.21, .27)	.09 (.05, .12)	-.04 (-.08, .00)	.07 (.03, .10)	-.06 (-.09,-.02)	.06 (.03, .10)	
Domestic Single+ch	-.03 (-.09, .02)	**-.16** ~ (-.23,-.09)	.06 (-.01, .13)	.03 (-.07, .13)	**-.24** ~ (-.32,-.14)	**.13** ~ (.06, .22)	.02 (-.05, .10)	-.05 (-.13, .03)	
Domestic Other	-.01 (-.11, .07)	**-.11** ~ (-.24, .02)	.02 (-.12, .16)	**.20** ~ (.04, .35)	-.02 (-.15, .12)	-.01 (-.16, .14)	.04 (-.09, .16)	-.05 (-.19, .10)	

~ → small effect; * → medium effect; # → large effect
* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table J.1 (continued). Cliff's Delta *δ* (and its 99% Confidence Interval) for the Behaviour **'Socially Desirable Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ AS	*δ FA*	*δ HE*	*δ IN*	*δ PE*	*δ PO*	*δ RE*	*δ WO*	*δ LF*
Occupation Paid	-.04 (-.06,-.02)	-.08 (-.11,-.06)	**.23** ~ (.20, .26)	.06 (.02, .09)	.04 (.00, .08)	.04 (.01, .08)	-.07 (-.10,-.03)	**.30** * (.27, .33)	
Occupation Famself	-.04 (-.10, .00)	.03 (-.03, .09)	.10 (.04, .16)	-.08 (-.16, .00)	-.03 (-.12, .05)	-.01 (-.09, .06)	-.06 (-.12, .01)	**.21** ~ (.14, .27)	
Occupation Seek	-.09 (-.17,-.02)	.02 (-.07, .10)	-.04 (-.12, .05)	**-.35** * (-.45,-.26)	**-.12** ~ (-.23,-.01)	.04 (-.07, .13)	.06 (-.03, .15)	**-.52** # (-.60,-.43)	**.35** * (.22, .48)
Occupation Other	-.05 (-.09,-.01)	**.11** ~ (.05, .16)	**-.51** # (-.56,-.45)	**-.38** * (-.44,-.31)	**-.20** ~ (-.27,-.13)	-.05 (-.12, .01)	.01 (-.05, .07)	**-.36** * (-.40,-.31)	
Occupation School	.05 (.02, .09)	**-.59** # (-.63,-.55)	**.32** * (.28, .36)	**.48** # (.43, .53)	**-.25** ~ (-.31,-.18)	**.24** ~ (.18, .30)	-.04 (-.09, .02)	**-.23** ~ (-.29,-.17)	-.10 (-.34, .12)
Occupation Housekeep	.06 (.03, .09)	**.20** ~ (.16, .24)	**-.14** ~ (-.20,-.09)	**-.16** ~ (-.22,-.10)	.04 (-.03, .11)	-.04 (-.11, .02)	.10 (.05, .16)	**-.20** ~ (-.24,-.15)	
Occupation Retired	.06 (.04, .08)	**.28** * (.25, .31)	**-.24** ~ (-.27,-.20)	.00 (-.04, .04)	**.16** ~ (.11, .20)	**-.13** ~ (-.17,-.09)	.07 (.03, .11)	-.07 (-.10,-.04)	
Income No income	.09 (.06, .11)	**-.26** ~ (-.31,-.21)	.09 (.03, .14)	**.21** ~ (.14, .27)	**-.18** ~ (-.24,-.11)	.08 (.02, .14)	.03 (-.02, .09)	**-.32** * (-.37,-.27)	
Income 0001_1000	.01 (-.01, .04)	.02 (-.02, .06)	**-.11** ~ (-.14,-.07)	**-.12** ~ (-.17,-.08)	-.06 (-.11,-.02)	.01 (-.03, .05)	.09 (.05, .13)	**-.13** ~ (-.16,-.09)	
Income 1001_2000	-.03 (-.05,-.01)	.07 (.04, .10)	-.04 (-.07, .00)	**-.14** ~ (-.18,-.11)	.03 (-.01, .07)	-.02 (-.06, .01)	.03 (-.01, .06)	.04 (.01, .08)	
Income 2001_3000	-.03 (-.06, .00)	.04 (.00, .07)	.07 (.04, .11)	**.16** ~ (.12, .20)	**.14** ~ (.09, .19)	-.03 (-.07, .02)	**-.14** ~ (-.18,-.09)	**.21** ~ (.17, .25)	
Income 3001_more	-.02 (-.07, .03)	.03 (-.03, .10)	.09 (.03, .15)	**.30** * (.23, .37)	**.14** ~ (.06, .22)	.02 (-.06, .09)	**-.16** ~ (-.23,-.09)	**.26** ~ (.19, .33)	
Income DK	-.01 (-.10, .08)	-.04 (-.14, .07)	.10 (-.03, .22)	-.09 (-.24, .07)	**-.12** ~ (-.27, .04)	.05 (-.09, .19)	.09 (-.04, .21)	-.03 (-.15, .09)	
Income WT	.05 (.01, .10)	-.03 (-.11, .05)	.08 (.00, .16)	.00 (-.09, .10)	-.04 (-.15, .06)	-.05 (-.15, .04)	.00 (-.08, .09)	.08 (.00, .17)	
Origin Dutch	.06 (.03, .10)	.00 (-.05, .05)	.02 (-.03, .07)	**.12** ~ (.06, .18)	**.14** ~ (.08, .20)	-.03 (-.09, .02)	-.06 (-.11,-.01)	.09 (.04, .14)	
Origin 1st West	-.01 (-.08, .05)	**.11** ~ (.04, .19)	-.07 (-.17, .03)	**-.16** ~ (-.27,-.04)	-.09 (-.20, .03)	.01 (-.10, .11)	**.11** ~ (.01, .21)	-.09 (-.19, .01)	
Origin 1st Nonw	-.10 (-.18,-.03)	.02 (-.05, .10)	-.04 (-.13, .06)	**-.27** ~ (-.38,-.16)	**-.24** ~ (-.35,-.12)	-.02 (-.13, .09)	**.11** ~ (.01, .20)	**-.14** ~ (-.24,-.04)	
Origin 2nd West	-.06 (-.12, .00)	.00 (-.08, .08)	-.04 (-.11, .03)	-.05 (-.15, .04)	-.02 (-.11, .08)	.04 (-.04, .12)	.00 (-.08, .07)	-.03 (-.11, .03)	
Origin 2nd Nonw	-.04 (-.15, .06)	**-.25** ~ (-.38,-.12)	**.21** ~ (.09, .33)	**.21** ~ (.03, .37)	**-.29** * (-.45,-.11)	**.14** ~ (-.01, .29)	-.01 (-.15, .13)	-.09 (-.24, .05)	
Sim PC no vs yes	-.02 (-.06, .02)	**-.24** ~ (-.29,-.19)	**.24** ~ (.17, .31)	**.29** * (.21, .36)	**.11** ~ (.03, .19)	**.19** ~ (.12, .26)	**-.13** ~ (-.20,-.06)	.09 (.03, .15)	

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

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table J.2. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Primacy Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *HE*	δ *IN*	δ *PE*	δ *PO*	δ *RE*	δ *WO*	δ *LF*
Gender ma vs fe	**-.12** ~ (-.15, -.09)	**.12** ~ (.09, .14)	.04 (.01, .07)	.02 (-.02, .06)	.03 (.00, .07)	**.12** ~ (.09, .16)	.06 (.03, .09)	
Age 15_24 yrs	**-.36 *** (-.40, -.32)	-.10 (-.14, -.06)	**-.31 *** (-.37, -.26)	**-.18** ~ (-.24, -.12)	**-.11** ~ (-.17, -.06)	-.09 (-.14, -.04)	-.05 (-.09, -.01)	
Age 25_34 yrs	**-.25** ~ (-.29, -.21)	.04 (.00, .08)	.02 (-.03, .08)	.04 (-.03, .11)	.02 (-.03, .07)	.03 (-.02, .08)	**.25** ~ (.21, .29)	
Age 35_44 yrs	**-.24** ~ (-.27, -.20)	.05 (.01, .08)	**.17** ~ (.12, .21)	.02 (-.03, .07)	.08 (.04, .13)	.04 (-.01, .08)	**.21** ~ (.18, .24)	
Age 45_54 yrs	-.10 (-.13, -.06)	.03 (.00, .06)	**.14** ~ (.10, .18)	.08 (.03, .12)	.06 (.02, .11)	.05 (.01, .09)	**.24** ~ (.21, .27)	
Age 55_64 yrs	**.26** ~ (.22, .29)	-.01 (-.05, .02)	.08 (.04, .12)	.04 (.00, .08)	.03 (-.01, .07)	.04 (.00, .07)	.05 (.02, .09)	
Age 65_older	**.36 *** (.33, .39)	-.01 (-.05, .02)	**-.16** ~ (-.20, -.12)	-.03 (-.07, .01)	-.09 (-.13, -.06)	-.06 (-.10, -.03)	**-.52 #** (-.54, -.49)	
Education Primary	.03 (-.03, .09)	**-.11** ~ (-.16, -.06)	**-.23** ~ (-.29, -.17)	**-.15** ~ (-.21, -.08)	-.08 (-.15, -.01)	**-.14** ~ (-.20, -.09)	-.09 (-.15, -.04)	
Education VMBO	**.14** ~ (.11, .18)	-.06 (-.09, -.03)	**-.13** ~ (-.16, -.09)	**-.13** ~ (-.17, -.09)	.01 (-.03, .05)	**-.11** ~ (-.15, -.08)	**-.18** ~ (-.22, -.15)	
Education HAVWO	**-.12** ~ (-.16, -.07)	-.04 (-.08, .00)	-.02 (-.08, .03)	.05 (.00, .11)	.02 (-.03, .07)	.03 (-.02, .08)	-.06 (-.11, -.02)	
Education MBO	-.04 (-.08, -.01)	.03 (.00, .06)	.03 (-.01, .08)	-.03 (-.07, .02)	-.02 (-.06, .02)	.01 (-.03, .04)	.06 (.02, .09)	
Education HBO	-.01 (-.04, .03)	.08 (.05, .10)	**.14** ~ (.10, .17)	**.13** ~ (.09, .17)	.01 (-.02, .05)	.08 (.05, .12)	.10 (.06, .13)	
Education WO	-.10 (-.14, -.05)	.06 (.02, .10)	**.18** ~ (.12, .24)	**.18** ~ (.12, .24)	.03 (-.02, .08)	**.16** ~ (.11, .21)	**.24** ~ (.19, .28)	
Domestic Single	-.04 (-.08, .00)	-.06 (-.09, -.02)	**-.15** ~ (-.19, -.11)	.01 (-.04, .06)	-.04 (-.08, .01)	**-.13** ~ (-.17, -.10)	-.10 (-.14, -.07)	
Domestic Cohab	**.34 *** (.31, .36)	.01 (-.02, .03)	.04 (.00, .07)	.02 (-.02, .05)	.02 (-.02, .05)	.03 (.00, .06)	**-.18** ~ (-.21, -.15)	
Domestic Cohab+ch	**-.25** ~ (-.27, -.22)	.04 (.02, .07)	.10 (.07, .14)	.01 (-.03, .05)	.02 (-.02, .05)	.08 (.04, .11)	**.23** ~ (.20, .26)	
Domestic Single+ch	**-.26** ~ (-.33, -.20)	-.08 (-.14, -.01)	**-.12** ~ (-.21, -.04)	**-.11** ~ (-.20, -.02)	-.04 (-.12, .04)	-.08 (-.15, .00)	.09 (.03, .15)	
Domestic Other	**-.16** ~ (-.28, -.04)	.02 (-.10, .14)	**-.19** ~ (-.36, -.01)	**-.11** ~ (-.28, .05)	-.05 (-.19, .09)	-.09 (-.22, .03)	.02 (-.10, .14)	

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table J.2 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Primacy Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ FA	δ HE	δ IN	δ PE	δ PO	δ RE	δ WO	δ LF
Occupation Paid	**-.19 ~** (-.22, -.16)	**.16 ~** (.14, .19)	**.18 ~** (.14, .21)	.10 (.06, .13)	.06 (.03, .10)	.10 (.07, .13)	**.52 #** (.49, .54)	
Occupation Famself	-.04 (-.10, .03)	.08 (.04, .13)	.09 (.02, .16)	.01 (-.07, .09)	.07 (.00, .15)	.06 (-.01, .12)	.10 (.05, .14)	
Occupation Seek	-.01 (-.09, .08)	-.07 (-.15, .00)	.06 (-.04, .17)	.01 (-.11, .11)	.07 (-.04, .17)	**-.11 ~** (-.21, -.02)	.02 (-.06, .10)	
Occupation Other	.09 (.03, .14)	**-.44 #** (-.50, -.39)	-.06 (-.13, .00)	-.02 (-.09, .05)	.07 (.00, .14)	-.05 (-.10, .01)	**-.40 *** (-.44, -.35)	
Occupation School	**-.38 *** (-.43, -.34)	**-.11 ~** (-.15, -.08)	**-.29 *** (-.35, -.24)	**-.18 ~** (-.24, -.11)	**-.12 ~** (-.18, -.06)	-.09 (-.15, -.04)	**-.11 ~** (-.15, -.06)	**.11 ~** (-.07, .28)
Occupation Housekeep	**.30 *** (.25, .35)	-.08 (-.12, -.03)	-.03 (-.10, .03)	-.03 (-.09, .03)	-.01 (-.07, .05)	-.07 (-.13, -.01)	**-.25 ~** (-.30, -.20)	
Occupation Retired	**.32 *** (.29, .35)	.01 (-.02, .04)	**-.12 ~** (-.16, -.08)	-.03 (-.07, .01)	-.10 (-.14, -.05)	-.04 (-.08, .00)	**-.47 #** (-.50, -.44)	
Income No income	-.08 (-.13, -.03)	-.10 (-.14, -.06)	**-.17 ~** (-.22, -.11)	**-.16 ~** (-.22, -.10)	-.03 (-.09, .03)	-.10 (-.15, -.04)	**-.13 ~** (-.17, -.08)	
Income 0001_1000	.09 (.06, .13)	**-.13 ~** (-.16, -.10)	-.08 (-.12, -.04)	-.04 (-.08, .01)	-.02 (-.06, .02)	-.10 (-.13, -.06)	**-.14 ~** (-.17, -.10)	
Income 1001_2000	-.01 (-.04, .02)	.03 (.01, .06)	.01 (-.02, .05)	.00 (-.04, .04)	-.01 (-.04, .03)	-.01 (-.04, .02)	.02 (-.01, .05)	
Income 2001_3000	-.02 (-.06, .02)	**.15 ~** (.12, .18)	**.16 ~** (.12, .20)	**.14 ~** (.09, .19)	.04 (-.01, .08)	**.15 ~** (.11, .20)	**.13 ~** (.09, .17)	
Income 3001_more	-.01 (-.07, .06)	**.11 ~** (.06, .16)	**.25 ~** (.18, .32)	**.18 ~** (.09, .25)	.02 (-.06, .09)	**.18 ~** (.11, .24)	**.20 ~** (.14, .26)	
Income DK	-.09 (-.21, .03)	-.06 (-.17, .05)	**-.12 ~** (-.29, .05)	.00 (-.25, .06)	-.02 (-.17, .13)	-.08 (-.21, .05)	.04 (-.07, .15)	
Income WT	-.06 (-.14, .02)	-.02 (-.09, .05)	**-.17 ~** (-.26, -.08)	-.10 (-.17, .02)	.07 (-.03, .17)	-.02 (-.10, .07)	.09 (.01, .17)	
Origin Dutch	.09 (.04, .14)	.06 (.02, .10)	.06 (.00, .12)	**.12 ~** (.05, .18)	.05 (-.01, .11)	**.12 ~** (.07, .17)	.02 (-.03, .06)	
Origin 1st West	-.03 (-.12, .07)	-.01 (-.09, .07)	-.03 (-.13, .07)	**-.14 ~** (-.25, -.02)	**-.12 ~** (-.24, .00)	**-.14 ~** (-.23, -.03)	-.07 (-.16, .03)	
Origin 1st Nonw	**-.22 ~** (-.30, -.13)	**-.14 ~** (-.23, -.06)	-.08 (-.19, .04)	**-.23 ~** (-.34, -.11)	-.06 (-.17, .06)	**-.23 ~** (-.31, -.14)	.08 (.00, .16)	
Origin 2nd West	.03 (-.04, .11)	-.01 (-.07, .05)	.00 (-.08, .08)	.03 (-.07, .12)	.01 (-.07, .10)	.02 (-.06, .09)	-.05 (-.13, .02)	
Origin 2nd Nonw	**-.26 ~** (-.37, -.13)	**-.11 ~** (-.24, .02)	**-.22 ~** (-.39, -.04)	**-.17 ~** (-.36, .01)	-.05 (-.19, .10)	**-.16 ~** (-.28, -.04)	-.01 (-.12, .09)	
Sim PC no vs yes	**-.13 ~** (-.20, -.07)	**.12 ~** (.05, .18)	**.20 ~** (.13, .27)	**.11 ~** (.04, .18)	.00 (-.08, .08)	**.14 ~** (.07, .21)	**.23 ~** (.16, .29)	

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

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

Table J.3. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Straightlining'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *HE*	δ *IN*	δ *PE*	δ *PO*	δ *RE*	δ *WO*
Gender ma vs fe	.02 (-.02, .05)	**.12** ~ (.08, .16)	.02 (-.02, .06)	-.02 (-.05, .01)	**-.13** ~ (-.16, -.09)	**.20** ~ (.17, .24)	.01 (-.01, .02)
Age 15_24 yrs	-.01 (-.12, .11)	**.14** ~ (.08, .20)	**-.26** ~ (-.39, -.12)	.05 (-.01, .10)	-.02 (-.09, .04)	**-.12** ~ (-.17, -.06)	.04 (.00, .08)
Age 25_34 yrs	.02 (-.04, .08)	**.25** ~ (.20, .31)	**-.12** ~ (-.19, -.05)	-.02 (-.07, .03)	.07 (.01, .13)	-.04 (-.09, .02)	-.01 (-.03, .01)
Age 35_44 yrs	**.16** ~ (.11, .21)	**.18** ~ (.13, .23)	.02 (-.04, .08)	.03 (-.02, .07)	**.13** ~ (.08, .18)	.03 (-.01, .08)	-.01 (-.02, .01)
Age 45_54 yrs	**.12** ~ (.08, .17)	.07 (.02, .11)	.03 (-.02, .09)	.01 (-.03, .05)	**.11** ~ (.06, .15)	.05 (.01, .10)	-.01 (-.02, .01)
Age 55_64 yrs	-.07 (-.11, -.03)	**-.13** ~ (-.17, -.08)	-.05 (-.10, -.01)	-.03 (-.07, .00)	-.07 (-.11, -.02)	.01 (-.03, .06)	-.01 (-.02, .01)
Age 65_older	**-.20** ~ (-.23, -.17)	**-.27** ~ (-.31, -.23)	.09 (.04, .13)	-.01 (-.04, .03)	**-.15** ~ (-.19, -.11)	.01 (-.04, .05)	.06 (.02, .11)
Education Primary	-.02 (-.08, .05)	**-.20** ~ (-.27, -.14)	-.03 (-.11, .05)	.04 (-.02, .09)	-.03 (-.10, .04)	-.03 (-.10, .03)	.00 (-.03, .04)
Education VMBO	-.01 (-.05, .03)	**-.15** ~ (-.19, -.11)	.02 (-.03, .06)	.08 (.05, .12)	.04 (.00, .09)	-.04 (-.08, .00)	.01 (-.01, .04)
Education HAVWO	-.02 (-.09, .04)	.03 (-.02, .09)	-.05 (-.12, .02)	-.05 (-.10, -.01)	-.09 (-.14, -.04)	-.01 (-.06, .05)	-.01 (-.03, .01)
Education MBO	.04 (.00, .08)	.02 (-.02, .07)	.01 (-.04, .06)	.02 (-.01, .06)	.02 (-.02, .06)	.01 (-.03, .06)	.00 (-.01, .02)
Education HBO	.00 (-.04, .05)	.10 (.06, .15)	.03 (-.02, .08)	-.07 (-.11, -.04)	-.01 (-.05, .03)	-.01 (-.05, .03)	-.01 (-.03, .00)
Education WO	-.04 (-.10, .01)	**.23** ~ (.17, .29)	-.05 (-.12, .03)	-.07 (-.11, -.02)	.01 (-.05, .07)	**.11** ~ (.04, .17)	.01 (-.01, .03)
Domestic Single		-.10 (-.15, -.05)	-.06 (-.11, -.01)	-.05 (-.09, -.01)	-.04 (-.08, .01)	**-.20** ~ (-.25, -.15)	.00 (-.02, .02)
Domestic Cohab	**-.27** ~ (-.31, -.24)	**-.14** ~ (-.18, -.10)	-.02 (-.07, .02)	-.01 (-.04, .02)	-.07 (-.11, -.04)	.03 (.00, .07)	.00 (-.02, .01)
Domestic Cohab+ch	**.29** * (.25, .32)	**.20** ~ (.16, .24)	.09 (.04, .13)	.04 (.00, .07)	.08 (.04, .12)	**.11** ~ (.08, .15)	.00 (-.01, .02)
Domestic Single+ch	.01 (-.10, .11)	.05 (-.04, .13)	-.04 (-.16, .08)	.01 (-.06, .09)	**.13** ~ (.04, .22)	-.05 (-.14, .03)	.01 (-.02, .05)
Domestic Other	-.10 (-.23, .07)	-.06 (-.23, .10)	-.10 (-.31, .11)	.05 (-.08, .20)	-.07 (-.23, .09)	-.05 (-.21, .11)	-.01 (-.04, .03)

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

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.3 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Straightlining'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *HE*	δ *IN*	δ *PE*	δ *PO*	δ *RE*	δ *WO*
Occupation Paid	**.11** ~ (.08, .15)	**.25** ~ (.21, .28)	.03 (-.02, .07)	-.01 (-.04, .02)	.10 (.06, .13)	.06 (.02, .09)	**-.11** ~ (-.13, -.09)
Occupation Famself	.06 (-.01, .14)	**.13** ~ (.06, .20)	.02 (-.07, .11)	.02 (-.05, .09)	.06 (-.02, .13)	.07 (.00, .14)	**.20** ~ (.18, .22)
Occupation Seek	.02 (-.10, .14)	.03 (-.07, .14)	**-.22** ~ (-.34, -.11)	-.04 (-.13, .05)	.00 (-.11, .11)	-.07 (-.18, .03)	-.01 (-.05, .05)
Occupation Other	.00 (-.06, .07)	**-.40 *** (-.46, -.34)	**-.14** ~ (-.22, -.06)	.04 (-.02, .10)	.01 (-.06, .08)	-.05 (-.12, .02)	.01 (-.04, .09)
Occupation School		**.11** ~ (.04, .17)	**-.29 *** (-.45, -.12)	.02 (-.03, .08)	-.04 (-.11, .04)	**-.13** ~ (-.20, -.07)	.05 (-.01, .12)
Occupation Housekeep	-.01 (-.07, .05)	**-.16** ~ (-.22, -.09)	-.03 (-.11, .04)	.02 (-.03, .08)	.04 (-.03, .10)	-.06 (-.12, .00)	-.02 (-.04, .00)
Occupation Retired	**-.19** ~ (-.22, -.16)	**-.23** ~ (-.27, -.18)	.09 (.04, .14)	-.02 (-.05, .02)	**-.17** ~ (-.21, -.12)	.02 (-.03, .06)	.01 (-.03, .06)
Income No income	-.02 (-.09, .05)	-.01 (-.08, .04)	-.02 (-.10, .07)	.03 (-.02, .08)	.02 (-.04, .08)	-.08 (-.14, -.03)	.05 (.00, .11)
Income 0001_1000	-.03 (-.07, .01)	**-.13** ~ (-.17, -.08)	-.08 (-.14, -.03)	.03 (-.01, .07)	.01 (-.03, .05)	-.07 (-.11, -.03)	.02 (.00, .04)
Income 1001_2000	.03 (-.01, .06)	-.02 (-.06, .02)	-.04 (-.08, .01)	-.03 (-.06, .01)	-.01 (-.05, .02)	-.02 (-.06, .01)	-.02 (-.03, .00)
Income 2001_3000	-.03 (-.07, .02)	**.11** ~ (.06, .16)	.06 (.01, .12)	-.06 (-.10, -.02)	-.07 (-.11, -.02)	**.12** ~ (.07, .17)	-.02 (-.04, -.01)
Income 3001_more	-.03 (-.10, .05)	**.16** ~ (.08, .24)	**.15** ~ (.06, .24)	-.04 (-.11, .03)	-.03 (-.11, .05)	**.18** ~ (.10, .26)	.01 (-.01, .03)
Income DK	**.17** ~ (.02, .31)	.05 (-.11, .21)	.05 (-.14, .24)	**.11** ~ (-.03, .25)	**.14** ~ (-.01, .29)	-.06 (-.20, .07)	.09 (.03, .16)
Income WT	.09 (-.01, .19)	**.12** ~ (.02, .22)	**.16** ~ (.05, .27)	**.18** ~ (.08, .27)	**.20** ~ (.10, .30)	.04 (-.05, .14)	.01 (-.02, .06)
Origin Dutch	-.03 (-.09, .03)	.01 (-.05, .07)	.06 (-.01, .13)	-.07 (-.12, -.01)	-.04 (-.09, .02)	.04 (-.02, .09)	-.01 (-.04, .01)
Origin 1st West	-.01 (-.11, .10)	-.03 (-.13, .09)	.03 (-.09, .16)	.05 (-.05, .16)	.04 (-.07, .15)	-.09 (-.19, .02)	.01 (-.04, .08)
Origin 1st Nonw	**.12** ~ (-.01, .25)	.01 (-.10, .12)	-.05 (-.17, .07)	**.18** ~ (.07, .28)	.05 (-.06, .16)	-.06 (-.17, .05)	.02 (-.02, .08)
Origin 2nd West	-.03 (-.11, .06)	-.05 (-.13, .04)	**-.12** ~ (-.22, -.02)	-.01 (-.09, .06)	.02 (-.08, .11)	-.02 (-.11, .06)	.00 (-.03, .04)
Origin 2nd Nonw	.09 (-.12, .30)	.10 (-.04, .24)	-.09 (-.34, .16)	.03 (-.10, .17)	.04 (-.12, .21)	.08 (-.07, .22)	.02 (-.04, .14)
Sim PC no vs yes	.09 (.02, .17)	**.31 *** (.23, .38)	.08 (.00, .17)	-.02 (-.09, .04)	.06 (-.01, .14)	**.12** ~ (.04, .19)	.01 (-.04, .04)

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.4. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Answering Don't Know'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *AS*	δ *FA*	δ *HO*	δ *IN*	δ *PO*	δ *RE*	δ *WO*
Gender ma vs fe		.05 (.02, .07)	**-.17** ~ (-.21, -.13)	**-.18** ~ (-.21, -.14)	**-.26** ~ (-.30, -.23)	-.01 (-.01, .00)	**-.13** ~ (-.17, -.10)
Age 15_24 yrs		.09 (.05, .12)		**.46** # (.41, .51)	**.28** * (.22, .34)	.05 (.03, .07)	**.24** ~ (.19, .30)
Age 25_34 yrs		.05 (.01, .09)	.09 (.01, .16)	**.13** ~ (.07, .18)	.10 (.04, .16)	.01 (.00, .03)	-.05 (-.09, .00)
Age 35_44 yrs		-.04 (-.07, -.01)	**.11** ~ (.05, .18)	.04 (.00, .09)	.09 (.04, .14)	-.01 (-.02, .00)	-.07 (-.11, -.03)
Age 45_54 yrs		-.07 (-.10, -.05)	.07 (.02, .13)	.01 (-.03, .06)	.01 (-.03, .06)	.01 (-.01, .02)	-.03 (-.07, .01)
Age 55_64 yrs			-.05 (-.10, -.01)	**-.16** ~ (-.20, -.12)	**-.15** ~ (-.19, -.10)	-.02 (-.03, -.01)	-.04 (-.09, .00)
Age 65_older			**-.13** ~ (-.17, -.09)	**-.20** ~ (-.24, -.16)	**-.14** ~ (-.18, -.10)	-.02 (-.03, -.01)	
Education Primary		**.15** ~ (.08, .22)	.08 (.00, .15)	**.16** ~ (.10, .23)	**.17** ~ (.11, .24)	.02 (.00, .05)	**.23** ~ (.15, .31)
Education VMBO		.03 (-.01, .07)	.02 (-.03, .07)	.07 (.03, .12)	.08 (.04, .12)	.00 (-.01, .02)	**.12** ~ (.07, .16)
Education HAVWO		-.02 (-.05, .02)	-.05 (-.12, .01)	.04 (-.02, .10)	.00 (-.05, .06)	.01 (-.01, .02)	.10 (.04, .16)
Education MBO		.00 (-.03, .03)	.07 (.02, .12)	.05 (.01, .09)	.06 (.01, .10)	.00 (-.01, .01)	-.04 (-.08, .00)
Education HBO		-.05 (-.08, -.02)	-.05 (-.10, .00)	**-.15** ~ (-.19, -.11)	**-.13** ~ (-.17, -.10)	-.01 (-.02, .00)	**-.12** ~ (-.16, -.09)
Education WO		-.03 (-.07, .01)	-.08 (-.15, -.01)	**-.17** ~ (-.23, -.11)	**-.19** ~ (-.24, -.13)	-.01 (-.02, .00)	-.08 (-.14, -.03)
Domestic Single		.06 (.02, .10)	-.08 (-.13, -.04)	-.02 (-.06, .02)	-.04 (-.08, .00)	-.01 (-.02, .01)	.01 (-.04, .06)
Domestic Cohab			-.07 (-.11, -.03)	**-.24** ~ (-.27, -.20)	**-.12** ~ (-.15, -.08)	-.02 (-.02, -.01)	-.04 (-.07, .00)
Domestic Cohab+ch		-.03 (-.06, -.01)	**.16** ~ (.11, .21)	**.19** ~ (.15, .22)	**.11** ~ (.07, .15)	.01 (.00, .02)	.00 (-.03, .04)
Domestic Single+ch		.02 (-.03, .07)	.00 (-.09, .09)	**.28** * (.20, .37)	**.16** ~ (.07, .25)	.04 (.01, .07)	.07 (-.01, .15)
Domestic Other		.04 (-.06, .14)	.09 (-.08, .27)	**.27** ~ (.12, .42)	.04 (-.10, .18)	-.01 (-.03, .02)	.10 (-.04, .24)

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

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.4 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Answering Don't Know'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *AS*	δ *FA*	δ *HO*	δ *IN*	δ *PO*	δ *RE*	δ *WO*
Occupation Paid		-.02 (-.05, .00)	.07 (.03, .11)	-.04 (-.08, .00)	.00 (-.03, .04)	-.01 (-.02, .00)	**-.26** ~ (-.30, -.22)
Occupation Famself		-.04 (-.09, .01)	.02 (-.07, .11)	**.12** ~ (.04, .20)	-.10 (-.17, -.02)	.00 (-.02, .02)	.03 (-.03, .09)
Occupation Seek		-.01 (-.07, .06)	-.02 (-.13, .09)	.04 (-.07, .16)	.03 (-.08, .14)	.01 (-.02, .04)	**.26** ~ (.16, .35)
Occupation Other		.06 (.00, .13)	-.01 (-.07, .07)	.04 (-.03, .11)	.04 (-.03, .11)	.01 (-.01, .02)	
Occupation School		.08 (.04, .12)	-.02 (-.13, .10)	**.46 #** (.41, .51)	**.25** ~ (.19, .32)	.04 (.02, .06)	**.27** ~ (.21, .33)
Occupation Housekeep		-.10 (-.15, -.05)	**.14** ~ (.07, .21)	.03 (-.03, .09)	**.14** ~ (.08, .21)	.01 (-.01, .02)	
Occupation Retired			**-.15** ~ (-.19, -.11)	**-.23** ~ (-.28, -.19)	**-.18** ~ (-.22, -.14)	-.02 (-.03, -.01)	
Income No income		.04 (.00, .08)	**.19** ~ (.09, .29)	**.28 *** (.23, .34)	**.19** ~ (.13, .25)	.03 (.02, .05)	**.34 *** (.27, .41)
Income 0001_1000		.00 (-.03, .04)	.09 (.04, .14)	**.13** ~ (.09, .17)	**.14** ~ (.10, .19)	.00 (-.01, .02)	**.20** ~ (.16, .25)
Income 1001_2000		-.01 (-.04, .02)	.01 (-.03, .06)	-.06 (-.10, -.03)	-.02 (-.05, .02)	-.01 (-.02, .00)	-.07 (-.10, -.03)
Income 2001_3000		-.06 (-.09, -.03)	**-.15** ~ (-.20, -.10)	**-.29 *** (-.33, -.24)	**-.24** ~ (-.28, -.21)	-.02 (-.03, -.01)	**-.23** ~ (-.27, -.20)
Income 3001_more		-.04 (-.10, .03)	**-.18** ~ (-.25, -.10)	**-.29 *** (-.36, -.22)	**-.30 *** (-.36, -.24)	-.02 (-.03, -.01)	**-.21** ~ (-.28, -.15)
Income DK		.05 (-.05, .16)	**.23** ~ (.01, .45)	**.35 *** (.21, .48)	**.22** ~ (.09, .36)	.05 (.00, .12)	**.19** ~ (.06, .32)
Income WT		**.12** ~ (.05, .19)	**.17** ~ (.05, .29)	**.38 *** (.30, .45)	**.16** ~ (.06, .26)	.03 (.00, .07)	.10 (.01, .20)
Origin Dutch		-.05 (-.09, .00)	-.08 (-.15, -.01)	-.09 (-.15, -.03)	**-.11** ~ (-.17, -.05)	-.04 (-.06, -.02)	**-.13** ~ (-.19, -.07)
Origin 1st West		.05 (-.04, .16)	**.12** ~ (.00, .25)	.00 (-.12, .11)	.09 (-.02, .20)	.03 (.00, .07)	**.15** ~ (.02, .27)
Origin 1st Nonw		.09 (.02, .17)	.09 (-.03, .22)	**.20** ~ (.10, .30)	**.21** ~ (.09, .32)	.07 (.03, .12)	**.22** ~ (.11, .33)
Origin 2nd West		-.04 (-.10, .03)	.07 (-.03, .18)	-.01 (-.10, .08)	.01 (-.07, .10)	.01 (-.01, .03)	.05 (-.04, .15)
Origin 2nd Nonw		.07 (-.02, .16)	**-.15** ~ (-.30, .05)	**.26** ~ (.08, .43)	**.13** ~ (-.03, .30)	.07 (.01, .14)	.03 (-.13, .19)
Sim PC no vs yes			-.01 (-.08, .06)	-.02 (-.10, .05)	-.08 (-.15, .00)	.00 (-.02, .02)	-.08 (-.21, .03)

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

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.5. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Acquiescence'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *PE*	δ *PO*	δ *RE*	δ *WO*
Gender ma vs fe	**-.12** ~ (-.14, -.10)	.04 (.01, .08)	.10 (.07, .14)	-.04 (-.07, -.01)	.00 (-.04, .03)
Age 15_24 yrs	**-.19** ~ (-.23, -.15)	-.01 (-.07, .05)	-.02 (-.08, .05)	.00 (-.05, .04)	-.05 (-.12, .02)
Age 25_34 yrs	**.13** ~ (.08, .17)	-.05 (-.11, .02)	.04 (-.02, .10)	.06 (.01, .11)	.00 (-.05, .06)
Age 35_44 yrs	**.14** ~ (.11, .17)	-.06 (-.11, -.01)	-.03 (-.08, .02)	.02 (-.02, .06)	.02 (-.03, .06)
Age 45_54 yrs	.05 (.03, .08)	-.07 (-.12, -.02)	-.05 (-.09, .00)	-.02 (-.06, .02)	.02 (-.02, .07)
Age 55_64 yrs	-.02 (-.05, .01)	.03 (-.02, .07)	.01 (-.03, .05)	.00 (-.04, .04)	.01 (-.04, .06)
Age 65_older	-.10 (-.13, -.08)	.10 (.05, .14)	.04 (-.01, .08)	-.03 (-.07, .01)	
Education Primary	**-.13** ~ (-.18, -.09)	-.03 (-.10, .04)	**-.17** ~ (-.23, -.10)	-.02 (-.07, .04)	-.07 (-.16, .02)
Education VMBO	-.06 (-.09, -.03)	-.04 (-.09, .00)	**-.23** ~ (-.27, -.19)	-.02 (-.06, .01)	**-.11** ~ (-.16, -.05)
Education HAVWO	-.01 (-.04, .03)	.01 (-.05, .06)	.08 (.02, .13)	-.01 (-.05, .03)	-.06 (-.13, .00)
Education MBO	.02 (-.01, .05)	-.01 (-.06, .03)	-.07 (-.11, -.02)	.01 (-.03, .05)	-.02 (-.06, .03)
Education HBO	.06 (.04, .09)	.06 (.02, .11)	**.20** ~ (.16, .24)	.03 (-.01, .06)	**.13** ~ (.09, .17)
Education WO	.08 (.04, .11)	.02 (-.04, .09)	**.31** * (.25, .36)	.00 (-.05, .04)	.05 (-.01, .10)
Domestic Single	.01 (-.02, .04)	-.04 (-.09, .01)	.00 (-.05, .05)	-.03 (-.07, .01)	-.03 (-.08, .03)
Domestic Cohab	-.05 (-.07, -.03)	.09 (.05, .13)	.05 (.01, .08)	-.01 (-.04, .02)	.00 (-.04, .04)
Domestic Cohab+ch	.04 (.01, .06)	-.05 (-.09, -.01)	-.02 (-.06, .02)	.02 (-.01, .05)	.03 (-.01, .07)
Domestic Single+ch	.04 (-.03, .10)	**-.12** ~ (-.21, -.03)	**-.15** ~ (-.24, -.06)	.03 (-.04, .10)	-.06 (-.15, .03)
Domestic Other	-.05 (-.16, .06)	.08 (-.07, .22)	**.12** ~ (-.03, .27)	.09 (-.05, .22)	-.02 (-.20, .15)

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.5 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Acquiescence'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *PE*	δ *PO*	δ *RE*	δ *WO*
Occupation Paid	.10 (.07, .12)	-.03 (-.07, .01)	.04 (.00, .08)	-.02 (-.05, .01)	**.14** ~ (.10, .19)
Occupation Famself	.05 (-.01, .10)	-.03 (-.11, .06)	.01 (-.07, .09)	.01 (-.05, .06)	-.03 (-.09, .03)
Occupation Seek	.03 (-.04, .10)	.02 (-.08, .13)	-.10 (-.21, .00)	-.03 (-.11, .06)	**-.31 *** (-.42, -.18)
Occupation Other	.01 (-.04, .06)	-.05 (-.12, .02)	**-.14** ~ (-.21, -.07)	.04 (-.02, .10)	
Occupation School	**-.17** ~ (-.22, -.13)	-.01 (-.07, .06)	.02 (-.04, .09)	.01 (-.04, .06)	-.04 (-.13, .04)
Occupation Housekeep	.02 (-.02, .06)	-.01 (-.07, .06)	**-.17** ~ (-.23, -.10)	.09 (.02, .15)	
Occupation Retired	-.10 (-.12, -.07)	.08 (.03, .12)	.08 (.03, .12)	-.04 (-.07, .00)	
Income No income	-.08 (-.12, -.03)	-.02 (-.08, .04)	**-.12** ~ (-.18, -.06)	.04 (-.01, .09)	**-.19** ~ (-.28, -.09)
Income 0001_1000	.00 (-.03, .03)	-.04 (-.09, .01)	**-.11** ~ (-.15, -.06)	.03 (.00, .07)	**-.13** ~ (-.19, -.08)
Income 1001_2000	.04 (.01, .06)	.02 (-.02, .06)	.03 (-.01, .07)	-.04 (-.07, .00)	.03 (-.01, .07)
Income 2001_3000	-.01 (-.04, .03)	.04 (.00, .09)	**.18** ~ (.14, .23)	.00 (-.03, .04)	**.11** ~ (.07, .16)
Income 3001_more	-.02 (-.08, .02)	.08 (-.01, .16)	**.28 *** (.20, .35)	-.02 (-.08, .04)	**.13** ~ (.07, .19)
Income DK	-.09 (-.18, .02)	**-.13** ~ (-.28, .03)	**-.21** ~ (-.35, -.08)	.02 (-.11, .14)	.04 (-.12, .20)
Income WT	.04 (-.03, .11)	-.07 (-.17, .03)	**-.33 *** (-.42, -.23)	-.03 (-.12, .06)	.00 (-.09, .09)
Origin Dutch	.00 (-.04, .04)	-.07 (-.13, .00)	**.11** ~ (.05, .17)	**-.17** ~ (-.22, -.12)	**.12** ~ (.05, .19)
Origin 1st West	.01 (-.07, .08)	.09 (-.03, .20)	-.09 (-.20, .02)	.01 (-.08, .10)	-.07 (-.22, .07)
Origin 1st Nonw	.04 (-.04, .13)	.07 (-.05, .19)	**-.18** ~ (-.28, -.07)	**.49** # (.39, .57)	**-.22** ~ (-.33, -.10)
Origin 2nd West	-.01 (-.07, .05)	.00 (-.10, .09)	-.04 (-.13, .05)	-.02 (-.09, .05)	-.07 (-.17, .04)
Origin 2nd Nonw	-.05 (-.17, .08)	**.16** ~ (-.03, .33)	**-.12** ~ (-.29, .05)	**.23** ~ (.08, .38)	-.02 (-.20, .15)
Sim PC no vs yes	.05 (.00, .10)	-.06 (-.15, .02)	**.14** ~ (.06, .22)	-.06 (-.13, .00)	

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

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

Table J.6. Cliff's Delta *δ* (and its 99% Confidence Interval) for the Behaviour **'Neutral Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ FA	*δ IN*	*δ PE*	*δ PO*	*δ WO*
Gender ma vs fe	.05 (.02, .08)	-.04 (-.07, .00)	-.06 (-.10, -.02)	-.05 (-.08, -.01)	-.01 (-.04, .02)
Age 15_24 yrs	-.08 (-.13, -.03)		.04 (-.02, .11)	-.02 (-.08, .04)	**.15** ~ (.10, .19)
Age 25_34 yrs	.09 (.04, .14)	-.06 (-.12, .00)	-.06 (-.12, .01)	-.10 (-.16, -.04)	-.04 (-.07, .00)
Age 35_44 yrs	**.11** ~ (.07, .15)	.03 (-.02, .07)	.02 (-.03, .08)	-.07 (-.12, -.02)	-.01 (-.05, .02)
Age 45_54 yrs	.02 (-.02, .06)	.04 (-.01, .08)	.01 (-.04, .06)	-.01 (-.06, .03)	-.04 (-.07, .00)
Age 55_64 yrs	-.04 (-.07, -.01)	-.01 (-.05, .03)	-.02 (-.06, .03)	.02 (-.02, .06)	.00 (-.04, .04)
Age 65_older	-.08 (-.11, -.04)	.00 (-.04, .04)	.00 (-.05, .04)	**.11** ~ (.07, .15)	
Education Primary	-.04 (-.10, .01)	.07 (.00, .13)	**.14** ~ (.08, .21)	**.14** ~ (.08, .21)	.09 (.02, .17)
Education VMBO	.01 (-.02, .05)	.07 (.03, .11)	**.15** ~ (.10, .19)	.10 (.06, .14)	.09 (.05, .13)
Education HAVWO	-.08 (-.12, -.03)	-.04 (-.09, .02)	-.08 (-.14, -.02)	-.08 (-.14, -.03)	.06 (.01, .11)
Education MBO	.02 (-.01, .05)	.04 (.00, .07)	.03 (-.02, .07)	.02 (-.02, .07)	-.02 (-.05, .02)
Education HBO	.00 (-.03, .04)	-.06 (-.09, -.02)	**-.15** ~ (-.19, -.10)	-.06 (-.11, -.02)	-.08 (-.11, -.05)
Education WO	.05 (.01, .10)	**-.14** ~ (-.20, -.09)	**-.16** ~ (-.23, -.09)	**-.18** ~ (-.23, -.13)	-.04 (-.09, .00)
Domestic Single	**-.19** ~ (-.23, -.16)	**-.11** ~ (-.16, -.07)	.03 (-.02, .08)	.07 (.02, .11)	.02 (-.02, .07)
Domestic Cohab	.03 (.00, .06)	.01 (-.02, .05)	-.05 (-.09, -.01)	.01 (-.03, .05)	-.02 (-.05, .01)
Domestic Cohab+ch	**.11** ~ (.08, .14)	.08 (.05, .12)	.00 (-.04, .04)	-.06 (-.10, -.03)	-.01 (-.04, .02)
Domestic Single+ch	-.07 (-.15, .00)		**.15** ~ (.05, .24)	.07 (-.02, .15)	.05 (-.01, .12)
Domestic Other	-.04 (-.17, .10)		.01 (-.16, .18)	-.05 (-.19, .11)	.01 (-.12, .16)

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table J.6 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Neutral Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *IN*	δ *PE*	δ *PO*	δ *WO*
Occupation Paid	**.12** ~ (.09, .14)	-.01 (-.04, .02)	-.06 (-.10, -.02)	-.09 (-.12, -.05)	**-.11** ~ (-.14, -.07)
Occupation Famself	.01 (-.05, .06)	.01 (-.07, .08)	-.01 (-.09, .08)	-.04 (-.12, .02)	.02 (-.03, .06)
Occupation Seek	-.06 (-.14, .02)	-.02 (-.12, .08)	-.02 (-.13, .10)	-.03 (-.13, .07)	
Occupation Other	-.05 (-.11, .01)	.01 (-.05, .08)	.06 (-.02, .13)	.07 (.00, .14)	
Occupation School	**-.11** ~ (-.17, -.05)		.03 (-.03, .11)	-.05 (-.11, .02)	**.17** ~ (.12, .22)
Occupation Housekeep	-.04 (-.09, .00)	.05 (-.01, .10)	.08 (.02, .15)	**.11** ~ (.05, .18)	
Occupation Retired	-.07 (-.11, -.04)	.00 (-.04, .03)	.01 (-.04, .05)	.09 (.05, .13)	
Income No income	-.08 (-.13, -.04)		.09 (.03, .16)	.04 (-.02, .10)	
Income 0001_1000	-.03 (-.07, .00)	.08 (.04, .12)	.06 (.01, .11)	.06 (.02, .10)	.09 (.05, .13)
Income 1001_2000	.04 (.01, .07)	.01 (-.03, .04)	-.01 (-.05, .03)	.00 (-.04, .03)	-.01 (-.04, .02)
Income 2001_3000	.00 (-.04, .04)	-.09 (-.13, -.05)	**-.12** ~ (-.17, -.07)	-.07 (-.12, -.03)	-.07 (-.11, -.04)
Income 3001_more	.01 (-.05, .07)	**-.16** ~ (-.23, -.10)	**-.17** ~ (-.26, -.07)	**-.16** ~ (-.23, -.08)	**-.12** ~ (-.17, -.07)
Income DK	.08 (-.05, .20)	.08 (-.08, .24)	**.17** ~ (.00, .32)	.10 (-.06, .26)	.02 (-.08, .13)
Income WT	.05 (-.03, .13)	.07 (-.01, .16)	**.11** ~ (.01, .22)	.06 (-.04, .17)	-.03 (-.10, .04)
Origin Dutch	-.01 (-.06, .04)	-.05 (-.10, .01)	-.05 (-.11, .02)	-.07 (-.13, -.01)	-.10 (-.15, -.05)
Origin 1st West	-.01 (-.09, .08)	.06 (-.04, .17)	.05 (-.08, .17)	.09 (-.02, .20)	.06 (-.04, .17)
Origin 1st Nonw	.06 (-.03, .15)	.08 (-.03, .18)	**.11** ~ (-.02, .23)	.07 (-.05, .18)	**.14** ~ (.05, .24)
Origin 2nd West	-.03 (-.10, .04)	.05 (-.04, .13)	.01 (-.09, .10)	.02 (-.06, .11)	.08 (-.01, .17)
Origin 2nd Nonw	.06 (-.08, .20)	**-.18** ~ (-.37, .02)	-.02 (-.20, .16)	**.11** ~ (-.05, .27)	.03 (-.09, .17)
Sim PC no vs yes	.02 (-.04, .08)	-.06 (-.13, .01)	-.05 (-.13, .03)	-.08 (-.15, .00)	

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table J.7. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Extreme Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *IN*	δ *PE*	δ *PO*	δ *WO*
Gender ma vs fe	-.08 (-.11, -.05)	.04 (.01, .07)	-.10 (-.14, -.06)	.00 (-.03, .04)	-.04 (-.08, .00)
Age 15_24 yrs	.00 (-.05, .06)		**-.16** ~ (-.22, -.09)	-.07 (-.13, -.01)	.01 (-.05, .07)
Age 25_34 yrs	**-.17** ~ (-.22, -.12)	-.02 (-.07, .04)	-.08 (-.14, -.01)	.02 (-.04, .08)	-.02 (-.08, .04)
Age 35_44 yrs	**-.15** ~ (-.19, -.11)	-.04 (-.09, .00)	-.04 (-.10, .01)	.06 (.01, .11)	-.03 (-.08, .02)
Age 45_54 yrs	-.07 (-.11, -.03)	-.01 (-.05, .03)	.03 (-.02, .08)	.06 (.01, .11)	-.03 (-.08, .02)
Age 55_64 yrs	.08 (.04, .12)	.04 (.00, .08)	.05 (.00, .10)	.03 (-.02, .07)	.02 (-.03, .08)
Age 65_older	**.19** ~ (.15, .23)	.01 (-.02, .05)	.08 (.03, .13)	-.09 (-.13, -.05)	
Education Primary	.10 (.03, .16)	-.01 (-.07, .05)	.06 (-.01, .13)	-.02 (-.08, .05)	.07 (-.03, .18)
Education VMBO	.09 (.06, .13)	-.06 (-.10, -.03)	.04 (-.01, .09)	-.04 (-.08, .01)	.01 (-.05, .06)
Education HAVWO	.00 (-.05, .05)	.01 (-.05, .07)	-.02 (-.09, .04)	.03 (-.02, .09)	.04 (-.03, .10)
Education MBO	-.02 (-.06, .02)	-.02 (-.06, .02)	-.04 (-.08, .01)	-.03 (-.08, .01)	-.05 (-.09, .00)
Education HBO	-.07 (-.11, -.03)	.03 (-.01, .07)	-.02 (-.06, .03)	.01 (-.03, .05)	-.03 (-.08, .01)
Education WO	**-.11** ~ (-.17, -.06)	**.14** ~ (.08, .20)	-.01 (-.08, .06)	**.11** ~ (.05, .17)	.08 (.02, .14)
Domestic Single	.08 (.04, .13)	.03 (-.01, .07)	.03 (-.02, .09)	.01 (-.04, .06)	-.03 (-.09, .03)
Domestic Cohab	.09 (.06, .12)	.04 (.00, .07)	.06 (.02, .10)	-.01 (-.05, .02)	.08 (.03, .13)
Domestic Cohab+ch	**-.13** ~ (-.17, -.10)	-.07 (-.11, -.04)	-.07 (-.11, -.03)	.00 (-.04, .04)	-.05 (-.10, -.01)
Domestic Single+ch	-.04 (-.12, .04)	.00 (-.09, .10)	-.04 (-.13, .06)	.05 (-.04, .14)	.01 (-.08, .09)
Domestic Other	.01 (-.13, .15)	.03 (-.15, .22)	-.06 (-.23, .10)	.00 (-.16, .15)	.09 (-.09, .28)

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table J.7 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Extreme Responding'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *FA*	δ *IN*	δ *PE*	δ *PO*	δ *WO*
Occupation Paid	**-.18** ~ (-.21, -.15)	-.02 (-.06, .01)	-.03 (-.07, .01)	.06 (.02, .09)	**-.13** ~ (-.18, -.08)
Occupation Famself	-.01 (-.08, .06)	.04 (-.03, .12)	-.02 (-.11, .06)	.08 (.00, .15)	**.12** ~ (.04, .20)
Occupation Seek	.04 (-.05, .13)	.07 (-.04, .17)	.00 (-.12, .12)	.08 (-.02, .19)	**.16** ~ (.03, .29)
Occupation Other	.05 (-.01, .11)	.05 (-.01, .12)	.10 (.03, .18)	**.11** ~ (.03, .18)	
Occupation School	.00 (-.06, .06)		**-.17** ~ (-.24, -.10)	-.08 (-.14, -.01)	.01 (-.05, .08)
Occupation Housekeep	**.17** ~ (.11, .22)	-.03 (-.09, .02)	**.13** ~ (.06, .20)	-.09 (-.15, -.02)	
Occupation Retired	**.16** ~ (.13, .20)	.01 (-.03, .05)	.03 (-.02, .08)	-.09 (-.13, -.05)	
Income No income	.10 (.05, .15)	-.04 (-.11, .03)	-.03 (-.10, .03)	-.04 (-.10, .02)	**.12** ~ (.04, .21)
Income 0001_1000	.06 (.03, .10)	-.04 (-.08, .00)	.04 (-.01, .09)	-.04 (-.08, .01)	.05 (.00, .10)
Income 1001_2000	-.04 (-.07, -.01)	-.05 (-.08, -.02)	.01 (-.03, .05)	.02 (-.02, .06)	-.05 (-.09, -.01)
Income 2001_3000	-.06 (-.10, -.02)	.06 (.01, .10)	-.02 (-.07, .04)	.02 (-.02, .07)	-.03 (-.08, .02)
Income 3001_more	-.06 (-.13, .02)	**.27** ~ (.19, .36)	-.07 (-.16, .03)	.02 (-.06, .11)	.07 (-.02, .16)
Income DK	-.05 (-.18, .07)	.01 (-.14, .17)	-.05 (-.21, .10)	-.02 (-.18, .13)	-.04 (-.21, .14)
Income WT	.01 (-.08, .10)	-.06 (-.15, .03)	.03 (-.08, .14)	.06 (-.04, .16)	-.03 (-.14, .09)
Origin Dutch	-.03 (-.08, .02)	.02 (-.03, .08)	-.06 (-.12, .01)	-.04 (-.10, .01)	-.06 (-.13, .01)
Origin 1st West	-.02 (-.12, .08)	-.02 (-.12, .09)	.06 (-.07, .19)	.04 (-.08, .15)	.05 (-.09, .19)
Origin 1st Nonw	-.01 (-.10, .08)	.02 (-.08, .11)	.06 (-.07, .18)	.06 (-.06, .17)	.08 (-.04, .21)
Origin 2nd West	.08 (.00, .17)	-.06 (-.14, .02)	.03 (-.07, .13)	.02 (-.06, .12)	-.01 (-.12, .09)
Origin 2nd Nonw	.01 (-.12, .14)		.08 (-.11, .26)	.06 (-.12, .23)	**.19** ~ (.00, .38)
Sim PC no vs yes	-.08 (-.15, -.01)	.00 (-.06, .06)	**-.14** ~ (-.22, -.05)	-.01 (-.09, .07)	

~ → small effect; * → medium effect; # → large effect
* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

Table J.8. Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Answering Won't Tell'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *AS*	δ *HO*	δ *IN*	δ *WO*	δ *LF*
Gender ma vs fe		-.05 (-.08, -.02)	-.04 (-.07, -.02)		-.02 (-.05, .01)
Age 15_24 yrs		**.14** ~ (.07, .22)	.02 (-.04, .09)		**-.13** ~ (-.22, -.03)
Age 25_34 yrs		.01 (-.04, .06)	.07 (.03, .12)		-.05 (-.10, .00)
Age 35_44 yrs		.05 (.00, .10)	.05 (.02, .09)		-.01 (-.05, .02)
Age 45_54 yrs		.00 (-.04, .04)	.06 (.03, .10)		-.02 (-.05, .01)
Age 55_64 yrs		-.03 (-.06, .00)	.00 (-.03, .03)		.00 (-.03, .03)
Age 65_older		-.02 (-.05, .00)	**-.12** ~ (-.14, -.09)		.06 (.03, .09)
Education Primary		.04 (.00, .09)	.02 (-.03, .07)		**.17** ~ (.12, .23)
Education VMBO		.03 (.00, .06)	.02 (-.01, .05)		.10 (.06, .13)
Education HAVWO		.01 (-.04, .06)	-.03 (-.06, .02)		-.02 (-.06, .03)
Education MBO		.05 (.01, .08)	.05 (.01, .08)		-.01 (-.04, .02)
Education HBO		-.06 (-.09, -.03)	-.04 (-.07, -.01)		-.10 (-.13, -.08)
Education WO		-.06 (-.11, -.02)	-.05 (-.09, -.01)		-.06 (-.09, -.03)
Domestic Single		.04 (.01, .07)	-.03 (-.06, .00)		.02 (-.01, .05)
Domestic Cohab		-.05 (-.08, -.02)	-.05 (-.07, -.02)		.00 (-.03, .02)
Domestic Cohab+ch		.01 (-.03, .04)	.08 (.05, .11)		-.02 (-.05, .01)
Domestic Single+ch		.04 (-.02, .11)	.02 (-.06, .10)		.04 (-.03, .11)
Domestic Other		.00 (-.10, .10)	.00 (-.14, .15)		-.02 (-.12, .10)

~ → small effect; * → medium effect; # → large effect
* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

Table J.8 (continued). Cliff's Delta δ (and its 99% Confidence Interval) for the Behaviour **'Answering Won't Tell'** for the Categories of the Characteristics Gender, Age, Education, Domestic Situation, Primary Occupation, Income, Origin, and Provided a PC (Sim PC) for All Applicable Surveys*.

	δ *AS*	δ *HO*	δ *IN*	δ *WO*	δ *LF*
Occupation Paid		-.03 (-.06, .00)	.05 (.02, .08)		-.06 (-.08, -.03)
Occupation Famself	**.11** ~ (-.01, .23)	.01 (-.06, .08)	.04 (-.02, .10)		.01 (-.04, .07)
Occupation Seek		.05 (-.03, .14)	.01 (-.07, .09)		-.06 (-.14, .01)
Occupation Other		.02 (-.02, .07)	-.01 (-.06, .04)		.05 (.00, .10)
Occupation School		**.18** ~ (.09, .26)	.02 (-.05, .10)		**-.15** ~ (-.25, -.03)
Occupation Housekeep		.06 (.01, .12)	.07 (.03, .12)		.07 (.02, .12)
Occupation Retired		-.03 (-.06, -.01)	**-.11** ~ (-.13, -.09)		.05 (.02, .08)
Income No income		.09 (.02, .16)	**.11** ~ (.06, .17)		.03 (-.03, .10)
Income 0001_1000		.03 (.00, .06)	-.02 (-.05, .01)		.03 (.00, .07)
Income 1001_2000		-.01 (-.04, .01)	-.06 (-.08, -.03)		.00 (-.03, .03)
Income 2001_3000		-.09 (-.12, -.05)	-.06 (-.08, -.03)		-.06 (-.09, -.03)
Income 3001_more		**-.11** ~ (-.16, -.05)	-.08 (-.12, -.03)		-.06 (-.10, -.03)
Income DK		**.12** ~ (-.04, .29)	**.19** ~ (.06, .33)		**.11** ~ (-.01, .25)
Income WT		**.43** # (.31, .54)	**.54** # (.44, .63)		**.14** ~ (.06, .23)
Origin Dutch		-.01 (-.06, .03)	-.03 (-.08, .01)		-.05 (-.09, .00)
Origin 1st West		-.04 (-.10, .03)	.05 (-.02, .13)		.08 (.01, .17)
Origin 1st Nonw		.06 (-.03, .16)	**.11** ~ (.02, .20)		**.14** ~ (.05, .24)
Origin 2nd West		.00 (-.07, .06)	-.04 (-.10, .01)		-.02 (-.08, .04)
Origin 2nd Nonw		.06 (-.09, .23)	.03 (-.12, .20)		-.03 (-.16, .13)
Sim PC no vs yes		-.07 (-.11, -.02)	.05 (.00, .10)		**-.12** ~ (-.18, -.06)

~ → small effect; * → medium effect; # → large effect
* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

APPENDIX K

See the respondent profiles for all respondent characteristics, all their categories, and all answer behaviours, taken over all surveys together and for each applicable separate survey, in Figures K.1 through K.64. For visual clarity, we did not show the total profiles for all categories of a characteristic taken together. Note that we did not show the profiles for subgroups that are not considered representative for their initial subgroups after excluding respondents who did not fill out at least five items. Also note that complete graphs are not shown in case all subgroups for a specific characteristic and survey are not considered representative for their initial subgroups. Occasionally, only one subgroup for a specific characteristic and survey appeared to be representative for its initial subgroup. In this case, the profiles for this subgroup and for the remaining categories combined were shown (see also Appendices I and J for the excluded subgroups).

SOCIALLY DESIRABLE RESPONDING: GENDER

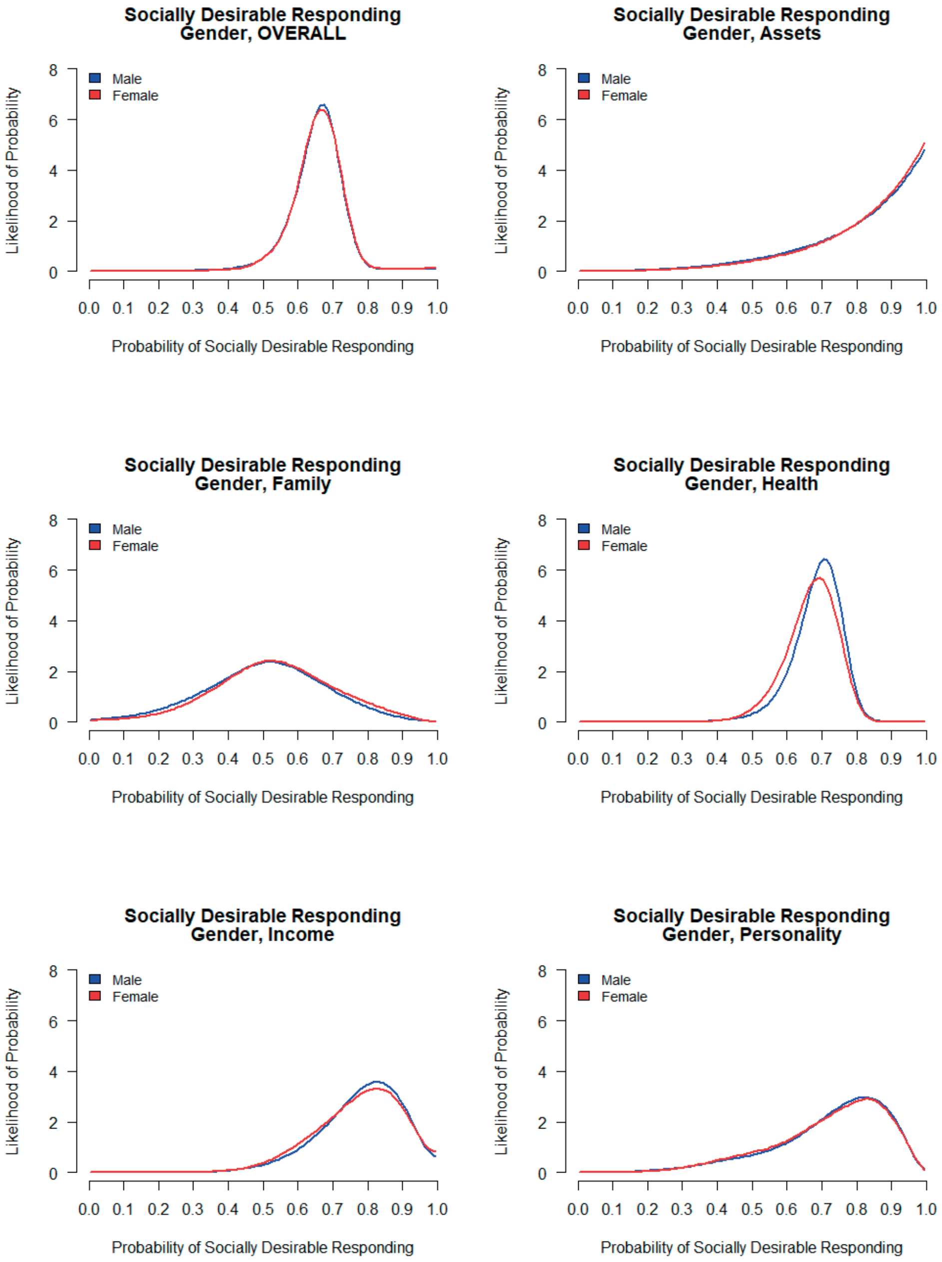
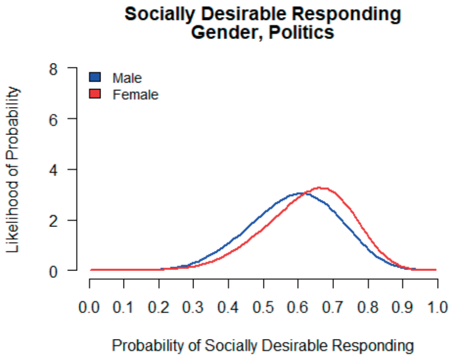
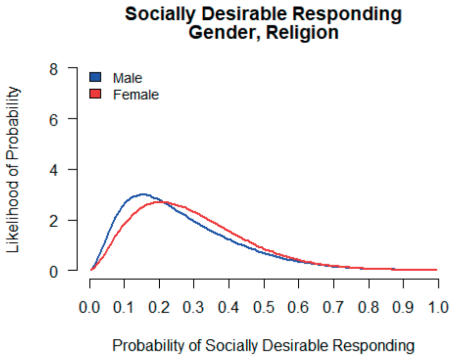


Figure K.1. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).





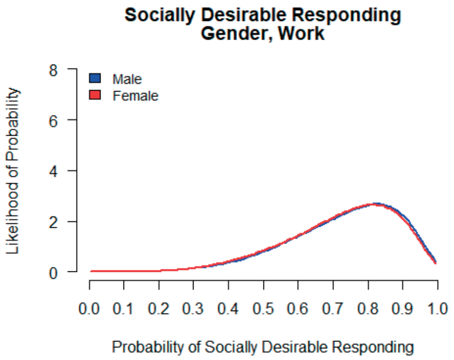


Figure K.1 (continued). Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: AGE

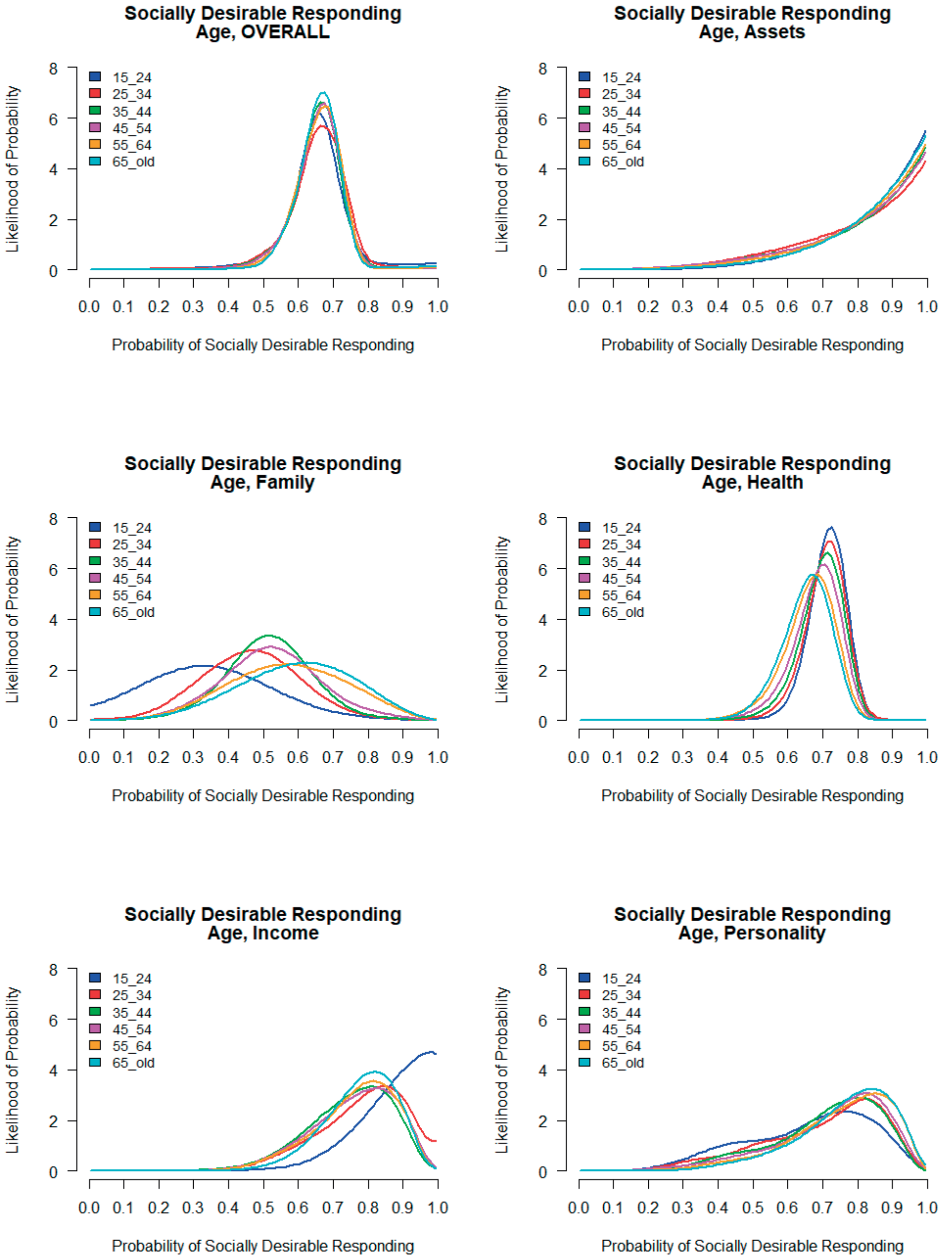


Figure K.2. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

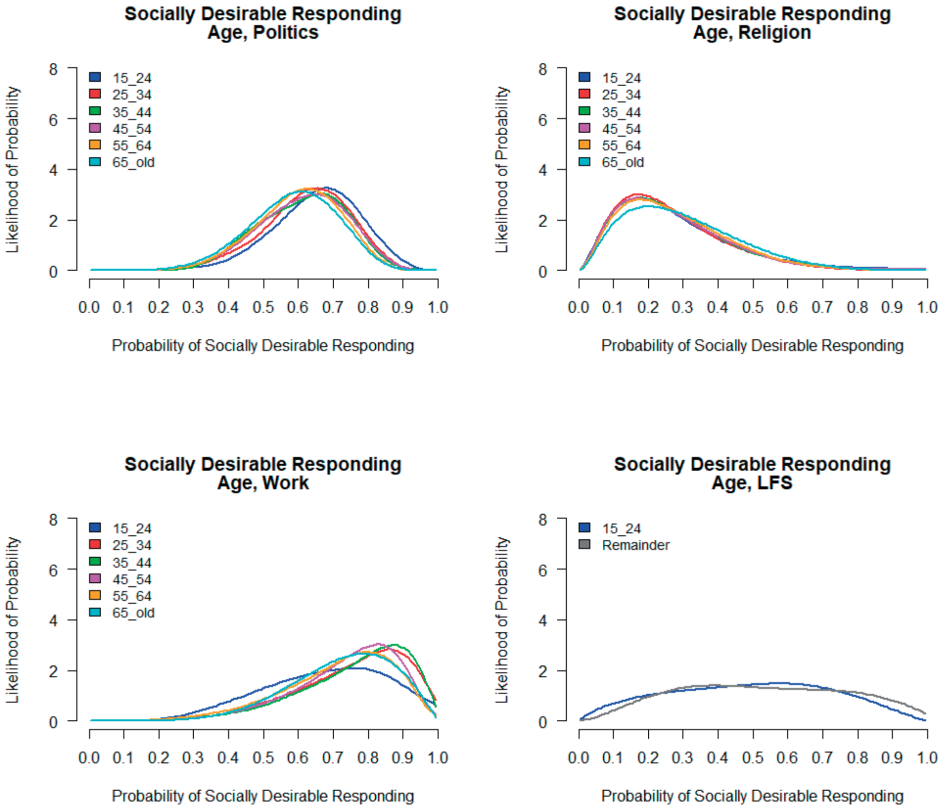


Figure K.2 (continued). Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: EDUCATION

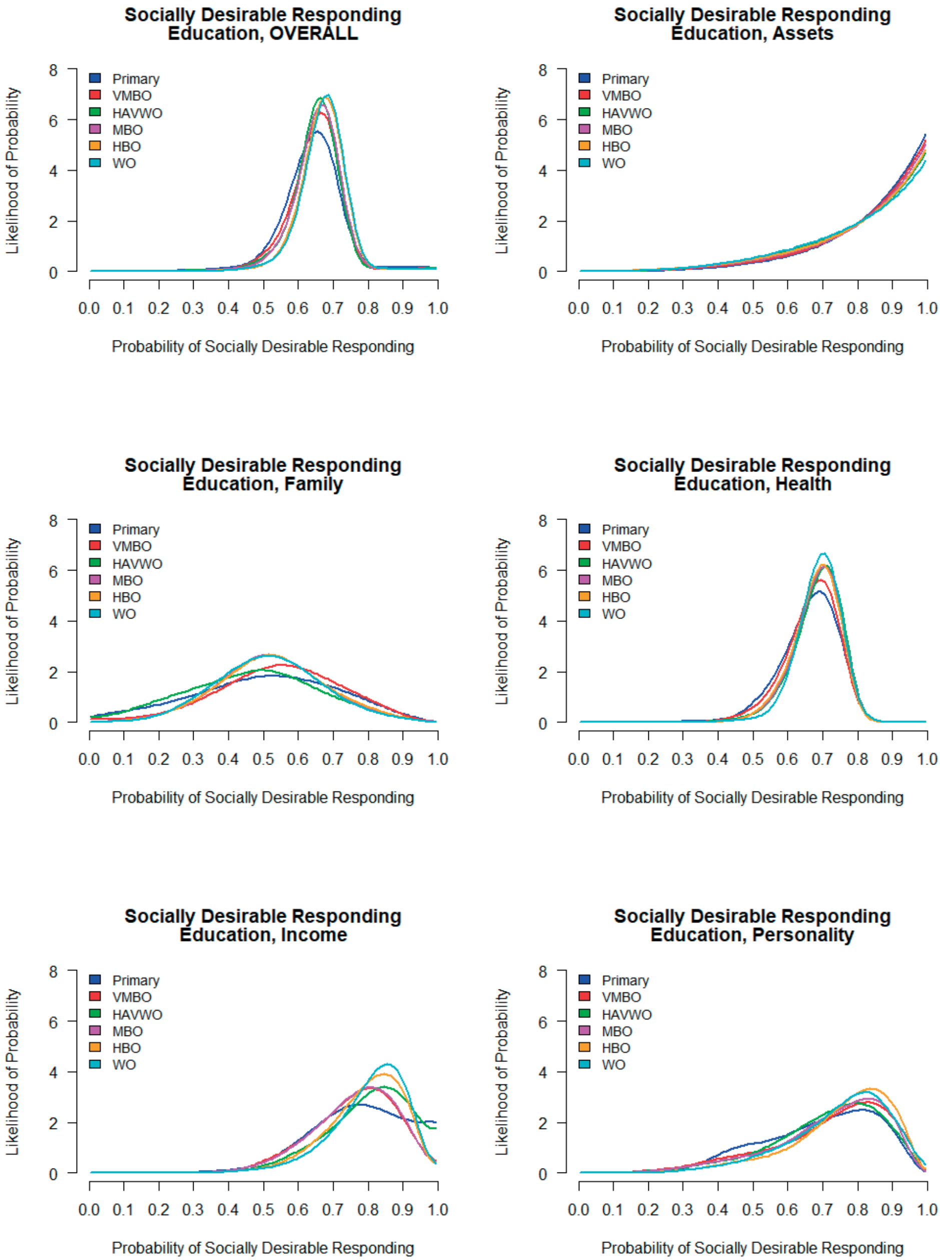


Figure K.3. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

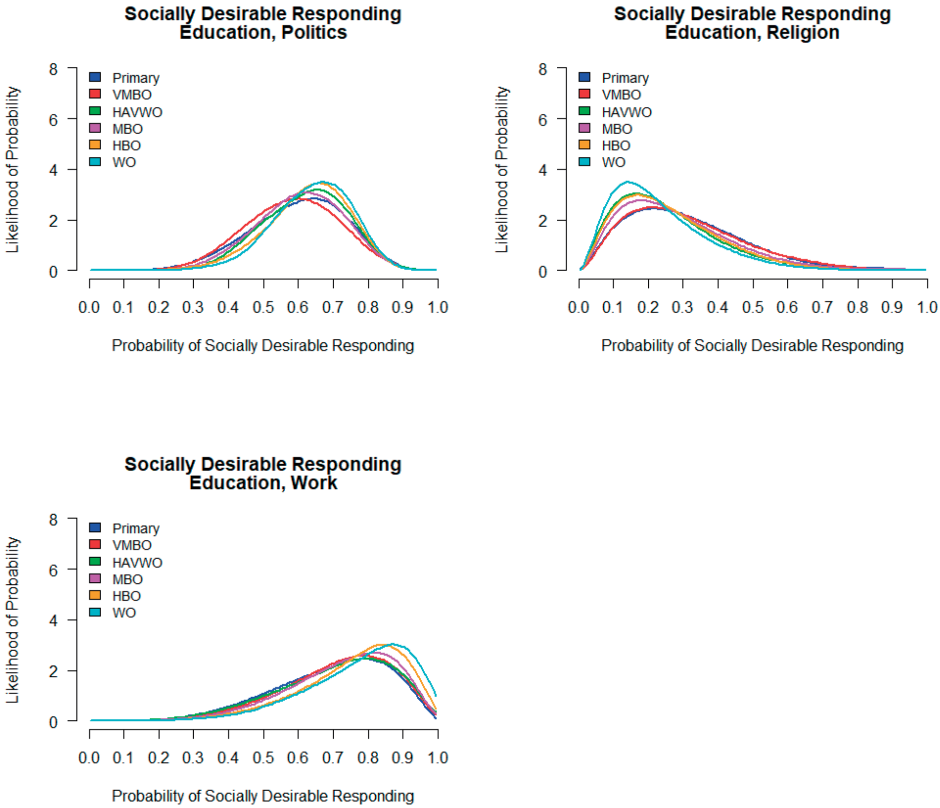


Figure K.3 (continued). Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: DOMESTIC SITUATION

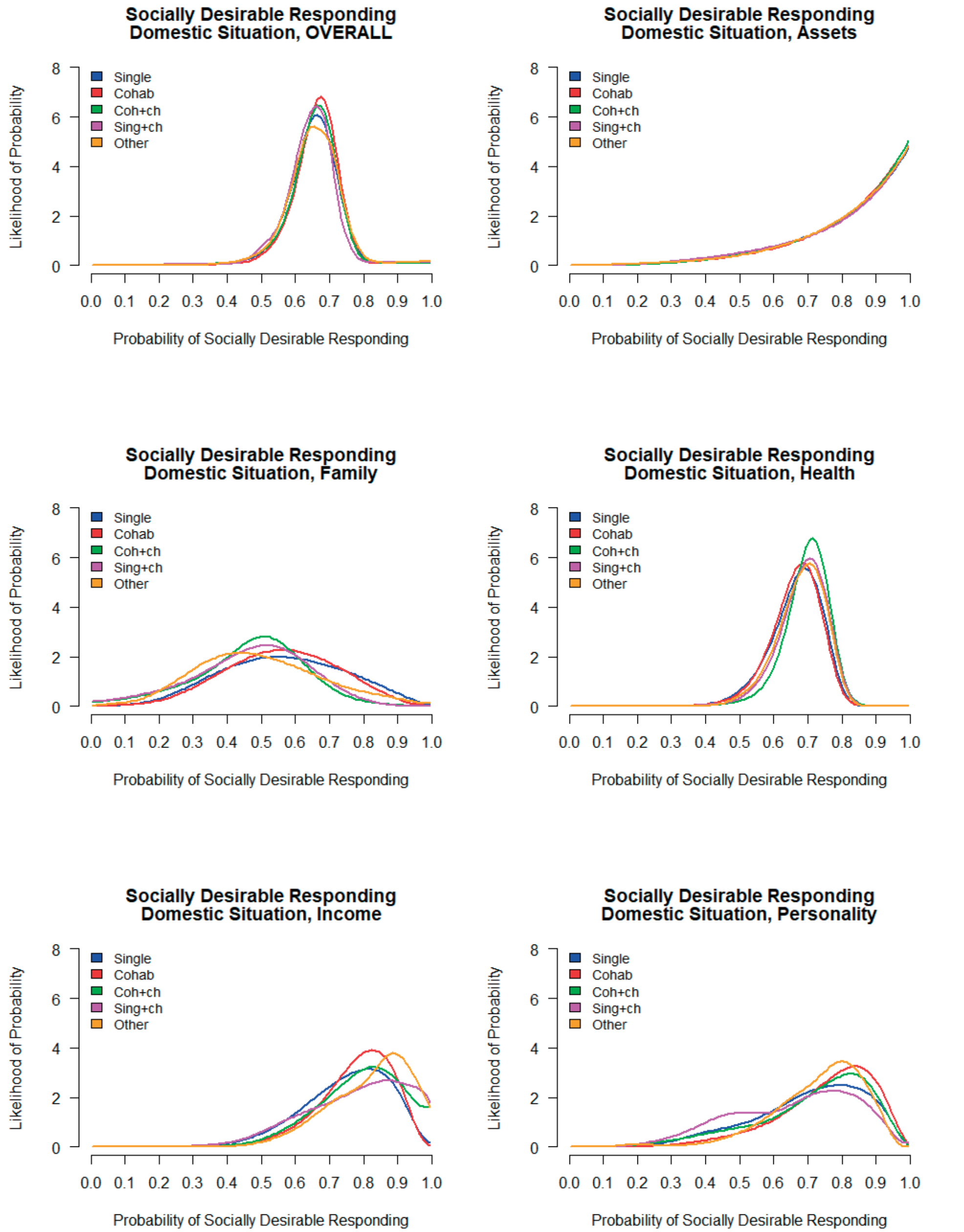


Figure K.4. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

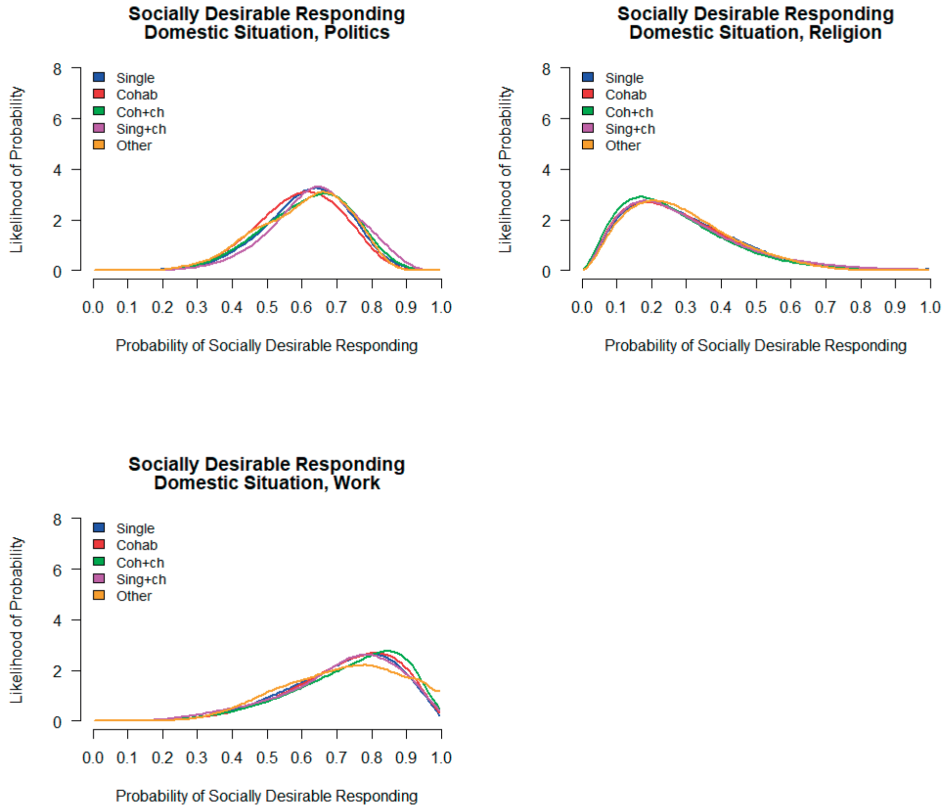


Figure K.4 (continued). Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: PRIMARY OCCUPATION

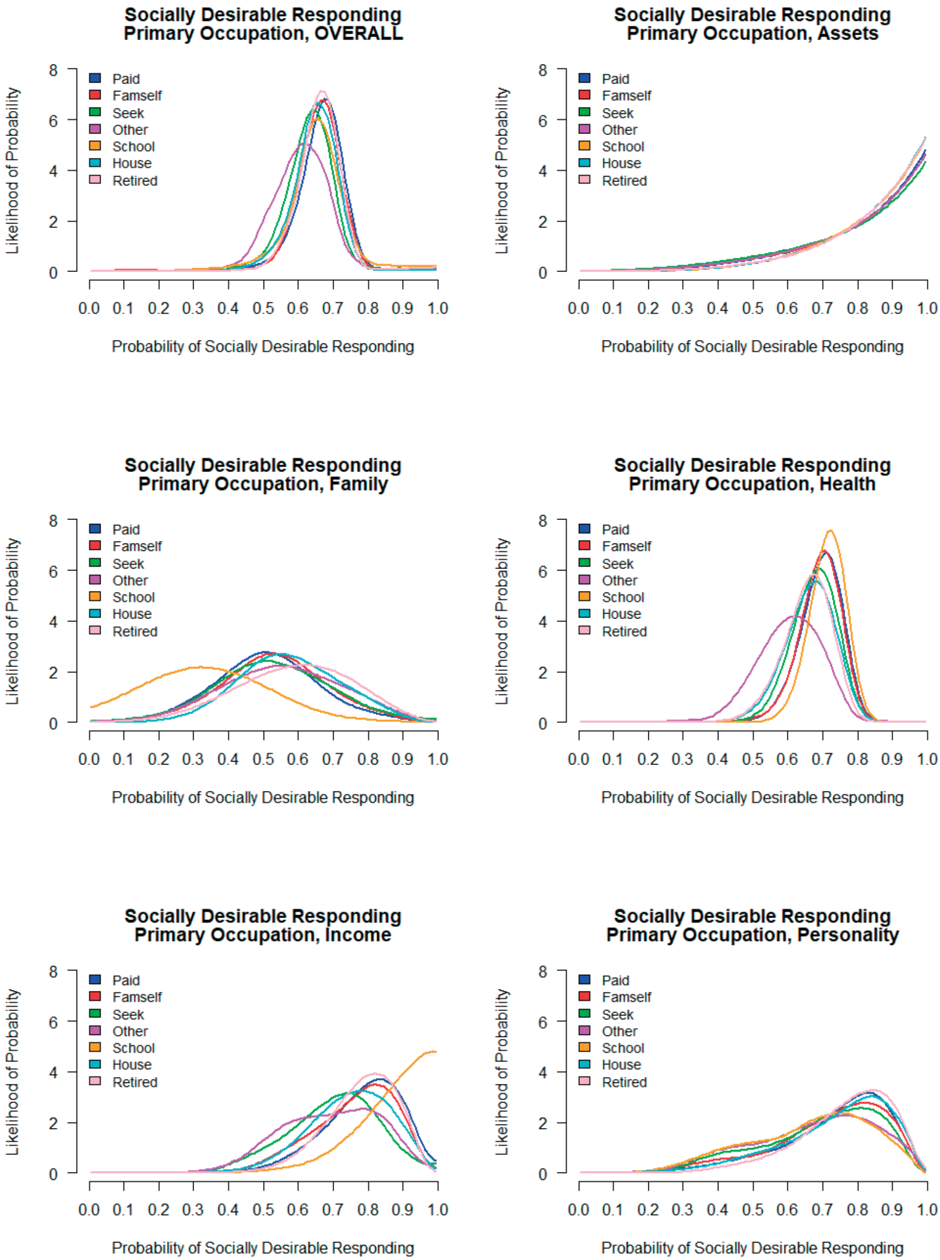


Figure K.5. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

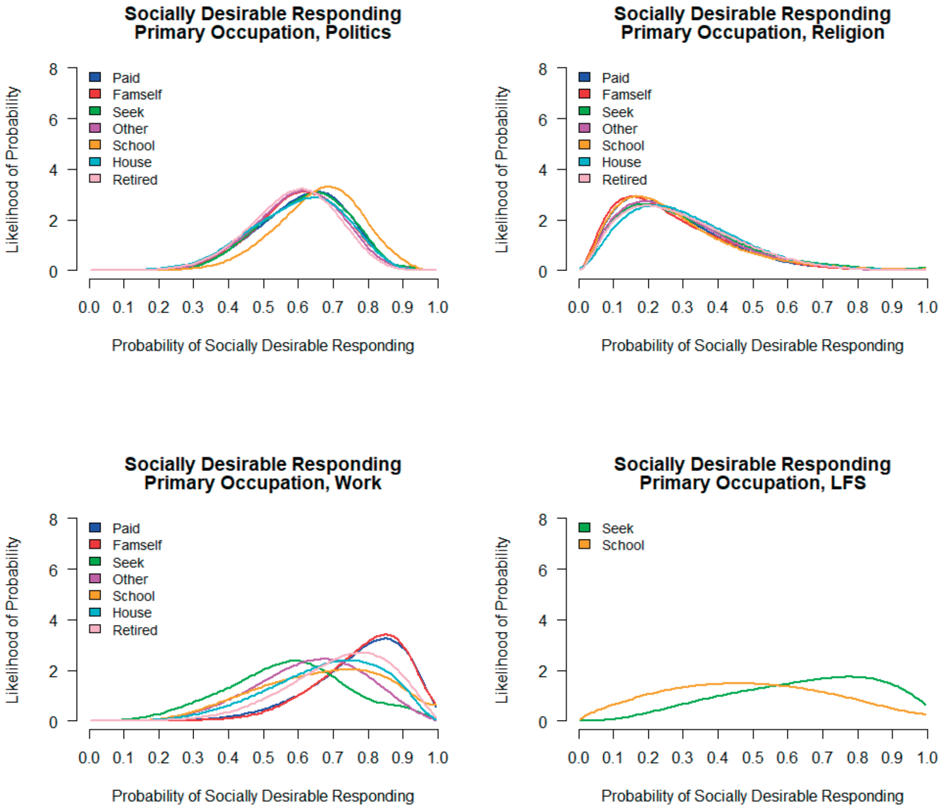


Figure K.5 (continued). Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: INCOME

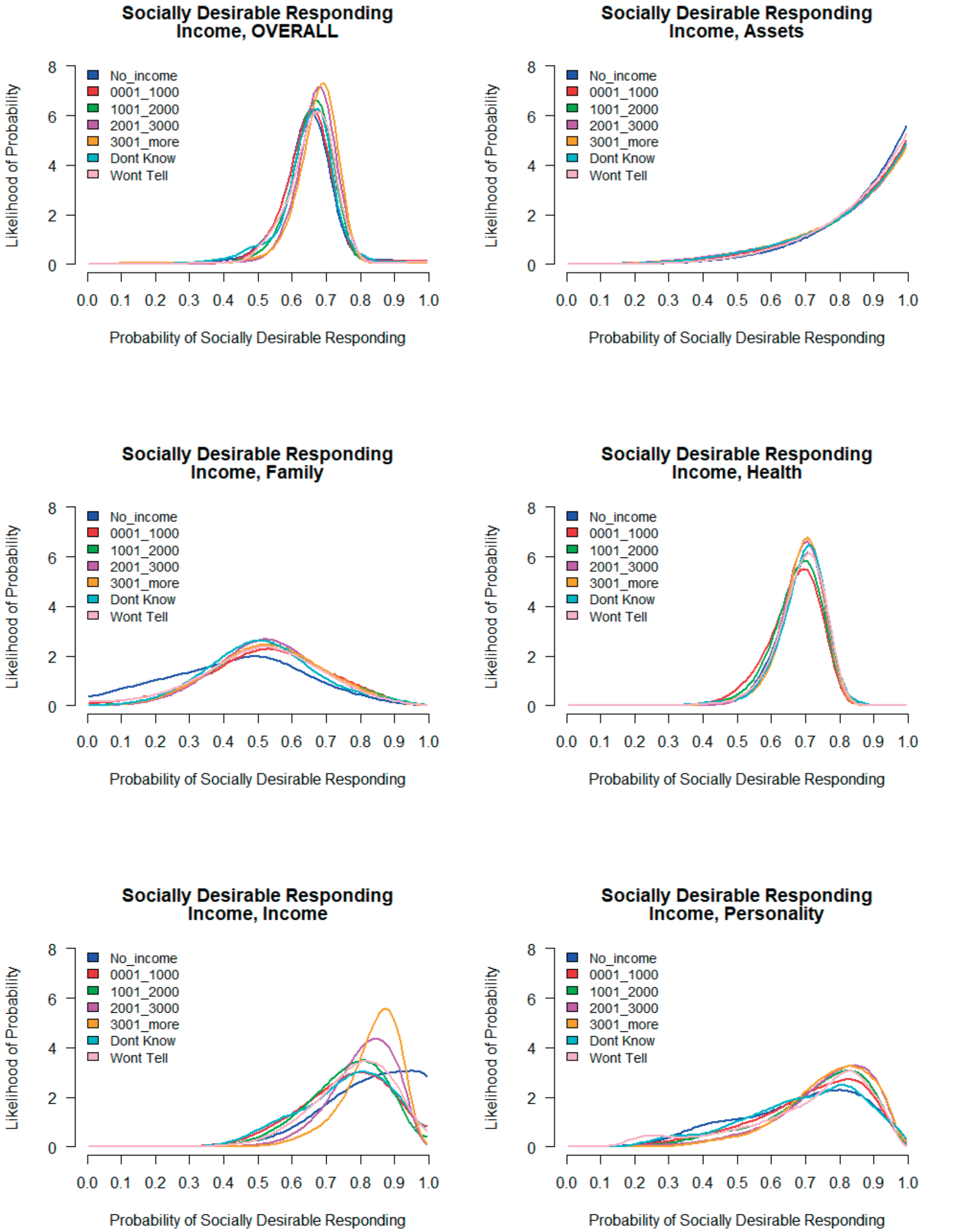
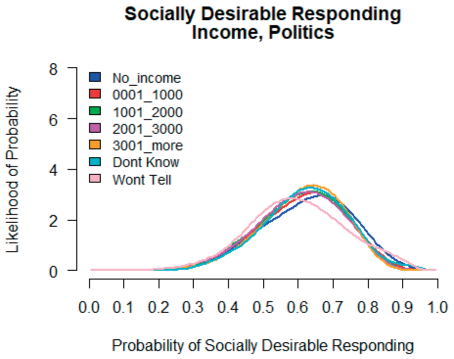
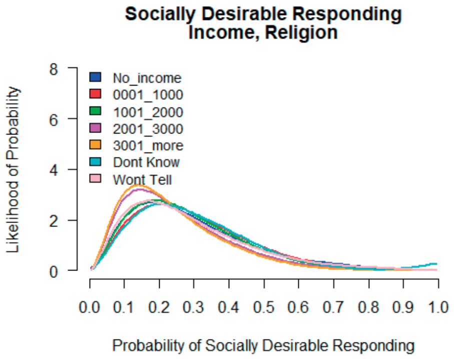


Figure K.6. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).





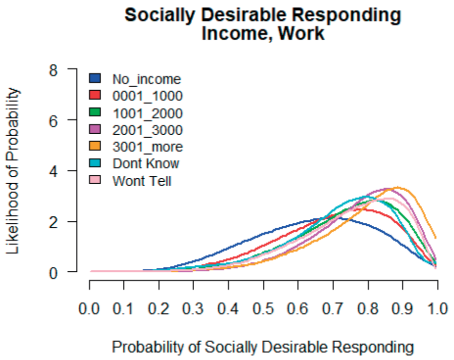


Figure K.6 (continued). Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: ORIGIN

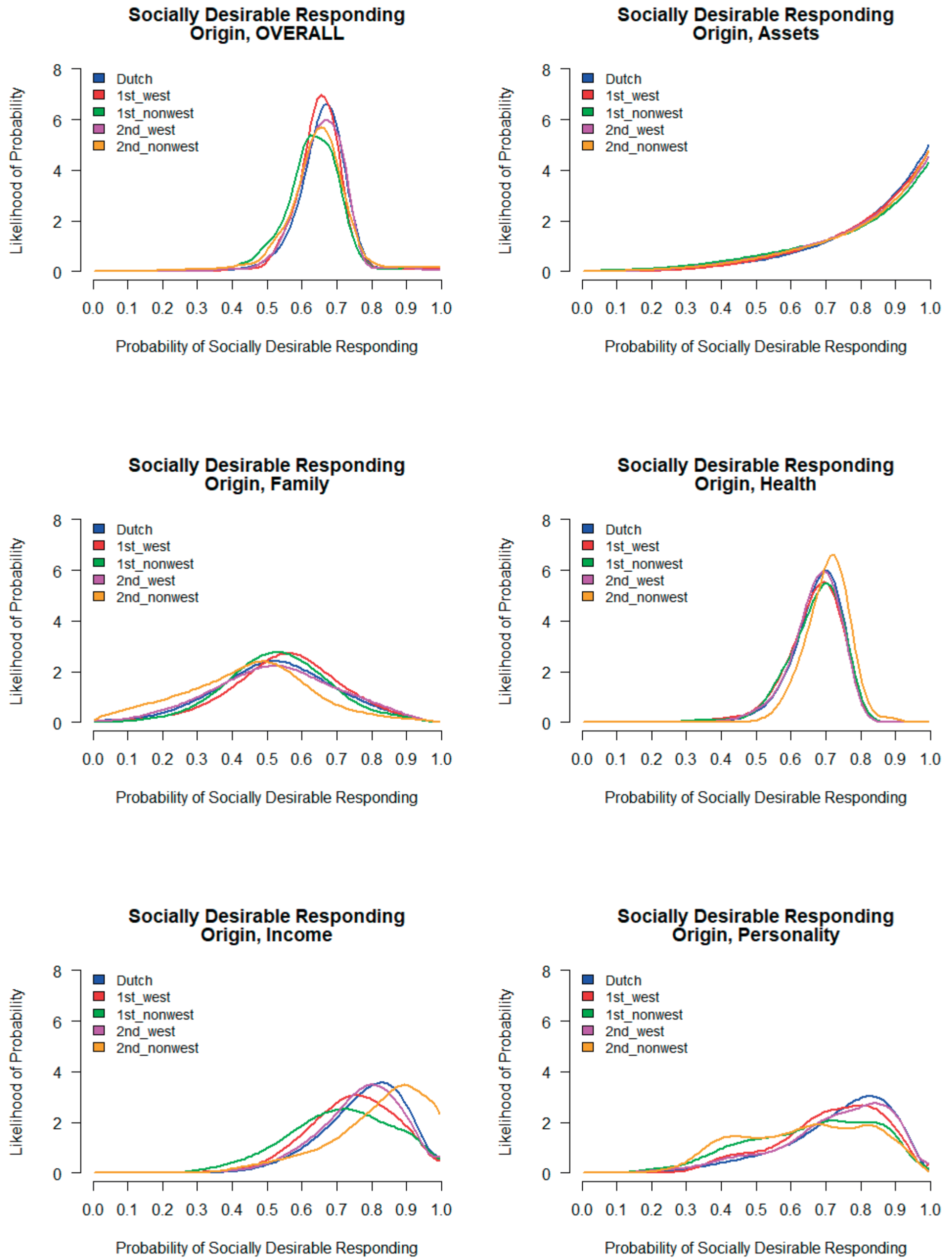
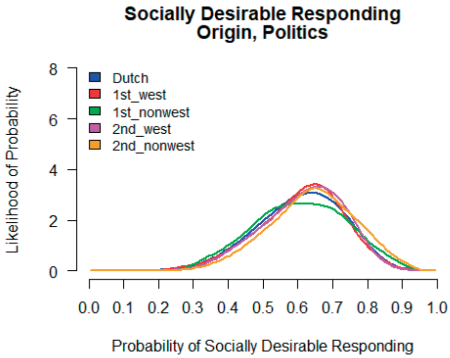
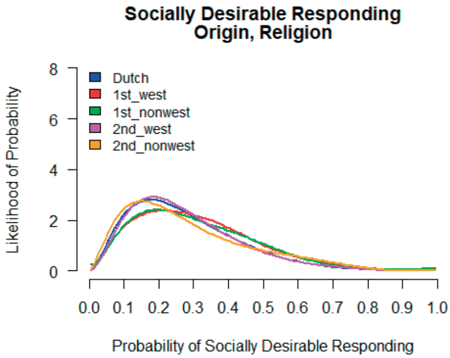


Figure K.7. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).





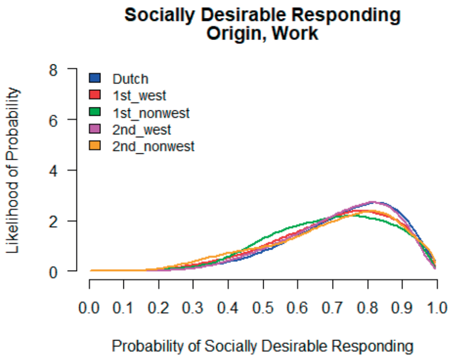


Figure K.7 (continued). Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

SOCIALLY DESIRABLE RESPONDING: PROVIDED A PC

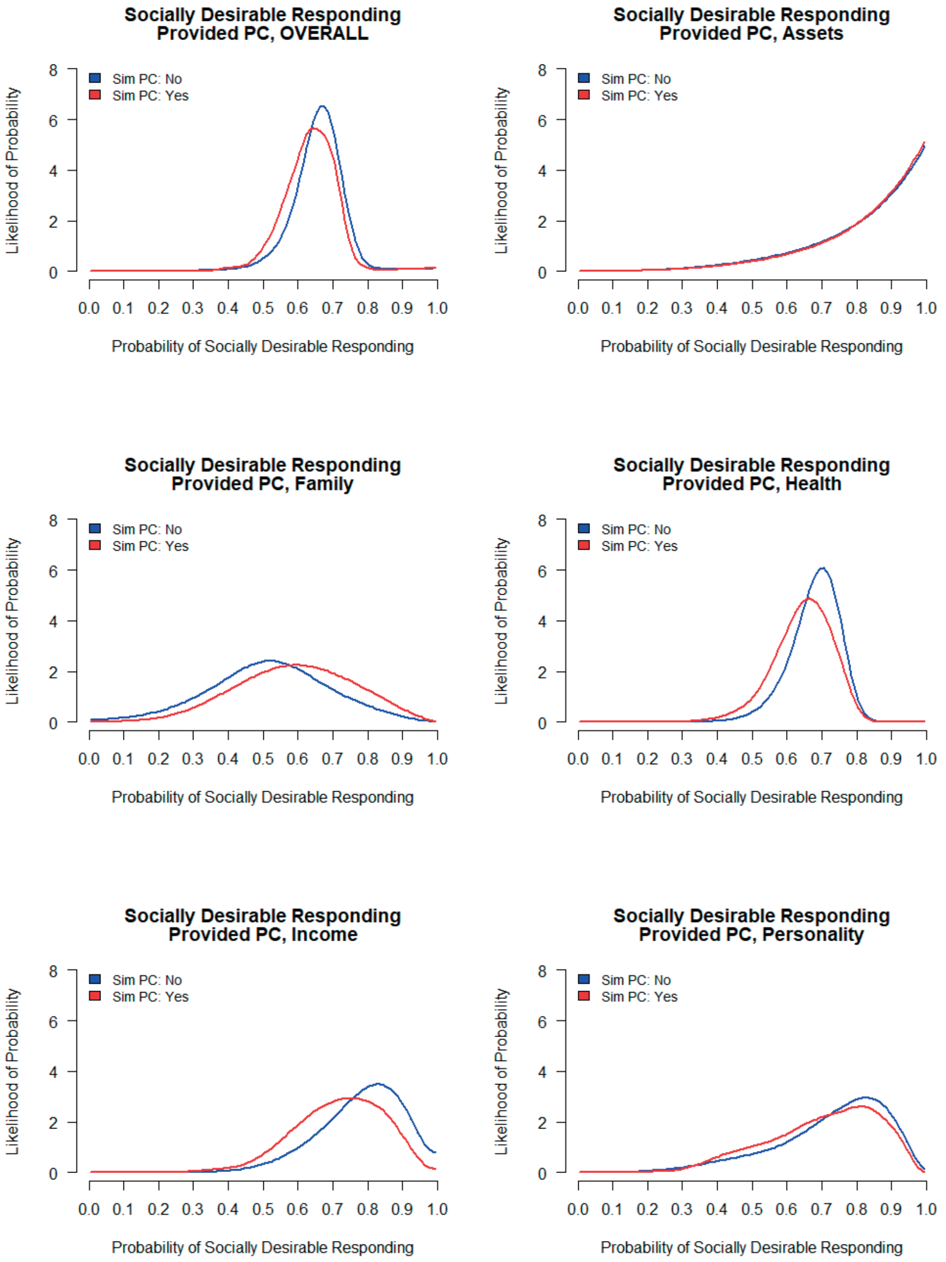


Figure K.8. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Socially Desirable Responding'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

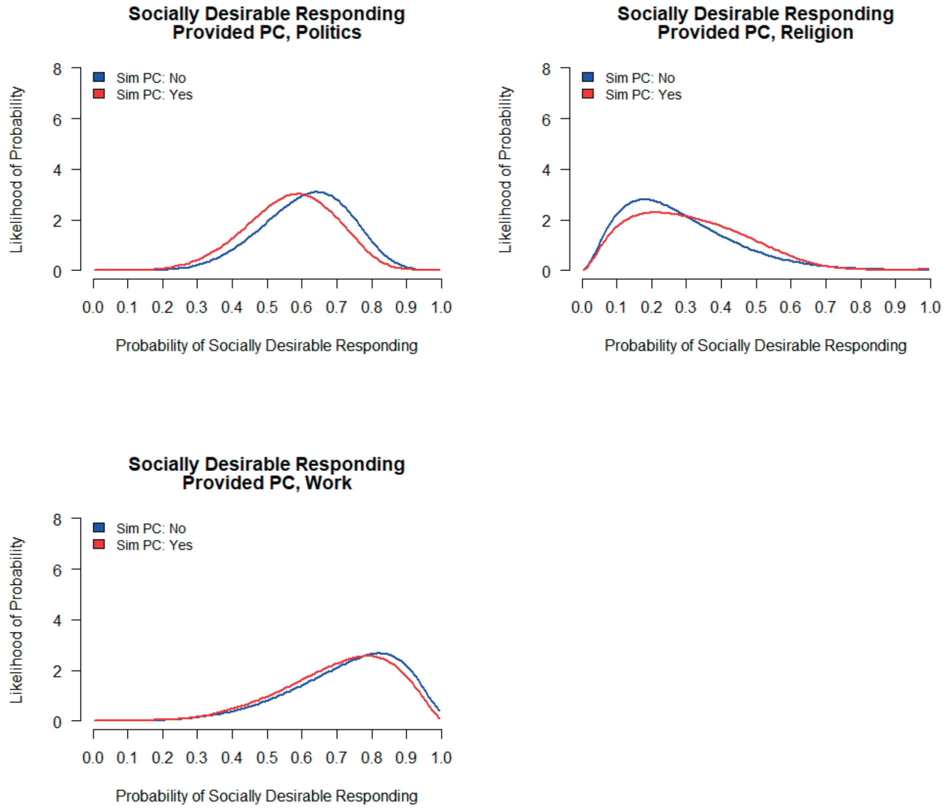


Figure K.8 (continued). Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **‘Socially Desirable Responding’** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: GENDER

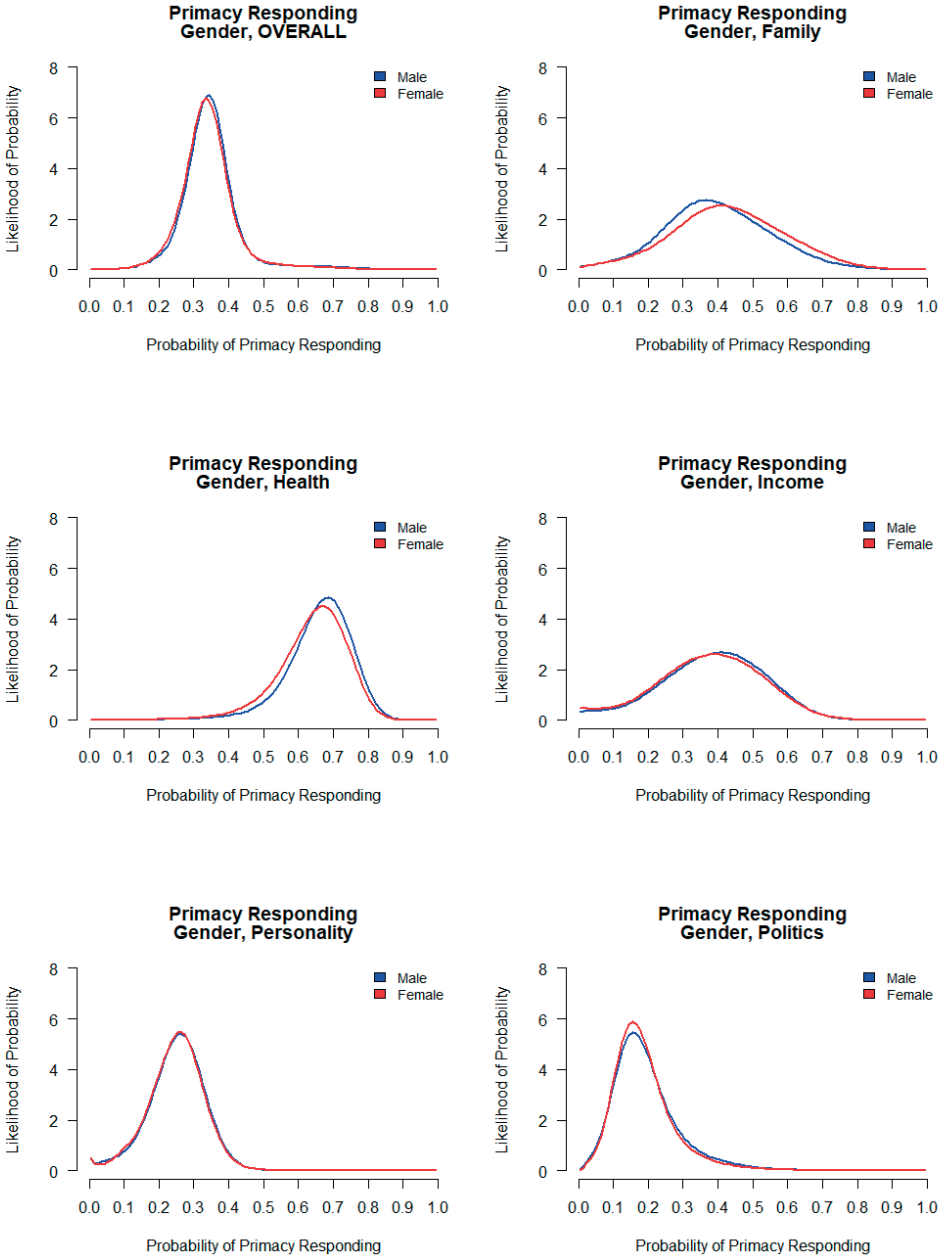
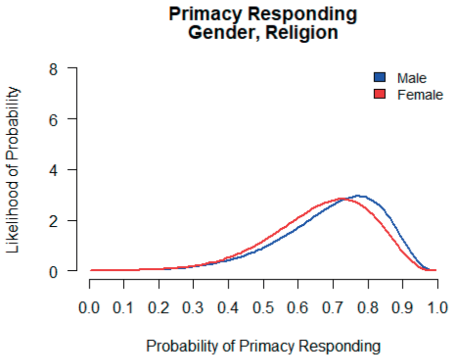


Figure K.9. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).



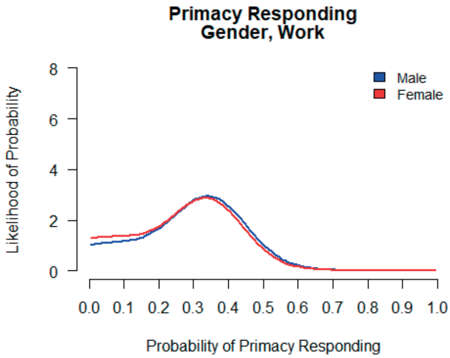


Figure K.9 (continued). Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: AGE

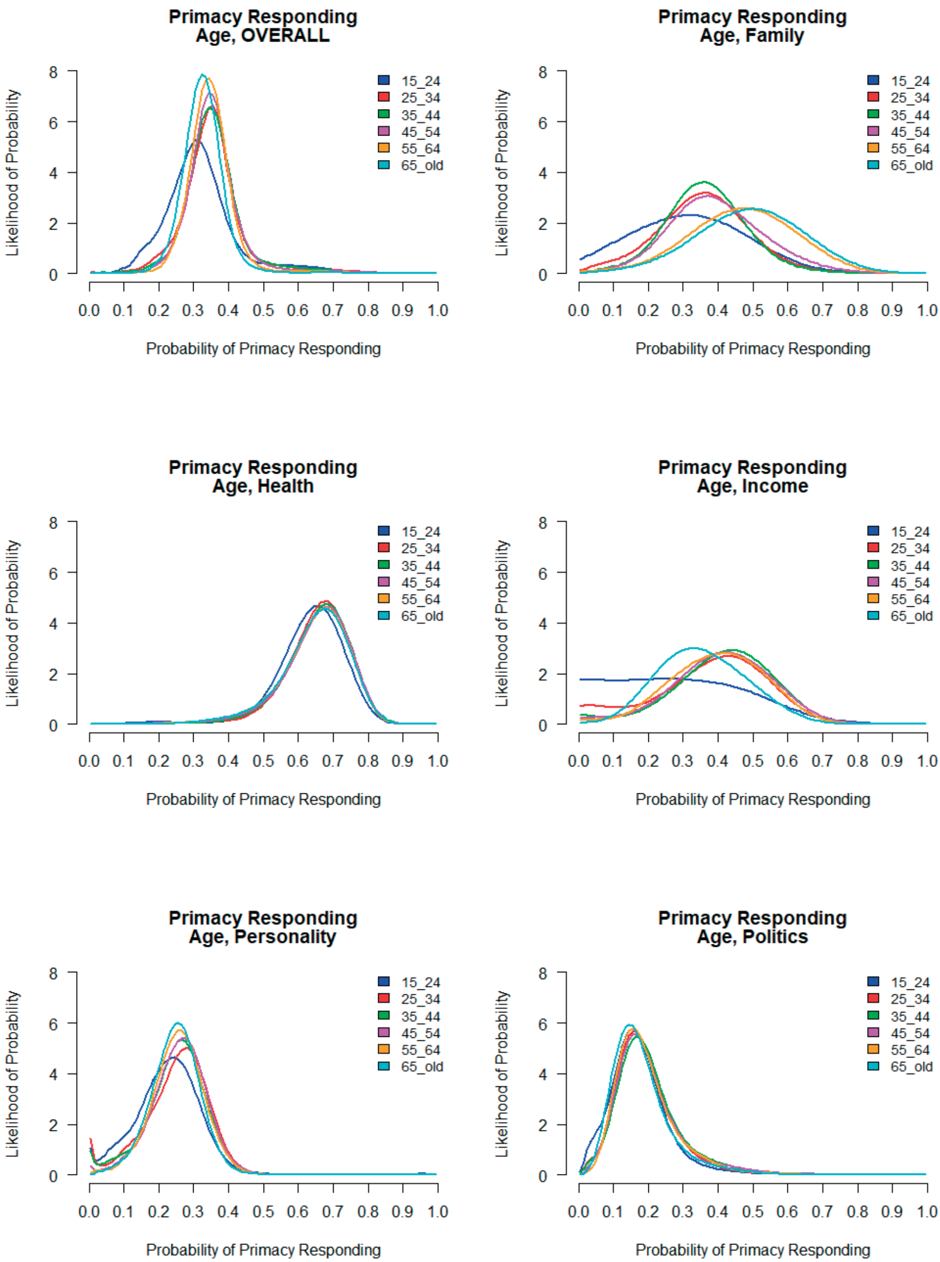


Figure K.10. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

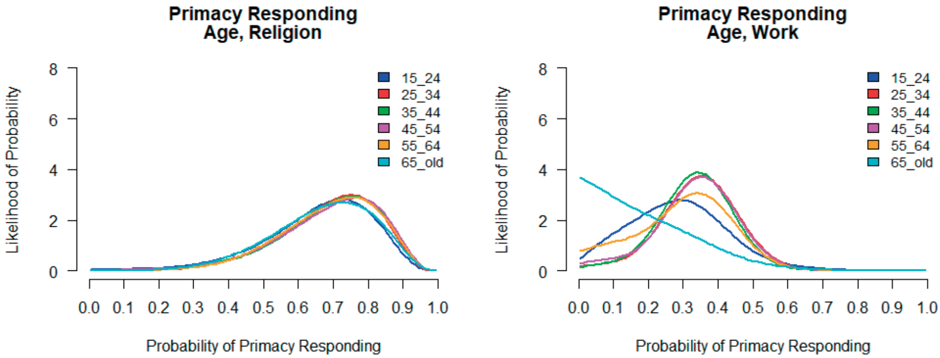


Figure K.10 (continued). Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: EDUCATION

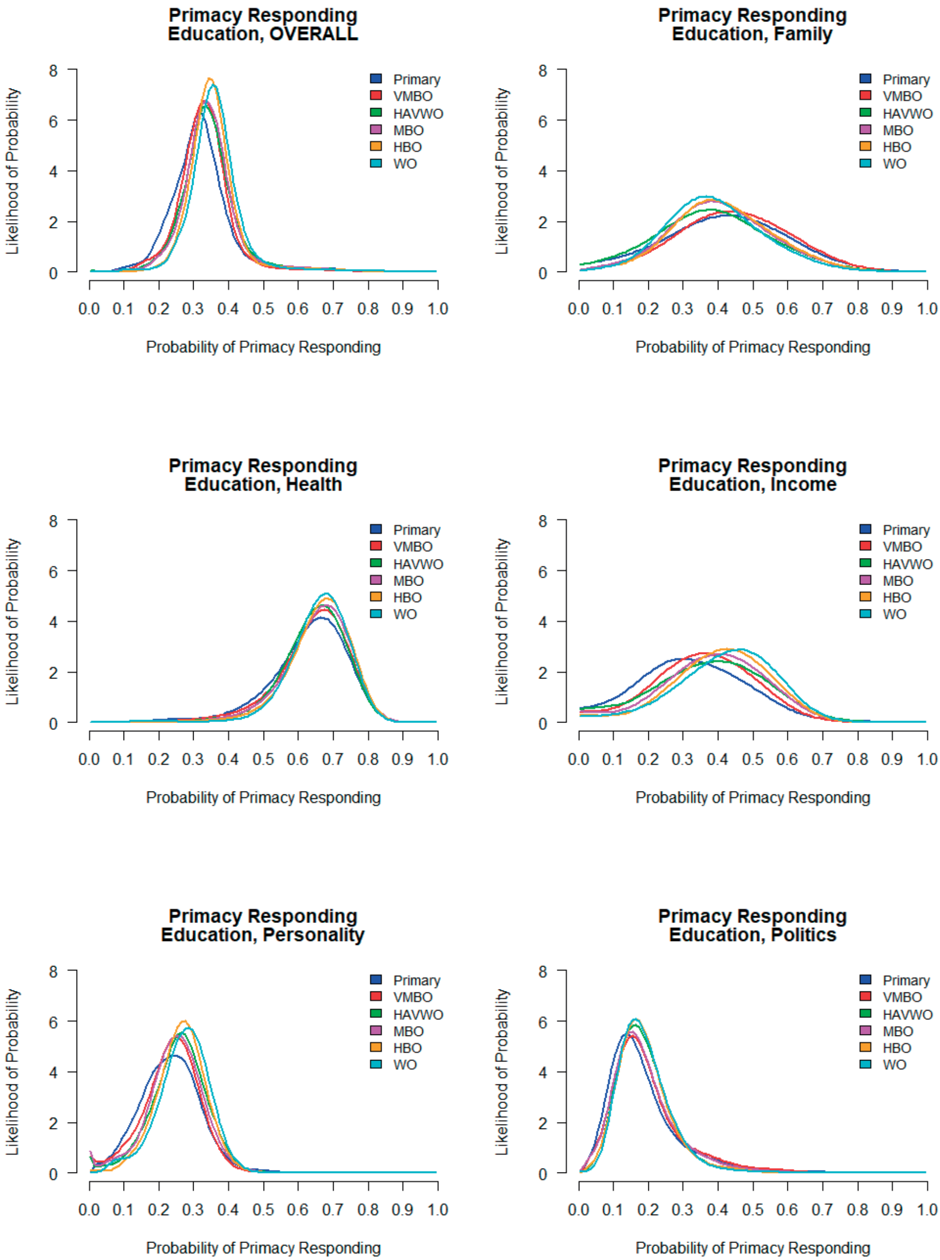


Figure K.11. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

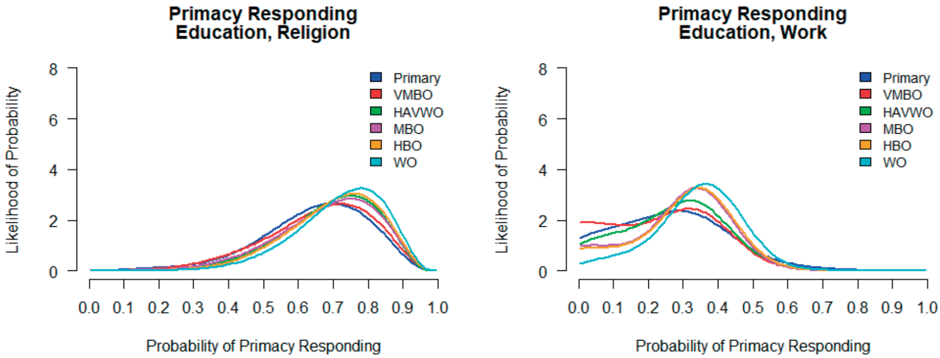


Figure K.11 (continued). Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: DOMESTIC SITUATION

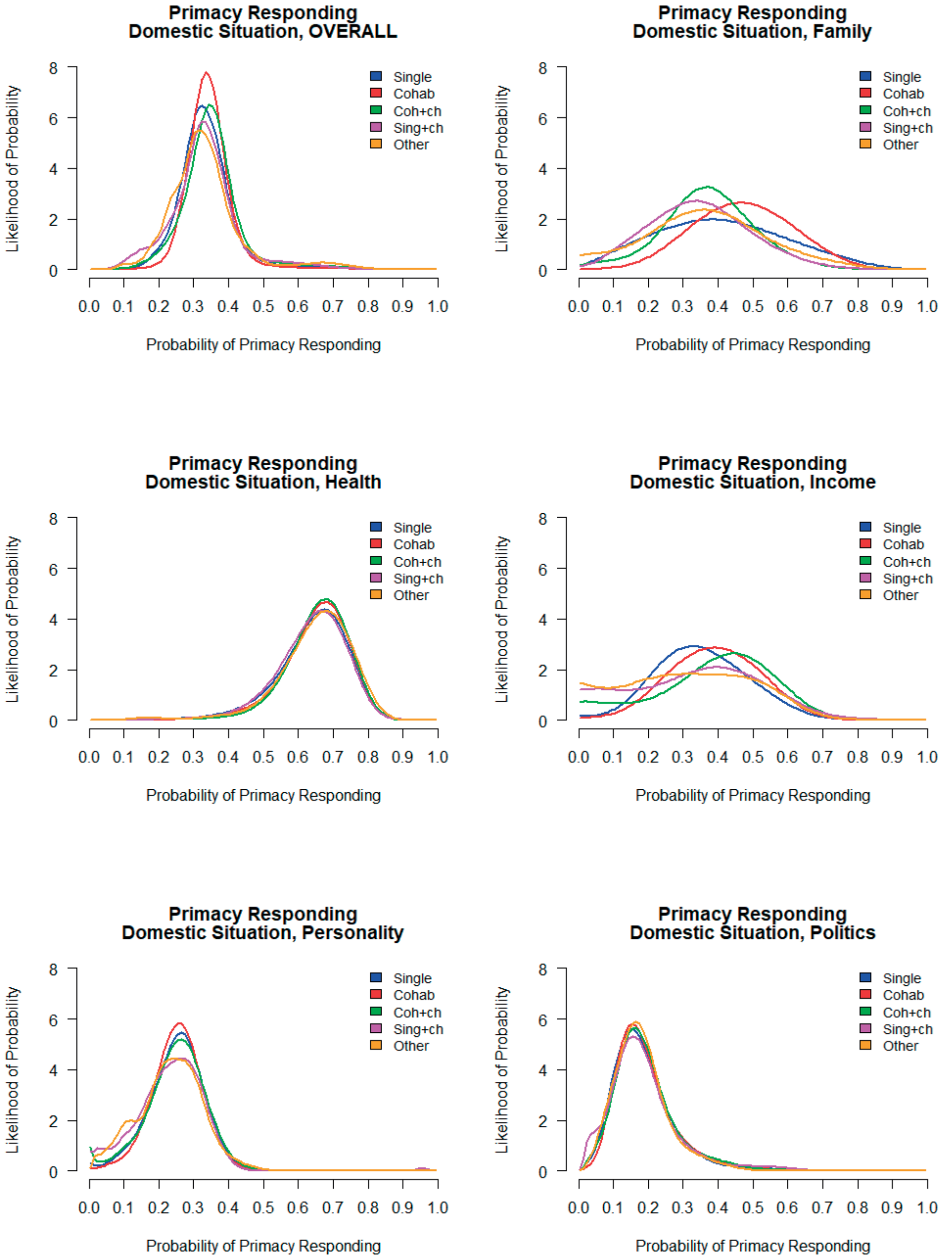


Figure K.12. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

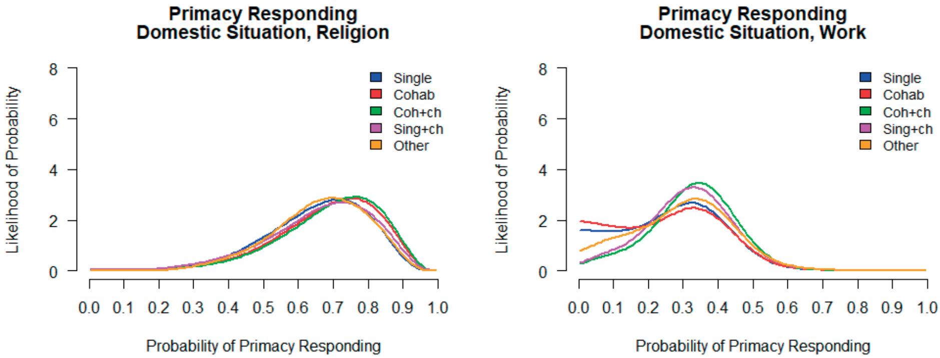


Figure K.12 (continued). Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: PRIMARY OCCUPATION

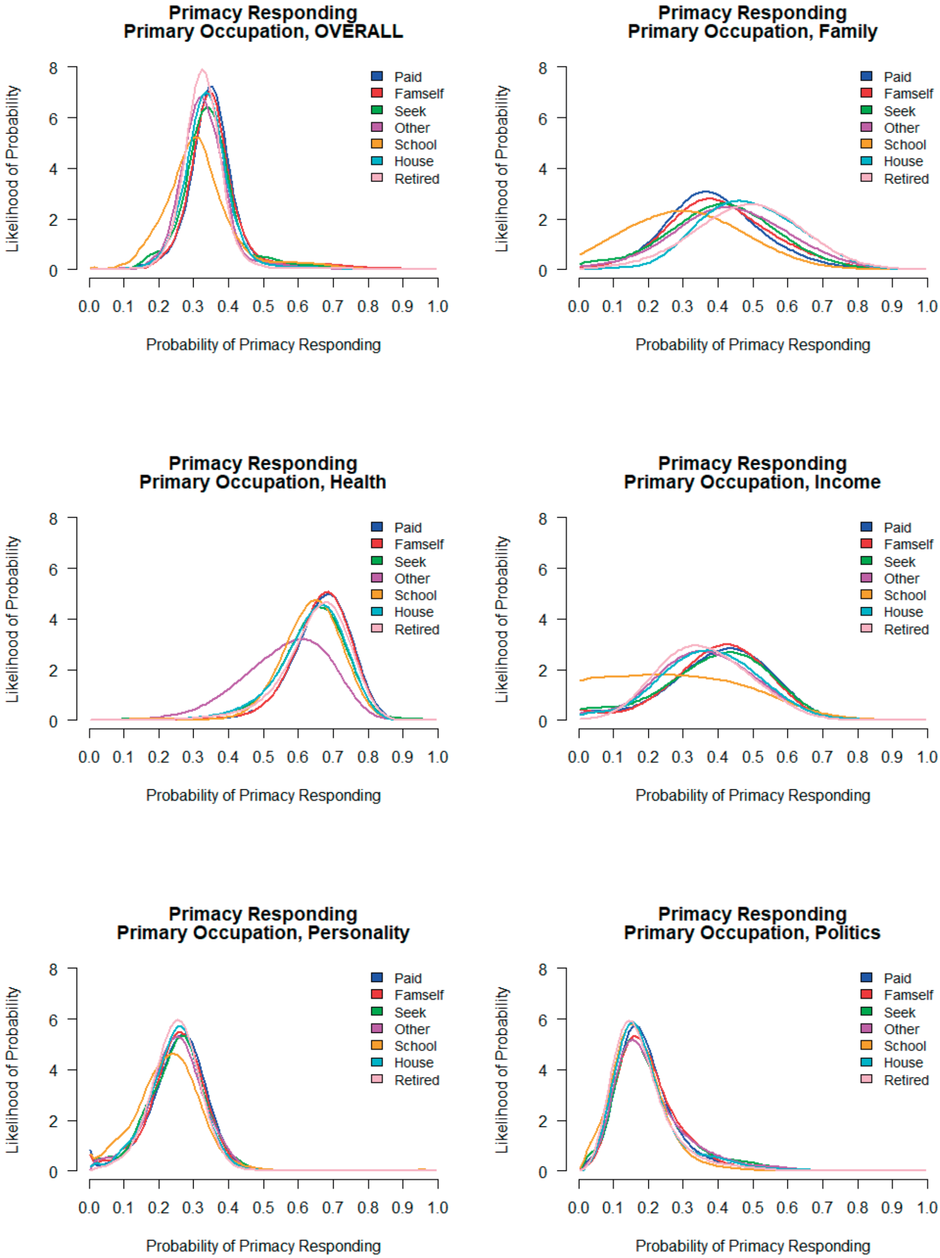


Figure K.13. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

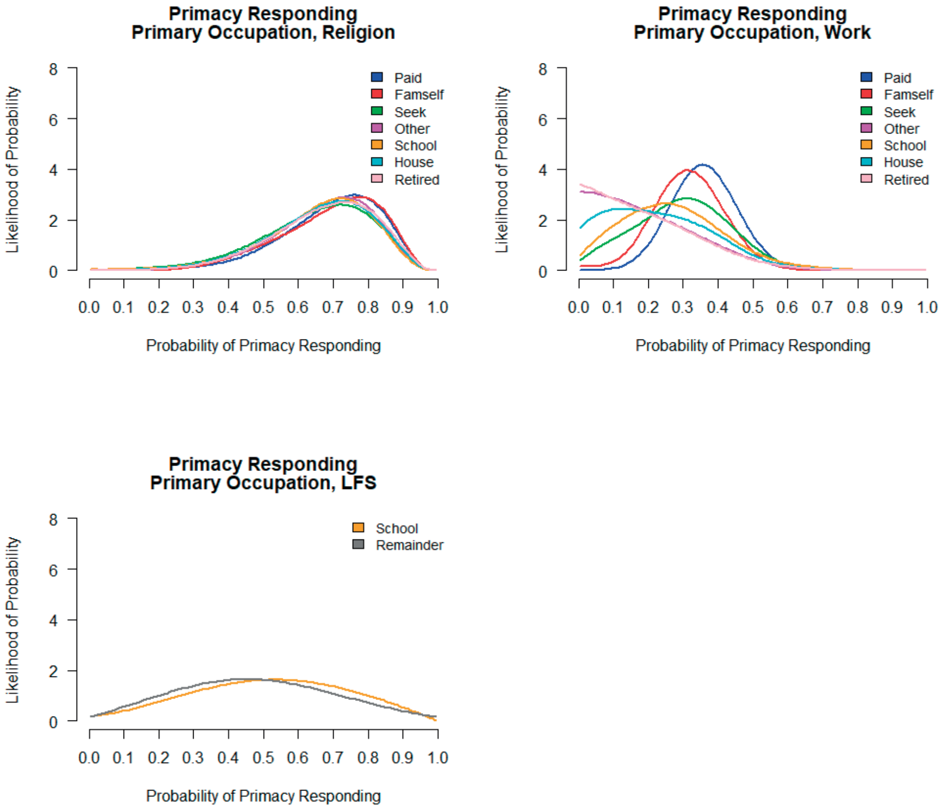


Figure K.13 (continued). Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: INCOME

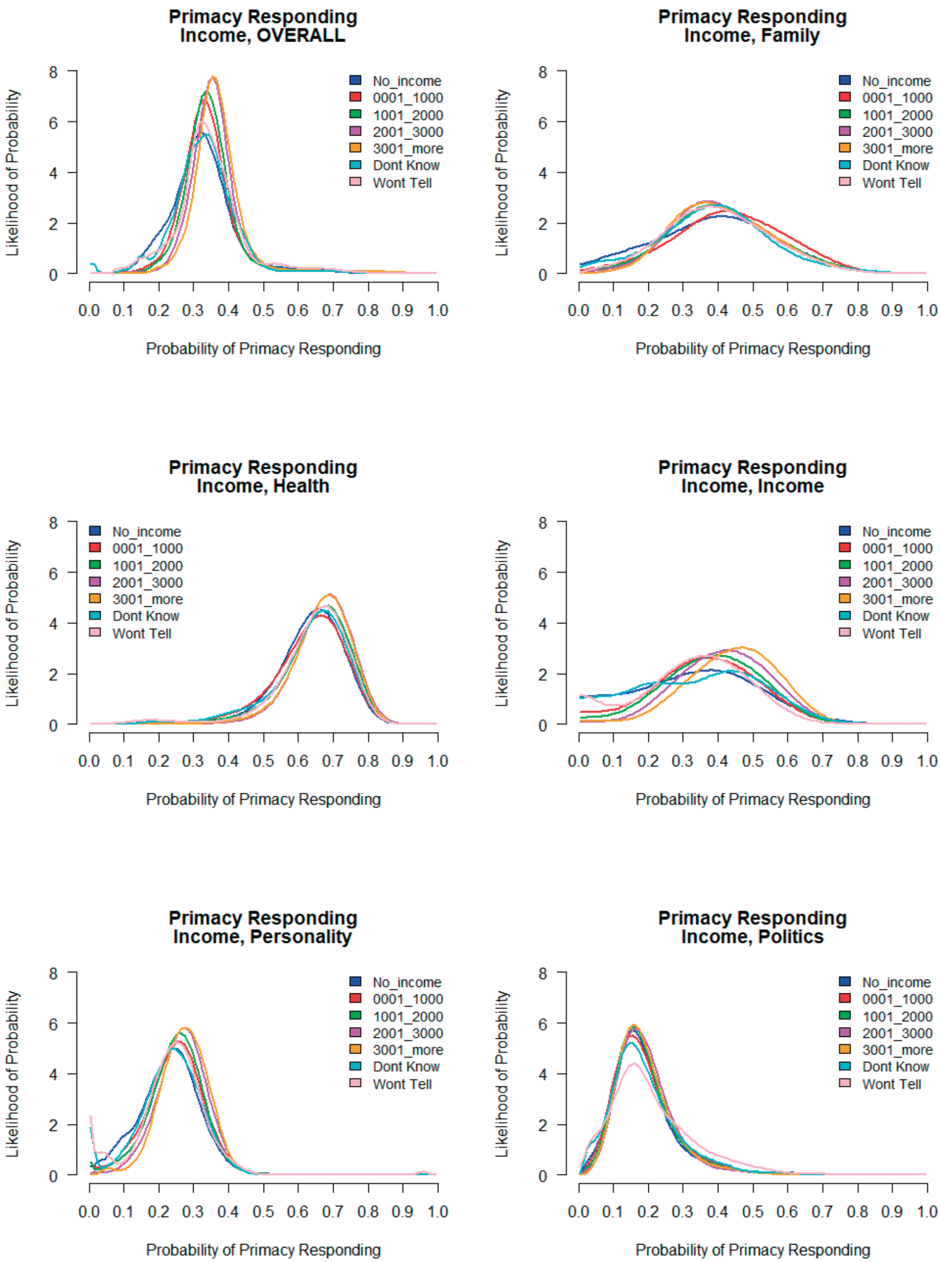


Figure K.14. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

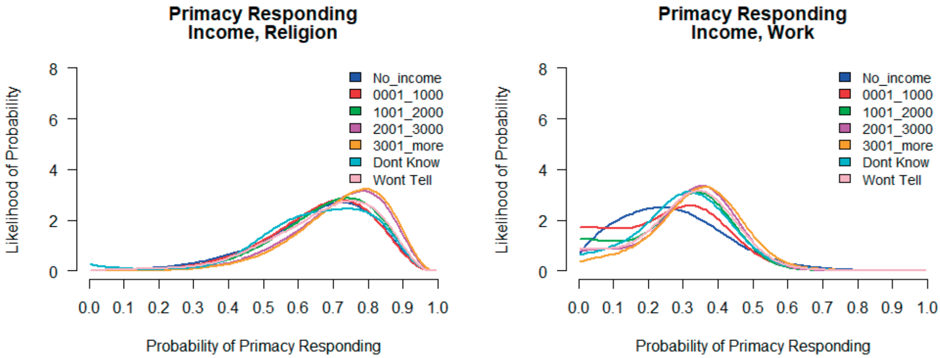


Figure K.14 (continued). Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: ORIGIN

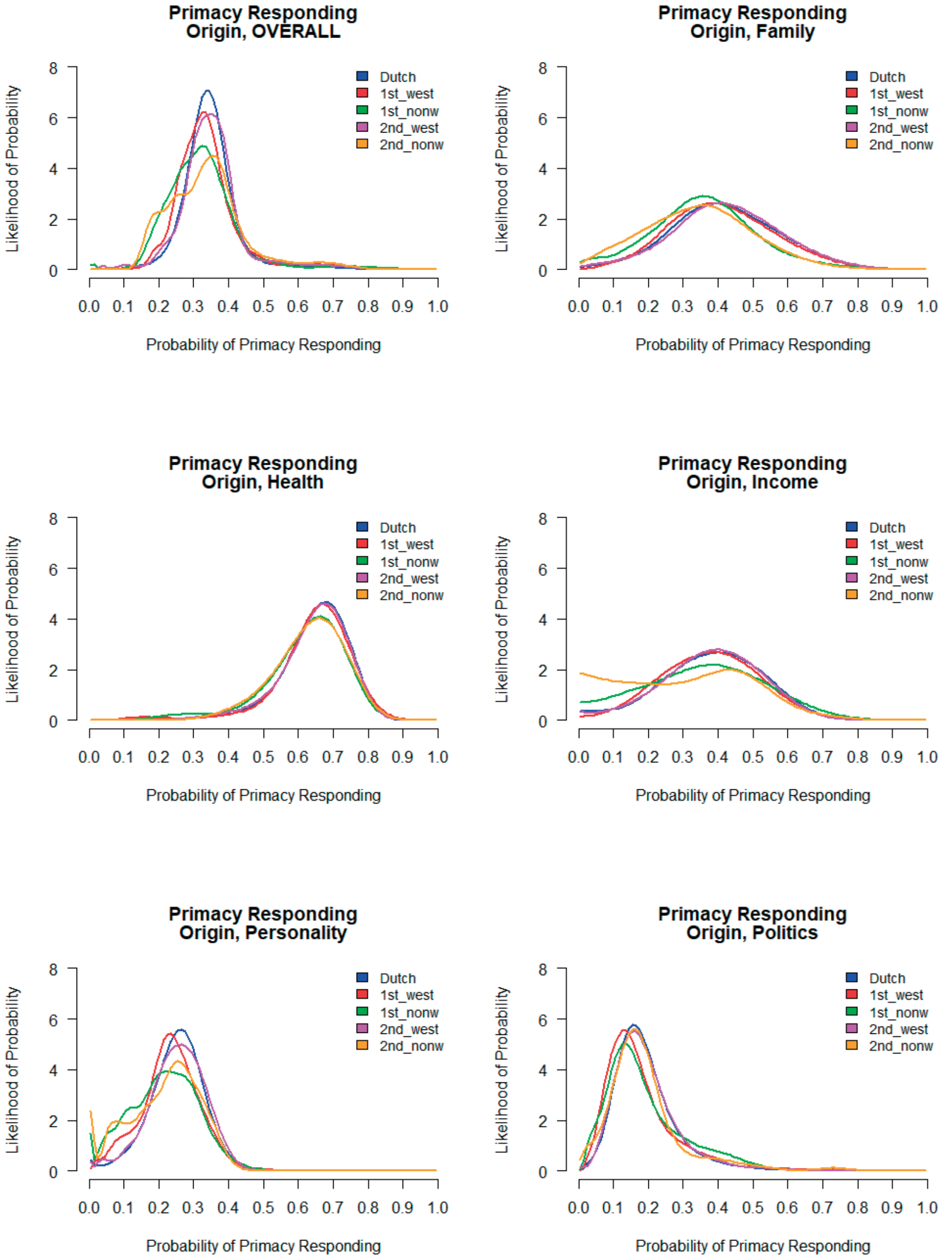


Figure K.15. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

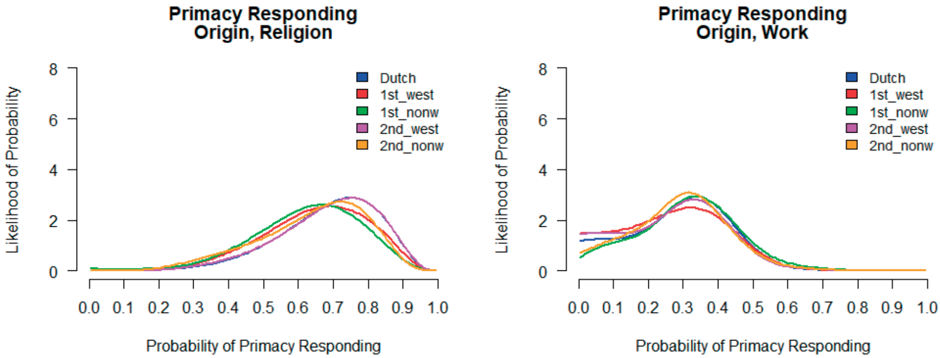


Figure K.15 (continued). Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

PRIMACY RESPONDING: PROVIDED A PC

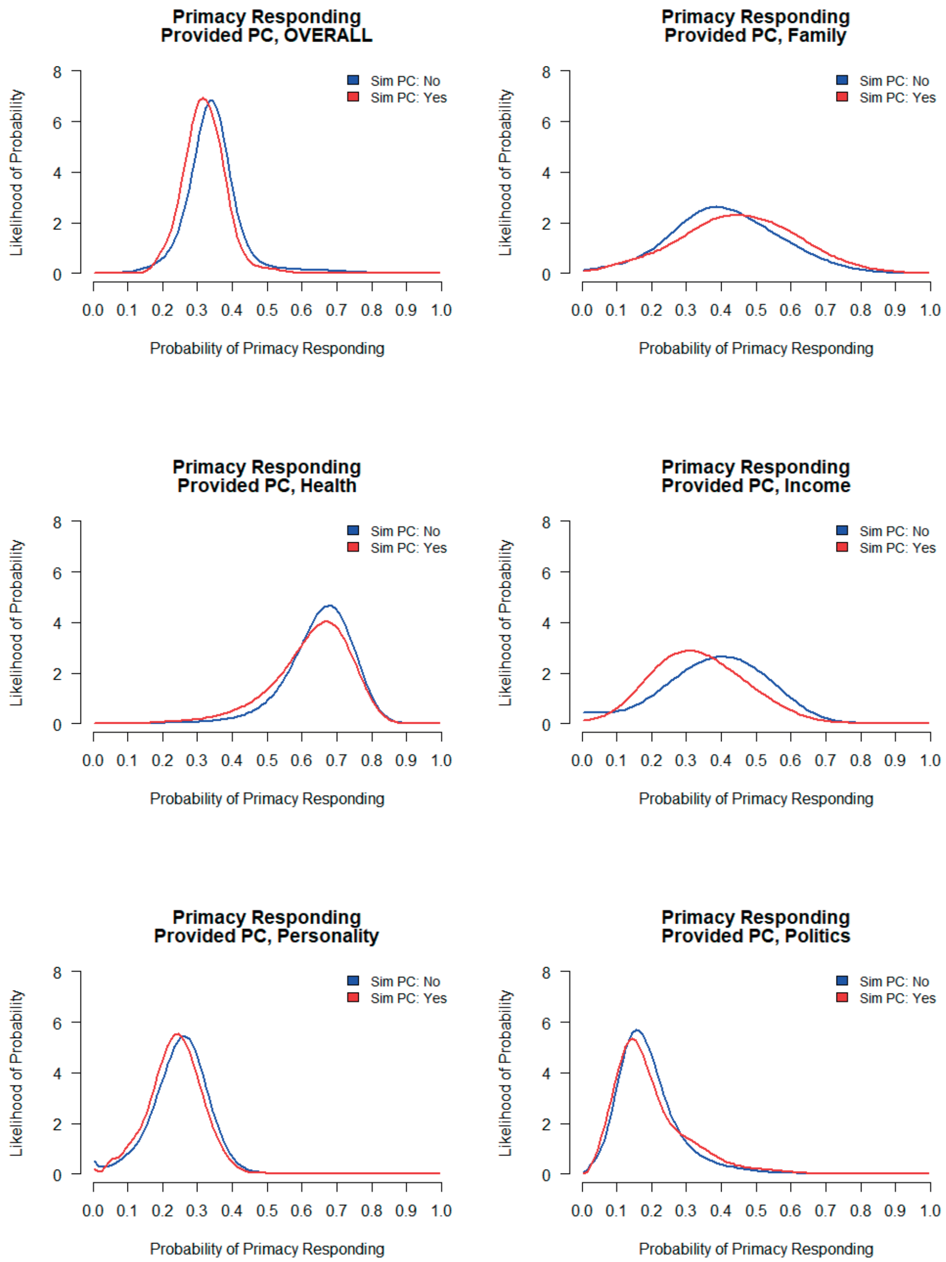


Figure K.16. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

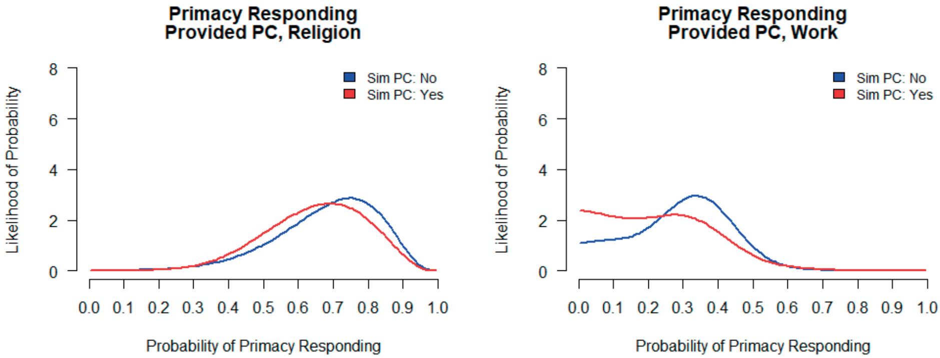


Figure K.16 (continued). Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Primacy Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO), Labour Force Survey (LF).

STRAIGHTLINING: GENDER

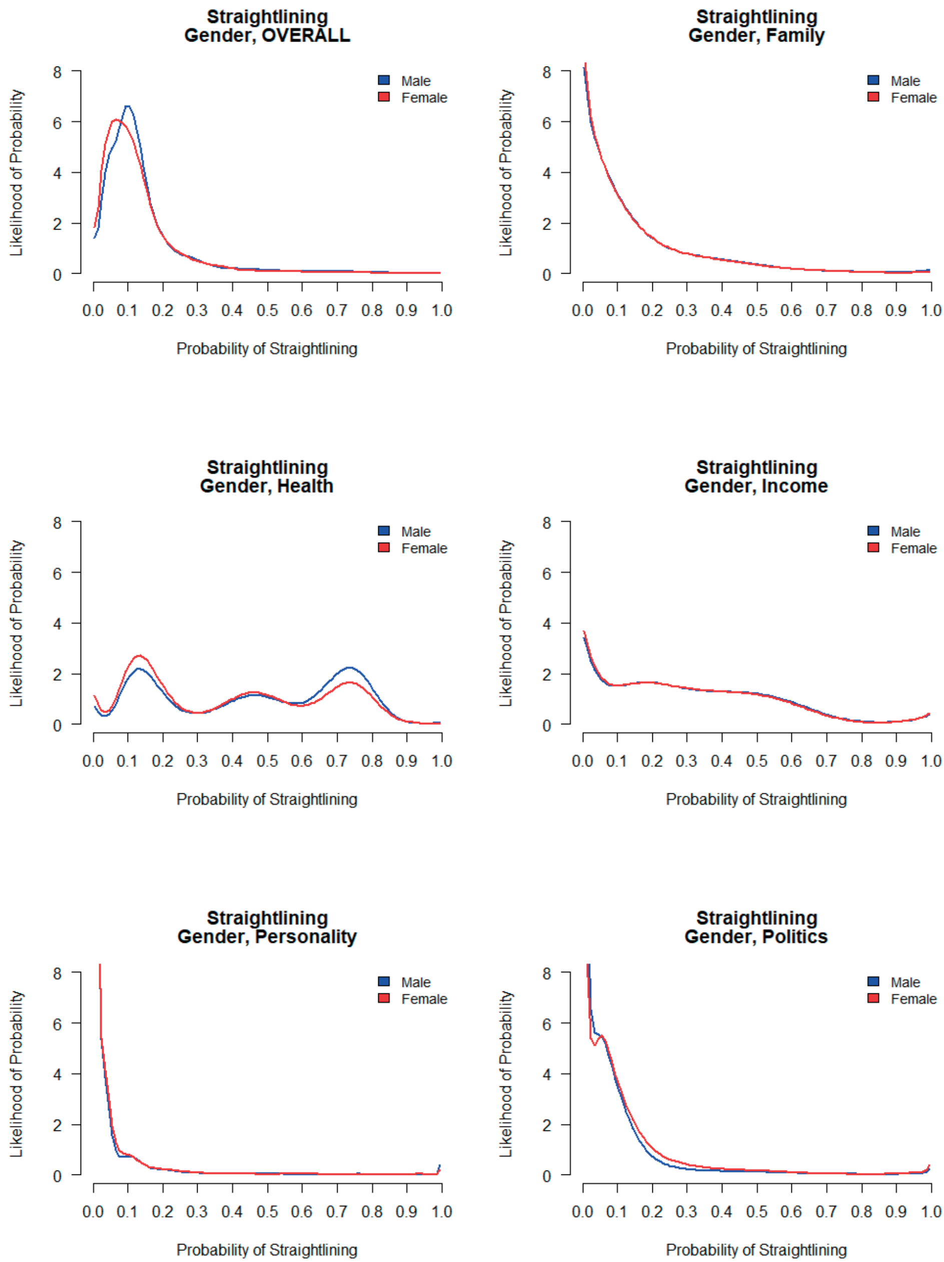
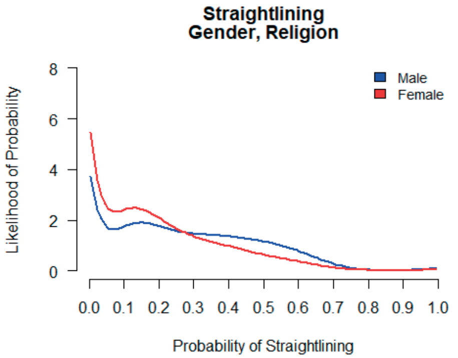


Figure K.17. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).



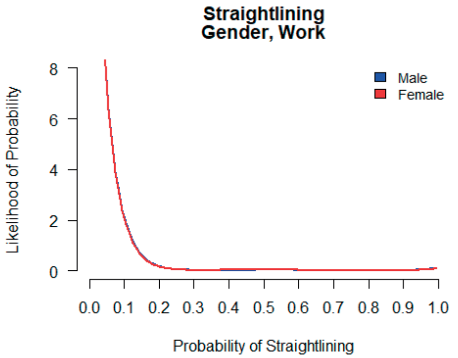


Figure K.17 (continued). Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: AGE

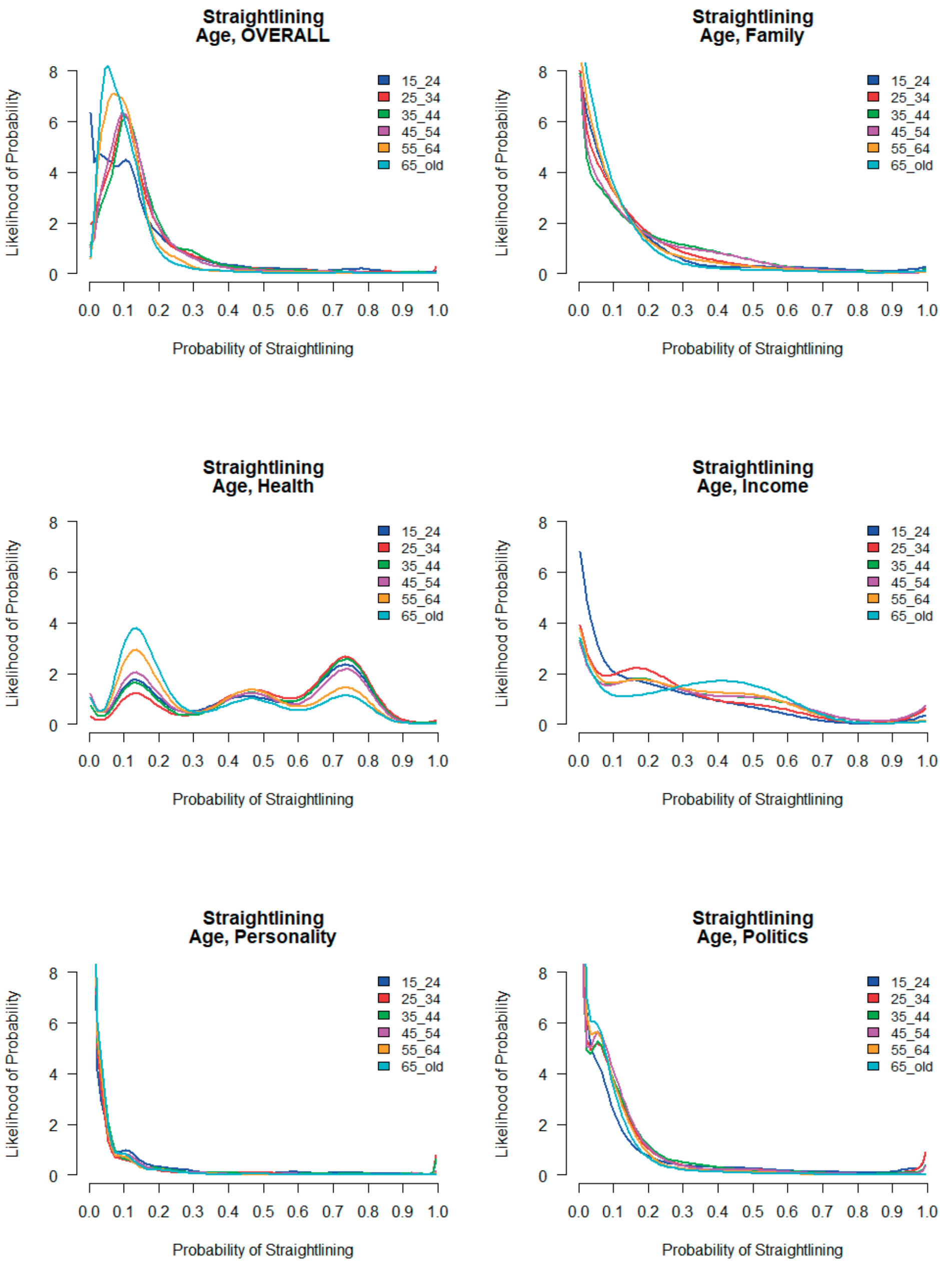


Figure K.18. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

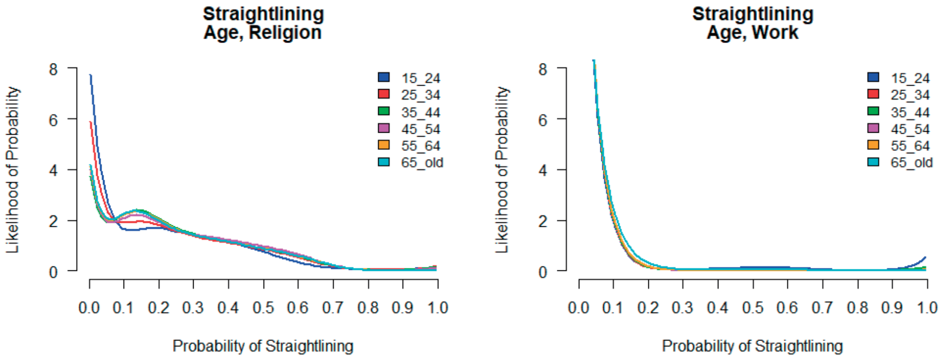


Figure K.18 (continued). Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: EDUCATION

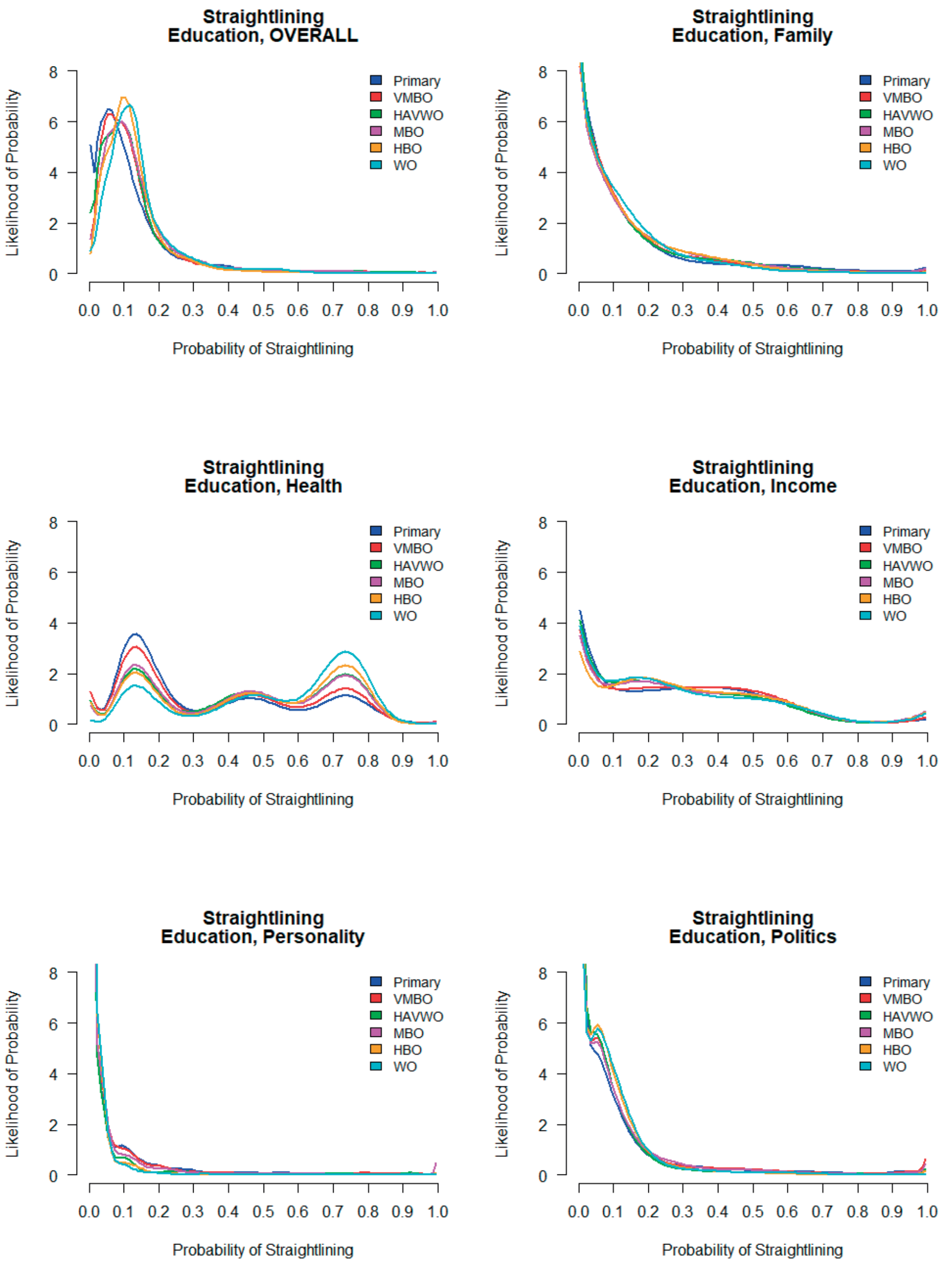


Figure K.19. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

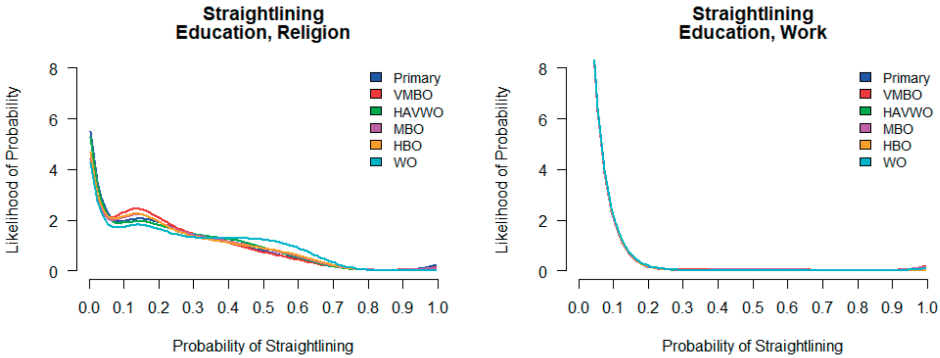


Figure K.19 (continued). Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: DOMESTIC SITUATION

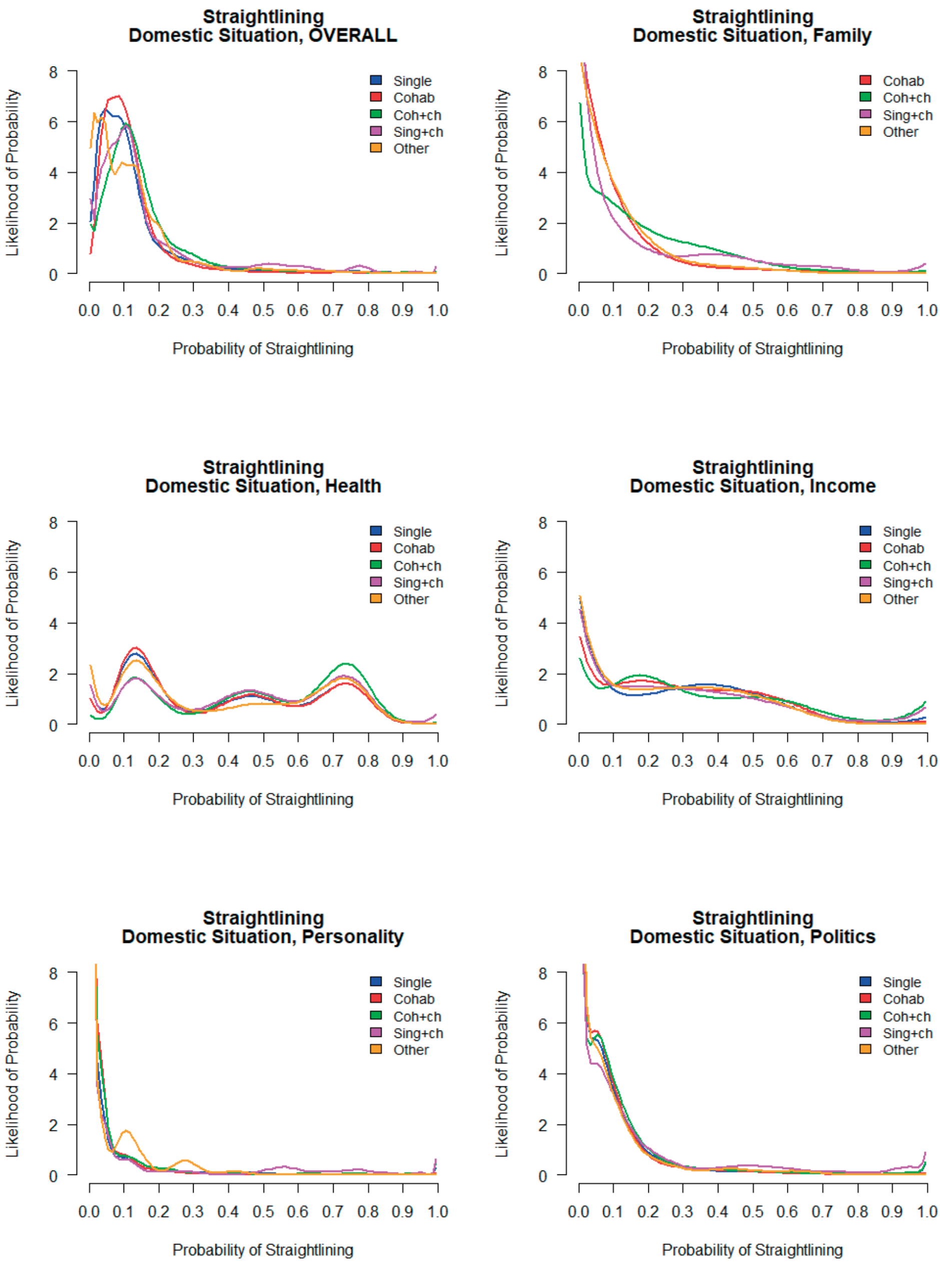


Figure K.20. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **‘Straightlining’** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

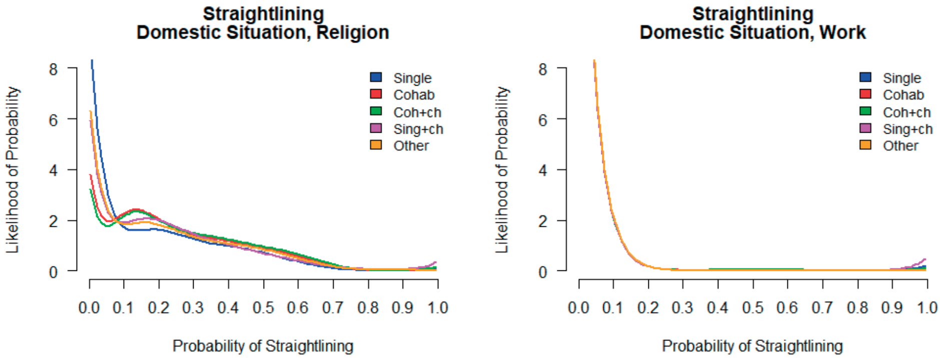


Figure K.20 (continued). Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: PRIMARY OCCUPATION

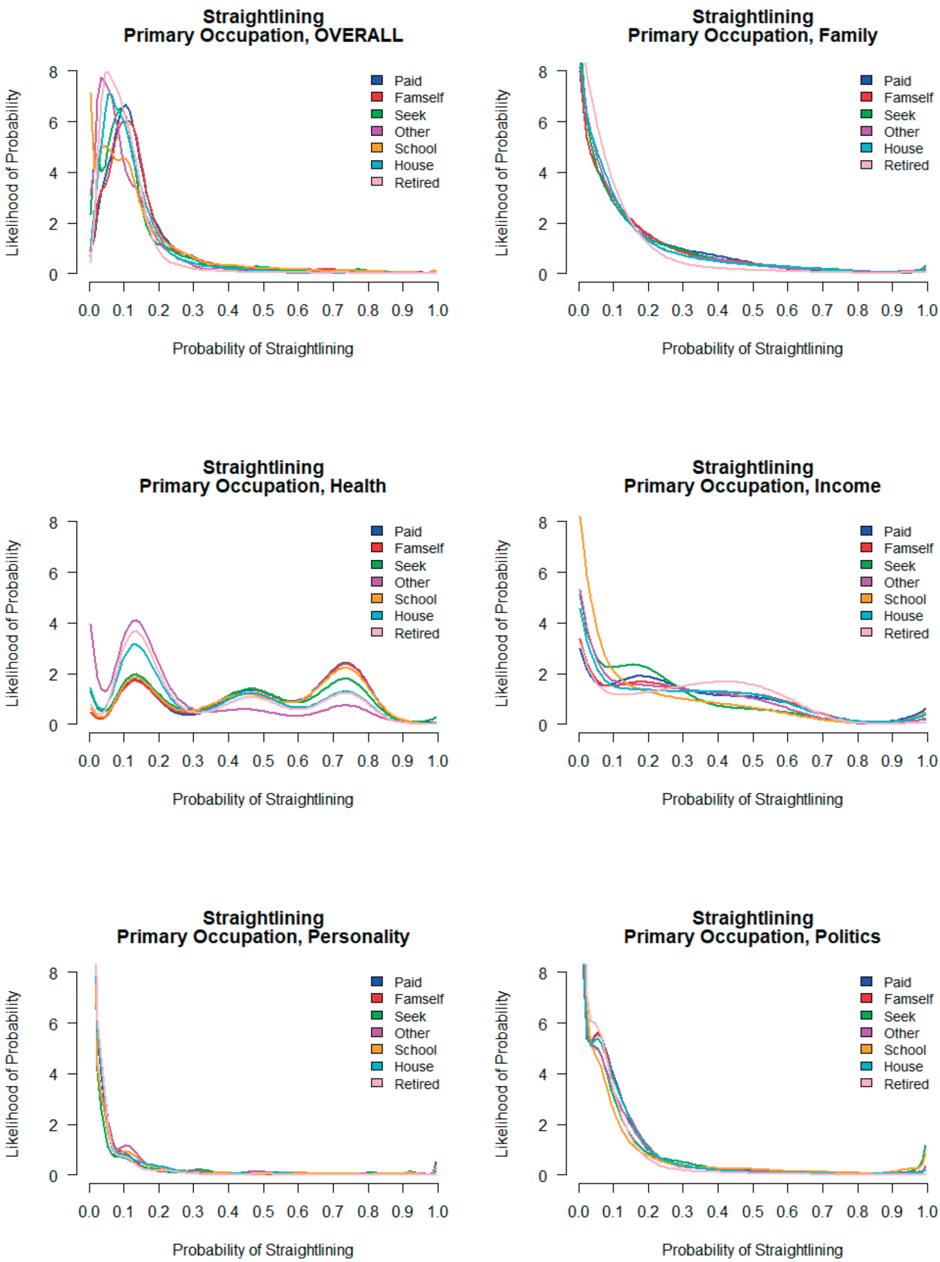


Figure K.21. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

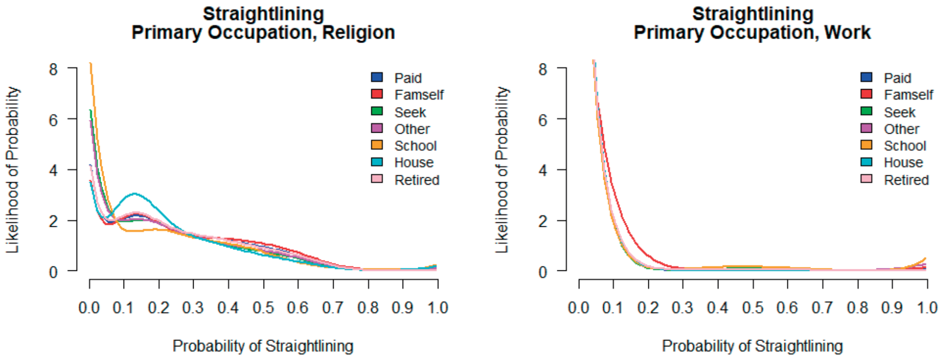


Figure K.21 (continued). Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: INCOME

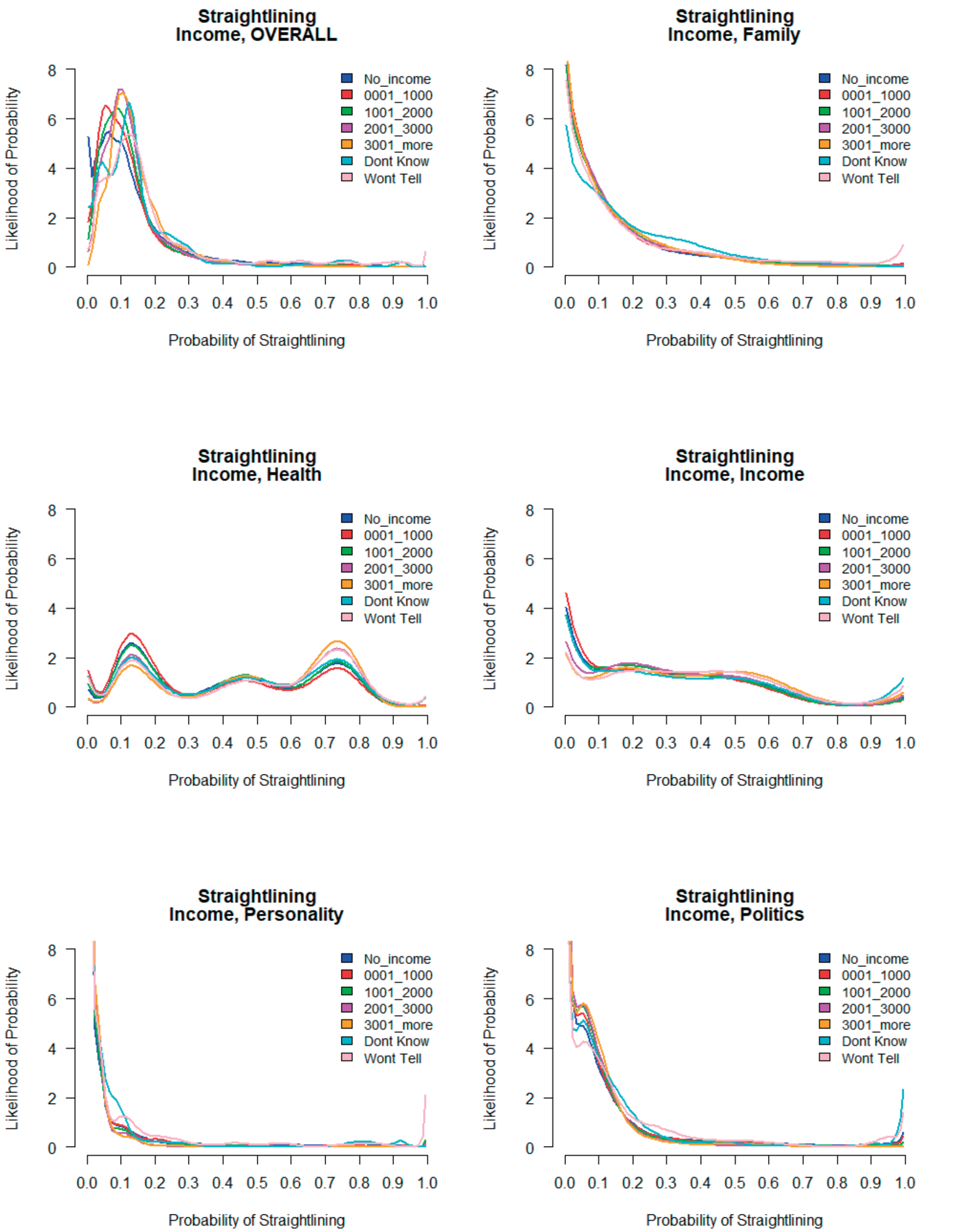
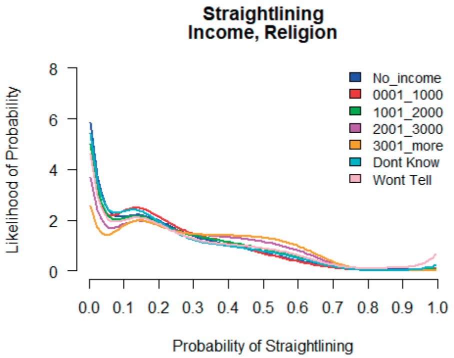


Figure K.22. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).



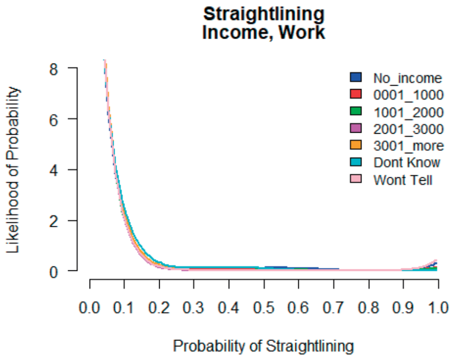


Figure K.22 (continued). Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: ORIGIN

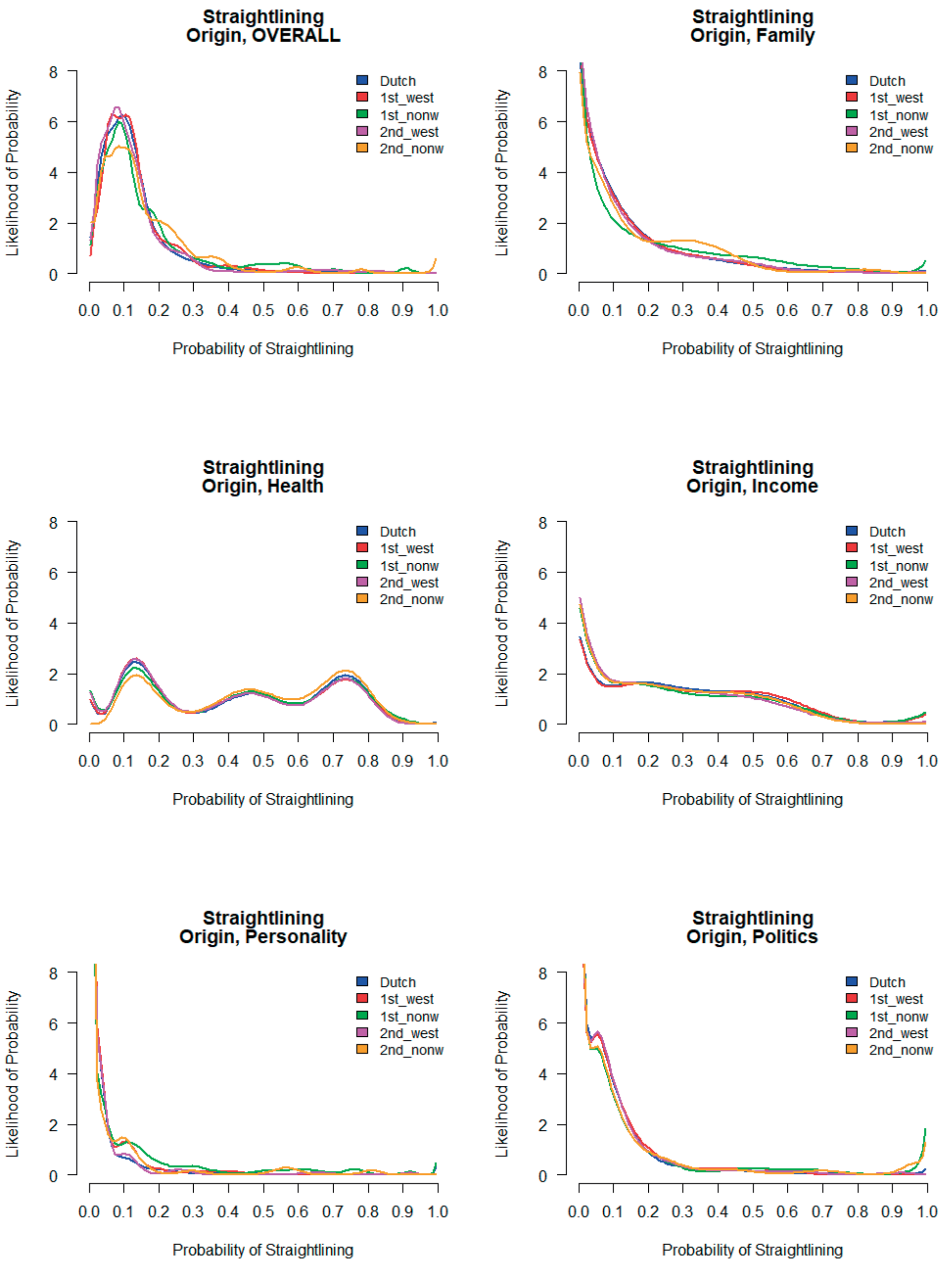


Figure K.23. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

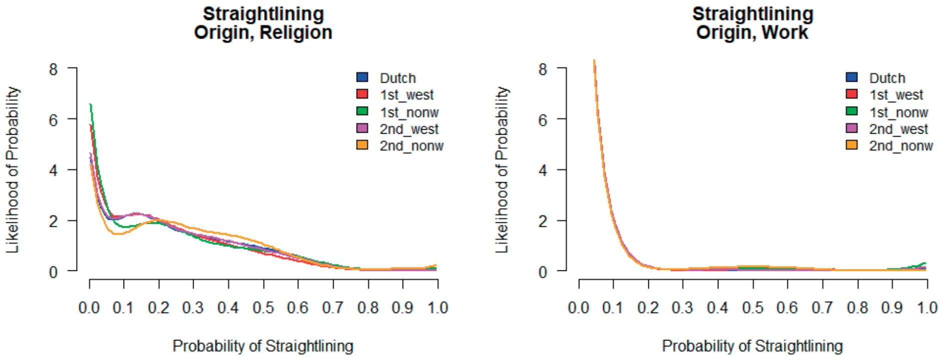


Figure K.23 (continued). Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

STRAIGHTLINING: PROVIDED A PC

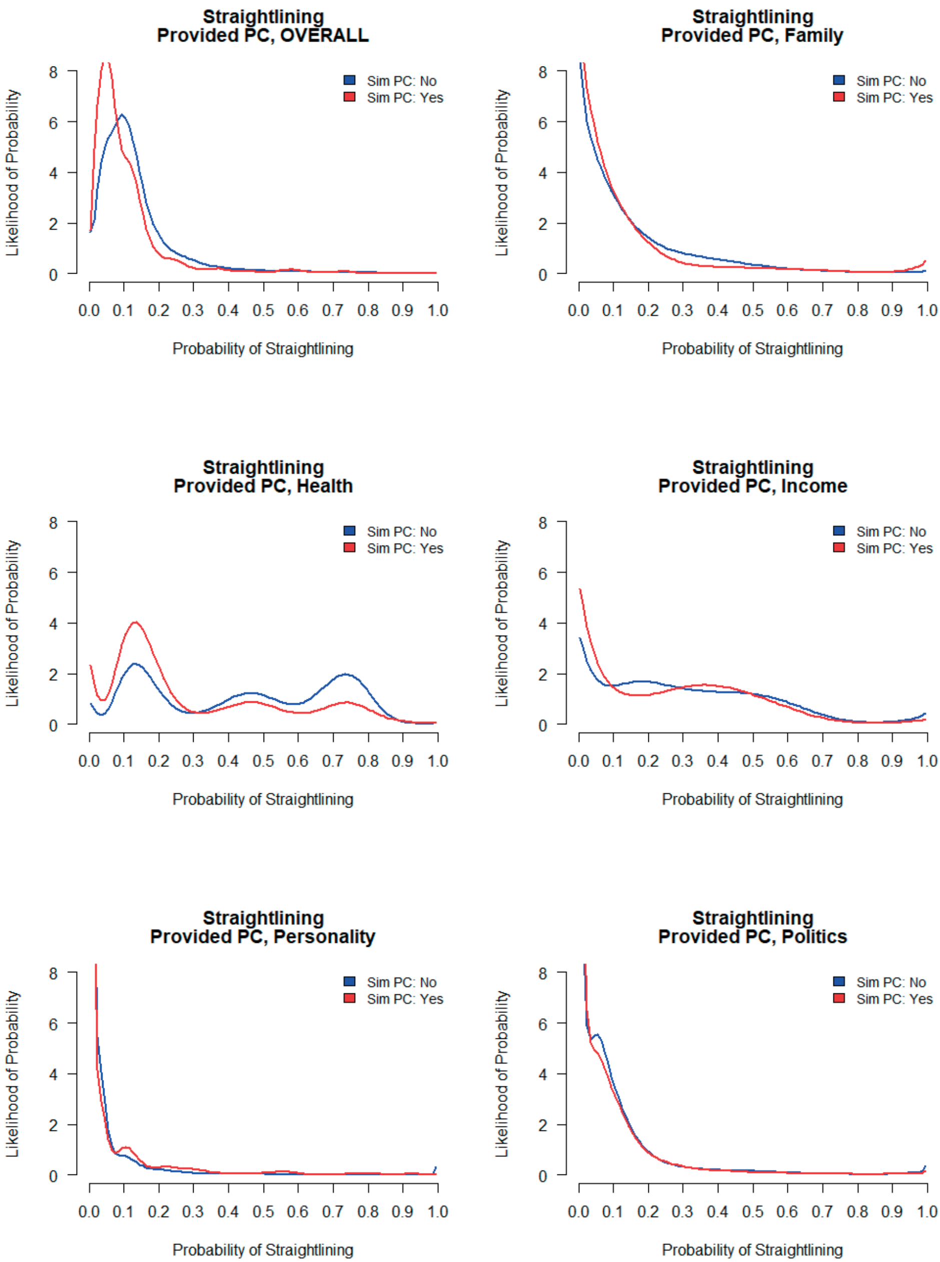


Figure K.24. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

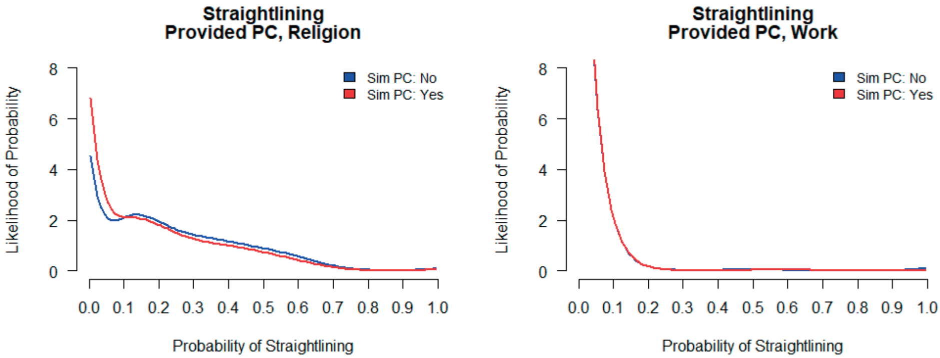


Figure K.24 (continued). Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Straightlining'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Health (HE), Income (IN), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: GENDER

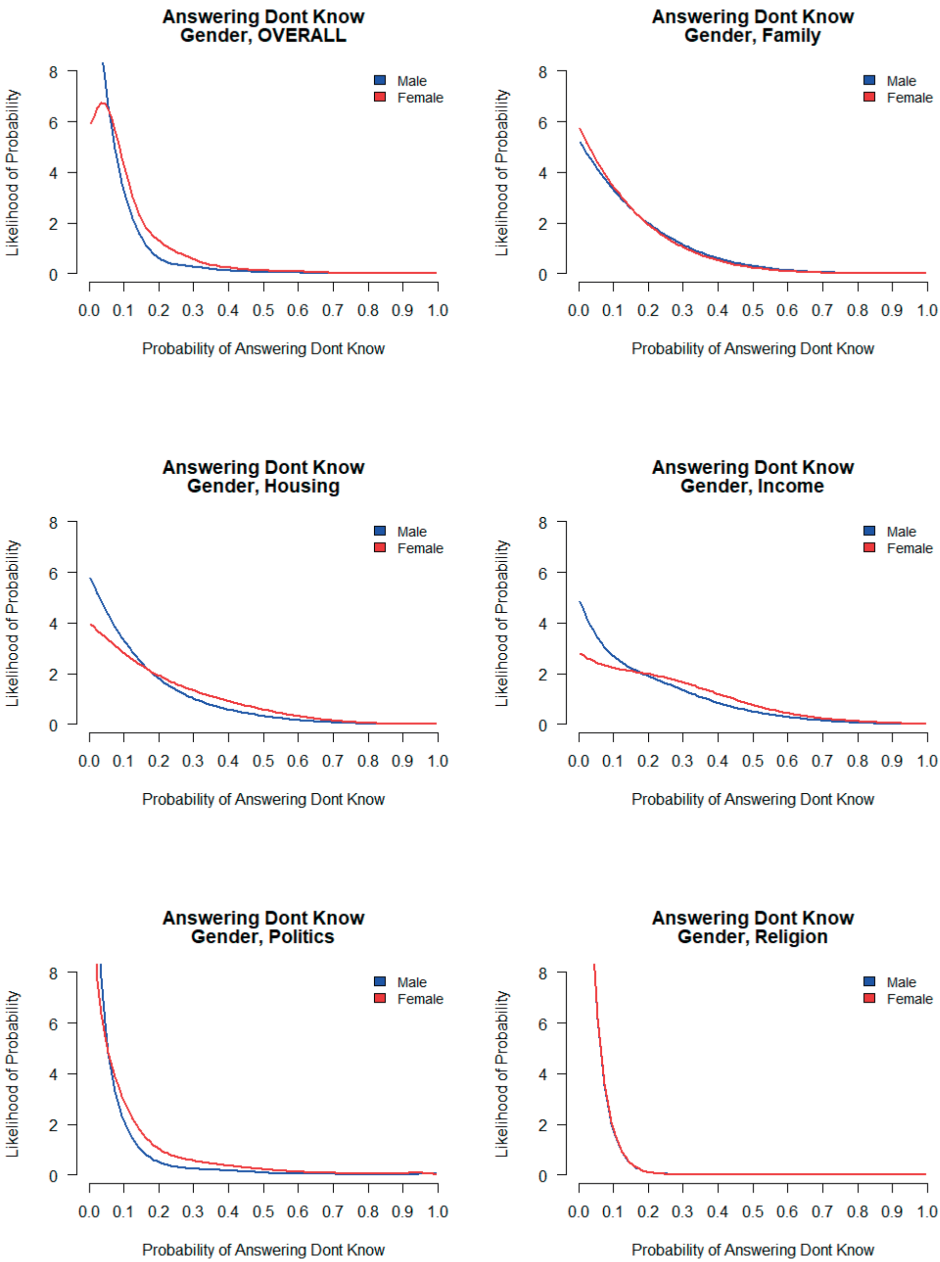


Figure K.25. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

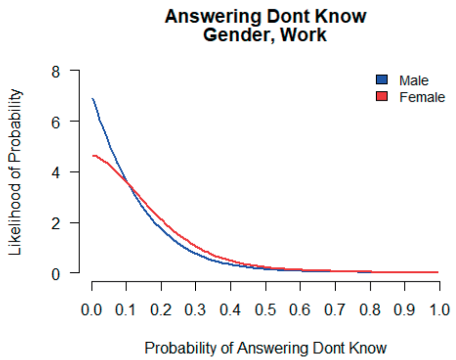


Figure K.25 (continued). Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: AGE

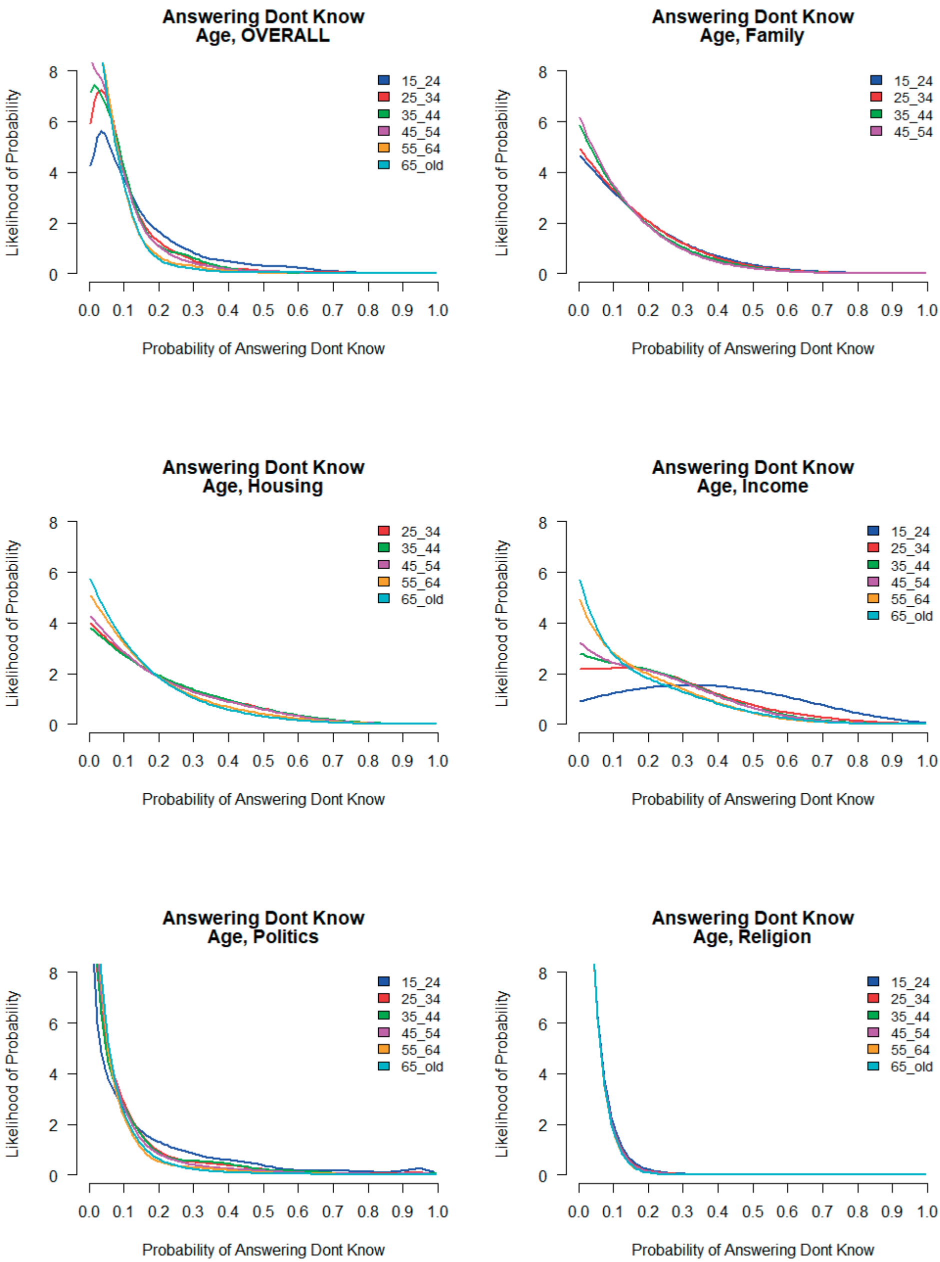


Figure K.26. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

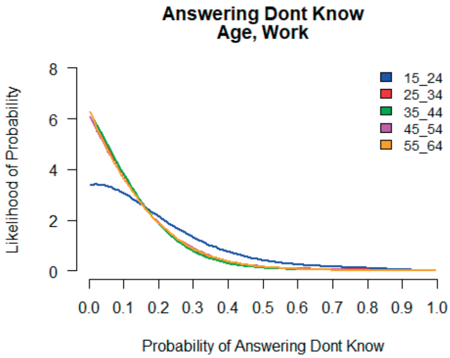


Figure K.26 (continued). Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: EDUCATION

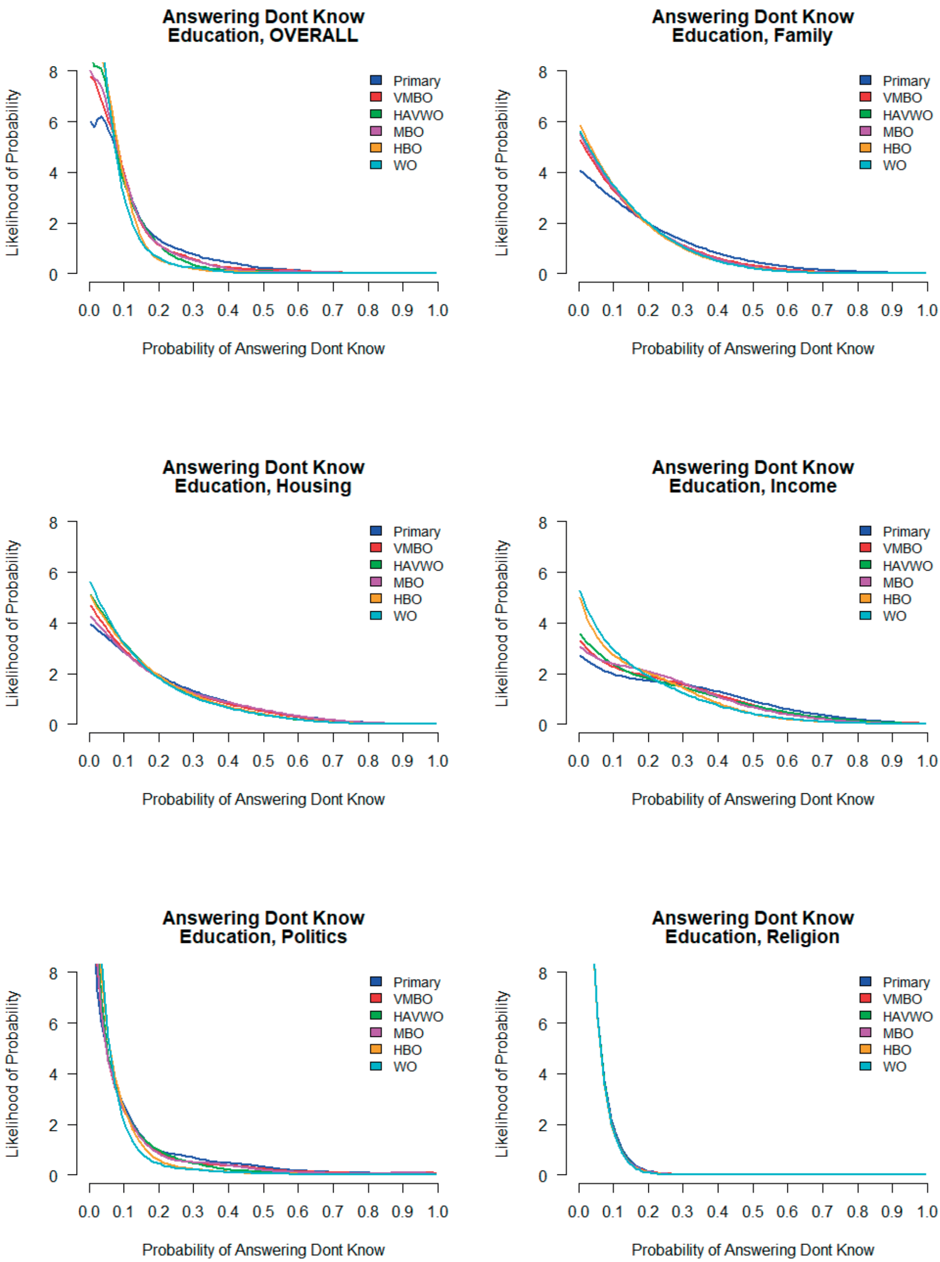


Figure K.27. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

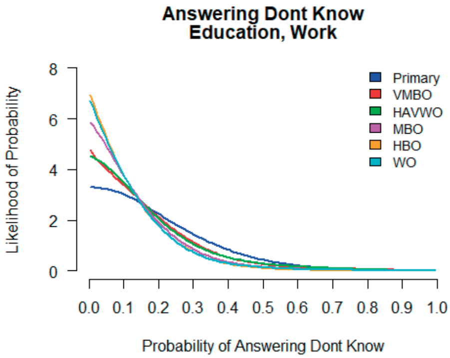


Figure K.27 (continued). Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: DOMESTIC SITUATION

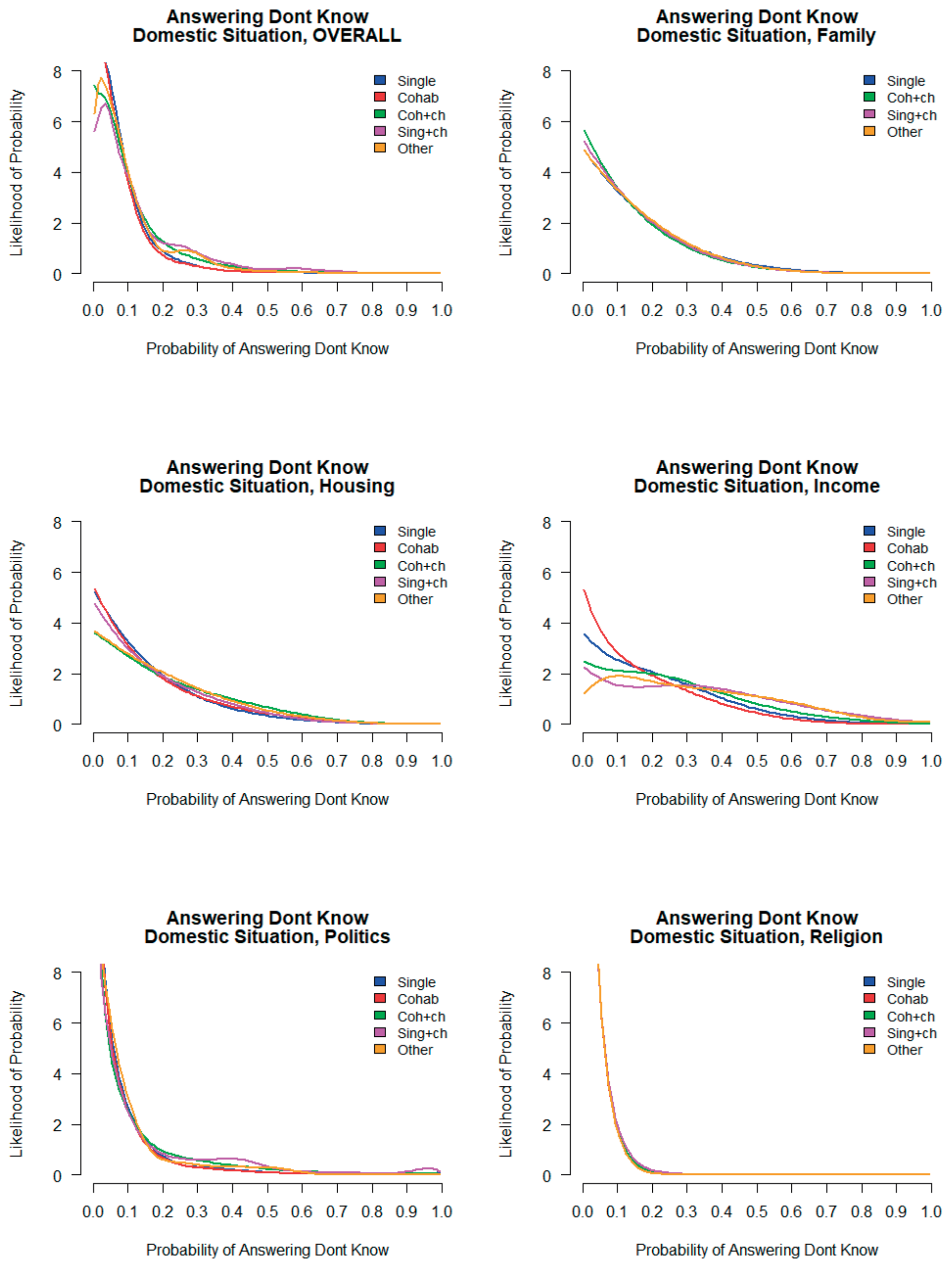


Figure K.28. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

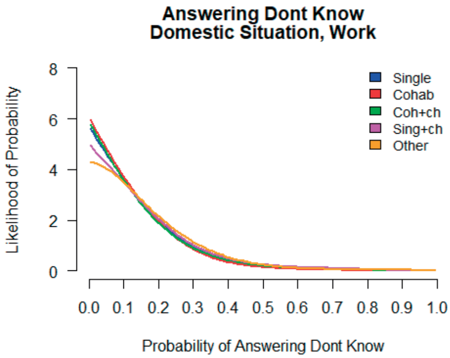


Figure K.28 (continued). Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: PRIMARY OCCUPATION

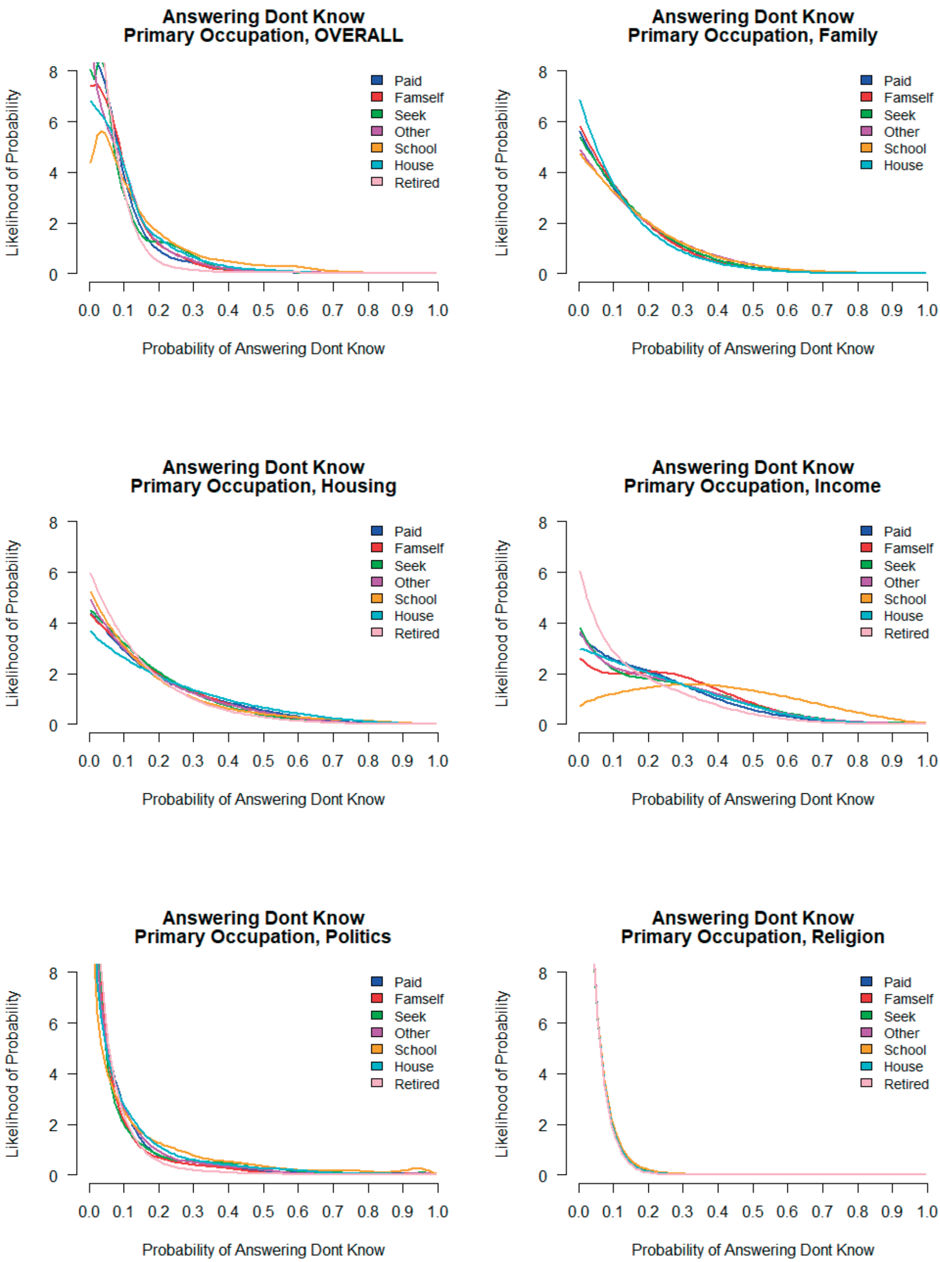


Figure K.29. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

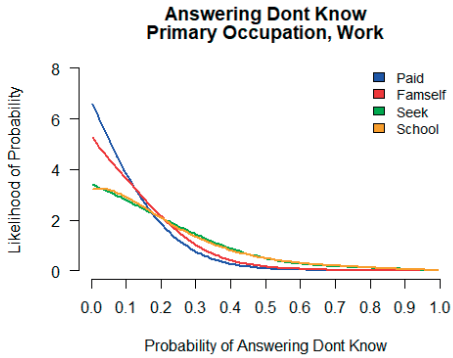


Figure K.29 (continued). Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON’T KNOW: INCOME

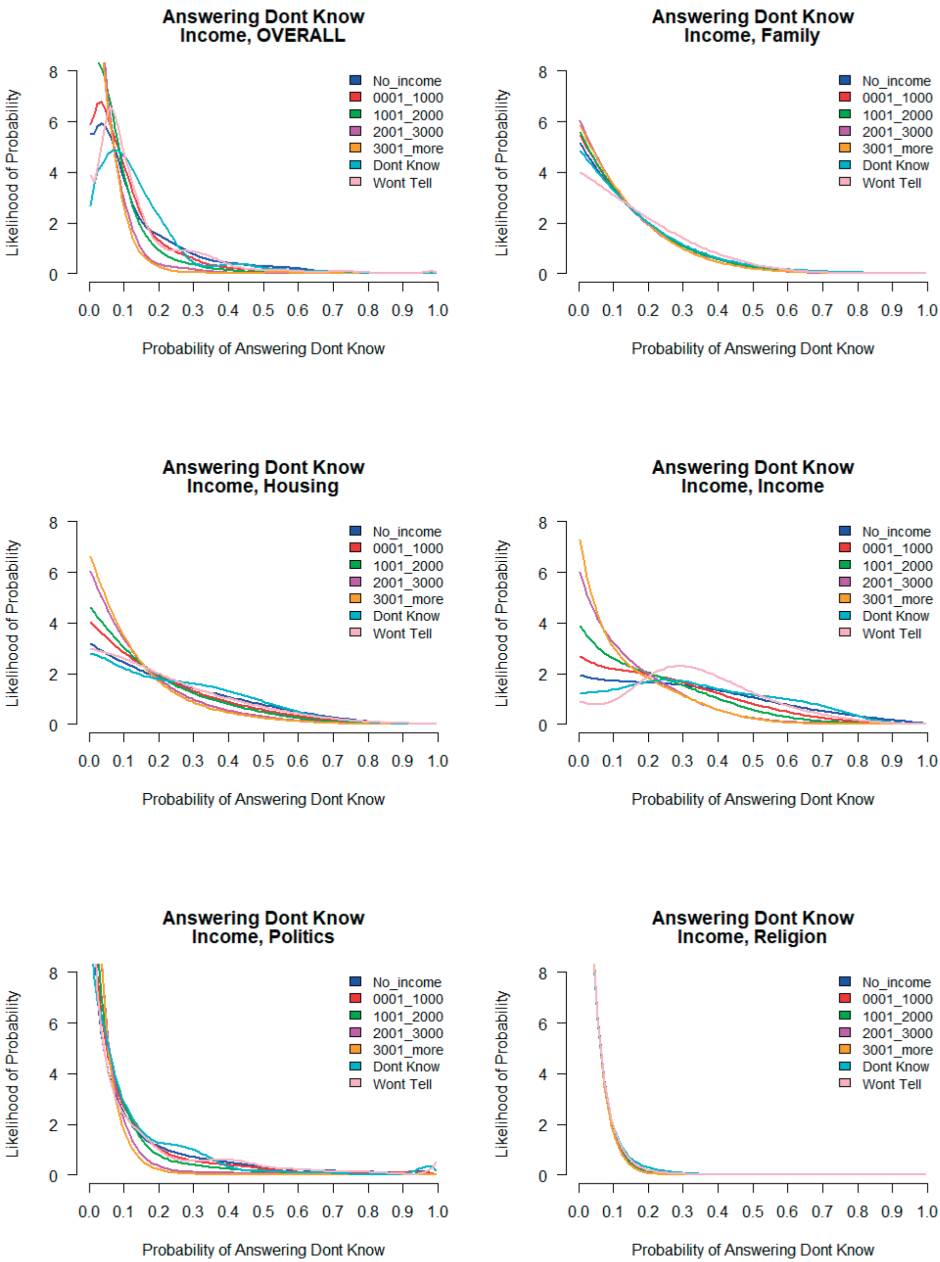


Figure K.30. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **‘Answering Don’t Know’** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

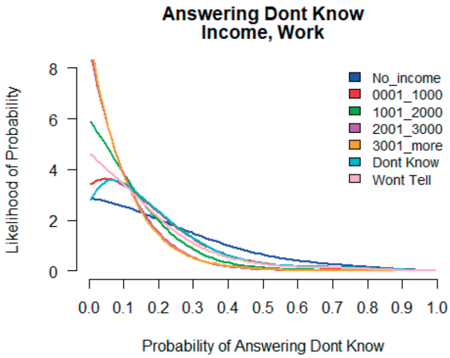


Figure K.30 (continued). Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON'T KNOW: ORIGIN

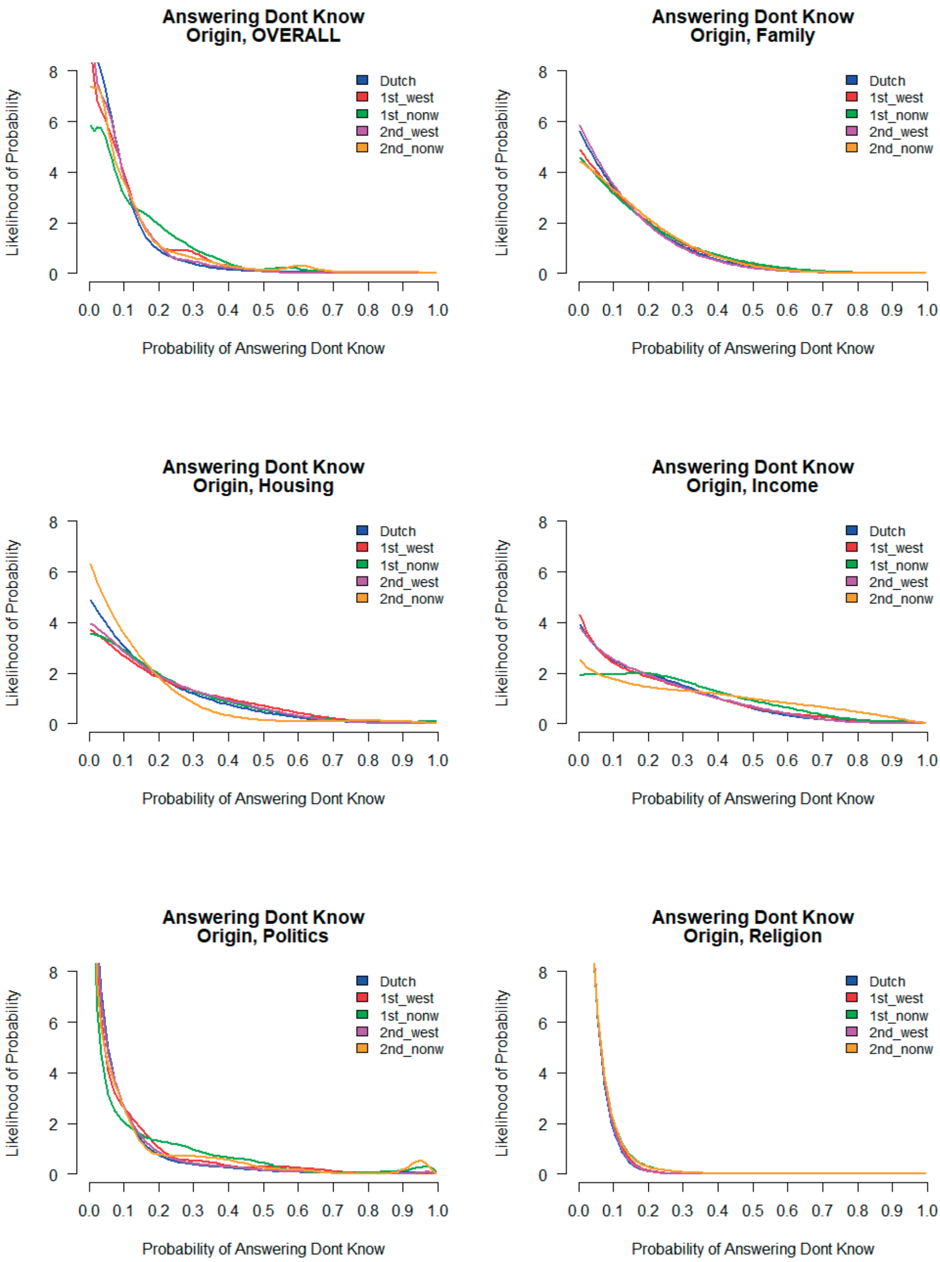


Figure K.31. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

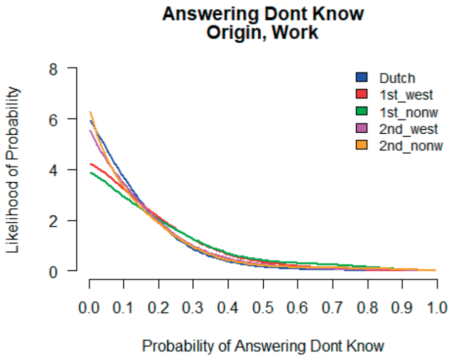


Figure K.31 (continued). Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Answering Don't Know'** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ANSWERING DON’T KNOW: PROVIDED A PC

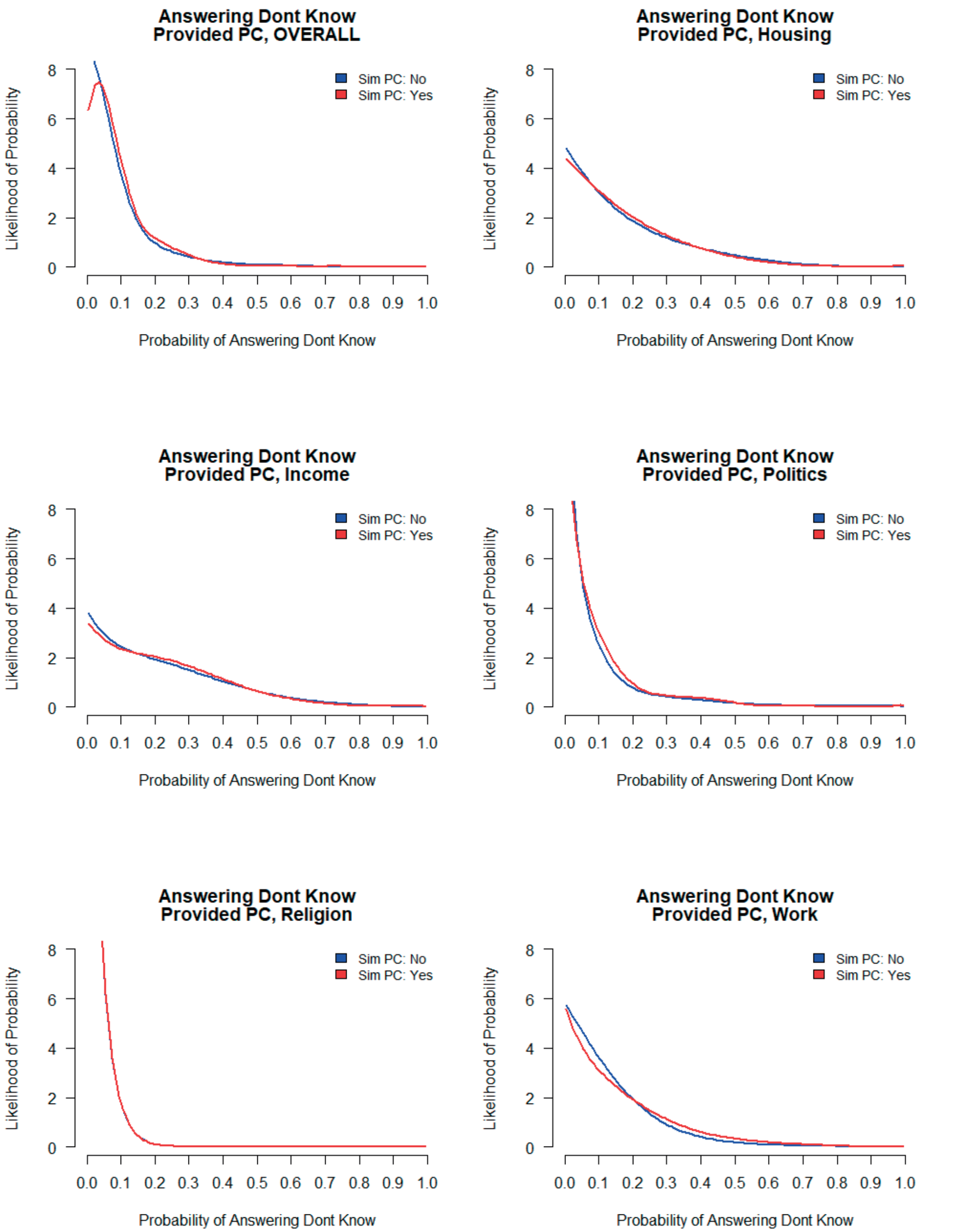


Figure K.32. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **‘Answering Don’t Know’** Overall and for All Applicable Surveys*.

* Assets (AS), Family and Household (FA), Housing (HO), Income (IN), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: GENDER

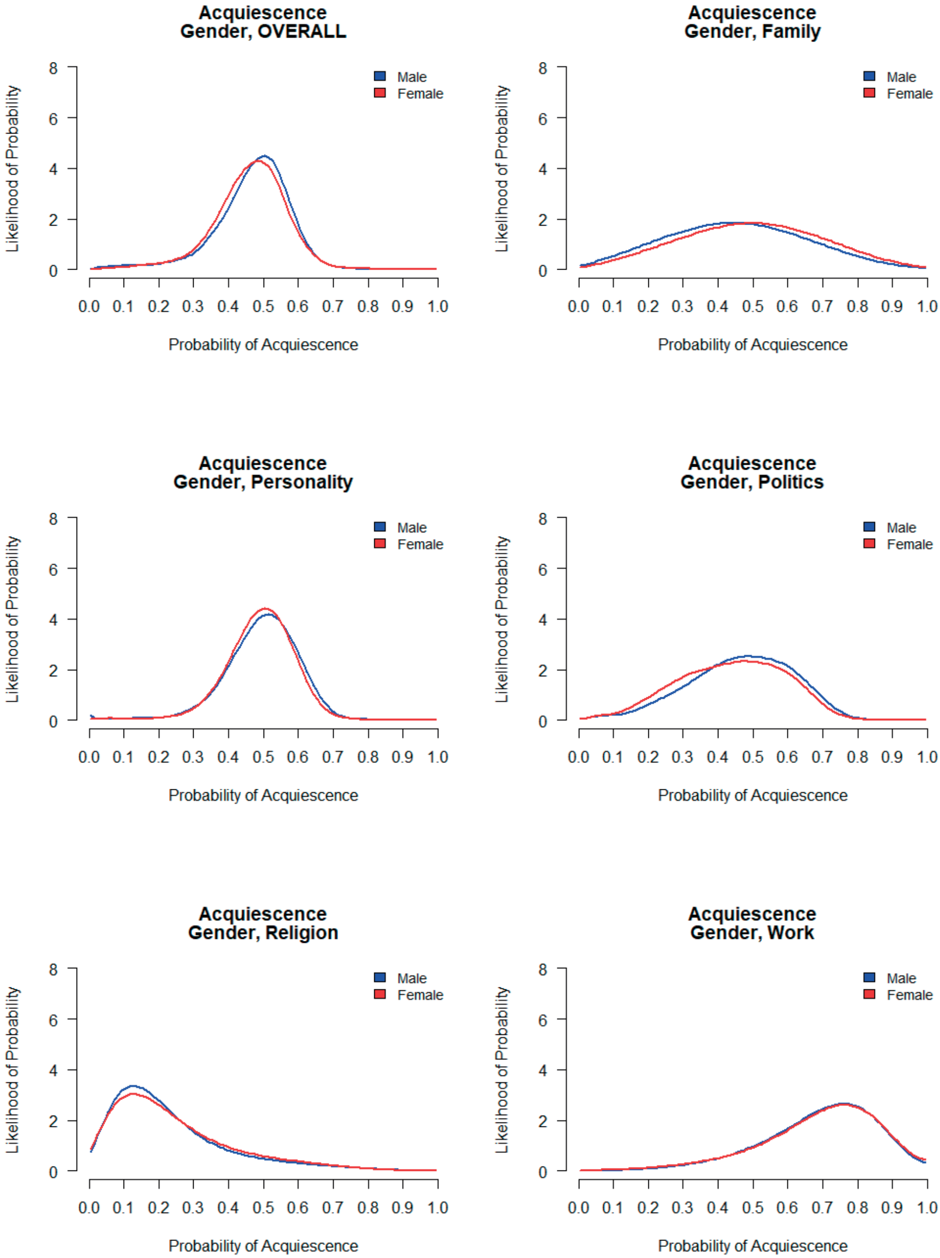


Figure K.33. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: AGE

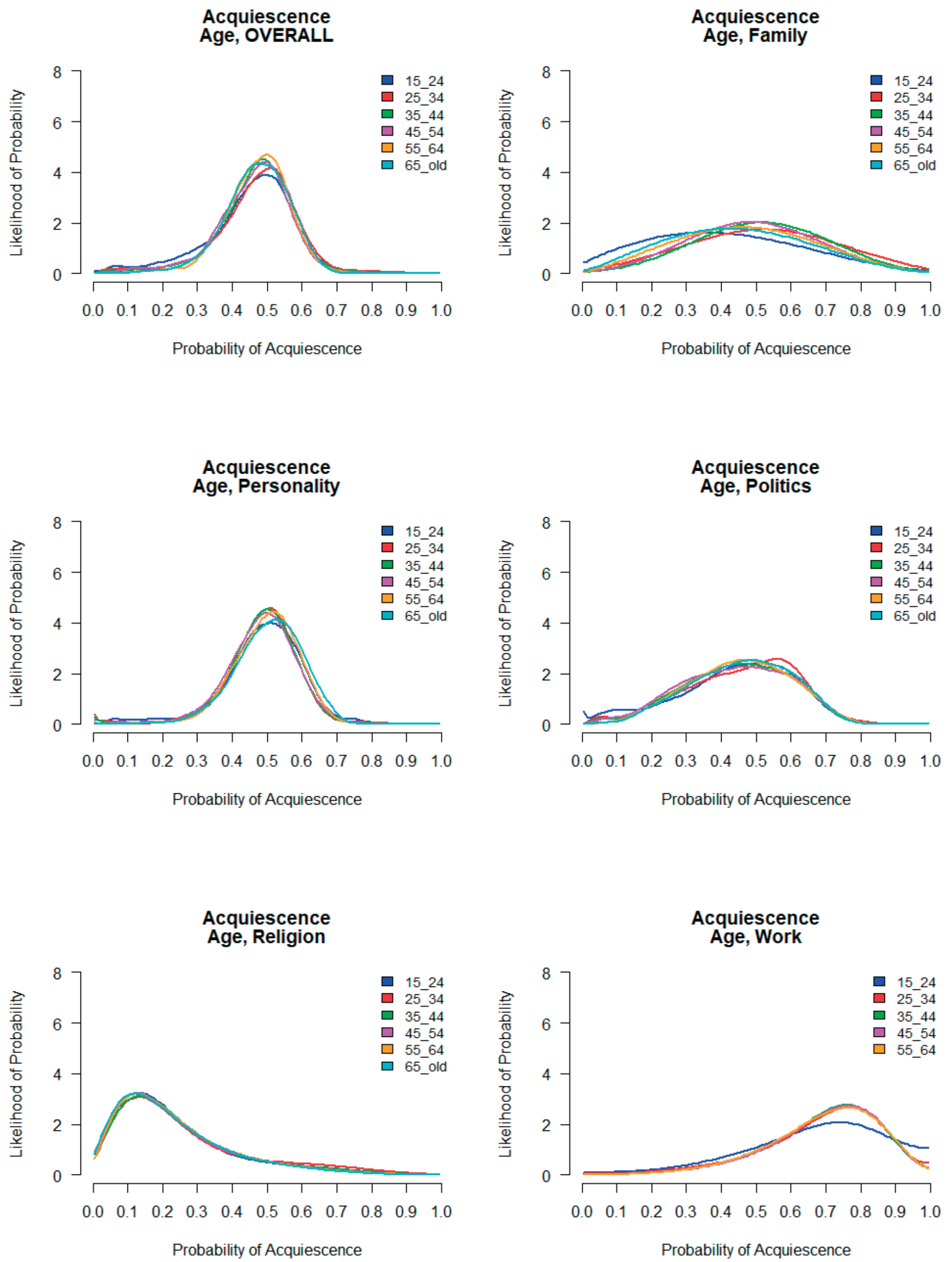


Figure K.34. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: EDUCATION

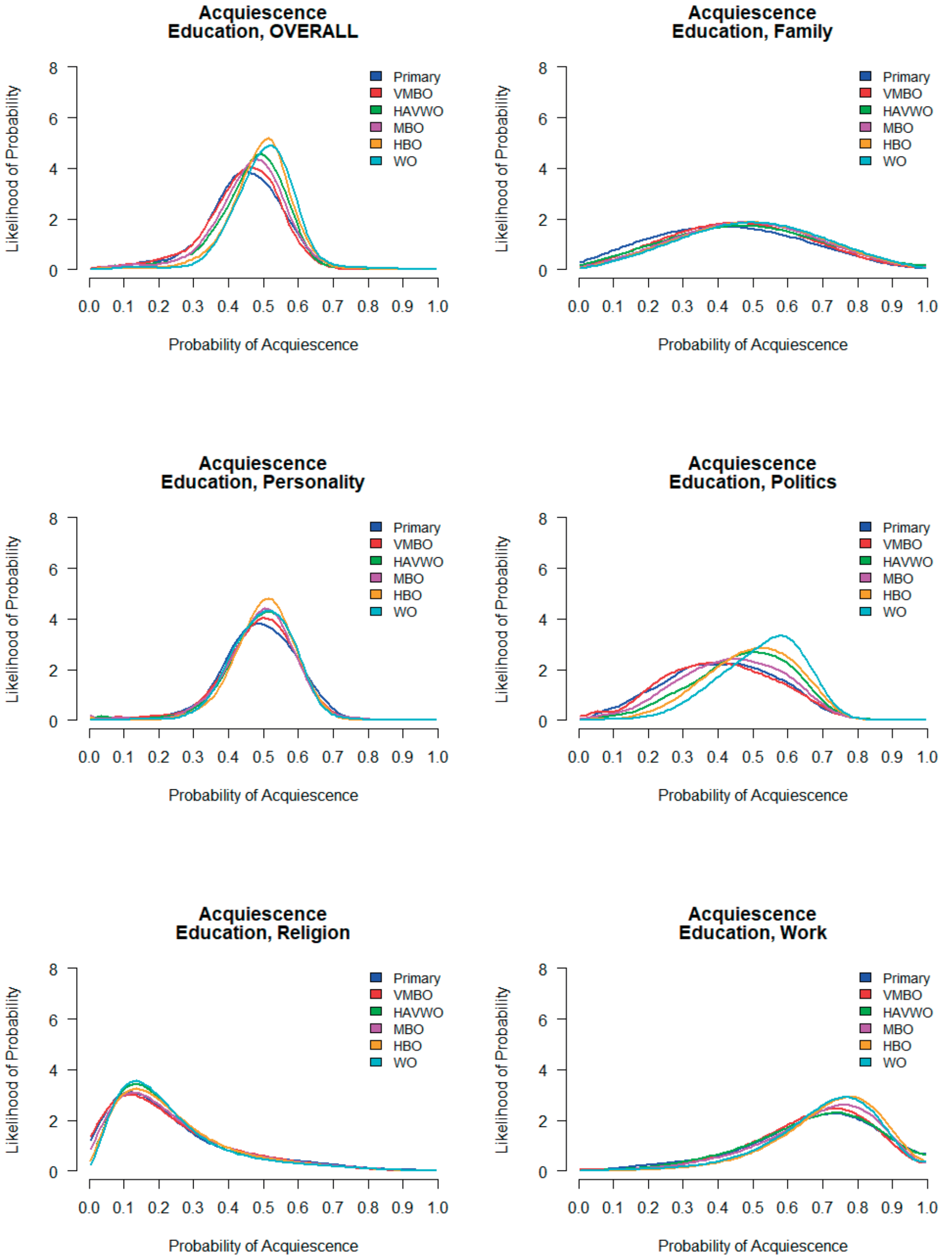


Figure K.35. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: DOMESTIC SITUATION

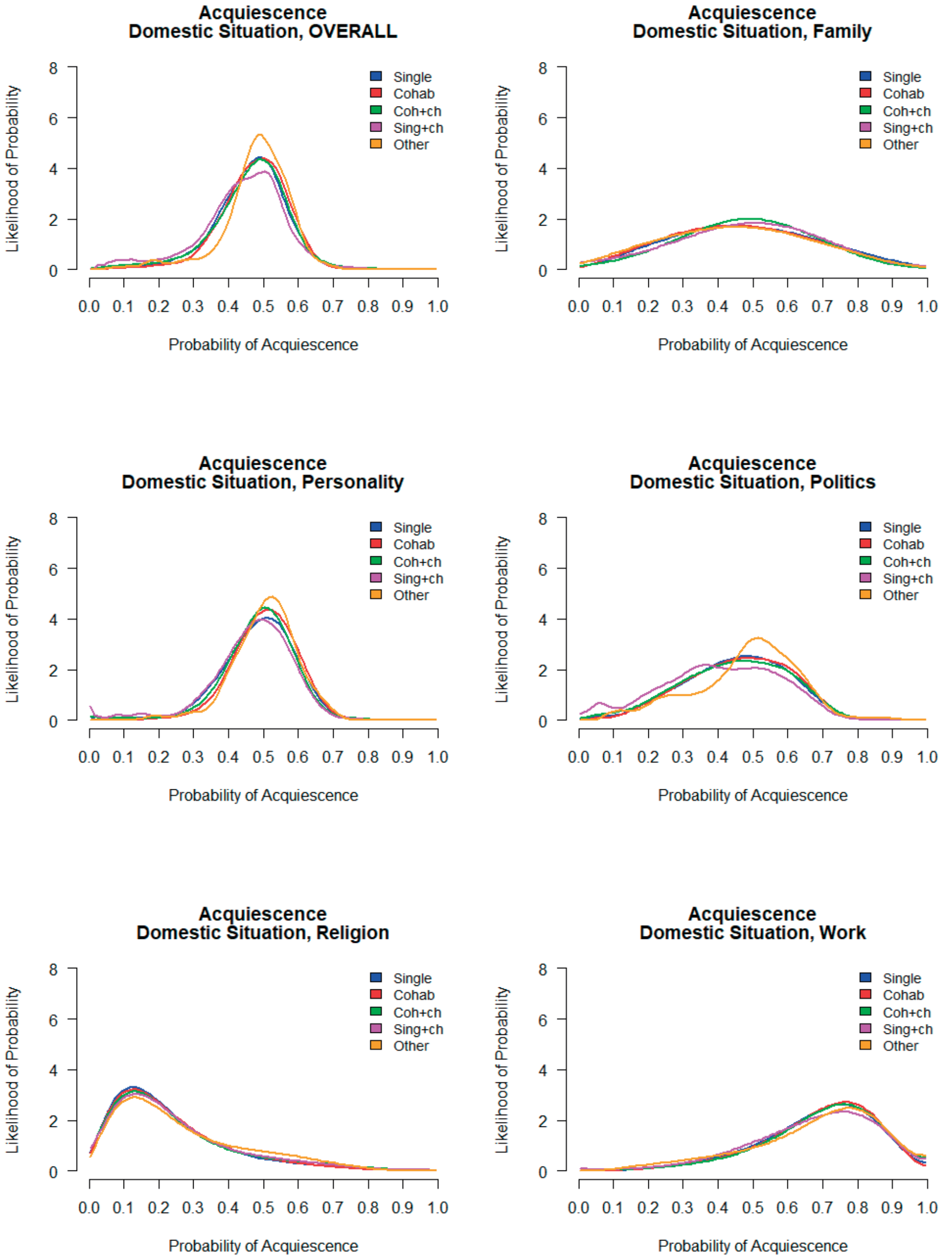


Figure K.36. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: PRIMARY OCCUPATION

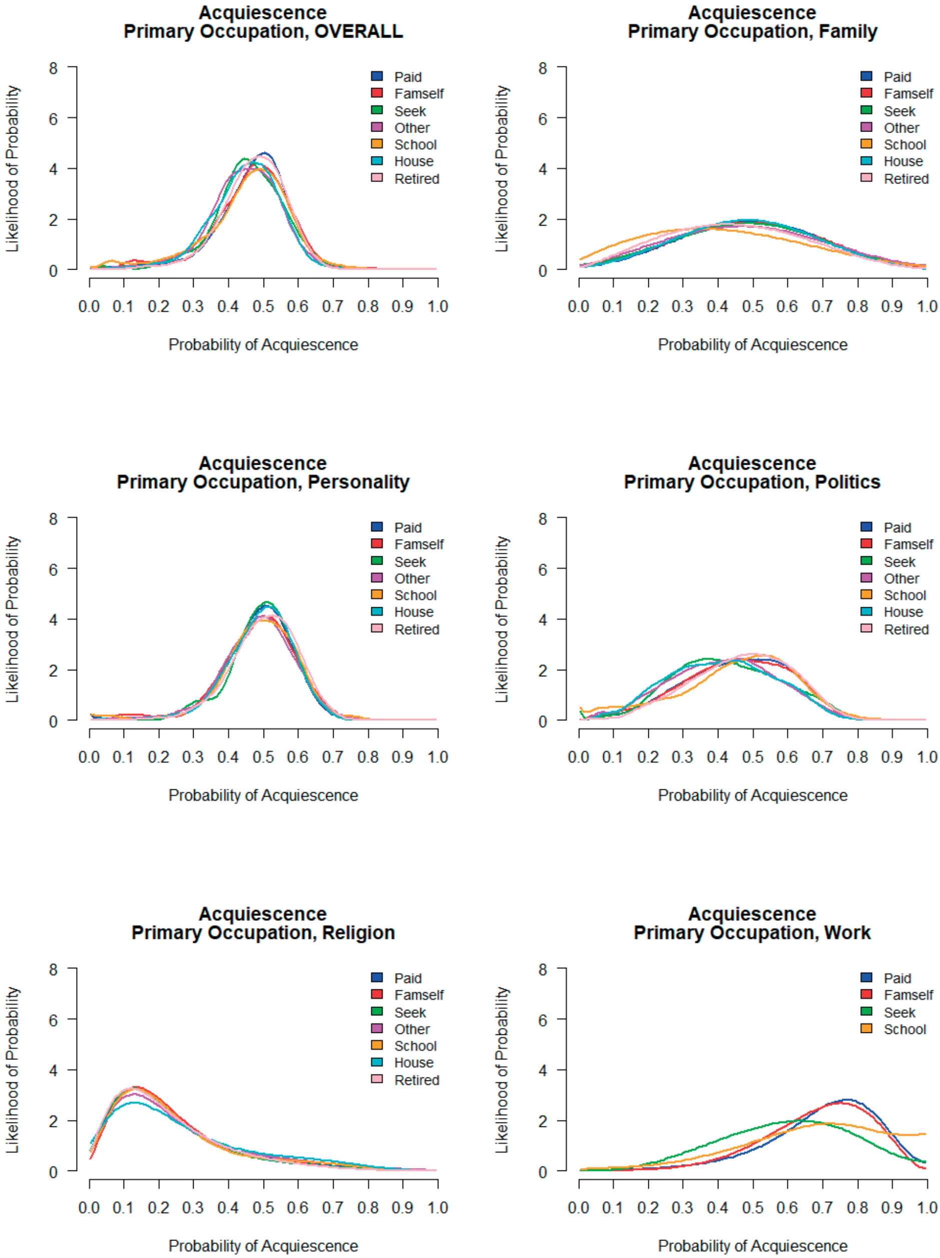


Figure K.37. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: INCOME

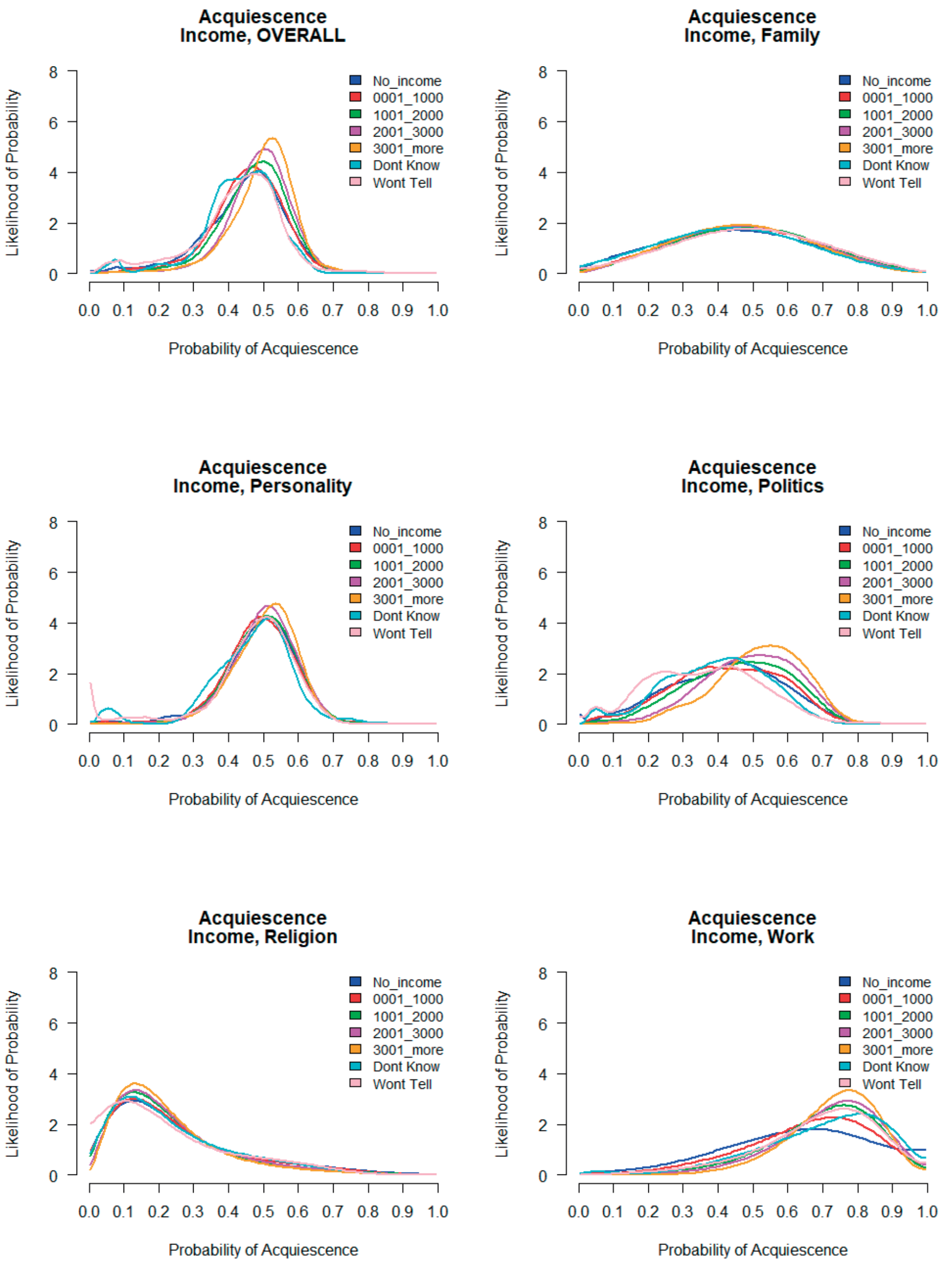


Figure K.38. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: ORIGIN

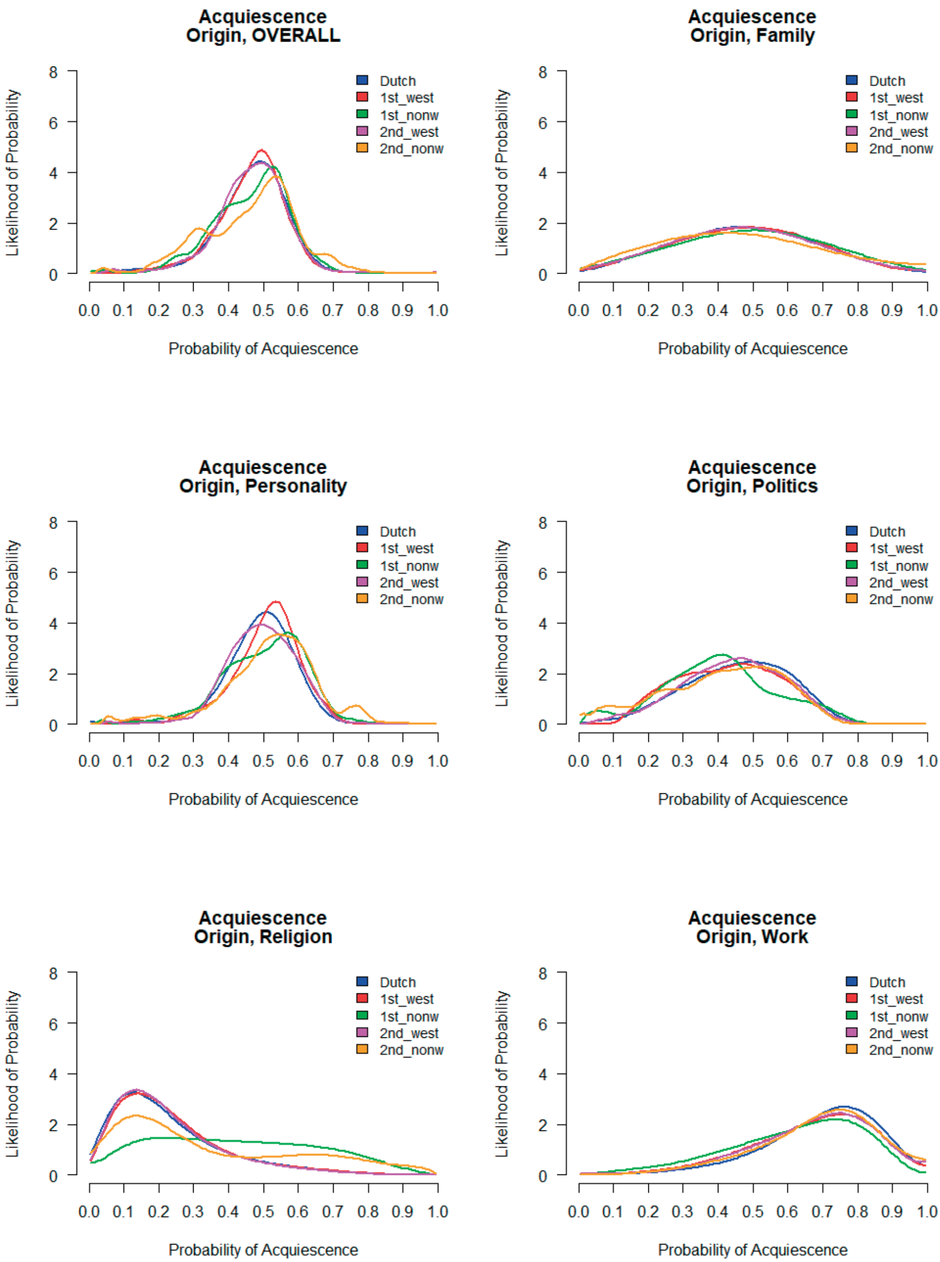


Figure K.39. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

ACQUIESCENCE: PROVIDED A PC

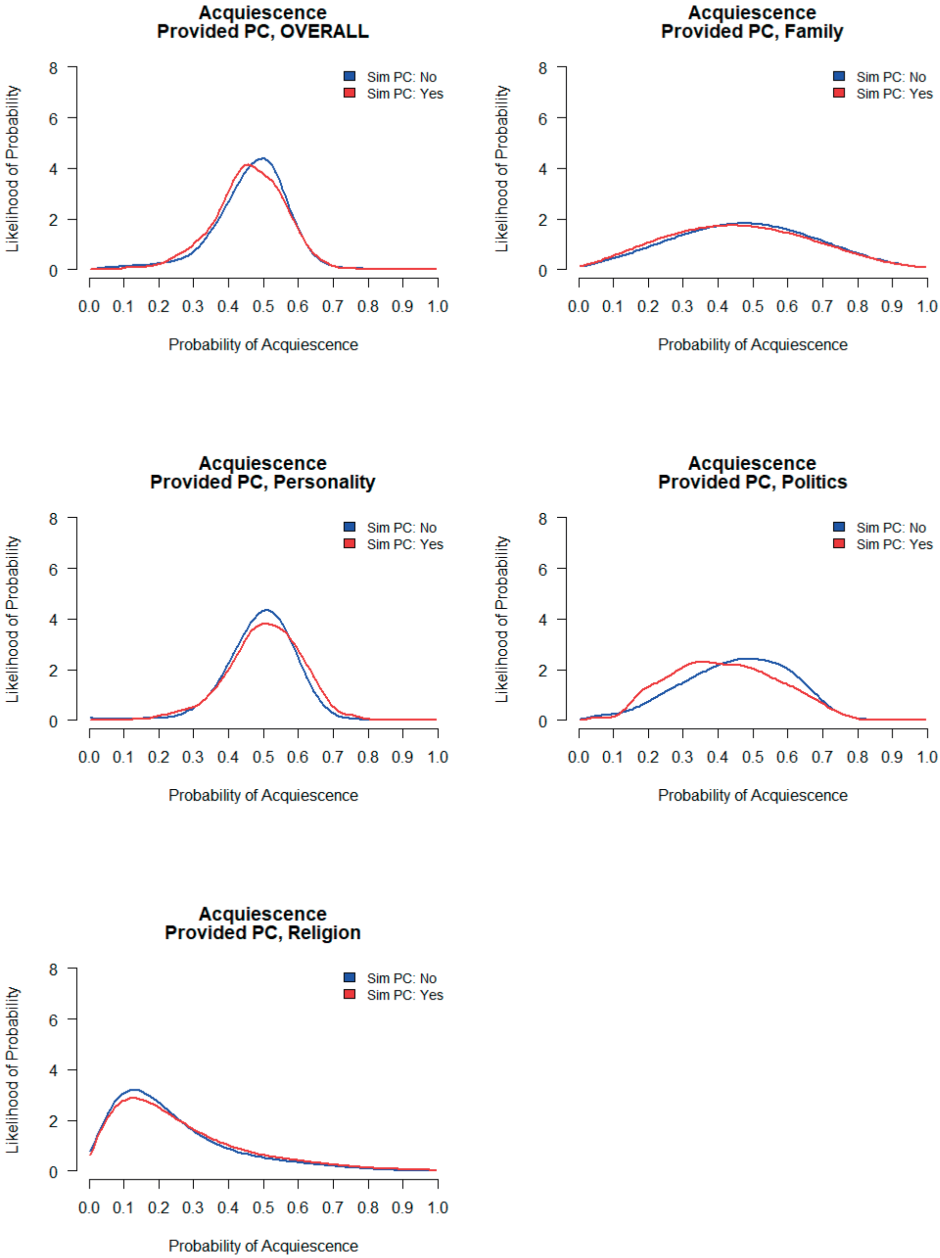


Figure K.40. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Acquiescence'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Personality (PE), Politics and Values (PO), Religion and Ethnicity (RE), Work and Schooling (WO).

NEUTRAL RESPONDING: GENDER

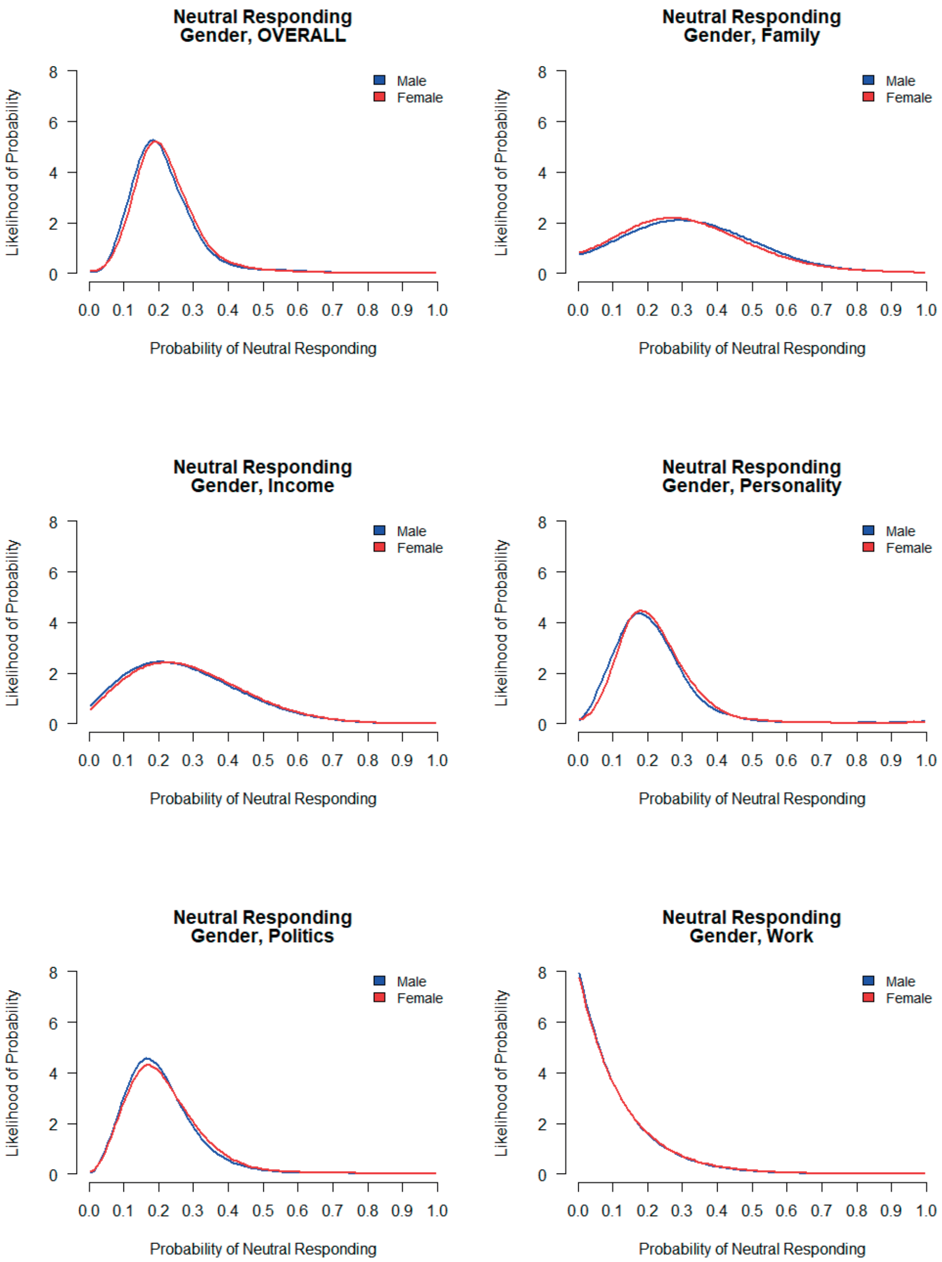


Figure K.41. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: AGE

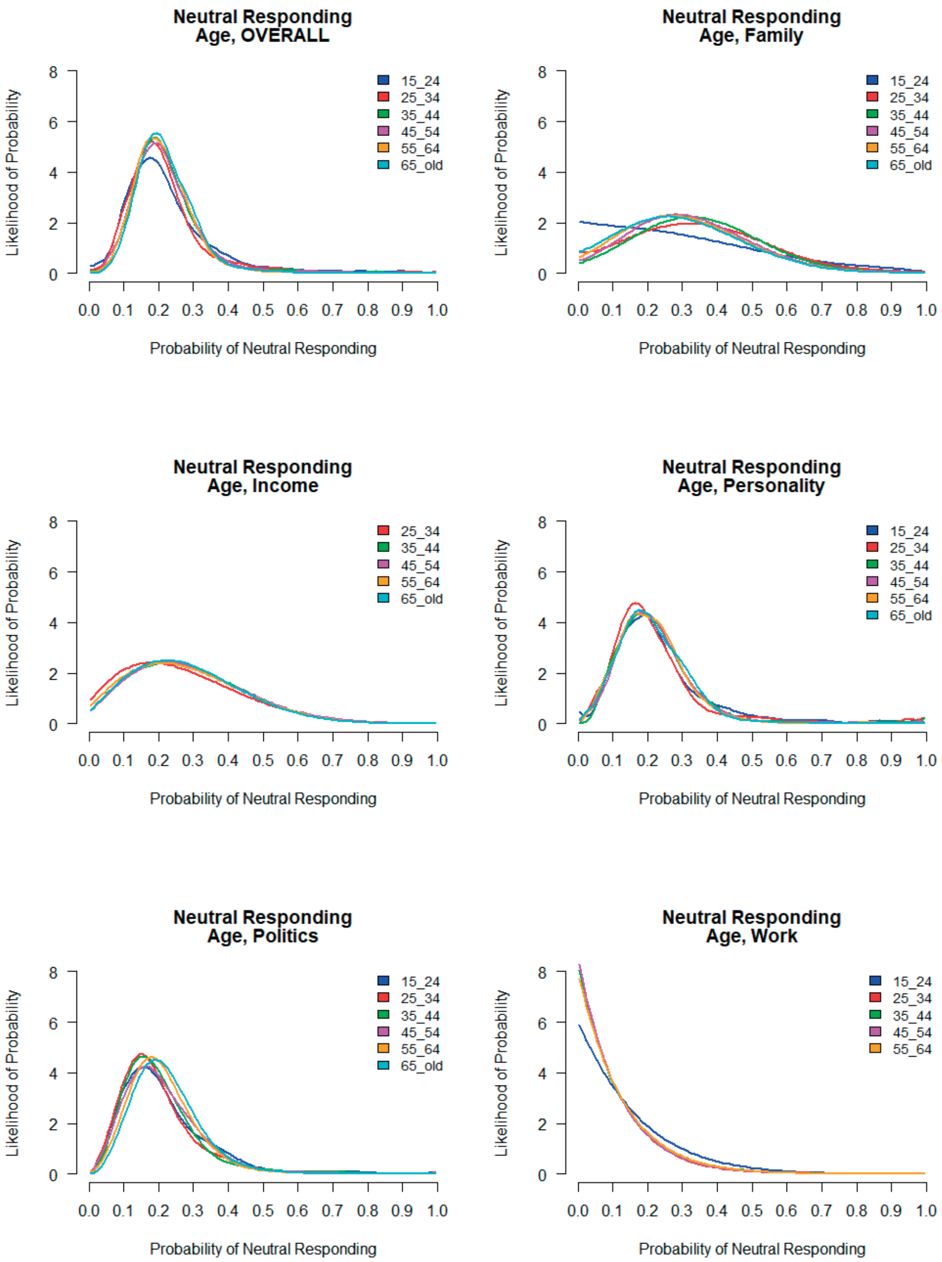


Figure K.42. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: EDUCATION

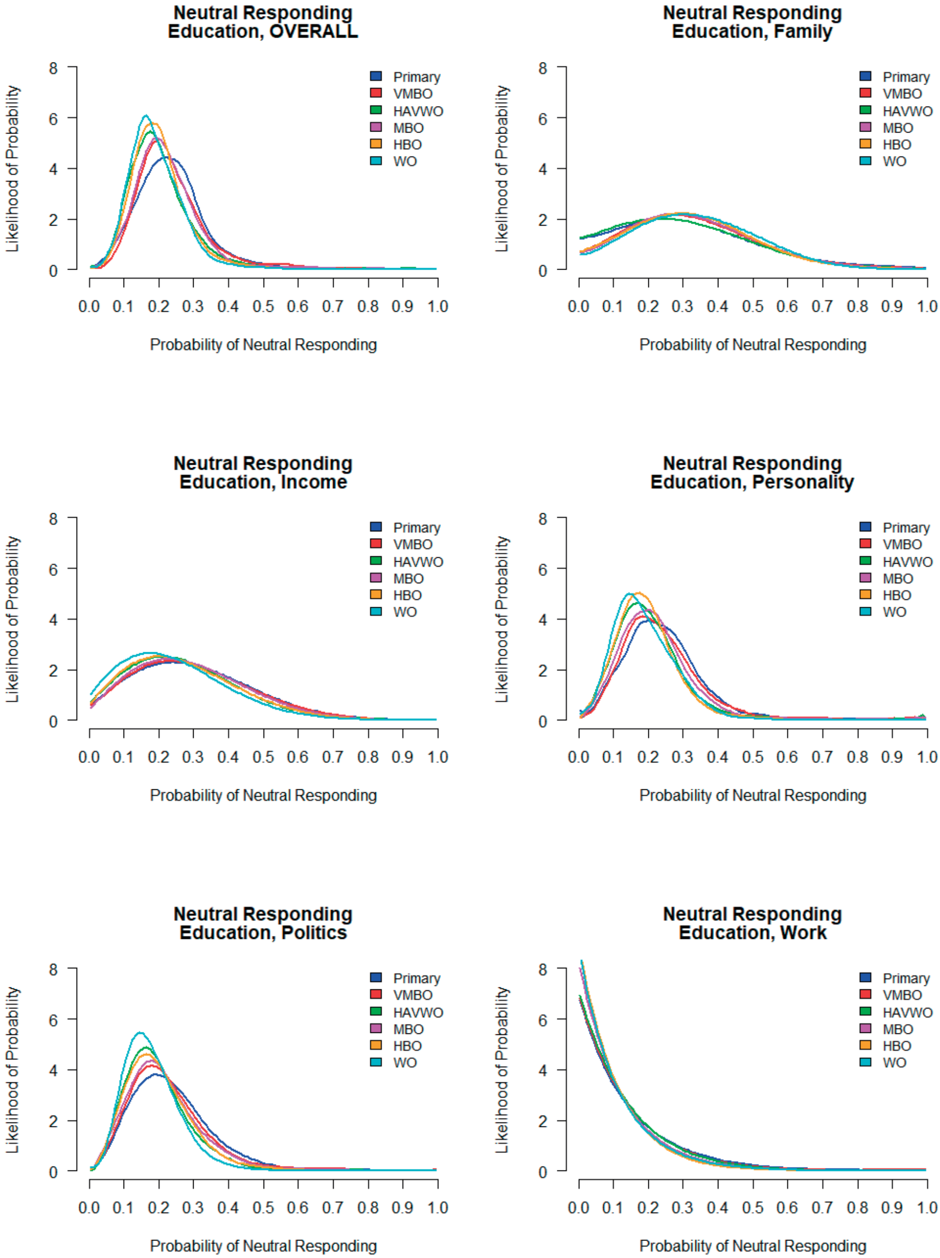


Figure K.43. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: DOMESTIC SITUATION

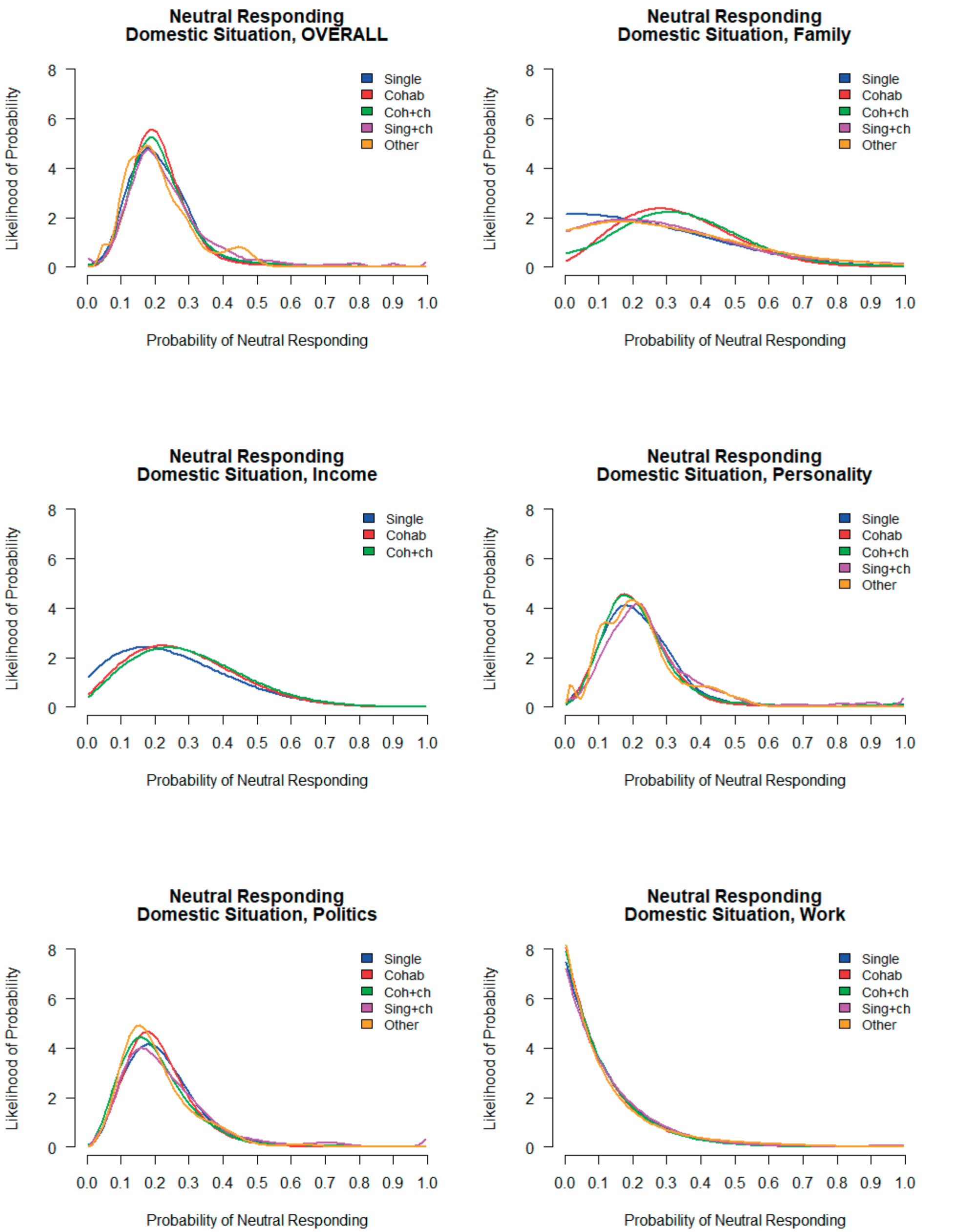


Figure K.44. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: PRIMARY OCCUPATION

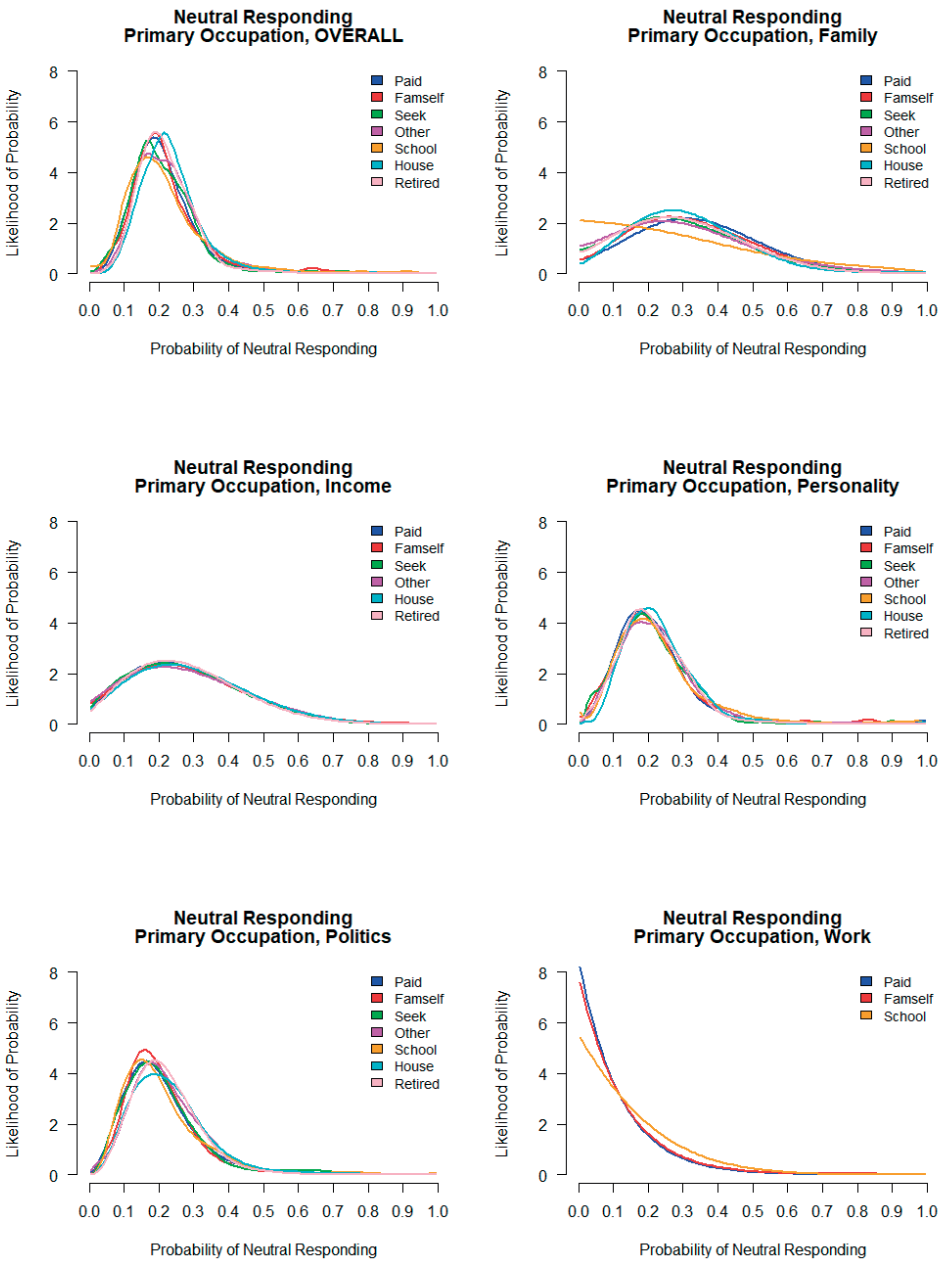


Figure K.45. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: INCOME

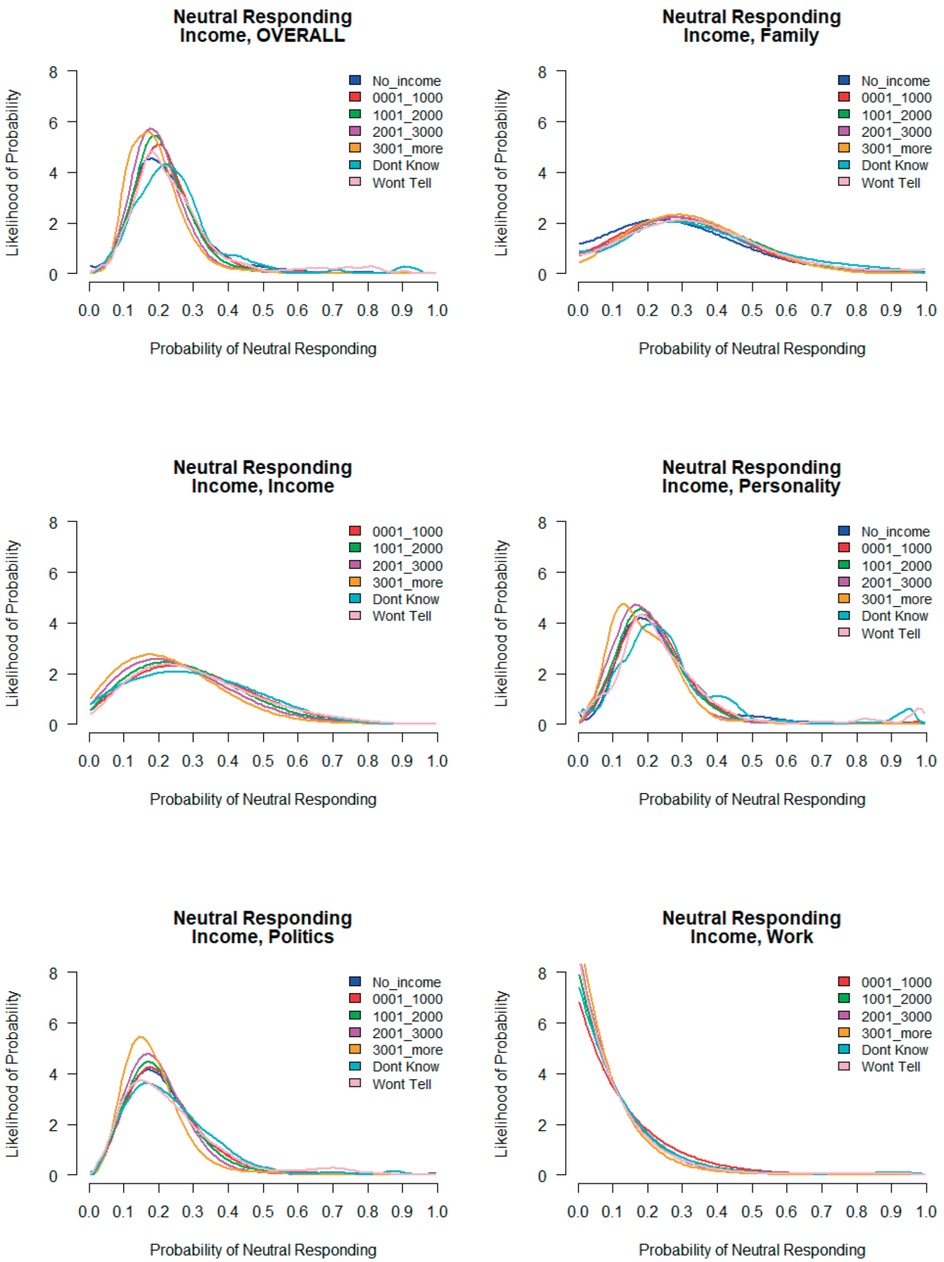


Figure K.46. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: ORIGIN

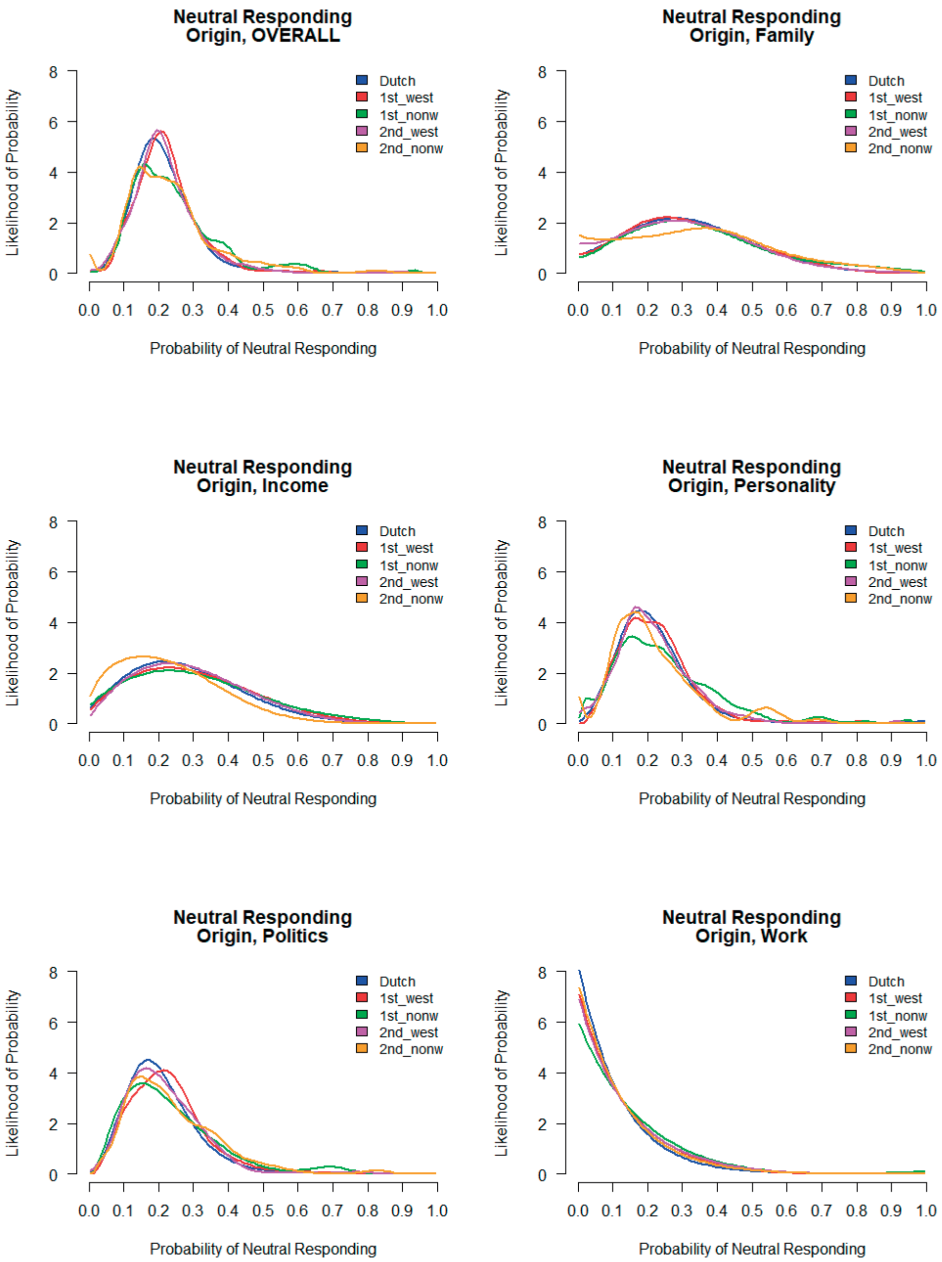


Figure K.47. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

NEUTRAL RESPONDING: PROVIDED A PC

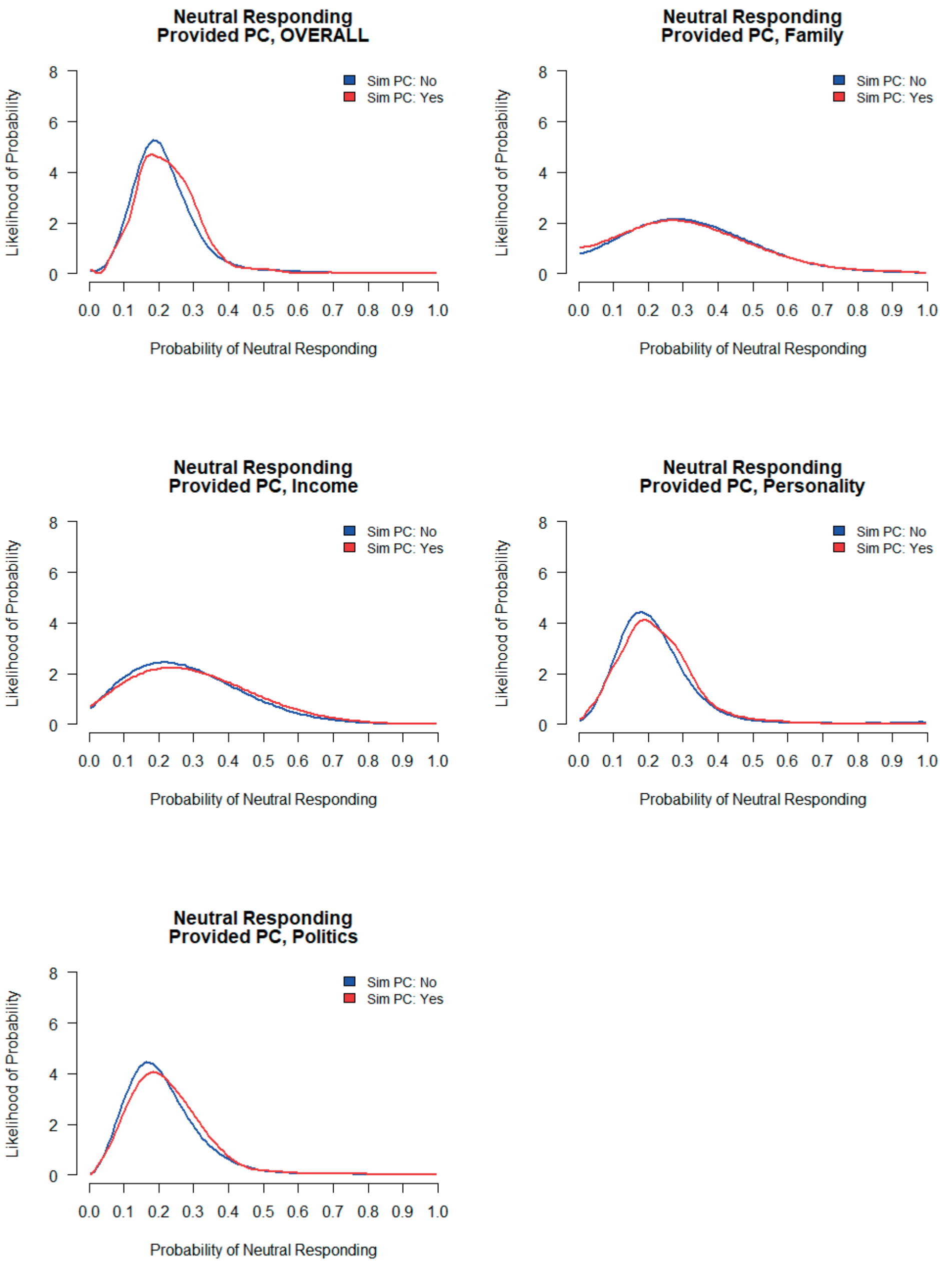


Figure K.48. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Neutral Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: GENDER

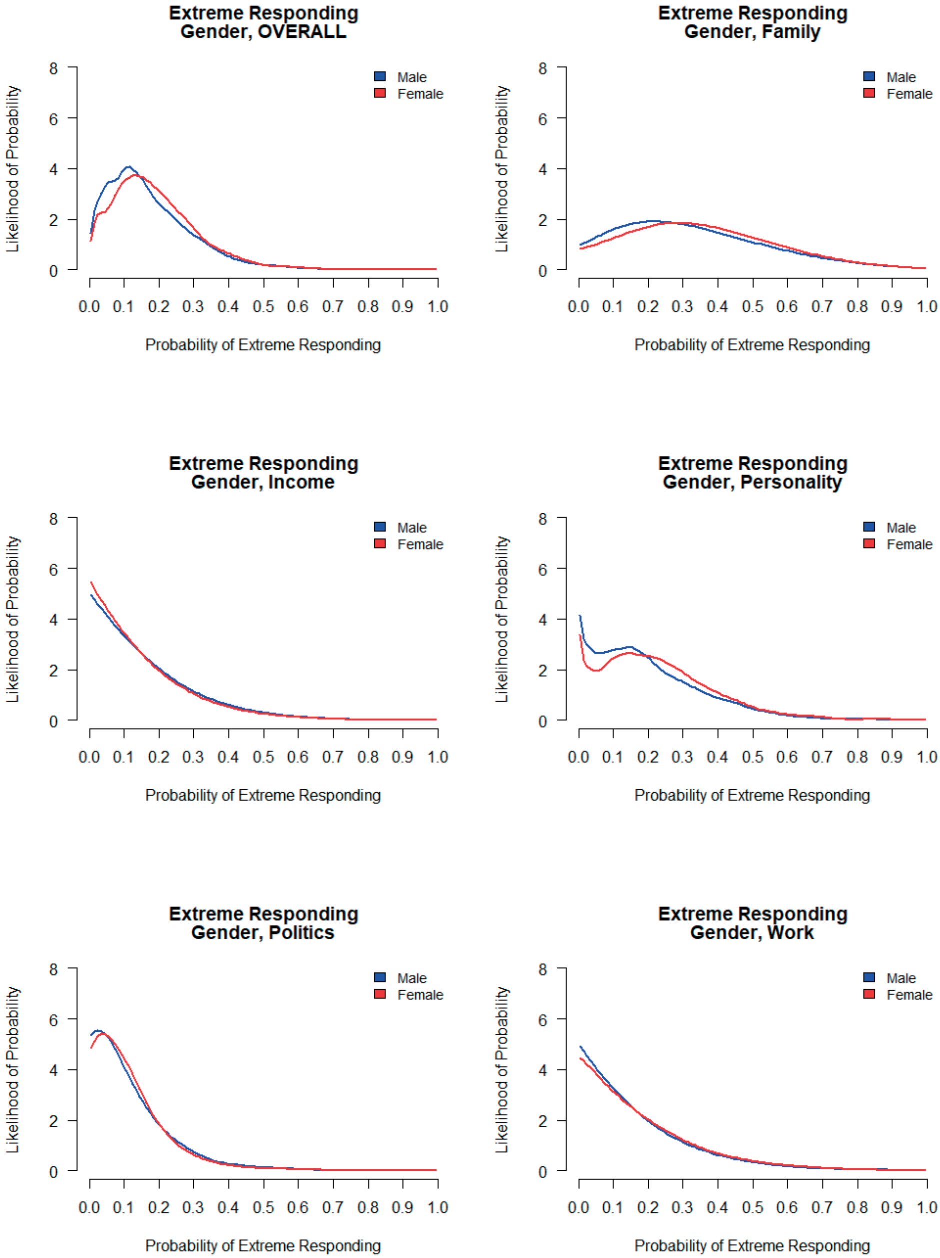


Figure K.49. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: AGE

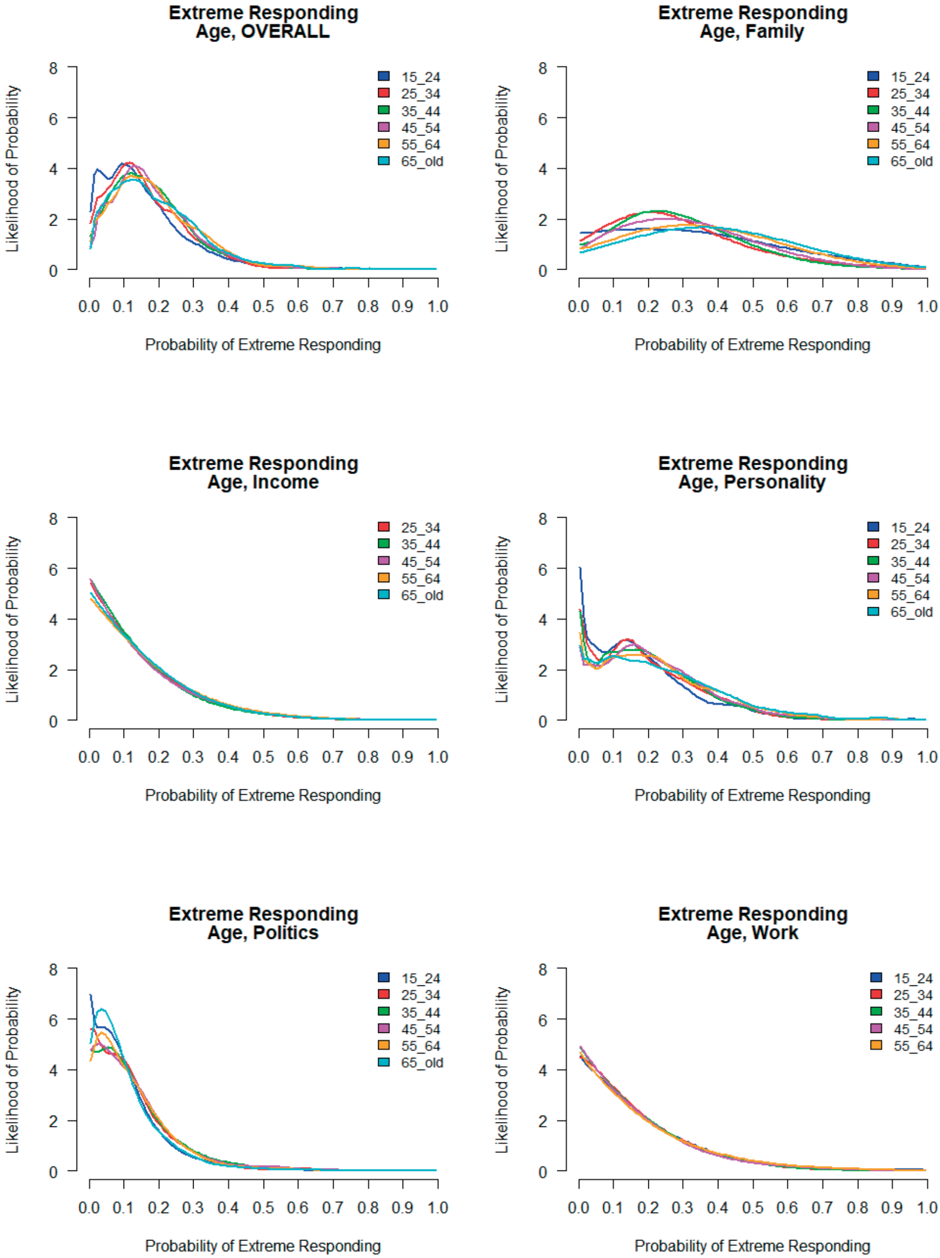


Figure K.50. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: EDUCATION

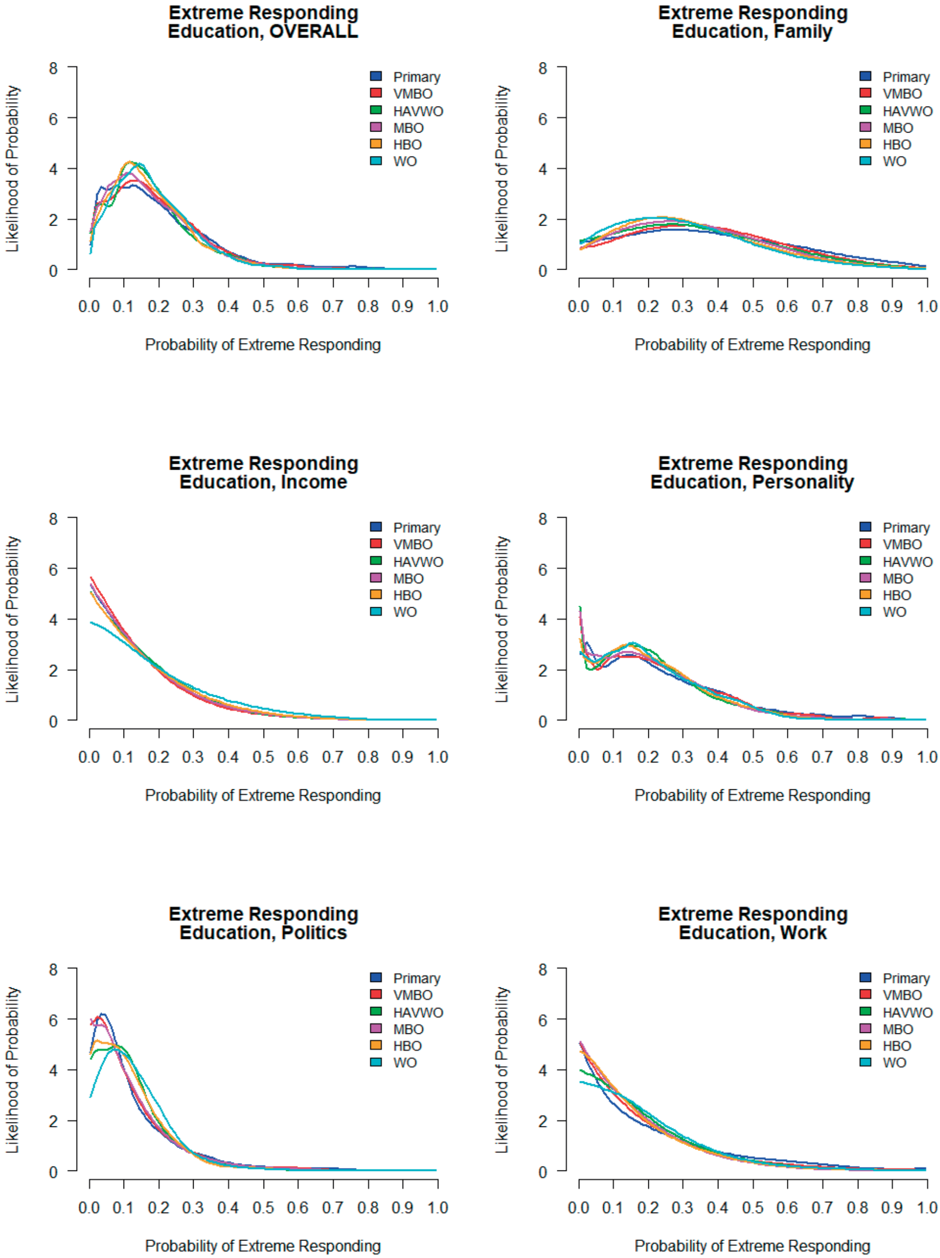


Figure K.51. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: DOMESTIC SITUATION

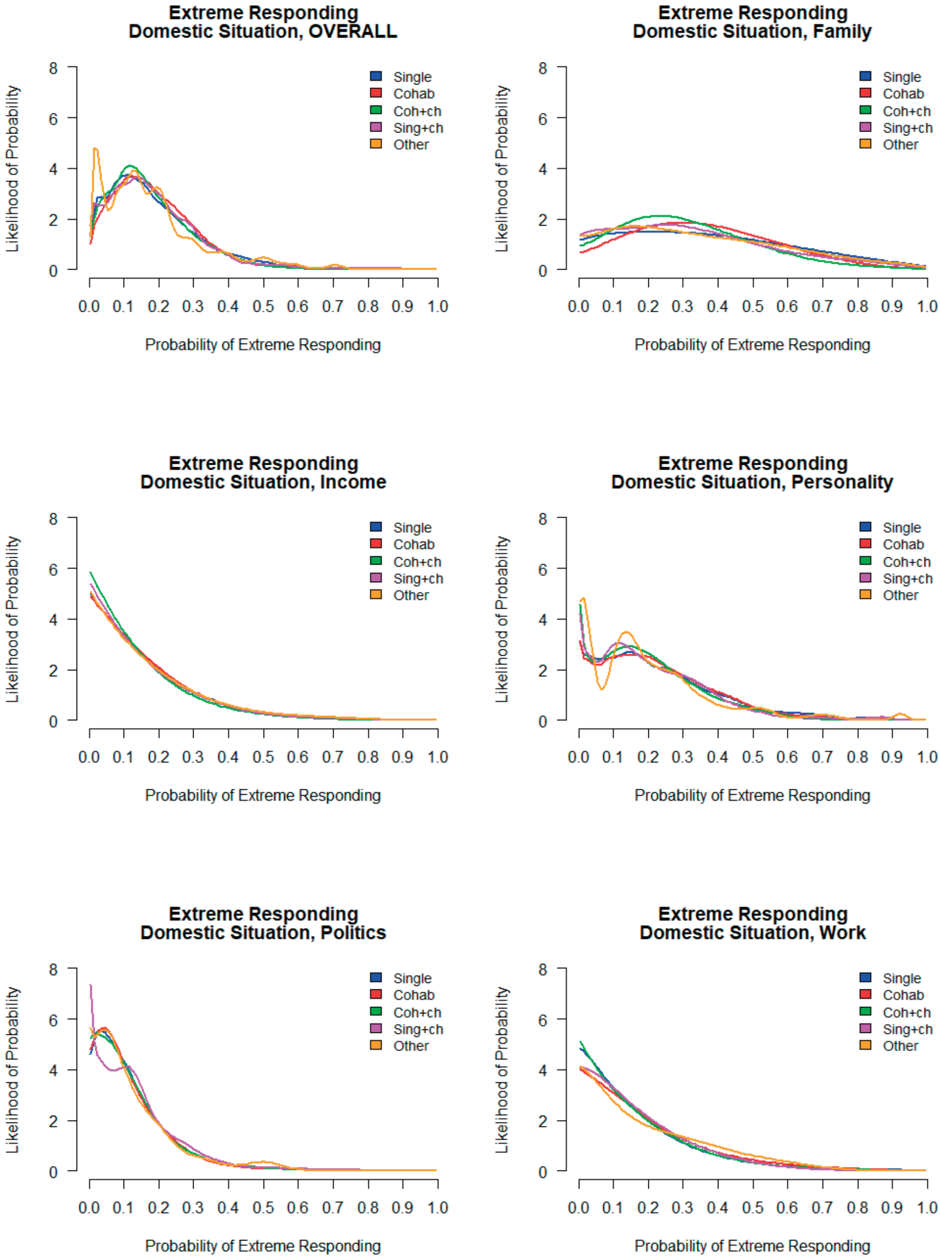


Figure K.52. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: PRIMARY OCCUPATION

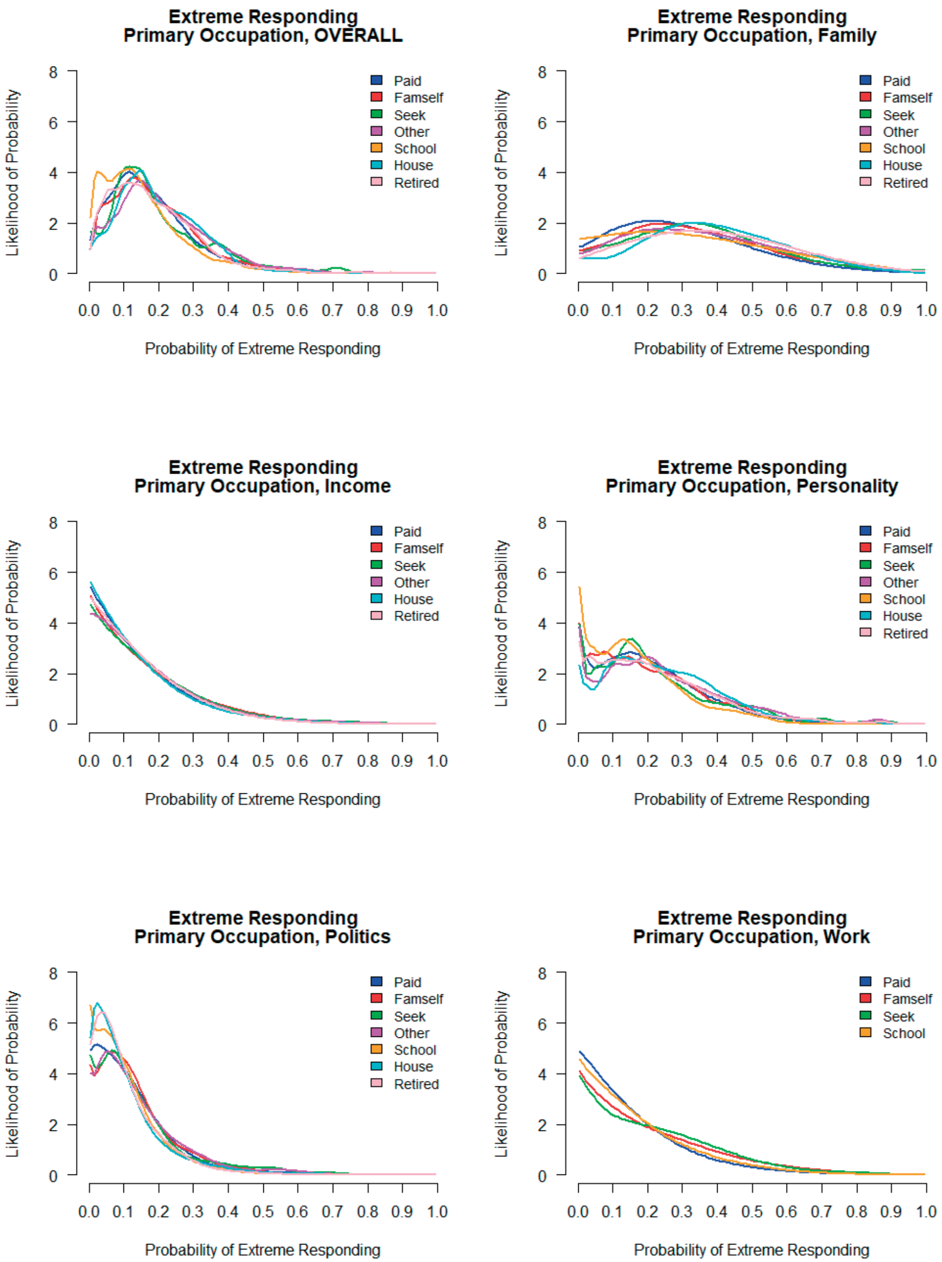


Figure K.53. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: INCOME

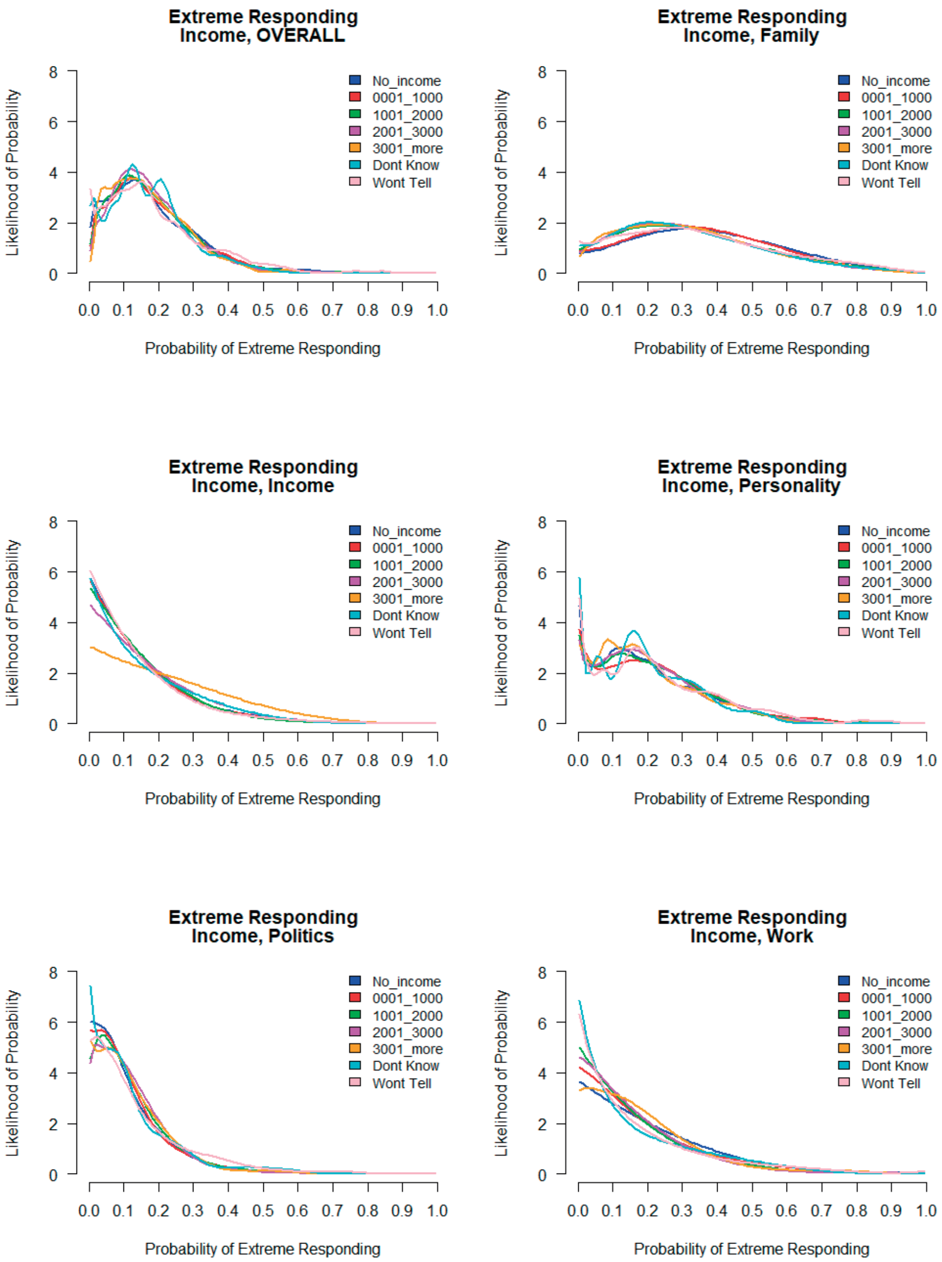


Figure K.54. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: ORIGIN

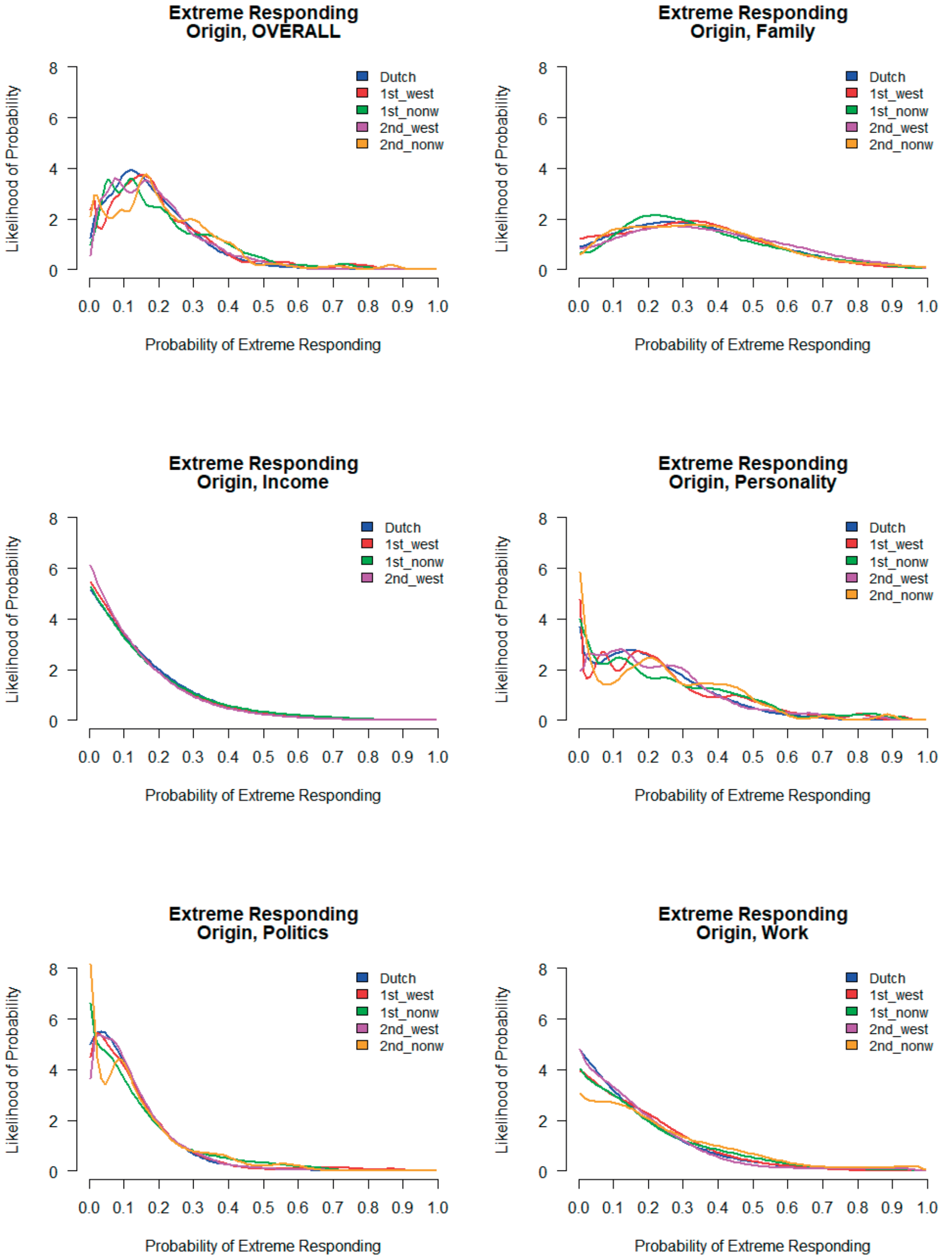


Figure K.55. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

EXTREME RESPONDING: PROVIDED A PC

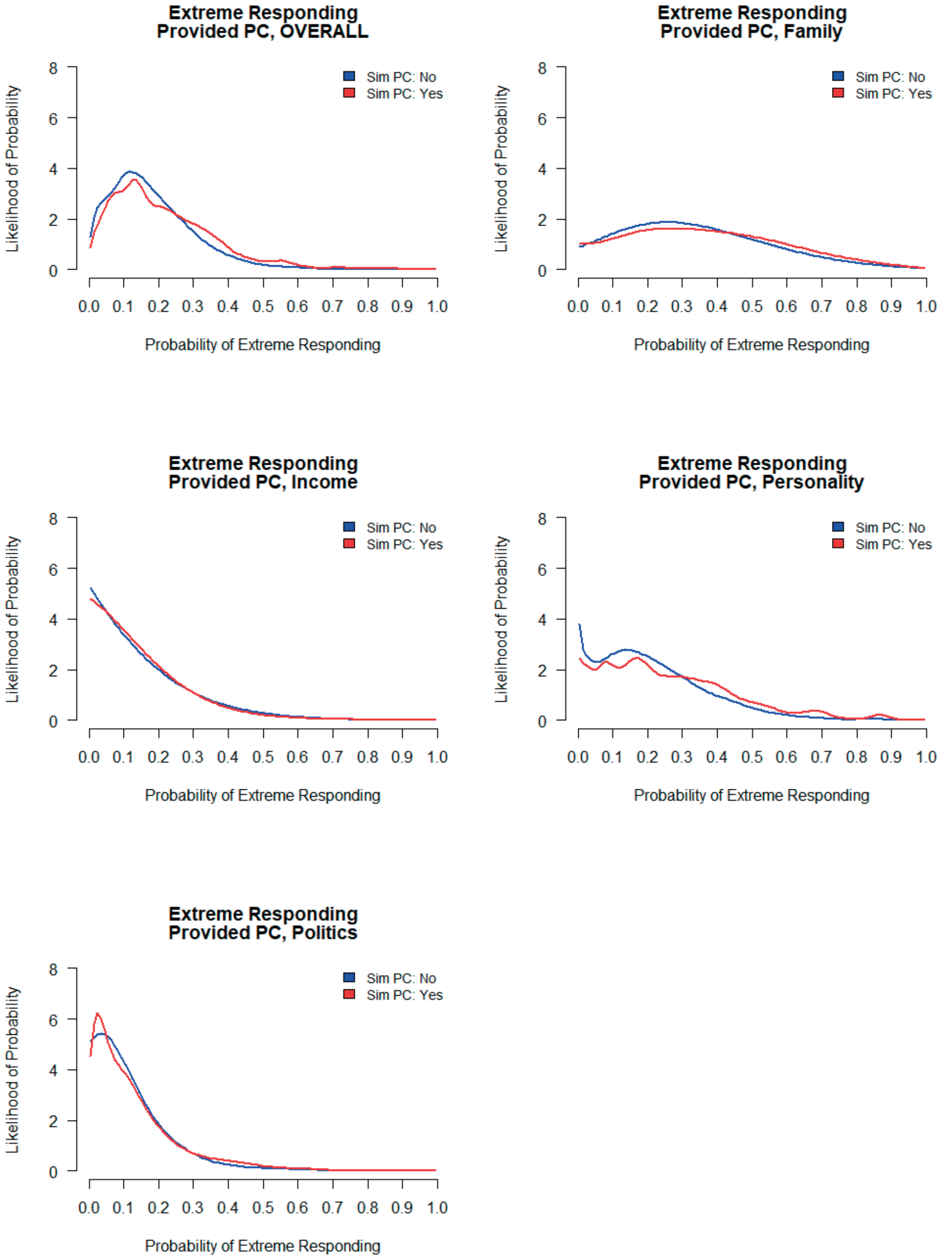


Figure K.56. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Extreme Responding'** Overall and for All Applicable Surveys*.

* Family and Household (FA), Income (IN), Personality (PE), Politics and Values (PO), Work and Schooling (WO).

ANSWERING WON'T TELL: GENDER

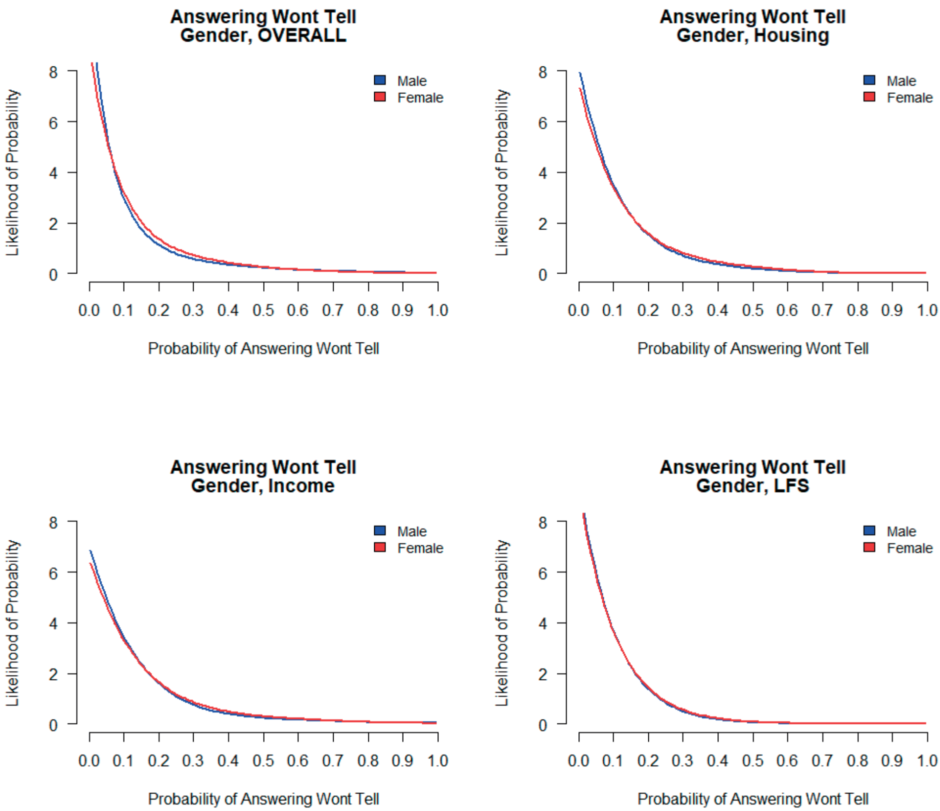


Figure K.57. Behaviour Profiles for the Categories of the Characteristic **Gender** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: AGE

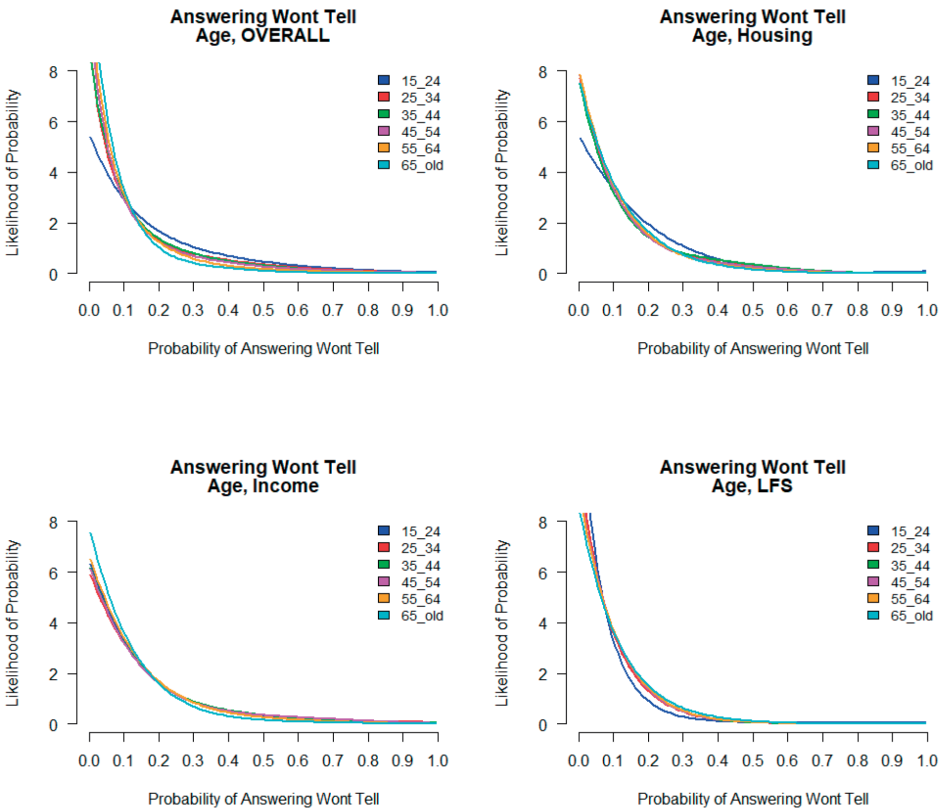


Figure K.58. Behaviour Profiles for the Categories of the Characteristic **Age** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: EDUCATION

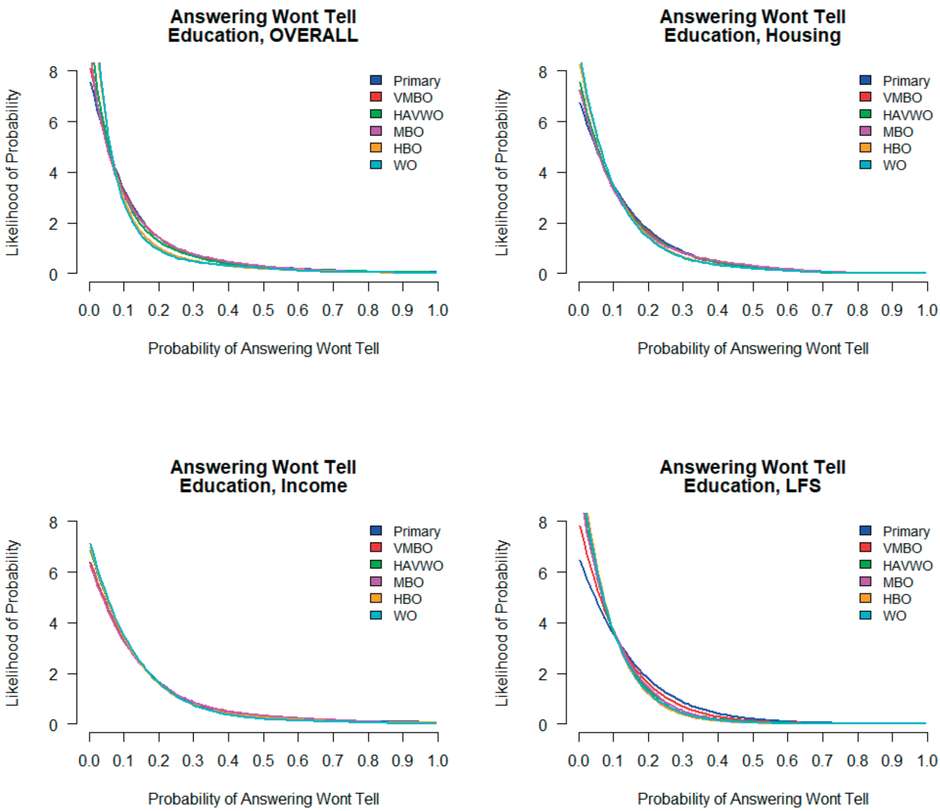


Figure K.59. Behaviour Profiles for the Categories of the Characteristic **Education** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: DOMESTIC SITUATION

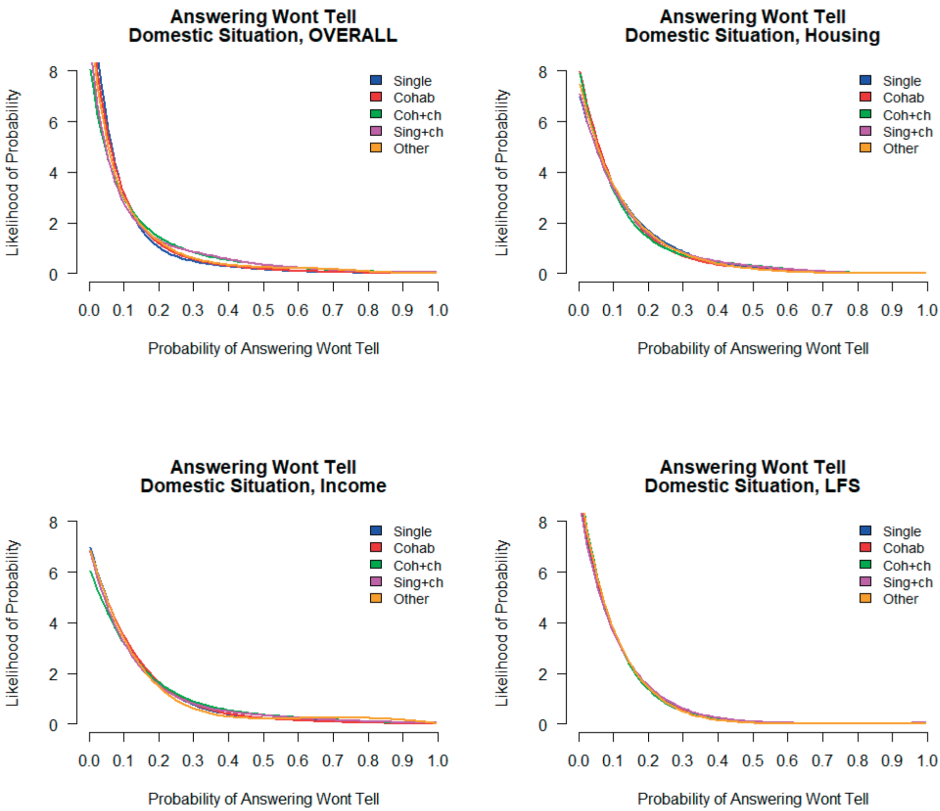


Figure K.60. Behaviour Profiles for the Categories of the Characteristic **Domestic Situation** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: PRIMARY OCCUPATION

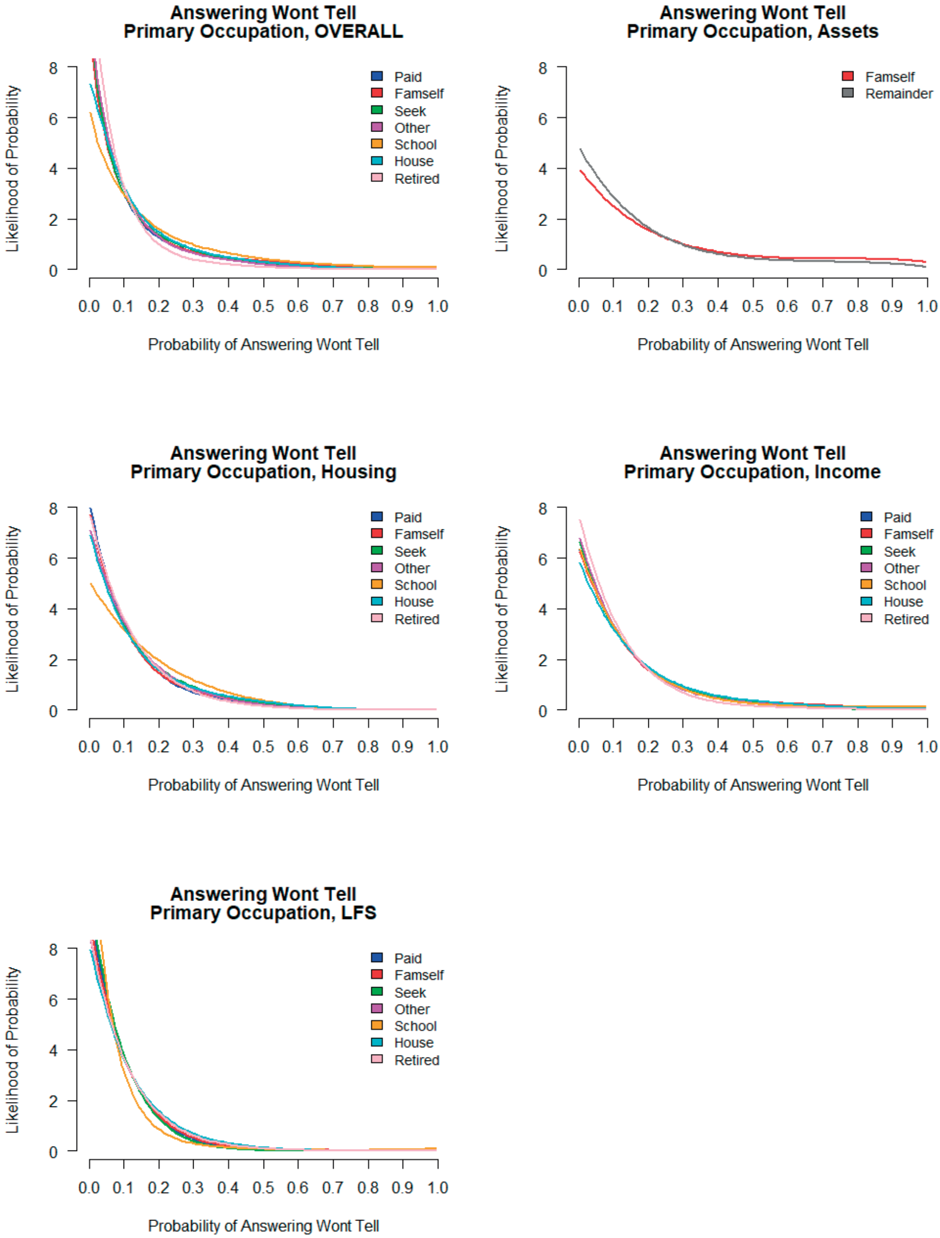


Figure K.61. Behaviour Profiles for the Categories of the Characteristic **Primary Occupation** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: INCOME

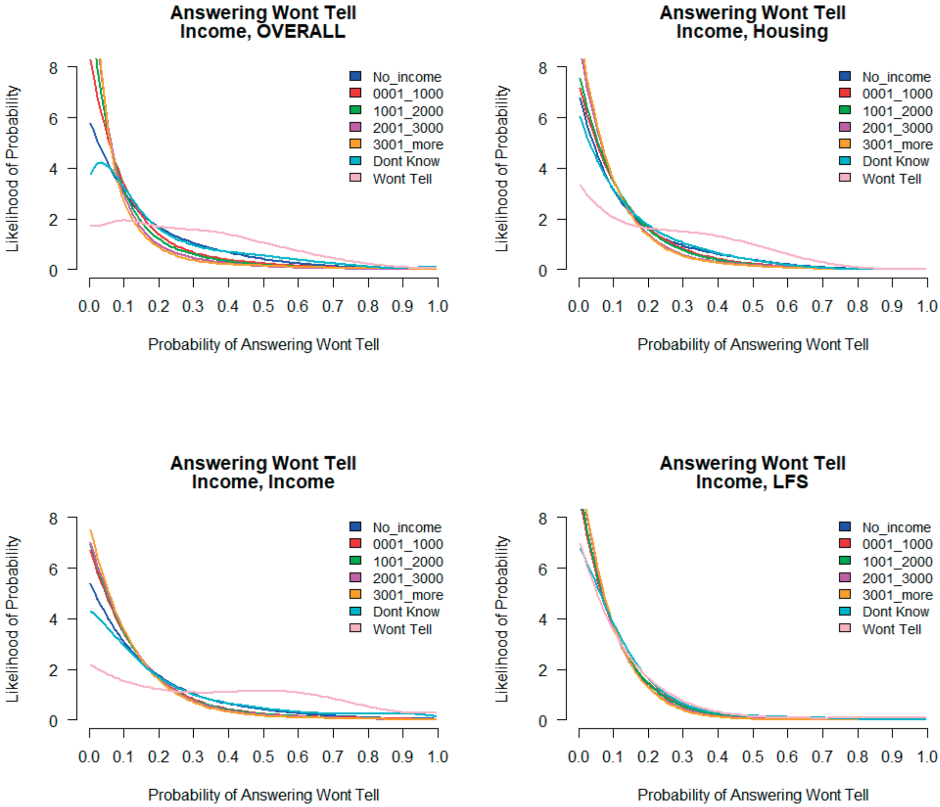


Figure K.62. Behaviour Profiles for the Categories of the Characteristic **Income** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: ORIGIN

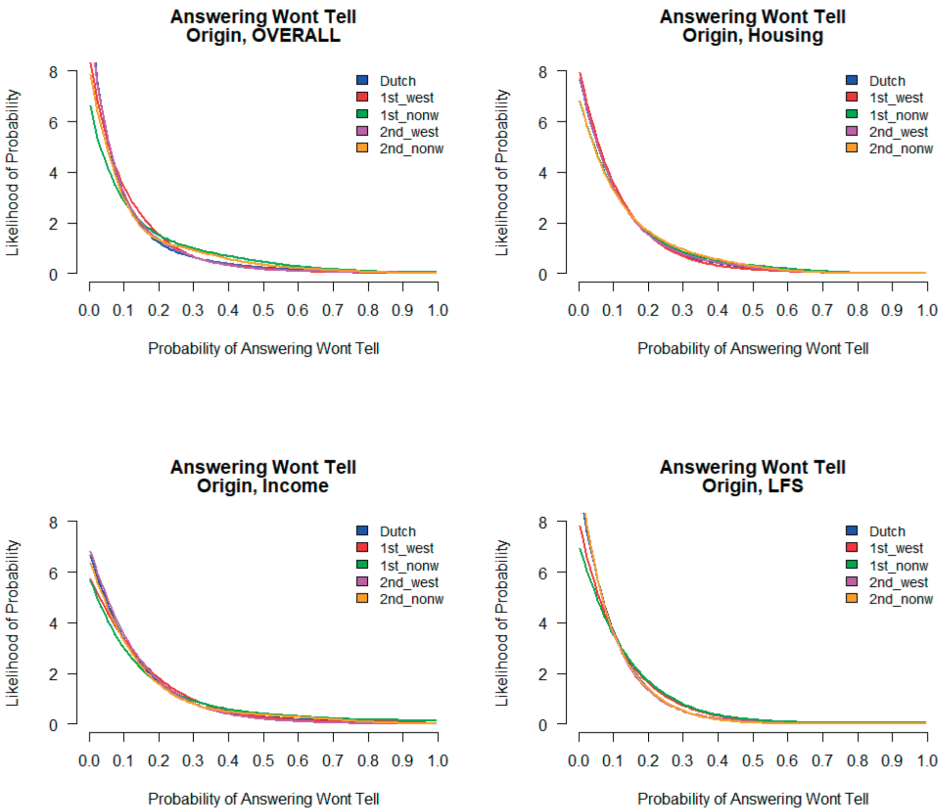


Figure K.63. Behaviour Profiles for the Categories of the Characteristic **Origin** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

ANSWERING WON'T TELL: PROVIDED A PC

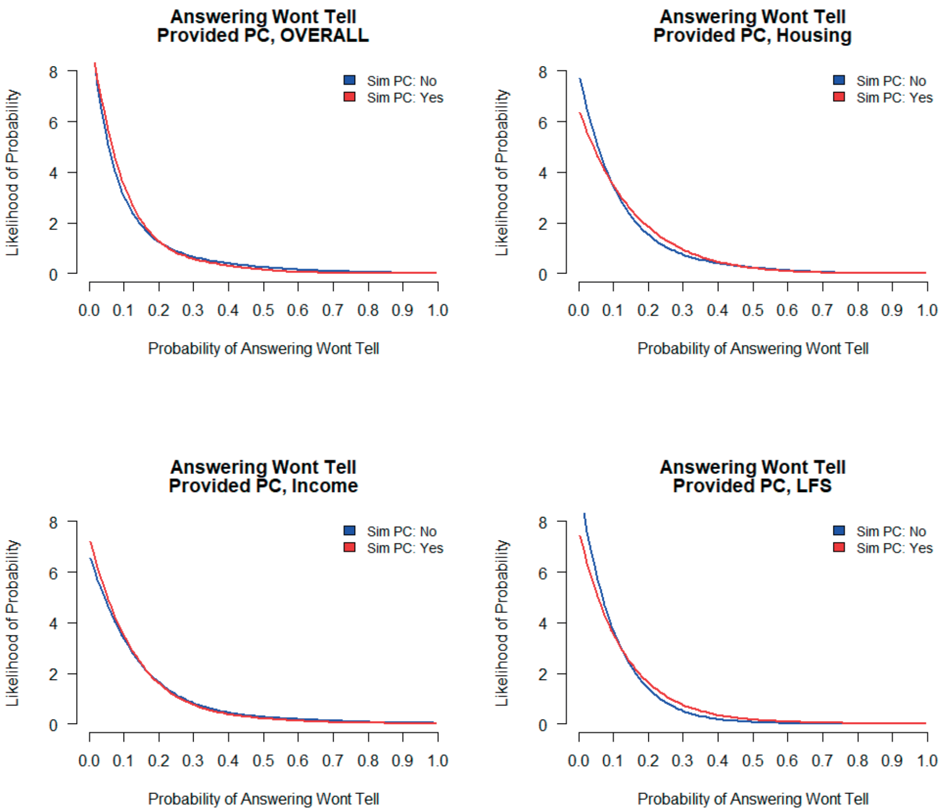


Figure K.64. Behaviour Profiles for the Categories of the Characteristic **Provided a PC** for the Answer Behaviour **'Answering Won't Tell'** Overall and for All Applicable Surveys*.

* Assets (AS), Housing (HO), Income (IN), Work and Schooling (WO), Labour Force Survey (LF).

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

Bibliography

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research, 42*(3), 957-970. https://doi.org/10.1016/j.ssresearch.2013.01.002

Achterberg, G. (1944). *Eurydice*. Uitgever: Het Spectrum.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods & Research, 20*(1), 139–181. https://doi.org/10.1177/0049124191020001005

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. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly, 48*(2), 409-442. https://doi.org/10.1086/268840

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. https://www.jstor.org/stable/2289229

Antoni, M., Bela, D., & Vicari, B. (2019). Validating earnings in the German national educational 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*(3), 624–639. https://doi.org/10.1086/228120

Bachman, J. G., & O'Malley, P. M. (1984b). Yea-saying, nay-saying, and going to extremes: Black–White differences in response styles. *Public Opinion Quarterly*, *48*(2), 491–509. https://doi.org/10.1086/268845

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M., & Vis, C. (2019). 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), 263-295. https://doi.org/10.1177/0049124117729692

Bais, F., Schouten, B., & Toepoel, V. (2020). Investigating response patterns across surveys: Do respondents show consistency in undesirable answer behaviour over multiple surveys? *Bulletin de Méthodologie Sociologique, 147-148*(1-2), 150-168. https://doi.org/10.1177/0759106320939891

Bais, F., Schouten, B., & Toepoel, V. (2021). Translating survey answer behaviour into meaningful summaries: The concept and estimation of survey behaviour profiles. CBS Discussion paper. Statistics Netherlands, The Hague.

Bais, F., Schouten, B., & Toepoel, V. (under review for Survey Methodology). Is undesirable answer behaviour consistent across surveys? An investigation into respondent characteristics.

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica, 66*(1), 8–17. https://doi.org/10.1111/j.1467-9574.2011.00504.x

Barnas, M. (2003). *Twee zonnen*. Uitgever: De Arbeiderspers.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390-399. https://doi.org/10.1086/297760

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

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* (pp. 71-86). Wiley.

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). Routledge.

Bernardi, R. A. (2006). Associations between Hofstedes cultural constructs and social desirability response bias. *Journal of Business Ethics, 65*(1), 43-53. https://doi.org/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*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Binswanger, J., Schunk, D., & Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly, 77*(3), 783–797. https://doi.org/10.1093/poq/nft030

Bishop, G. F., Oldendick, R. W., & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly, 47*(4), 528-546. https://doi.org/10.1086/268810

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: the pressure to answer survey questions. *Public Opinion Quarterly, 50*(2), 240–250. https://www.jstor.org/stable/2748887

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*(5), 335-352. https://doi.org/10.1177/0146621608329891

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., & Associates. (1979). *Improving interview method and questionnaire design.* Jossey-Bass.

Bradburn, N., Sudman, S., Blair, E., & Stocking, C. (1978). Question threat and response bias. *Public Opinion Quarterly, 42*(2), 221-234. https://doi.org/10.1086/268444

Buelens, B., & Van den Brakel, J. A. (2011). Inference in surveys with sequential mixed-mode data collection. CBS Discussion paper 2011-21. Statistics Netherlands, DMH, Heerlen.

Butler R. J., & McDonald, J. B. (1987). Interdistributional income inequality. *Journal of Business & Economic Statistics, 5*(1), 13-18. https://doi.org/10.2307/1391210

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131. https://doi.org/10.1037/0022-3514.42.1.116

Calinescu, M., & Schouten, B. (2016). Adaptive survey designs for nonresponse and measurement error in multi-purpose surveys. *Survey Research Methods, 10*(1), 35–47. https://doi.org/10.18148/srm/2016.v10i1.6157

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.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Cannell, C., Marquis, K., & Laurent, A. (1977). A summary of studies of interviewing methodology. *Vital and Health Statistics, 2*(69), 1–78.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 389-437). Jossey-Bass.

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*(3), 170–175. https://doi.org/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*(2), 187-212. https://doi.org/10.1177/0022022100031002003

Cho, Y. I., Fuller, A., File, T., Holbrook, A. L., & Johnson, T. (2006). Culture and survey question answering: A behavior coding approach. In Proceedings of the Survey Research Methods Section of the American Statistical Association, 4082-4089.

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*(4), 465-480. https://doi.org/10.1177/002202217400500407

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494-509. https://doi.org/10.1037/0033-2909.114.3.494

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*(3), 331-350. DOI: 10.1207/s15327906mbr3103_4

Cliff, N. (1996b). *Ordinal methods for behavioral data analysis*. New Jersey: Lawrence Erlbaum Associates.

Cole, J., McCormick, A. C., & Gonyea, R. M. (2012). Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications. Paper presented at the annual meeting of the American Educational Research Association, April 17, Vancouver, Canada.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys: A total survey error perspective. In P. Biemer, S. Eckman, B. Edwards, E. D. de Leeuw, F. Kreuter, L. Lyberg, C. Tucker, & B. West (Eds.), *Total survey error in practice* (pp. 133-154). Wiley.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 271-286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Daikeler, J., Bach, R. L., Silber, H., & Eckman, S. (2020). Motivated misreporting in smartphone surveys. *Social Science Computer Review*. February. https://doi.org/10.1177/0894439319900936

De Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research, 45*(1), 104-115. https://doi.org/10.1509/jmkr.45.1.104

De Leeuw, E. D. (1992). Data quality in mail, telephone, and face-to-face surveys. Amsterdam: TT-Publicaties.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(2), 233–255.

De Leeuw, E. D., Hox, J. J., & Dillman, D. (2008). *International handbook of survey methodology*. Taylor & Francis Group.

Dekkers, M. (2018). *Volledige vergunning*. Uitgever: Bas Lubberhuizen.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics – Understanding why and how*. Springer. https://doi.org/10.1007/1-84628-168-7

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

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*(4), 337–348. https://doi.org/10.1080/13645579.2014.895289

Dijkstra, W. (1994). Sequence: A program for analysing sequential data. *Bulletin de Méthodologie Sociologique, 43*(1), 134-144. https://doi.org/10.1177/075910639404300111

Dolnicar, S., & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review, 24*(2), 127-143. https://doi.org/10.1108/02651330710741785

Duan, N., Alegria, M., Canino, G., McGuire, T. G., & Takeuchi, D. (2007). Survey conditioning in self-reported mental health service use: Randomized comparison of alternative instrument formats. *Health Services Research, 42*(2), 890-907. https://doi.org/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*(3), 721-733. https://doi.org/10.1093/poq/nfu030

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review. Research report RR791. London: Department for Education and Skills. URL: http://www.bristol.ac.uk/cmm/team/cross-classified-review.pdf

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382. https://doi.org/10.1037/h0031619

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511518201

Fowler, F. J., Jr. (1995). *Applied social research methods series, Vol. 38. Improving survey questions: Design and evaluation*. Sage Publications, Inc.

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. Sage Publications, Inc.

Fricker S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392. https://www.jstor.org/stable/3521546

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

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*(5), 892–913. https://doi.org/10.1093/poq/nfn059

Gallhofer, I. N., & Saris, W. E. (2000). Formulierung und Klassifikation von Fragen. *ZUMA Nachrichten, 46*, 43-72. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-211156

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

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*(3), 328–351. https://doi.org/10.1086/269326

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing ‘too fast’ responses and respondents from web surveys. *Public Opinion Quarterly, 79*(2), 471–503.

Gummer T., & Roßmann, J. (2014). Explaining interview duration in web surveys: A multilevel approach. *Social Science Computer Review, 33*(2), 217–234. https://doi.org/10.1177/0894439314533479

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243-266. https:doi.org/10.1177/1470595806066332

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*(7), 794–800. https://doi.org/10.1016/j.paid.2013.06.017

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*(3), 306-322. DOI: 10.1177/1470595814541424

Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics, 27*(1), 49-63.

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., Cho, Y. I., & Johnson, T. P. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. *Public Opinion Quarterly, 70*(4), 565-595. https://doi.org/10.1093/poq/nfl027

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*(1), 79–125. https://www.jstor.org/stable/3521667

Hox, J. J. (1995). Covariance structure modeling in Windows: A multitrait-multimethod analysis using Amos, Eqs, and Lisrel. *Bulletin de Méthodologie Sociologique, 46*, 71–87. https://doi.org/10.1177/075910639504600106

Hox, J. J . (2010). *Quantitative methodology series. Multilevel analysis: Techniques and applications*. Routledge/Taylor & Francis Group.

Hox, J. J., De Leeuw, E. D., & 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). Wiley.

Hui, C. H., & Triandis, H. C. (1989), Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review, 78*(1), 3–20. https://doi.org/10.1111/j.1751-5823.2010.00102.x

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. doi: 10.7892/boris.133689

Jenkins, C. R., & D. A. Dillman (1997). Towards a theory of self-administered questionnaire design. In L. E. Lyberg, P. P. Biemer, M. Collins, L. Decker, E. D. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 165-196). Wiley.

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*(6), 439–445. https://doi.org/10.1023/A:1021927909027

Johnson, T. P., & Van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. Harness, F. J. R. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). Wiley.

Joshi, A. (2015). Why and when to use profile likelihood based confidence intervals. Retrieved from https://www.cytel.com/blog/why-and-when-to-use-profile-likelihood-based-confidence-intervals

Juster, F. T., & Smith, J. P. (1997). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association, 92*(440), 1268-1278. https://doi.org/10.1080/01621459.1997.10473648

Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician, 29*(1), 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*(5), 956-984. https://doi.org/10.1093/poq/nfq062

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution.* Oxford University Press.

Kellogg, R. T. (2007). *SAGE cognitive psychology program. Fundamentals of cognitive psychology*. 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 (NCS). *International Journal of Methods in Psychiatric Research, 7*(1), 33-55. https://doi.org/10.1002/mpr.33

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality & Quantity: International Journal of Methodology, 47*(1), 193-211. doi: 10.1007/s11135-011-9511-4

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research, 42*(3), 227-263. https://doi.org/10.1177/0049124113500480

Knäuper, B. (1998). Filter questions and question interpretation: Presuppositions at work. *Public Opinion Quarterly, 62*(1), 70–78. https://doi.org/10.1086/297832

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research, 40*(1), 88-104. https://doi.org/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*(5), 847-865. https://doi.org/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*(3), 213-236. https://doi.org/10.1002/acp.2350050305

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). Springer.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201–219. https://doi.org/10.1086/269029

Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology, 57*(3), 416-425. https://doi.org/10.1037/0022-3514.57.3.416

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. D. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141-164). Wiley.

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*(3), 371–403. https://doi.org/10.1086/341394

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 70*, 29–44. https://doi.org/10.1002/ev.1033

Kwak, N., & Radler, B. T. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18*(2), 257–273.

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 https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/2010AAPORCellPhoneTFReport.pdf

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.

Leckie, G. (2013). Cross-classified multilevel models - Concepts. LEMMA VLE Module 12, 1-60. http://www.bristol.ac.uk/cmm/learning/course.html

Leckie, G. (2019). Multiple membership multilevel models. arXiv:1907.04148 [stat.ME]. https://arxiv.org/ftp/arxiv/papers/1907/1907.04148.pdf

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*(4), 418–424. https://www.jstor.org/stable/3151390

Lensvelt-Mulders, G. J. L. M. (2008). Surveying sensitive topics. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman, *International handbook of survey methodology* (pp. 461-478). Taylor & Francis, Psychology Press, EAM series.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology, 24*(7), 1003-1020. https://doi.org/10.1002/acp.1602

Lozar Manfreda, K., & Vehovar, V. (2002). Mode effect in web surveys. In Proceedings from the American Association for Public Opinion Research (AAPOR), 57th Annual Conference.

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. https://doi.org/10.1023/A:1021975824957

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review, 34*(1), 78-94. https://doi.org/10.1177/0894439315574248

Lynn, P., & Kaminska, O. (2013). The impact of mobile phones on survey measurement error. *Public Opinion Quarterly, 77*(2), 586–605. https://doi.org/10.1093/poq/nfs046

Maas, C. J. M., Lensvelt-Mulders, G. J. L. M., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*(3), 72–77. https://doi.org/10.1027/1614-2241.5.3.72

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly, 72*(5), 914-934. https://doi.org/10.1093/poq/nfn050

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology, 23*(4), 498–509. https://doi.org/10.1177/0022022192234006

Marshall, R., & Lee, C. (1998). A cross-cultural, between-gender study of extreme response style. *European Advances in Consumer Research*, *3,* 90-95.

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 & Research, 20*(1), 60-103. https://doi.org/10.1177/0049124191020001003

Medway, R., & Tourangeau, R. (2015). Response quality in telephone surveys. Do prepaid cash incentives make a difference? *Public Opinion Quarterly, 79*(2), 524–543. https://doi.org/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*(7), 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 (Ed.), *Response set in personality assessment* (pp. 115–145). Aldine.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60*(1), 58–88. https://doi.org/10.1086/297739

Oksenberg, L., & Cannell, C. F. (1977). Some factors underlying the validity of response in self-report. *Bulletin of the International Statistical Institute*, 325-346.

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

Olson, K., Smyth, J. D., & Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology, 7*(2), 275- 308. https://doi.org/10.1093/jssam/smy006

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). Middle alternatives, acquiescence, and the quality of questionnaire data. Working Papers 0103, Harris School of Public Policy Studies, University of Chicago.

Ongena, Y. P., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics, 22*(3), 419-451.

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). 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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Sage Publications.

Redline, C. D., & 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* (pp. 179-193). Wiley.

Revilla, M., & Ochoa, C. (2014). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review, 33*(1), 97–114. https://doi.org/10.1177/0894439314531214

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 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. University of Essex.

Roßmann, J. (2017). *Satisficing in Befragungen. Theorie, Messung und Erklärung*. Springer.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology, 6*(3), 376-400. doi: 10.1093/jssam/smx020

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*(9), 2647–2651. https://doi.org/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, 46*(2), 1738-1748. https://doi.org/10.1111/ejn.13610

Saris, W. E., & Gallhofer, I. N. (2007a). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley.

Saris, W. E., & Gallhofer, I. N. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods, 1*(1), 29-43. https://doi.org/10.18148/srm/2007.v1i1.49

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. https://doi.org/10.18148/srm/2010.v4i1.2682

Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology, 34*, 311–347. https://doi.org/10.1111/j.0081-1750.2004.00155.x

Schaeffer, N. C. (2000). Asking questions about threatening topics: A selective overview. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 105-121). Lawrence Erlbaum Associates Publishers.

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

Scherpenzeel, A. C. (1995). A question of quality. Evaluating survey questions by multitrait-multimethod studies [Doctoral dissertation, University of Amsterdam]. Leidschendam: PTT Nederland NV, KPN Research. https://hdl.handle.net/11245/1.113408

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research, 25*(3), 341-383. https://doi.org/10.1177/0049124197025003004

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*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., & Berry, S. H. (2004). A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review, 22*(1), 128-138. https://doi.org/10.1177/0894439303256551

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 (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 231-258). Wiley. https://doi.org/10.1002/9781118596869.ch10

Schouten, B., Calinescu, M., & Luiten, A. (2013a). Optimizing quality of response through adaptive survey designs. *Survey Methodology, 39*(1), 29–58.

Schouten, B., A. Peytchev, and J. Wagner (2017). *Adaptive survey design*. Chapman and Hall.

Schouten, B., Van den Brakel, J. A., Buelens, B., Van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*(6), 1555-1570. https://doi.org/10.1016/j.ssresearch.2013.07.005

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 & Research, 33*(1), 118–156. https://doi.org/10.1177/0049124103262689

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. 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*(3), 193–212. https://doi.org/10.1002/acp.2350050304

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*(2), 193-201. https://doi.org/10.1093/ijpor/14.2.193

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*(3), 218-230. https://doi.org/10.1108/eb028885

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall/CRC.

Simon, H. A. (1957). *Models of man.* Wiley.

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

Smyth, J. D., & Olson, K. (2019). The effects of mismatches between survey question stems and response options on data quality and responses. *Journal of Survey Statistics and Methodology, 7*(1), 34-65. https://doi.org/10.1093/jssam/smy005

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *The Journal of Social Psychology, 122*(2), 151–156. https://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*(3), 121-138. doi: 10.18148/srm/2007.v1i3.600

Stricker, L. J. (1963). Acquiescence and social desirability response styles, item characteristics, and conformity. *Psychological Reports, 12*, 319–341. https://doi.org/10.2466/pr0.1963.12.2.319

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. Jossey-Bass.

Tarnai, J., & Dillman, D. A. (1992). Questionnaire context as a source of response differences in mail versus telephone surveys. In N. Schwarz & S. Sudman, *Context effects in social and psychological research*. Springer Verlag.

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods, 50*, 2325-2344. https://doi.org/10.3758/s13428-017-0997-0

Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of design in web surveys: Comparing trained and fresh respondents. *Public Opinion Quarterly, 72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). National Academy Press.

Tourangeau, R. (1987). Attitude measurement: A cognitive perspective. In H. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 149-162). Springer-Verlag. https://doi.org/10.1007/978-1-4612-4798-2_8

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314. https://doi.org/10.1037/0033-2909.103.3.299

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

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. https://doi.org/10.1037/0033-2909.133.5.859

Van der Ploeg, C. (2008). A comparison of different estimation methods of voting transitions with an application in the Dutch national elections. Vrije Universiteit, Amsterdam.

Van der Ploeg, C., Van de Pol, F., & Kampen, J. (2008). Verschuivingen in partijvoorkeur tussen 2003 en 2006. In H. Schmeets & R. van der Bie (Eds.), *Het Nationaal Kiezersonderzoek 2006. Opzet, uitvoering en resultaten* (pp. 157-167), Statistics Netherlands, Voorburg/Heerlen.

Van der Ploeg, C., Van de Pol, F., & Kampen, J. (2009). A comparison of different estimation methods with an application in the Dutch national elections of 2003 and 2006. Discussion paper 09030, Statistics Netherlands, The Hague/Heerlen.

Van der Vaart, W. (1996). *Inquiring into the past. Data quality of responses to retrospective questions* [Doctoral dissertation, Vrije Universiteit Amsterdam]. Universal Press.

Van der Vaart, W., Van der Zouwen, J., & Dijkstra, W. (1995). Retrospective questions: Data quality, task difficulty, and the use of a checklist. *Quality and Quantity, 29*(3), 299-315. https://doi.org/10.1007/BF01101976

Van der Zouwen, J. (2000). An assessment of the difficulty of questions used in the ISSP-questionnaires, the clarity of their wording, and the comparability of the responses. *ZA-Informationen, 46*, 96-114. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-199369

Van der Zouwen, J., & Dijkstra, W. (1996). The impact of the question on the interactions in survey-interviews. Paper presented at the Fourth International ISA Conference on Social Science Methodology, University of Essex, Colchester, UK.

Van der Zouwen, J., & Dijkstra, W. (1998). Het vraaggesprek onderzocht. Wat zegt het verloop van de interactie in survey-interviews over de kwaliteit van de vraagformulering? *Sociologische gids, 45*, 387-403.

Van der Zouwen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behavior codes and transcripts of question-answer sequences: A diagnostic approach. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 109-130). Wiley. https://doi.org/10.1002/0471654728.ch6

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*(3), 346-360. https://doi.org/10.1177/0022022104264126

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research, 47*(1), 157–172. https://doi.org/10.1509/jmkr.47.1.157

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly, 74*(5), 1027-1045. https://doi.org/10.1093/poq/nfq059

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 and Behavioral Statistics, 25*(2), 101–132. doi: http://doi.org/10.2307/1165329

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society Series C, 37*(1), 87-94.

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, DMH, Heerlen.

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*(5), 490–504. https://doi.org/10.1177/0022022195265003

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73-95). Wiley & Sons. https://doi.org/10.1002/9781118596869.ch4

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*(1), 51–68. https://doi.org/10.1002/acp.1331

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. https://doi.org/10.1093/poq/nfr009

Zax, M., & Takahashi, S. (1967). Cultural influences on response style: Comparisons of Japanese and American college students. *The Journal of Social Psychology*, *71*(1), 3-10. https://doi.org/10.1080/00224545.1967.9919760

Zhang, C. (2013). Satisficing in web surveys: Implications for data quality and strategies for reduction [Doctoral dissertation, 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*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*(2), 168–190. https://doi.org/10.1177/1094428104263674

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

ENGLISH SUMMARY

Surveys are administered to respondents to describe and analyze facts and attitudes about various personal and societal topics for the population. Answers to survey questions do not necessarily correspond to reality and may deviate from the truth for several reasons. For instance, respondents who have a lower cognitive ability may not give optimal responses to items that demand some level of cognitive processing. Or items that ask for sensitive information may not result in substantive answers, as respondents are not willing to reveal such information. We name such suboptimal or non-substantive responses ‘undesirable answer behaviour’. The extent to which responses deviate from the ‘true’ value is referred to as ‘measurement error’. This means that undesirable answer behaviour is an indicator of measurement error. As these examples illustrate, both characteristics of respondents and characteristics of items may lead to undesirable answer behaviour and hence measurement error, resulting in lower data quality. In the central part of this thesis, we introduced the concept of ‘behaviour profiles’ and used such profiles to investigate the relation between respondent characteristics and undesirable answer behaviour. The purpose of this thesis was to detect respondent characteristics that are structurally related to measurement error.

In chapter 2, we anticipated on examining the association between item characteristics and undesirable answer behaviour. We investigated to what degree characteristics of items could be determined reliably. For this purpose, more than 2000 items from 11 population surveys were coded on a large selection of item characteristics by survey researchers. Nine item characteristics were coded by two coders: ‘Time reference’, ‘conditions’, ‘memory’, ‘hypothetical situation’, ‘calculations’, ‘ambiguity’, ‘mismatch’, ‘formulation’, and ‘clarification’. The other seven item characteristics appeared to be difficult to code during a pilot study and were therefore coded by three coders: ‘Content of the question’, ‘difficult language usage’, ‘emotional charge’, ‘presumption of a filter question’, ‘sensitive information’, ‘centrality’, and ‘response complexity’. For most item characteristics, the coders needed to determine whether the characteristic was either absent or present in eligible items. The results showed that intercoder reliability was low for various item characteristics: Formulation, clarification, content of the question, difficult language usage, emotional charge, presumption of a filter question, sensitive information, and centrality. We attributed this low intercoder reliability for a substantial part to the subjective personal interpretation of determining the presence of the item characteristic. To cope with low intercoder reliability, we made four suggestions: 1) Excluding survey items for which the applicable category (mostly absent or present) could not be determined unambiguously; 2) redefining and refining the item characteristics that could not be coded reliably; 3) computerizing the definition of item characteristics and the applicable category for specific item characteristic content, and; 4) for item characteristics coded by three coders, taking into account the number of coders who determined the characteristic as present.

In chapter 3, we investigated a large selection of answer behaviours on their occurrence and variance between respondents, surveys, and their interaction to explore their consistency potential across ten population surveys. The selection consisted of 12 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’, ‘straightlining’, ‘fast responding’, and ‘slow responding’. Regarding the occurrences, the outcomes showed that respondents avoided follow-up questions and responded socially desirable frequently overall. Respondents answered ‘don’t know’ or ‘won’t tell’ only occasionally overall. Regarding the variances, we theorized that large respondent variance combined with small respondent-survey interaction variance indicated behaviour consistency across surveys. Especially when respondents gave a ‘won’t tell’-answer frequently, this was indicated to occur consistently across surveys. To a lower degree, such consistency may be indicated for respondents who answered fast or slowly, or who gave ‘don’t know’-answers or extreme answers frequently. Based on the varying occurrences and variances for the different surveys and behaviours in this study, we chose to switch to a method that summarizes and visualizes answer behaviour for each survey and behaviour in detail. This method is new in analyzing survey answer behaviour: The construction of ‘answer behaviour profiles’.

In chapter 4, we introduced and explored our method of constructing behaviour profiles to anticipate on examining the relation between respondent characteristics and answer behaviour. We defined the behaviour profile as a density distribution that summarizes answer behaviour for a group of respondents (the ‘respondent profile’) or a group of items (the ‘item profile’). The behaviour profile shows the likelihood of each behaviour probability from 0 to 1. The respondent profile is the averaged profile for individual respondent profiles that are each based on a specific number of items. The item profile is the averaged profile for individual item profiles that are each based on a specific number of respondents. By means of simulations, we examined the statistical properties of the behaviour profiles. We showed that a respondent profile becomes more informative and precise as each respondent fills out more items. And we showed that a respondent profile becomes more stable and certain as the group consists of a larger number of respondents. When a respondent profile is both precise and certain, the respondent profile is accurate. The very same applies to item profiles, with the concepts of 'respondent' and 'item' interchanged. Respondent and item profiles can be based on both respondent and item characteristics. When a behaviour profile is precise but not certain, the profile may be peaked. We showed a simple form of smoothing in order to smoothen a peaked profile. Subsequently, we used real survey data to illustrate the properties of respondent profiles, item profiles, and smoothing in practice. The next step is to investigate the relation between respondent characteristics and undesirable answer behaviour by using respondent profiles.

In chapter 5, we aimed at answering our main research question by examining the potentially consistent relation between respondent characteristics and undesirable answer behaviour across ten population surveys. For this purpose, we constructed respondent profiles for subgroups of respondents that are based on the categories of eight respondent characteristics: 'Age', 'educational level', 'gender', 'domestic situation', 'primary occupation', 'income', 'origin', and 'borrowing a sim PC'. These characteristics were investigated on their relation to eight specific undesirable answer behaviours: 'Socially desirable responding', 'answering don't know', 'answering won't tell', 'acquiescence', 'neutral responding', 'extreme responding', 'primacy responding', and 'straightlining'. The profile for each characteristics' category was compared to the profile for all other categories of that characteristic taken together. These comparisons were executed by an adaptation of the effect size measure Cliff's Delta. Although various categories of various characteristics appeared to be related to various behaviours, the occurrence and direction of these relations were substantially dependent on the specific characteristic, survey, and behaviour. This means that a subgroup of respondents frequently showed more of some behaviour than the other subgroups for one survey, while less of that behaviour than the other subgroups for another survey. Thus, we did not find consistent relations between any respondent characteristic and any answer behaviour.

In chapter 6, we resume the foundation and main results of this thesis, give suggestions for follow-up research, and discuss the limitations and strengths of our behaviour profile method. As we did not find consistent relations between respondent characteristics and answer behaviour, we cannot recommend the construction of respondent schemes and adaptive survey design from the perspective of consistent behaviour. We consider it likely that item characteristics have a more substantial and consistent influence on answer behaviour than respondent characteristics. Therefore, we recommend the use of item profiles to investigate this influence to detect characteristics of items that may lead to measurement error. In case of consistent relations between item characteristics and measurement error, item schemes could be constructed to guide adaptive survey design. We also make two suggestions for follow-up research on how to cope with item interdependency within surveys with respect to our method. We close chapter 6 by elaborating on the advantages of our method. Behaviour profiles 1) summarize and visualize answer behaviour for complete groups of respondents or items for the full probability range; 2) take into account the degree of precision of individual group members and the degree of certainty of the group, and; 3) can detect small subgroups of respondents or items that deviate from the main body of a larger group.

NEDERLANDSE SAMENVATTING
(Summary in Dutch)

Surveys worden afgenomen bij respondenten om feiten en attituden over verschillende persoonlijke en maatschappelijke onderwerpen voor de populatie te beschrijven en analyseren. Antwoorden op vragen van surveys komen niet per se overeen met de werkelijkheid en kunnen om verschillende redenen afwijken van de waarheid. Respondenten met een lager cognitief vermogen kunnen bijvoorbeeld mogelijk geen optimale antwoorden geven op items die een bepaalde cognitieve verwerking vereisen. Of items die om gevoelige informatie vragen leiden mogelijk niet tot inhoudelijke antwoorden, omdat respondenten niet welwillend zijn om zulke informatie te onthullen. We noemen zulke suboptimale of niet-inhoudelijke antwoorden ‘ongewenst antwoordgedrag’. De mate waarin antwoorden afwijken van het ‘ware’ antwoord wordt aangeduid als ‘meetfout’. Dit betekent dat ongewenst antwoordgedrag een indicator is van een meetfout. Zoals deze voorbeelden illustreren, kunnen zowel kenmerken van respondenten als kenmerken van items leiden tot ongewenst antwoordgedrag en dus meetfouten, met als gevolg lagere data-kwaliteit. In het centrale deel van deze these introduceren we het concept ‘gedragsprofielen’ en gebruiken we zulke profielen om de relatie tussen respondent-kenmerken en ongewenst antwoordgedrag te onderzoeken. Het doel van deze these was om kenmerken van respondenten te detecteren die een structurele relatie hebben met meetfouten.

In hoofdstuk 2 anticipeerden we op het onderzoeken van de associatie tussen kenmerken van items en ongewenst antwoordgedrag. We onderzochten in welke mate item-kenmerken betrouwbaar bepaald konden worden. Voor dit doeleinde werden meer dan 2000 items van 11 populatie-surveys gecodeerd door survey-onderzoekers op een grote selectie van item-kenmerken. Negen item-kenmerken werden gecodeerd door twee beoordelaars: ‘Tijd’, ‘condities’, ‘geheugen’, ‘hypothetische situatie’, ‘berekeningen’, ‘dubbelzinnigheid’, ‘mismatch’, ‘formulering’, en ‘toelichting’. De andere zeven item-kenmerken bleken moeilijk te coderen tijdens een proefstudie en werden daarom gecodeerd door drie beoordelaars: ‘Inhoud van de vraag’, ‘moeilijk taalgebruik’, ‘emotionele lading’, ‘vermoeden van een filtervraag’, ‘gevoelige informatie’, ‘centraliteit’, en ‘antwoord-complexiteit’. Voor de meeste item-kenmerken moesten de beoordelaars bepalen of het kenmerk al dan niet aanwezig is in desbetreffende items. De resultaten laten zien dat interbeoordelaar-betrouwbaarheid laag was voor verschillende item-kenmerken: ‘Formulering’, ‘toelichting’, ‘inhoud van de vraag’, ‘moeilijk taalgebruik’, ‘emotionele lading’, ‘vermoeden van een filtervraag’, ‘gevoelige informatie', en 'centraliteit'. We schrijven deze lage interbeoordelaar-betrouwbaarheid voor een belangrijk deel toe aan de subjectieve persoonlijke interpretatie van het bepalen van de aanwezigheid van het item-kenmerk. We deden vier suggesties om met lage interbeoordelaar-betrouwbaarheid om te gaan: 1) Het excluderen van survey items waarvoor de van toepassing zijnde categorie (meestal 'afwezig' of 'aanwezig') niet ondubbelzinnig bepaald kon worden; 2) het herdefiniëren en verfijnen van de item-kenmerken die niet betrouwbaar gecodeerd konden worden; 3) het automatiseren van de definitie van item-kenmerken en de van toepassing zijnde categorie voor specifieke inhoud ten aanzien van elk item-kenmerk, en; 4) het aantal beoordelaars dat het item-kenmerk als aanwezig heeft gecodeerd als uitgangspunt nemen voor de item-kenmerken die zijn gecodeerd door drie beoordelaars.

In hoofdstuk 3 onderzochten we een grote selectie van antwoordgedrag op het vóórkomen ervan en op hun variatie tussen respondenten, surveys, en de interactie tussen respondenten en surveys, om potentiële gedragsconsistentie over tien populatie-surveys heen te verkennen. De selectie bestond uit 12 antwoordgedragingen: 'Het vermijden van vervolgvragen', 'sociaal wenselijk antwoorden', 'het antwoorden van 'ik weet het niet'', 'het antwoorden van 'ik wil het niet zeggen'', 'meegaand antwoorden', 'neutraal antwoorden', 'extreem antwoorden', 'het kiezen van één van de eerste antwoordcategorieën', 'het kiezen van één van de laatste antwoordcategorieën', 'het kiezen van dezelfde antwoordcategorie voor meerdere bij elkaar horende vragen', 'snel antwoorden', en 'traag antwoorden'. Voor wat betreft het vóórkomen van antwoordgedrag lieten de uitkomsten zien dat respondenten in het algemeen vaak vervolgvragen vermeden en sociaal wenselijk antwoordden. Respondenten gaven in het algemeen slechts af en toe aan het antwoord niet te weten of niet te willen zeggen. Voor wat betreft de variaties was onze theorie dat grote respondent-variatie gecombineerd met kleine interactie-variatie op gedragsconsistentie over surveys zou wijzen. Vooral wanneer respondenten vaak aangaven het antwoord niet te willen zeggen wees dit op het consistent vóórkomen ervan over surveys heen. In mindere mate leek er ook sprake van dergelijke gedragsconsistentie te kunnen zijn voor respondenten die vaak snel of traag antwoordden, aangaven het antwoord niet te weten, of extreme antwoorden gaven. Gebaseerd op het wisselende vóórkomen en de wisselende variaties ten aanzien van de verschillende surveys en gedragingen in deze studie, kozen we ervoor om over te stappen naar een methode die antwoordgedrag gedetailleerd samenvat en visualiseert voor elke survey en elk gedrag. Deze methode is nieuw in het analyseren van antwoordgedrag: De constructie van 'antwoordgedrags-profielen'.

In hoofdstuk 4 introduceerden en verkenden we onze methode van het construeren van gedragsprofielen om te anticiperen op het onderzoeken van de relatie tussen respondent-kenmerken en antwoordgedrag. We definieerden het gedragsprofiel als een dichtheidsverdeling dat antwoordgedrag samenvat voor een groep respondenten (het 'respondent-profiel') of een groep items (het 'item-profiel'). Het gedragsprofiel laat de waarschijnlijkheid zien van elke gedragskans van 0 tot 1. Het respondent-profiel is het gemiddelde profiel van individuele respondent-profielen die alle zijn gebaseerd op een specifiek aantal items. Het item-profiel is het gemiddelde profiel van individuele item-profielen die alle zijn gebaseerd op een specifiek aantal respondenten. Met behulp van simulaties onderzochten we de statistische eigenschappen van de gedragsprofielen. We toonden aan dat een respondent-profiel informatiever en preciezer wordt naarmate elke respondent meer items invult. En we toonden aan dat een respondent-profiel stabieler en zekerder wordt naarmate de groep uit een groter aantal respondenten bestaat. Wanneer een respondent-profiel zowel precies als zeker is, is het respondent-profiel accuraat. Hetzelfde geldt voor de item-profielen als de concepten 'respondent' en 'item' met elkaar worden uitgewisseld. Respondent-profielen en item-profielen kunnen gebaseerd zijn op zowel respondent-kenmerken als item-kenmerken. Wanneer een gedragsprofiel precies is maar niet zeker, kan het profiel gepiekt zijn. We lieten een eenvoudige vorm van vlakmaken zien om een gepiekt profiel vlak te maken. Vervolgens gebruikten we echte survey data om de eigenschappen van respondent-profielen, item-profielen, en vlakmaken in de praktijk te illustreren. De volgende stap is om de relatie tussen respondent-kenmerken en ongewenst antwoordgedrag te onderzoeken door respondent-profielen te gebruiken.

In hoofdstuk 5 richtten we ons op het beantwoorden van onze voornaamste onderzoeksvraag door de potentieel consistente relatie tussen respondent-kenmerken en ongewenst antwoordgedrag over tien populatie-surveys te onderzoeken. Voor dit doeleinde construeerden we respondent-profielen voor subgroepen respondenten die zijn gebaseerd op de categorieën van acht respondent-kenmerken: 'Leeftijd', 'onderwijsniveau', 'geslacht', 'huishoudelijke samenstelling', 'voornaamste bezigheid', 'inkomen', 'herkomst', en 'het lenen van een computer'. Deze kenmerken werden onderzocht op hun relatie tot acht specifieke ongewenste gedragingen: 'Sociaal wenselijk antwoorden', 'het antwoorden van 'ik weet het niet'', 'het antwoorden van 'ik wil het niet zeggen'', 'meegaand antwoorden', 'neutraal antwoorden', 'extreem antwoorden', 'het kiezen van één van de eerste antwoordcategorieën', en 'het kiezen van dezelfde antwoordcategorie voor meerdere bij elkaar horende vragen'. Het profiel voor elke categorie van een kenmerk werd vergeleken met het profiel voor alle andere categorieën van dat kenmerk samengenomen. Deze vergelijkingen werden uitgevoerd met behulp van een aanpassing van de effectmaat Cliff's Delta. Alhoewel verschillende categorieën van verschillende kenmerken gerelateerd bleken aan verschillende gedragingen, waren het vóórkomen en de richting van deze relaties voor een belangrijk deel afhankelijk van het specifieke kenmerk, de specifieke survey, en het specifieke gedrag. Dit betekent dat een subgroep respondenten vaak meer van een bepaald gedrag liet zien dan de andere subgroepen voor een bepaalde survey, maar minder van dat gedrag dan de andere subgroepen voor een andere survey. Kortom, we hebben geen enkele consistente relatie tussen respondent-kenmerken en antwoordgedrag gevonden.

In hoofdstuk 6 vatten we het fundament en de voornaamste resultaten van deze these samen, geven we suggesties voor vervolgonderzoek, en bespreken we de tekortkomingen en kracht van onze gedragsprofielen-methode. Omdat we geen consistente relaties tussen respondent-kenmerken en antwoordgedrag hebben gevonden, kunnen we de constructie van respondent-schema's en aangepast survey-ontwerp vanuit het perspectief van consistent gedrag niet aanbevelen. We beschouwen het aannemelijk dat item-kenmerken een belangrijkere en consistentere invloed hebben op antwoordgedrag dan respondent-kenmerken. Daarom bevelen we aan om item-profielen te gebruiken om deze invloed te onderzoeken, om item-kenmerken te detecteren die kunnen leiden tot meetfouten. In het geval van consistente relaties tussen item-kenmerken en meetfouten zouden item-schema's geconstrueerd kunnen worden om aangepast survey-ontwerp te begeleiden. We doen ook twee suggesties voor vervolgonderzoek over hoe om te gaan met afhankelijkheid tussen items binnen surveys ten aanzien van onze methode. We sluiten hoofdstuk 6 af door uit te weiden over de voordelen van onze methode. Gedragsprofielen 1) vatten antwoordgedrag samen en visualiseren antwoordgedrag voor complete groepen respondenten of items voor het volledige kansbereik; 2) houden rekening met de mate van precisie van individuele groepsleden en de mate van zekerheid van de groep, en; 3) kunnen kleine subgroepen respondenten of items detecteren die afwijken van de meerderheid van een grotere groep.

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

DANKWOORD

(Acknowledgements in Dutch *and English*)

Allereerst wil ik mijn promotoren Barry Schouten en Vera Toepoel bedanken. Barry, ik heb jou ervaren als een sympathieke, geduldige en precieze begeleider die me waardevolle handvatten heeft gegeven om mezelf te kunnen ontwikkelen en verrijken op survey-inhoudelijk, methodologisch en statistisch vakgebied. Vera, ik heb jou ervaren als een sympathieke en pragmatische begeleidster die me waardevolle handvatten heeft gegeven om een meer effectieve promovendus te kunnen worden. In het bijzonder ben ik jullie beiden dankbaar dat jullie mij langdurig de kans zijn blijven geven om dit door een complex aan factoren deels moeizame en slepende traject tot een succesvol einde te brengen. Uiteindelijk is dat succesvolle einde mede aan jullie begeleiding en volhouden te danken.

De betreffende mensen en afdelingen van Universiteit Utrecht (afdeling Methoden & Statistiek), het Centraal Bureau voor de Statistiek (afdeling Methodologie) en CentERdata bedank ik graag voor het beschikbaar stellen van alle data, en voor alle inhoudelijke en/of administratieve hulp. Graag wil ik tevens de volgende mensen bedanken: Nina Ambachtsheer, Judit Arends-Tóth, Jeldrik Bakker, Jan van den Brakel, Madelon Cremers, Salima Douhou, Anne Elevelt, Deirdre Giesen, Marieke Haan, Hester van Herk, Manon Houtveen, Joop Hox, Els van Kampen, Natalia Kieruj, Edith de Leeuw, Hidde Leplaa, Peter Lugtig, Annemieke Luiten, Vivian Meertens, Mirjam Moerbeek, Mattijn Morren, Kees Mulder, Rense van Nieuwenhuis, Daniel Oberski, Rodinde Pauw, Karlijn Soppe, Bella Struminskaya, Corrie Vis, Rachel Vis-Visschers, Robert van de Walle, en de CBS-veldinterviewers uit de regio Amsterdam, voor uitgevoerde coderingen en/of deelname aan waardevolle discussies en/of het reviewen van papers en/of het geven van nuttig advies en/of het simpelweg zijn van een plezierige en/of inspirerende collega.

Graag bedank ik verder Jeltje Wassenberg-Severijnen en Marieke Westeneng-den Otter voor hun steun en geruststelling bij het geven van onderwijs; Vanessa Torres van Grinsven voor het geven van de gelegenheid om een onderzoekstekst in het tijdschrift STAtOR te laten publiceren; Marieke Noordam voor de plezierige coachings-gesprekken waarbij ik fijn mijn hart heb kunnen en mogen luchten; en Bart Bakker voor het hartelijke welkom al die jaren op de afdeling Methodologie van het Centraal Bureau voor de Statistiek. Een speciaal woord van dank gaat uit naar Wil Legemaat die haar vakantiehuis op De Bloemert te Midlaren alles bij elkaar vele

maanden hartelijk ter beschikking heeft gesteld opdat ik efficiënt aan het project heb kunnen werken.

I would like to thank Stephanie Eckman, Anton Örn Karlsson, Jeffrey Kromrey, George Leckie, Kristen Olson, and Bodo Winter for the valuable discussions and/or their useful suggestions; Norman Cliff for introducing Cliff's Delta to the world of statistics; and Guillaume Rousselet for the availability of an accessible online document about Cliff's Delta that was the inspiration to create an adaptation of the statistic and to use this adaptation for this thesis.

Leonie en Susanna, mijn dank voor alle gesprekken tijdens lunches, koffies en borrels die altijd als plezierig en niet zelden ook als een noodzakelijke uitlaatklep voelden. Sanne, mijn dank voor al je aandacht, interesse en steun jegens mij en het voor mij moeizaam verlopen traject. Noémi, mijn dank voor zowel je inhoudelijke hulp ten aanzien van het project en wanneer ik er in de consultatiewinkel met een bepaald vraagstuk niet goed uit kwam, als voor je steun en vriendschap. Aruna, mijn dank voor al je steun en vriendschap tijdens een verrijkende maar ambivalente periode in m'n leven. Dave, mijn dank voor de momenten van gezelligheid en ontspanning die ik altijd goed heb kunnen gebruiken.

Mam, veel dank voor al je onvoorwaardelijke steun en liefde tijdens een periode in m'n leven waarin ik voor m'n persoonlijke beleving niet altijd vrij, plezierig en mezelf kon en mocht zijn. Jij hebt al die tijd volledig geaccepteerd en omarmd wie en hoe ik was. Dankzij jou heb ik me geborgen kunnen voelen in een tijd waarin ik me dikwijls niet geborgen voelde.

Joost, veel dank voor al je onvoorwaardelijke steun en liefde tijdens een periode in m'n leven waarin ik dikwijls geen zelfvertrouwen en autonomie heb ervaren. Dankzij jou heb ik me begrepen en erkend kunnen voelen. Zonder jouw inhoudelijke hulp en emotionele steun had ik dit project nooit tot een succesvol einde kunnen brengen.

Midlaren, Januari 2021 *Frank Bais*

Zijde

Morgen van zijde en zonder vrees,
dat het des nachts geschiede
in leugen zal ontvlieden.
Er is geen donker tevergeefs.

Ik kan u niet bedieden
dan in een wit ontwijken
van elk beslissend teeken,
waaraan gij zoudt bezwijken;

hoe ook de woorden smeeken
om bloed en vleesch.

Maar langs mijn oogen strijken
en in mijn keel zijn heesch
onuitgesproken blijken
dat gij hier zijt geweest.

(Gerrit Achterberg)

Zo overzichtelijk

Hij schenkt witte wolken in mijn glas
blauwe lucht. IJsblokjes rinkelen in het landschap
en de dag is een tafel voor twee.

(Maria Barnas)

0									
1	0								
0	0	1							
1	0	1	1						
0	1	1	1	1					
1	0	1	1	0	1				
1	1	0	1	1	1	1			
0	0	1	1	1	0	1	0		
0	1	0	0	0	0	1	0	1	
0	0	0	1	0	0	1	0	0	1

About the author

After not finishing secondary school, Frank Bais (1979) combined his job as a healthcare worker with obtaining his propaedeutics in psychology at the Open University in 2006. He finished a bachelor in clinical psychology ('cum laude') in 2010 and a research master psychology in psychological methods and clinical psychology ('with merit') in 2013 at the University of Amsterdam. During this period, he investigated the transdiagnostic similarity between generalized anxiety disorder and obsessive-compulsive disorder, decision making and overconfidence in the general knowledge domain, and the relation between intolerance of uncertainty and future-oriented decision making. Additionally, he was an intern at Statistics Netherlands to examine voting transitions between two Dutch national elections.

Between 2014 and 2020, Frank was a PhD candidate in survey methodology at Utrecht University and Statistics Netherlands. For this project, he constructed behaviour profiles and an adaptation of the statistical effect size Cliff's Delta to investigate the relation between answer behaviour and respondent characteristics across multiple surveys. Additionally, he has several publications in international journals, presented his research at several conferences (in Utrecht, The Hague, and Cologne), was a member of the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS), was a reviewer for Journal of the Royal Statistical Society Series A, followed a broad selection of methodological and statistical post-graduate courses (in Utrecht and London), guided students in working groups and practicals in methodology and statistics, and provided methodological and statistical consultation to students and researchers.

Over de auteur (About the author in Dutch)

Na het niet hebben afgerond van de middelbare school combineerde Frank Bais (1979) zijn baan als thuiszorgmedewerker met het behalen van zijn propedeuse psychologie aan de Open Universiteit Nederland in 2006. Hij behaalde een bachelor in klinische psychologie ('cum laude') in 2010 en een onderzoeksmaster psychologie in psychologische methoden en klinische psychologie ('met genoegen') in 2013 aan de Universiteit van Amsterdam. Tijdens deze periode onderzocht hij de transdiagnostische gelijkenis tussen de gegeneraliseerde angststoornis en de obsessieve-compulsieve stoornis, beslissen en overschat vertrouwen in het algemene kennisdomein, en de relatie tussen intolerantie voor onzekerheid en toekomst-geöriënteerd beslissen. Ook was hij stagiair bij het Centraal Bureau voor de Statistiek om kiezersstromen tussen twee landelijke Nederlandse verkiezingen te onderzoeken.

Tussen 2014 en 2020 was Frank promovendus in survey-methodologie aan Universiteit Utrecht en het Centraal Bureau voor de Statistiek. Voor dit project construeerde hij gedragsprofielen en een aanpassing van de statistische effectmaat Cliff's Delta om de relatie tussen antwoordgedrag en respondentkenmerken over meerdere surveys te onderzoeken. Ook heeft hij verschillende publicaties in internationale journals, presenteerde hij zijn onderzoek op verschillende conferenties (in Utrecht, Den Haag en Keulen), was hij lid van de 'Interuniversity Graduate School of Psychometrics and Sociometrics' (IOPS), was hij een reviewer voor het tijdschrift 'Journal of the Royal Statistical Society Series A', volgde hij een brede selectie van methodologische en statistische post-master cursussen (in Utrecht en Londen), begeleidde hij werkgroepen en practica in methodologie en statistiek, en gaf hij methodologisch en statistisch advies aan studenten en onderzoekers.

This thesis is about using survey data for the construction of answer behaviour profiles. Behaviour profiles are density distributions that summarize and visualize the occurrence of answer behaviour for respondents or items. Behaviour profiles can be constructed for individual or groups of respondents (the respondent profile) or items (the item profile), for single surveys or across multiple surveys, for any survey mode or device, and can be based on multiple respondent and/or item characteristics. In this thesis, we explore the statistical properties of behaviour profiles and construct respondent profiles to compare subgroups of respondents on their answer behaviour across multiple surveys. For the analyses, an adaptation of the non-parametric effect size Cliff's Delta is used.