Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

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

by Frank Bais, Barry Schouten and Vera Toepoel

Release date: June 21, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© Her Majesty the Queen in Right of Canada as represented by the Minister of Industry, 2022

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 191-224
Statistics Canada, Catalogue No. 12-001-X

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

Frank Bais, Barry Schouten and Vera Toepoel[1]

Abstract

In this study, we investigate to what extent the respondent characteristics age and educational level may be associated with undesirable answer behaviour (UAB) 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 UAB and an inventive adaptation of a non-parametric effect size measure are used. The occurrence of UAB 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 UAB across surveys. Taking all surveys together, the degree of UAB varies by age and education. The results do *not* show consistent UAB across individual surveys: Age and educational level are associated with a relatively *higher* occurrence of UAB for some surveys, but a relatively *lower* occurrence for other surveys. We conclude that the occurrence of UAB across surveys may be more dependent on the survey and its items than on respondent's cognitive ability.

Key Words: Respondent profiles; Answer behaviour consistency; Adapted Cliff's Delta; Measurement error; Cognitive ability; Satisficing.

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 and Dillman, 2008). The occurrence and size of measurement error and hence response data quality can be influenced by respondent characteristics (Olson and Smyth, 2015; Tourangeau, Rips and Rasinski, 2000). Respondent characteristics can be thought of as fixed tendencies of a respondent that may lead to undesirable answer behaviour (UAB), like satisficing (Holbrook, Green and Krosnick, 2003; Kaminska, McCutcheon and Billiet, 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, Narayan and Smith, 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 and 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

1. Frank Bais, PO Box 1034, 6801 MG Arnhem. E-mail: frank.bais@cito.nl; Barry Schouten, PO Box 24500; 2490 HA Den Haag. E-mail: bstn@cbs.nl; Vera Toepoel, PO Box 80140, 3508 TC Utrecht. E-mail: v.toepoel@uu.nl.

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 not 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 paper, 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. From here, we use the abbreviation "UAB" for the term "undesirable answer behaviour" throughout the paper.

Investigating the relation between cognitive ability and UAB 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 UAB, 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 UABs. In our study, we want to distinguish respondent UAB that is survey-specific from UAB that occurs consistently across surveys. In order for respondent consistency to appear, UAB needs to occur across topics and survey designs. In other words, we need the full presence of topic and design variability to investigate UAB consistency across various surveys. We consider this topic and design variability as given and do not take into account survey and item characteristics for this study.

This study aims at linking cognitive ability to measurement error by using our method of constructing behaviour profiles. In case cognitive ability appears to have a consistent relation to specific UABs, surveys can be adapted according to the age or educational level of respondents in order to minimize measurement error. In case of such structural associations, the adaptation can be done globally, regardless of the survey. This also implies that our method could be used to predict measurement error. This means that time-consuming and expensive tests that examine the risk of measurement error could initially be omitted. If our method shows an increased risk of measurement error for specific respondents, setting up such tests could be valuable. If our method does not find such an increased risk, we could conclude that survey-independent adaptive survey design based on cognitive ability may not be useful.

For the purpose of our study, the specific survey topic or design would not even have to be taken into account. We realize that examining item characteristics and other respondent characteristics on their relation to measurement error across surveys is relevant as well. However, we consider our study a first step into investigating characteristics of respondents and items in their potentially consistent relation to UAB and measurement error across surveys. For this first step, we chose to examine the obvious respondent characteristics age and educational level in relation to eight relevant UABs (see Section 2).

Note that the undesirability of answer behaviour is potential by definition as we cannot validate its truthfulness (see Bais, Schouten and Toepoel, 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 across 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 paper. In summary, our foundation of ten large different surveys to detect potential behaviour consistency and to indicate the extent to which behaviour may be undesirable is solid and powerful.

This paper reads as follows: In Section 2 of this paper, we briefly elaborate on the theoretical framework on which our main research question is based. In Section 3, we describe the data, methods, and statistics that were used to compare the different age and educational categories for each UAB across surveys. As a method to detection of consistent UAB, we use so-called "respondent profiles", as suggested and explored by Bais (2021). In Section 4, we show all statistical results and give answers to our main research question. In Section 5, we conclude with a discussion of these results and make suggestions on how to proceed.

2. Theoretical framework

Cognitive ability may be considered a stable personal characteristic that has its influence on UAB (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 UAB. Both age and educational level have been shown to be related to UAB and hence survey data quality (Krosnick, 1991, 1999; Krosnick et al., 1996). Older and lower educated respondents show less accurate UAB than younger respondents (Andrews and Herzog, 1986) and higher educated respondents (Antoni, Bela and Vicari, 2019), and a less stable attitude reliability measurement than younger and higher educated respondents (Alwin and Krosnick, 1991). See Table 2.1 for an overview of the age and educational categories as used in this study, and relevant literature.

In this study, we include two overarching kinds of UAB: Satisficing behaviour, and behaviour that is based on sensitive content. Satisficing behaviour refers to taking short-cuts in the question-answering process. Satisficing is positively related to item difficulty and can be the outcome of low cognitive ability (Heerwegh and Loosveldt, 2011; Krosnick, 1991, 1999; Krosnick et al., 1996). As a result of satisficing, respondents may show one of the following six specific UABs: Answering "don't know", acquiescence, neutral responding, extreme responding, primacy responding, and straightlining. See Table 2.2 for the meaning of these UABs and their relevant literature.

UAB can also be the result of sensitive survey content. Such UAB 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, Sudman, Blair and Stocking, 1978; Shoemaker, Eichholz and Skewes, 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 and Yan, 2007). As a result of sensitive content, respondents may give one of the following two specific UABs: Socially desirable responding and answering "won't tell". Note that "socially desirable responding" is in fact undesirable because of its relation to measurement error (see for instance DeMaio, 1984; Heerwegh and Loosveldt, 2011). See Table 2.2 for the meaning of the UABs and relevant literature. See Figure 2.1 for the complete theoretical framework.

We need to emphasize that most of the specific UABs in this study are referred to in some literature as "response styles" (see for instance He and Van de Vijver, 2013; He, Van de Vijver, Espinosa and Mui, 2014; Van Herk, Poortinga and Verhallen, 2004; Van Rosmalen, Van Herk and Groenen, 2010). We deliberately do not use the concept of response style throughout this paper. The goal of this study is to investigate whether groups of respondents express a stable and consistent pattern or style of specific UAB across surveys. This means that we need to avoid confusing "response style" as a UAB with "style" as a consistent pattern that groups may show across surveys. Therefore, we distinguished between the UAB itself and the pattern or style of UAB across surveys that we are actually expecting to find.

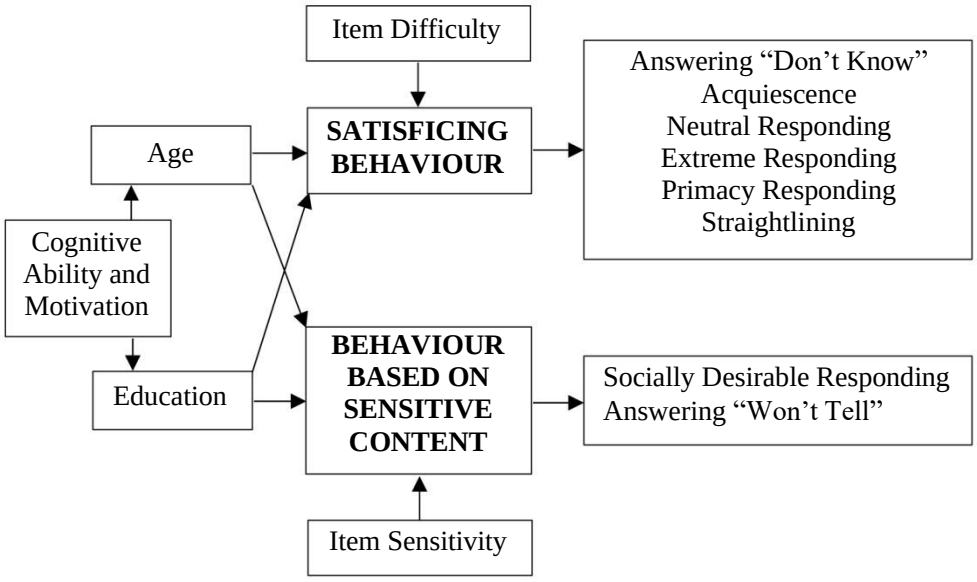
Table 2.1
Respondent characteristics, their categories, and selected relevant literature

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 and Krosnick (1991); Andrews and Herzog (1986); Greenleaf (1992); He, Van de Vijver, Espinosa and Mui (2014); Hox et al. (1991); Kieruj and Moors (2013); Meisenberg and Williams (2008); O'Muircheartaigh, Krosnick and Helic (2000); Pickery and Loosveldt (1998); Schonlau and Toepoel (2015); Zhang and Conrad (2014)
Education	1. primary school 2. vmbo: intermediate secondary education 3. havo/vwo: higher secondary education 4. mbo: intermediate vocational education 5. hbo: higher vocational education 6. wo: university	Aichholzer (2013); Alwin and Krosnick (1991); Greenleaf (1992); He et al. (2014); Krosnick (1991); Krosnick and Alwin (1987); Krosnick, Holbrook, Berent, Carson, Hanemann, Kopp, Mitchell, Presser, Ruud, Smith, Moody, Green and Conaway (2002); Marín, Gamba and Marín (1992); McClendon (1986, 1991); Narayan and Krosnick (1996); O'Muircheartaigh et al. (2000); Pickery and Loosveldt (1998); Schuman and Presser (1981); Zhang and Conrad (2014)

Table 2.2
The answer behaviours, their meaning, and selected relevant literature

Answer Behaviour	Meaning of the Answer Behaviour	Selected Relevant Literature for the Answer Behaviour
Socially Desirable Responding	The tendency to minimize showing socially undesirable behaviour.	Andersen and Mayerl, 2019; Campanelli, Nicolaas, Jäckle, Lynn, Hope, Blake and Gray, 2011; DeMaio, 1984; Heerwegh and Loosveldt, 2011; Holbrook et al., 2003; Jann, Krumpal and Wolter, 2019; Johnson and Van de Vijver, 2003; Kreuter, Presser and Tourangeau, 2008; Krosnick, 1999; Paulhus, 2002; Roberts, 2007; Roberts and Jäckle, 2012; Tourangeau et al., 2000; Tourangeau and 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 and Herrmann, 2002; Binswanger, Schunk and Toepoel, 2013; Bishop, Tuchfarber and Oldendick, 1986; Bradburn et al., 1978; Fricker, Galesic, Tourangeau and Yan, 2005; Krosnick et al., 2002; Leigh and Martin, 1987; Roberts, 2007; Roßmann, Gummer and Silber, 2017; Schuman and Presser, 1981; Shoemaker et al., 2002; Tourangeau et al., 2000; Vis-Visschers, Arends-Tóth, Giesen and Meertens, 2008
Acquiescence	The tendency to answer affirmatively, regardless of the content of the question.	Billiet and McClendon, 2000; De Leeuw, 1992; Díaz de Rada and Domínguez, 2015; Heerwegh and Loosveldt, 2011; McClendon, 1991; Messick, 1966; O'Muircheartaigh et al., 2000; Saris, Revilla, Krosnick and Shaeffer, 2010; Schaeffer and Presser, 2003; Stricker, 1963
Neutral Responding	The tendency to choose the neutral midpoint category from a bipolar answering scale.	He and Van de Vijver, 2013; Kalton, Roberts and Holt, 1980; Krosnick and Fabrigar, 1997; O'Muircheartaigh et al., 2000; Si and Cullen, 1998; Stern, Dillman and Smyth, 2007; Tarnai and Dillman, 1992
Extreme Responding	The tendency to choose an extreme category from the answering scale.	Aichholzer, 2013; De Leeuw, 1992; Díaz de Rada and Domínguez, 2015; Ye, Fulton and Tourangeau, 2011
Primacy Responding	The tendency to choose an option at the beginning of an answering list.	Galesic, Tourangeau, Couper and Conrad, 2008; Krosnick, 1991; Krosnick, 1992; Krosnick and 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 and Domínguez, 2015; Fricker et al., 2005; Krosnick, 1991; Krosnick and Alwin, 1989; Roßmann et al., 2017; Schonlau and Toepoel, 2015; Zhang, 2013; Zhang and Conrad, 2014

Figure 2.1 Literature-based theoretical framework.



Literature overview: Age and education

Age and education seem to be related to non-substantive UAB, giving neutral, extreme, and acquiescent answers, and straightlining. Some studies found more acquiescence for older than for younger respondents (Meisenberg and Williams, 2008; O'Muircheartaigh, Krosnick and Helic, 2000), while other studies found the opposite (Hox, De Leeuw and Kreft, 1991) or no effect (He, Van de Vijver, Espinosa and Mui, 2014). Older respondents are found to give more extreme answers (Greenleaf, 1992; He et al., 2014; Meisenberg and Williams, 2008), including across questionnaires (Kieruj and 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 another study did not find a relation between age and straightlining for respondents who give answers at a high pace (Zhang and Conrad, 2013). Older respondents are found to give more "no opinion"-answers (Pickery and Loosveldt, 1998) or "don't know"-answers (O'Muircheartaigh et al., 2000) than younger respondents.

Lower educated respondents are found to give more "no opinion"-answers (Narayan and Krosnick, 1996; Krosnick et al., 2002; Pickery and Loosveldt, 1998) and "don't know"-answers (O'Muircheartaigh et al., 2000; Schuman and Presser, 1981) than higher educated respondents. Most studies found a negative relation between education and acquiescence (McClendon, 1991; Narayan and Krosnick, 1996; O'Muircheartaigh et al., 2000), although some research did not find a relation (Bachman and O'Malley, 1984; He et al., 2014; Hox et al., 1991). Also a negative relation between education and extreme responding is found (Aichholzer, 2013; Greenleaf, 1992; He et al., 2014; Marín, Gamba and Marín, 1992 – but see Bachman and O'Malley, 1984 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 give answers at a high pace, more straightlining was found for lower than for higher educated respondents (Zhang and Conrad, 2013). Evidence for the relation between education and primacy responding was mixed; see Krosnick and Alwin (1987) versus McClendon (1991).

As summarized above, the literature shows that the relation between age or education and UAB is not unambiguous. The literature needs to be complemented by results that are based on a fixed panel of respondents filling out multiple surveys. Existing findings from different studies are often mixed and may not be comparable because of different respondent samples. This means that it is hard to make literature-based predictions for our panel study and consistent UAB across surveys. Therefore, we do not construct hypotheses and merely explore to what degree UAB for different age and educational groups is consistent across surveys. By using a fixed panel and large set of ten surveys, our aim is to obtain an overarching overview of the relation of age and education to eight relevant UABs.

3. Method

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 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. Also note the relatively high response rates for all surveys, ensuring comparable samples across the surveys. Considering these high and comparable response rates, we do not expect them to have a substantial relation to the occurrence of UAB within the context of this study.

The LISS Panel consists of about 7,000 individuals from about 4,500 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 nonresponse, 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 2,074 items that were used to cover the UABs as presented in Section 2.

Table 3.1
Overview of all surveys, a description of their content, and their response rate (and the number of respondents)

Survey (administration period, nr. of items)	Topics of the content	Response rate (and nr. of respondents)
Economic Situation Assets (AS) (Jun/Jul '12, i = 50)	Income, property and investment	75.2% (5,588)
Family and Household (FA) (Mar/Apr '13, i = 409)	Housing and household; social behaviour	88.8% (5,826)
Health (HE) (Nov/Dec '12, i = 243)	Health and well-being	85.4% (5,780)
Economic Situation Housing (HO) (Jun/Jul '13, i = 73)	Housing and household; income, property and investment	58.2% (3,199)
Economic Situation Income (IN) (Jun/Jul '13, i = 286)	Employment, labour, retirement; income, property, investment; social security, welfare	78.4% (5,015)
Personality (PE) (May/Jun '13, i = 200)	Psychology	90.6% (5,169)
Politics and Values (PO) (Dec '12/Jan '13, i = 148)	Politics; social attitudes and values	85.7% (5,732)
Religion and Ethnicity (RE) (Jan/Feb '13, i = 71)	Religion; social stratification and groupings	88.6% (5,908)
Work and Schooling (WO) (Apr/May '13, i = 471)	Education; employment, labour and retirement	86.5% (5,585)
Labour Force Survey (LF) (Dec '13, i = 123)	Education; employment and labour	81.2% (3,166)

3.2 Coding the undesirable answer behaviours

Each item (the total of the question and all answering options together) of all surveys was investigated on whether it was eligible for the selected UABs separately. The answering categories of the eligible items were coded for each UAB. In case a category was filled out for which the UAB occurred, the response was coded as 1; in case a category was filled out for which the UAB did not occur, the response was coded as 0. For all UABs, 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 UABs that matched our research. See Table 3.2 for an overview of the UABs and their eligible kind of items. See Table 3.3 for the proportions of items for which the UABs are applicable per survey and in total. From here, we discuss the coding process of the UABs that need more elaboration: Socially desirable responding, acquiescence, and straightlining.

Table 3.2
The answer behaviours and their eligible kind of items

Answer Behaviour	Eligible items
Defined on Item Level	
Socially Desirable Responding	All items coded as asking for sensitive information, containing at least one answer category coded as possibly being socially desirable and at least one category coded as not being socially desirable.
Answering "Don't Know"	All items containing a "don't know" answer category.
Answering "Won't Tell"	All items containing a "won't tell" answer category.
Acquiescence	All more or less subjective (battery) items in the form of an ordinal agree/disagree or yes/no answer scale.
Neutral Responding	All (battery) items with an odd and minimum number of five answer categories on an ordinal scale, containing a neutral middle answer category.
Extreme Responding	All (battery) items with a minimum number of four answer categories on an ordinal scale, containing non-neutral first and last answer categories.
Primacy Responding	All (battery) items containing at least four response options.
Defined on Battery Level	
Straightlining	The items of all batteries containing at least 3 items and at least 4 answer categories, only in case all items of the battery were actually filled out.

Table 3.3
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*

	AS	FA	HE	HO	IN	PE	PO	RE	WO	LF	TO
Nr. of items	50	409	243	73	286	200	148	71	471	123	2,074
Nr. of batteries	-	11	5	-	3	16	12	4	2	-	53
Ave. nr. of items/battery	-	5.5	7.6	-	5.7	11.1	6.0	5.8	12.0	-	7.8
Soc. Des. responding	0.20	0.12	0.62	0.01	0.25	0.30	0.51	0.42	0.19	0.32	0.28
Answering "don't know"	0.52	0.08	0.01	0.33	0.47	0.02	0.45	0.49	0.11	0.01	0.18
Answering "won't tell"	0.28	-	-	0.30	0.31	-	0.01	-	0.04	0.81	0.12
Acquiescence	-	0.03	-	-	0.01	0.96	0.68	0.24	0.05	0.03	0.17
Neutral responding	-	0.10	-	-	0.05	0.93	0.66	-	0.04	-	0.17
Extreme responding	-	0.13	-	-	0.05	0.93	0.66	-	0.06	-	0.18
Primacy responding	-	0.37	0.23	-	0.24	0.93	0.73	0.55	0.19	0.27	0.35
Straightlining	-	0.15	0.16	-	0.06	0.89	0.49	0.32	0.05	-	0.20

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

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, Schouten, Lugtig, Toepoel, Arends-Tóth, Douhou, Kieruj, Morren and Vis, 2019). Next, the answering categories of these items were coded by an independent fourth 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 in a socially desirable manner across surveys from those who are not.

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 and Tourangeau, 2015). Both positively and negatively worded items were present throughout the surveys to measure acquiescence. Both battery (a set of related items sharing the same answering options) 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 into 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 certain 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 UABs 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 UABs simply needs to be placed at some point. By the current way of coding these UABs, enough variability between respondents is present in order to distinguish age and educational subgroups that may differ in the occurrence of UAB.

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 and 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 received the code "1" five times. This means that we took into account the length of the battery for this UAB. In other words, the more items a battery consists of, the stronger the UAB 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 UABs is transformed into meaningful respondent behaviour summaries.

3.3 Respondent profiles

In order to compare respondents on consistent UAB across surveys, a few aspects need to be taken into account regarding the UAB. First, the number of items that is applicable to the UAB per survey can be relatively small. This means that uncertainty exists around the actual occurrence of UAB, since it is based on, by definition, a limited 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 UAB for respondents may be more uncertain 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 UAB is indicated with varying uncertainty across different respondents within a survey. Hence, to compare respondents sharing the same characteristic on their UAB across surveys, simply using individual UAB 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 (2021) for an extensive statistical elaboration on this method.

The respondent profile

The respondent profile is a summary of UAB for a group of respondents. It represents the relative proportions of a specified population group (for instance lower educated respondents) in showing a specified UAB (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. Note that when we speak of a "respondent profile", we refer to a group of respondents by definition. When we discuss a profile for a single respondent, we explicitly speak of an "individual respondent profile".

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 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 UAB 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 UAB occurrence and is calculated for each probability from 0 to 1. Hence, to construct the individual profile for respondent r, the likelihood of the UAB occurrence is calculated on the basis of 10 actual "don't know"-answers out of 50 possible occasions for all 100 probabilities:

$$\lambda_r(p) = \binom{I_r}{G_r} p^{G_r} (1-p)^{I_r - G_r}, \tag{3.1}$$

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 UAB 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:

$$\tilde{\lambda}_r(p) = \frac{\lambda_r(p)}{\int_{p=0}^{1} \lambda_r(p)\,dp}, \tag{3.2}$$

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 UAB 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:

$$E_r = \int_{p=0}^{1} p\tilde{\lambda}_r(p)\,dp. \tag{3.3}$$

The likelihood curve resulting from formula's (3.1) and (3.2) is an individual respondent profile. The profile delineates the expected UAB 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 UAB value of 0.50, but who filled out a different number of items for which the UAB was possible. For instance, respondents A and B showed UAB for 10 out of 20 items and for 30 out of 60 items respectively. See Graph 1 in Figure 3.1. 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 UAB occurrence is relatively more certain.

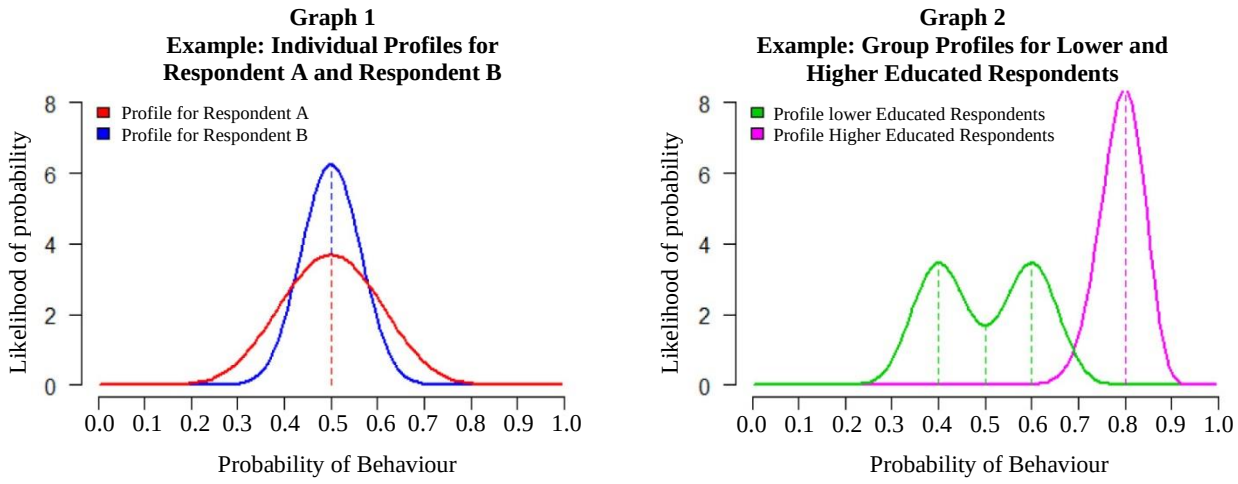
By considering all respondents who meet the condition of a specific category for a characteristic (for instance lower educated respondents for educational level), the average respondent group profile can be calculated by simply summing their comparable individual profiles and dividing the outcome by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

$$\bar{\lambda}(p) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\lambda}_r(p), \tag{3.4}$$

where $\bar{\lambda}$ is the respondent profile of the group UAB 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 UAB occurrence for this group of respondents can be calculated as follows:

$$\bar{E} = \int_{p=0}^{1} p\bar{\lambda}(p)dp. \tag{3.5}$$

Figure 3.1 Examples of respondent profiles with similar expected values (Graph 1) and different expected values (Graph 2).

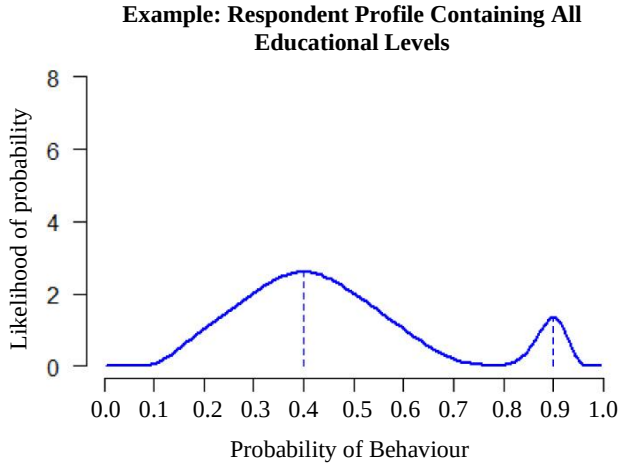


The likelihood curve resulting from formula (3.4) is a group respondent profile. To illustrate the uncertainty on the group level, consider the two groups of lower and higher educated respondents showing a specific UAB. See Graph 2 in Figure 3.1. The expected values for the groups are 0.50 and almost 0.80 respectively. Our method shows that the expected UAB occurrence is more precisely estimated for higher than for lower educated respondents. It is also visible that for lower educated respondents, the UAB occurrence is not centered around the expected group value of 0.50, but around the values of 0.40 and 0.60. Although formula (3.4) refers to a profile for a group of respondents, it does give an indication of

individual UAB. Consider the respondent profile in Figure 3.2 containing individuals on all educational levels. The majority of individuals does not show a specific UAB very often considering the large bump left of the center. On the right, a small peak is visible that refers to a subgroup of individuals showing the UAB very often. These respondents may be either lower or higher educated respondents, or they may share another characteristic that is associated with a high UAB occurrence. The point here is that the respondent profile takes into account the individual UAB and that subgroups of individuals showing a specific occurrence of UAB may be identified in the profile.

Note that by using this method of constructing respondent profiles, we assume that individual UAB 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 paper. We refer to Bais (2021) for suggestions on how to cope with interdependence across items in future research using respondent profiles.

Figure 3.2 Example of a respondent profile containing all educational levels.



Also note that we choose not to use a more traditional model like multilevel analysis to analyze our data. We do not follow identified individual respondents across surveys, but we analyze subgroups of respondents sharing the same characteristic by our profile method for several reasons. Besides taking into account the uncertainty that comes along with the delimited and varying number of respondents and/or items, respondent profiles fully summarize and graphically visualize UAB for subgroups of respondents. And by means of full respondent profiles, relatively small subgroups that deviate from the main body of a larger group may be detected. Throughout this paper, note that a *category* of respondents refers to respondents in a specific single age or educational category (see Table 2.1), while a *(sub)group* of respondents may also refer to respondents from several age or educational categories.

In summary, the expected values of two groups with different characteristics indicate the average UAB occurrences for the groups as a whole. In this way, an idea is obtained about the difference of the occurrences of specific UAB (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 UAB occurrences of two groups.

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 adapted version of the statistic, as we are not considering data observations but density distributions.

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 and Kromrey, 2004; Rousselet, Foxe and Bolam, 2016; Rousselet, Pernet and 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 and Kromrey, 2004). The sample estimate of Cliff's Delta $\hat{\delta}$ is

$$\hat{\delta} = \frac{\sum_{a=1}^{A} \sum_{b=1}^{B} \operatorname{sgn}\left(X_a - X_b\right)}{AB}, \tag{3.6}$$

where $\left(X_a - X_b\right)$ 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.

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. When we transpose this idea to respondent profiles, we may consider the UAB probabilities from 0 to 1 (with a specific step size interval) our "observations" and the likelihoods for each probability their "frequencies" or "weights".

$$\hat{\delta} = \frac{\sum_{a=1}^{A}\sum_{b=1}^{B} \operatorname{sgn}(P_a - P_b)\,\bar{\lambda}(P_a)\,\bar{\lambda}(P_b)}{\sum_{a=1}^{A}\sum_{b=1}^{B} \bar{\lambda}(P_a)\bar{\lambda}(P_b)}, \tag{3.7}$$

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.

As a brief illustration, we calculate the adapted Cliff's Delta by means of formula (3.7) for the respondent profiles in Figure 3.1. Consider Graph 1. When comparing the profiles for respondent A to respondent B, Cliff's Delta is 0. Although the two profiles slightly differ, their shapes are symmetrically formed around the shared expected value of 0.50. This means that the various values in the denominator of formula (3.7) cancel each other out. Consider Graph 2. When comparing the profiles for lower to higher educated respondents, Cliff's Delta is -0.99. The profiles hardly overlap and the higher educated respondents clearly show more of some UAB than the lower educated respondents. The reason that Cliff's Delta is not exactly 1 can be explained by the very small part of overlap around the probability of 0.70 (see Graph 2). Note that the sign would change and Cliff's Delta would be 0.99 when we would compare higher to lower (instead of lower to higher) educated respondents.

For our study, we use the adaptation of Cliff's Delta in order to compare respondent profiles. The respondent profiles and this adaptation take into account the fact that each respondent fills out a delimited and different number of items (see Section 3.3). 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 and 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 4.2 for these categories). For our adapted Cliff's Delta, relatively small or unequal sample sizes are no issue.

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 10,000 category profiles and 10,000 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, Kraaikamp, Lopuhaä and Meester, 2005 for more on this bootstrap method). For each profile, respondents are randomly sampled with replacement and their individual profiles are averaged by means of formula (3.4). 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 10,000 Cliff's Delta's and rank them from low to high. Because of the large number of Cliff's Delta's in our study, we choose to use 99% confidence intervals. This means that we use the 51st and the 9,950th 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 UABs. Each Cliff's Delta is accompanied by its 99% confidence interval.

4. Results

In this section, we first show the Cliff's Delta's for all surveys together as if they were one large survey. Second, we consider the Cliff's Delta's per survey to give an indication about UAB consistency across surveys to answer our research question. All Cliff's Delta's 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 UAB to characteristics that are known from the literature to affect UAB, 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 UAB for a subgroup after excluding respondents may differ from the initial occurrence of UAB 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 UAB occurrence. Therefore, we used two criteria to guarantee the representativeness of each original subgroup: 1) Each subgroup consists of more than 30% of the number of respondents in the original group, and; 2) the UAB occurrence in each subgroup does not differ more than 0.02 from the original group's UAB occurrence.

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 4.1. 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.

Table 4.1
Overall survey participation in total and per subgroup in average number of surveys (and absolute number of respondents)

	TOT	15_24	25_34	35_44	45_54	55_64	> 64
Age	7.6	6.0	6.8	7.3	7.7	8.3	8.5
	(6,700)	(838)	(803)	(1,083)	(1,223)	(1,289)	(1,464)
		Primary	VMBO	HAVW	MBO	HBO	WO
Education	7.6	7.3	7.7	7.3	7.6	7.7	7.6
	(6,688)	(601)	(1,634)	(791)	(1,549)	(1,504)	(609)

We used respondent profiles and Cliff's Delta to explore whether the degree of participation made a difference in the occurrence of the specific UABs taking all surveys together. We split up the complete sample of panel respondents into a group who filled out at most eight surveys and a group who filled out at least nine surveys. See Table 4.2. It is clear that participation rate did not affect the occurrence of most UABs. 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 4.2
Cliff's Delta for Low (Filled out at most eight surveys) versus High (Filled out at least nine surveys) survey participation per answer behaviour[1]

	SD	PR	DK	ST	WT	AC	NE	EX
At most eight vs. at least nine surveys	-0.09	0.07	0.08	*0.14* ~	*0.29* *	-0.06	0.02	-0.10

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

[1]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 UAB than respondents aged 65-74. Therefore, we compared respondents aged 65-74 to respondents aged 75 or older on their group UAB proportion. See Table A.1 in Appendix A. Age subgroups did not or hardly differ for most UABs 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.

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 $|\delta| < 0.11$ indicates no effect, $0.11 \leq |\delta| < 0.28$ a small effect, $0.28 \leq |\delta| < 0.43$ a medium effect, and $|\delta| \geq 0.43$ a large effect, as investigated by Vargha and Delaney (2000), see also Goedhart (2016). A subgroup is always compared to the aggregated total of all remaining applicable subgroups regarding the specific characteristic. See Table 4.3 for the Cliff's Delta's for all surveys taken together.

From Table 4.3, 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 4.1), 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 4.3, 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 4.1), 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.

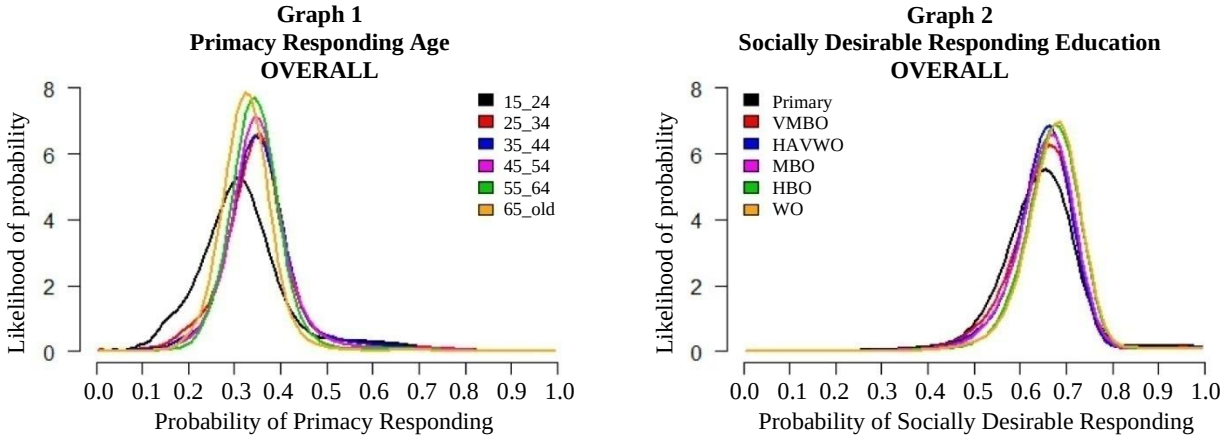
A present overall effect size for a specific category and UAB does not by definition mean a present effect size for various surveys; an overall effect size may exist without effect sizes for any surveys. The opposite may be true as well; an overall effect size may be absent, as positive and negative effect sizes for various surveys cancel each other out. In the following section, we investigate to what extent either positive or negative effect sizes consistently exist across surveys and answer our main research question.

Table 4.3
Overall Cliff's Delta (and its 99% confidence interval) taken over all surveys, for all age categories[1] and all educational categories[2] for all answer behaviours[3]

	Satisficing Behaviour					Behaviour Based on Sensitive Content		
	DK	**AC**	**NE**	**EX**	**PR**	**ST**	**SD**	**WT**
Age 1524	*0.30* * (0.25, 0.35)	-0.06 (-0.12, -0.00)	-0.02 (-0.08, 0.04)	*-0.15* ~ (-0.21, -0.10)	*-0.24* ~ (-0.30, -0.18)	0.00 (-0.06, 0.07)	-0.04 (-0.09, 0.01)	*0.25* ~ (0.20, 0.31)
Age 2534	*0.11* ~ (0.05, 0.16)	0.05 (-0.00, 0.11)	-0.06 (-0.12, -0.00)	-0.08 (-0.14, -0.02)	0.08 (0.03, 0.14)	*0.12* ~ (0.06, 0.17)	0.02 (-0.03, 0.08)	0.09 (0.04, 0.14)
Age 3544	0.08 (0.04, 0.13)	-0.01 (-0.06, 0.04)	0.03 (-0.02, 0.07)	0.01 (-0.04, 0.06)	*0.13* ~ (0.08, 0.17)	*0.19* ~ (0.15, 0.24)	-0.02 (-0.07, 0.02)	0.08 (0.03, 0.12)
Age 4554	0.02 (-0.02, 0.07)	-0.04 (-0.09, 0.00)	0.01 (-0.04, 0.05)	0.04 (-0.01, 0.08)	*0.13* ~ (0.08, 0.17)	*0.11* ~ (0.07, 0.16)	-0.01 (-0.05, 0.03)	0.02 (-0.02, 0.06)
Age 5564	*-0.15* ~ (-0.19, -0.11)	0.03 (-0.01, 0.07)	-0.02 (-0.06, 0.02)	0.06 (0.01, 0.10)	0.06 (0.03, 0.10)	*-0.12* ~ (-0.16, -0.08)	0.02 (-0.02, 0.06)	-0.06 (-0.10, -0.02)
Age 65Ol	*-0.20* ~ (-0.24, -0.16)	0.02 (-0.02, 0.06)	0.04 (0.00, 0.08)	0.05 (0.01, 0.09)	*-0.17* ~ (-0.20, -0.13)	*-0.22* ~ (-0.26, -0.18)	0.02 (-0.02, 0.06)	*-0.17* ~ (-0.20, -0.14)
Edu PRI	*0.20* ~ (0.14, 0.26)	*-0.13* ~ (-0.19, -0.06)	*0.14* ~ (0.08, 0.20)	0.03 (-0.04, 0.10)	*-0.21* ~ (-0.27, -0.15)	*-0.14* ~ (-0.20, -0.07)	*-0.13* ~ (-0.20, -0.08)	0.08 (0.02, 0.14)
Edu VM	0.10 (0.06, 0.14)	*-0.18* ~ (-0.22, -0.14)	*0.14* ~ (0.10, 0.18)	0.04 (-0.00, 0.08)	*-0.13* ~ (-0.17, -0.09)	-0.04 (-0.08, 0.00)	-0.08 (-0.12, -0.04)	0.07 (0.04, 0.11)
Edu HA	0.00 (-0.05, 0.06)	0.01 (-0.04, 0.06)	-0.10 (-0.16, -0.05)	-0.02 (-0.08, 0.03)	0.00 (-0.05, 0.06)	-0.04 (-0.09, 0.02)	-0.06 (-0.10, -0.01)	0.02 (-0.03, 0.07)
Edu MB	0.07 (0.03, 0.11)	-0.04 (-0.08, 0.00)	0.05 (0.01, 0.09)	-0.02 (-0.07, 0.02)	0.02 (-0.02, 0.06)	0.05 (0.00, 0.09)	-0.02 (-0.06, 0.02)	0.08 (0.04, 0.11)
Edu HB	*-0.17* ~ (-0.21, -0.13)	*0.18* ~ (0.14, 0.22)	*-0.12* ~ (-0.16, -0.08)	-0.03 (-0.07, 0.01)	*0.12* ~ (0.09, 0.16)	0.02 (-0.02, 0.06)	*0.13* ~ (0.10, 0.17)	*-0.13* ~ (-0.16, -0.09)
Edu WO	*-0.21* ~ (-0.27, -0.16)	*0.22* ~ (0.16, 0.27)	*-0.18* ~ (-0.23, -0.12)	0.02 (-0.04, 0.08)	*0.19* ~ (0.14, 0.24)	*0.12* ~ (0.07, 0.18)	*0.14* ~ (0.08, 0.19)	*-0.13* ~ (-0.18, -0.08)

~→ small effect; *→ medium effect; #→ large effect.
[1]15-24 Years (Age 1524), 25-34 Years (Age 2534), 35-44 Years (Age 3544), 45-54 Years (Age 4554), 55-64 Years (Age 5564), 65 Years and Older (Age 65Ol).
[2]Primary Education (Edu PRI), VMBO (Edu VM), HAVWO (Edu HA), MBO (Edu MB), HBO (Edu HB), WO (Edu WO).
[3]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).

Figure 4.1 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 (purple) 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.



4.3 Consistency outcomes for Cliff's Delta

These results for Cliff's Delta concern the consistency of subgroups across surveys. To reveal consistency, we considered the number of surveys for which at least a small effect $\left(|\delta| \geq 0.11\right)$ was the result. Considering consistency conservatively, as an at least small effect for a specific UAB and category for *all or almost all* applicable surveys, 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*. See Table 4.4 containing all results for the UABs and categories for which more than half of the applicable surveys showed either positive or negative effect sizes: There is no category that shows an effect for all or almost all surveys for any UAB.

Table 4.4
Cliff's Delta (and its 99% confidence interval) for the behaviours Answering don't know, Primacy responding, and Neutral responding, for the applicable Age categories[1] and Educational categories[2] for the Applicable surveys[3]

	FA	HE	HO	IN	PE	PO	RE	WO
Answering "Don't Know"								
Age 1524	0.09 (0.05, 0.12)			*0.46 #* (0.41, 0.51)		*0.28 ** (0.22, 0.34)	0.05 (0.03, 0.07)	*0.24 ~* (0.19, 0.30)
Age 65Ol			*-0.13 ~* (-0.17, -0.09)	*-0.20 ~* (-0.24, -0.16)		*-0.14 ~* (-0.17, -0.10)	-0.02 (-0.03, -0.01)	
Edu PRI	*0.15 ~* (0.08, 0.23)		0.08 (-0.00, 0.15)	*0.16 ~* (0.10, 0.23)		*0.17 ~* (0.11, 0.24)	0.02 (0.00, 0.05)	*0.23 ~* (0.15, 0.31)
Primacy Responding								
Age 1524	*-0.36 ** (-0.40, -0.32)	-0.10 (-0.13, -0.06)		*-0.31 ** (-0.37, -0.26)	*-0.18 ~* (-0.24, -0.12)	*-0.11 ~* (-0.17, -0.06)	-0.09 (-0.14, -0.04)	-0.05 (-0.09, -0.01)
Edu PRI	0.03 (-0.03, 0.09)	*-0.11 ~* (-0.16, -0.06)		*-0.23 ~* (-0.29, -0.17)	*-0.15 ~* (-0.22, -0.08)	-0.08 (-0.15, -0.01)	*-0.14 ~* (-0.20, -0.09)	-0.09 (-0.15, -0.04)
Edu WO	-0.10 (-0.14, -0.05)	0.06 (0.02, 0.10)		*0.18 ~* (0.12, 0.24)	*0.18 ~* (0.12, 0.24)	0.03 (-0.02, 0.09)	*0.16 ~* (0.11, 0.21)	*0.24 ~* (0.19, 0.28)
Neutral Responding								
Edu WO	0.05 (0.01, 0.10)			*-0.14 ~* (-0.20, -0.09)	*-0.16 ~* (-0.23, -0.09)	*-0.18 ~* (-0.23, -0.13)		-0.04 (-0.09, -0.00)

~→ small effect; *→ medium effect; #→ large effect
[1]15-24 Years (Age 1524), 65 Years and Older (Age 65Ol).
[2]Primary Education (Edu PRI), WO (Edu WO).
[3]Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO).

Therefore, for each UAB and category, we considered the number of surveys for which at least a small either positive or negative effect was found. See Table 4.5. It is striking that relatively many cells or category-UAB pairs showed both positive and negative effects (marked by "2" in Table 4.5). This means that a category may show *more* of a specific UAB for some surveys, while *less* for other surveys. For instance, consider the category 15-24 years for the UAB answering "won't tell" (WT) in Table 4.5. 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 4.4). Strikingly, this is applicable to only seven out of the 96 possible cases (as we have results for eight UABs and twelve categories) and at a maximum of only 75% of the applicable surveys.

Table 4.5
The Categories for Age and Education (Edu) with either at Least Two Positive *or* Two Negative Effect Sizes Receiving a "1" (Unidirectional results) and the Categories with at Least One Positive *and* One Negative Effect Size Receiving a "2" (Contrasting results) for All Behaviours*

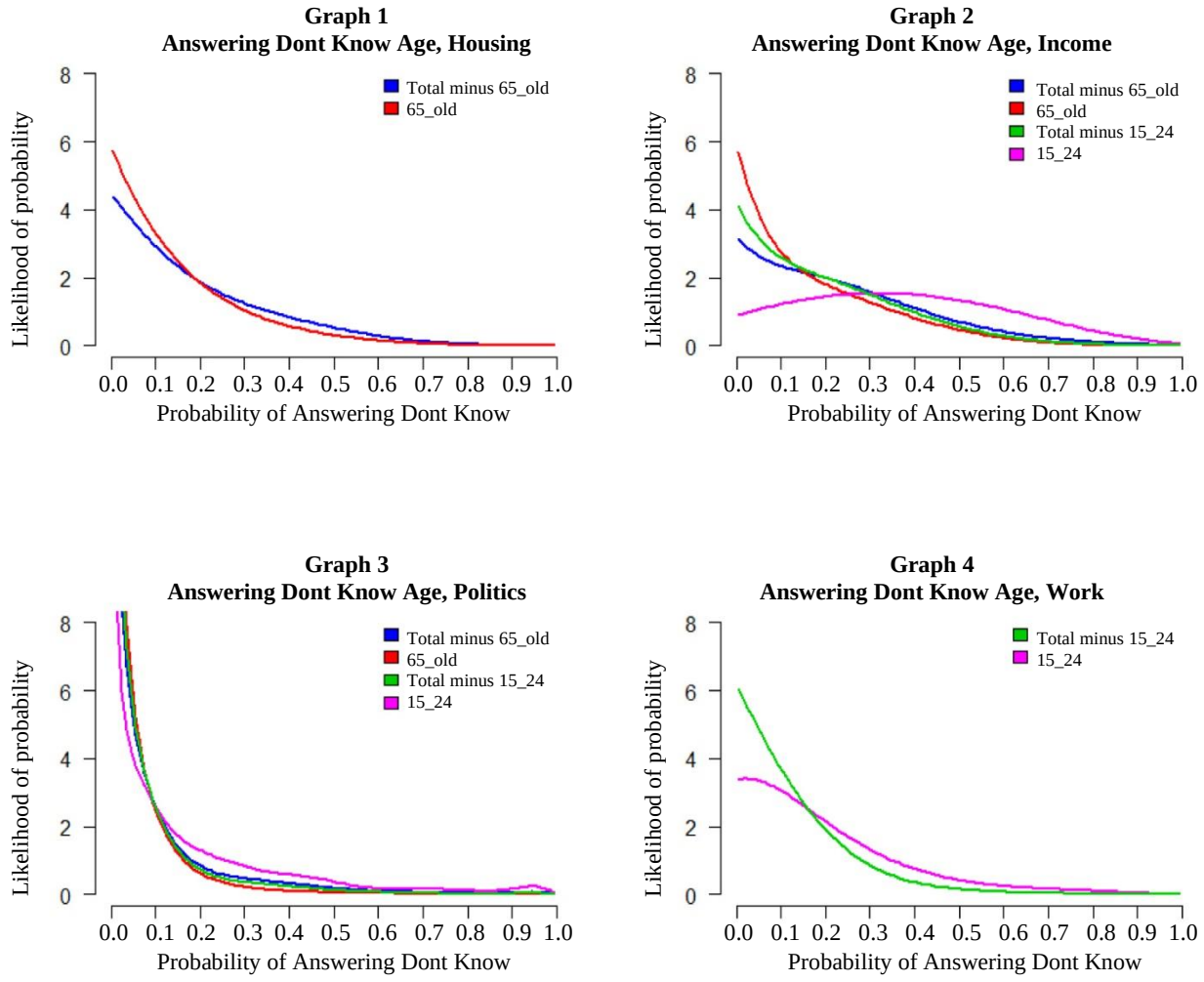
	Number of Surveys	**3**	**5**	**4/5**	**4/5**	**4/5/6**	**6/7**	**7**	**8**
	Answer Behaviour	**WT**	**AC**	**NE**	**EX**	**DK**	**ST**	**PR**	**SD**
Age	15-24 years	2				1	2	1	2
	25-34 years						2	2	2
	35-44 years						1	2	2
	45-54 years						1	1	
	55-64 years					1			2
	65 years or older					1	1	2	2
Edu	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

The empty cells refer to either no effects, or one positive effect, or one negative effect.
*Answering Won't Tell (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Answering Don't Know (DK), Straightlining (ST), Primacy Responding (PR), Socially Desirable Responding (SD).

For the UAB answering "don't know", Table 4.4 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 4.2). 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.

In summary, the results refer to an absence of UAB consistency across all or almost all surveys: Both satisficing and sensitivity-based UABs did not emerge consistently across surveys. We conclude that respondents' UAB across surveys may be more influenced by the survey and its topic and items than solely by the age or educational level of the respondent. We close with a discussion in the following section.

Figure 4.2 Consistently *More* "Don't Know"-Answers for Respondents Aged 15-24 (purple) 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).



5. Conclusion and discussion

In this study, we investigated to what extent cognitive ability is associated with a high occurrence of undesirable answer behaviour (UAB) *consistently* across different surveys. For cognitive ability, we used the respondent characteristics age and educational level. The occurrence of UAB is indicated by varying uncertainty, as every respondent filled out a different number of the items that were applicable to each behaviour. To take this varying uncertainty into account, we used an adaptation of the robust effect size statistic Cliff's Delta to compare groups of respondents in the form of density distributions or *respondent profiles*. The UAB of respondents from a specific category (for instance "15-24 years" for the characteristic "age") was compared to the UAB 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 UABs*. This study used response data from a panel consisting of the same respondents. In general, if UAB 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 UAB 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 UAB consistently across *all or almost all* surveys.

Note that within a single survey, items are clustered around a central topic and may also be similar in their characteristics. This means that some item interdependency may occur within surveys. If we would have found consistent response patterns across surveys, these patterns may have been influenced by such item interdependency. Obviously, some respondents may be more sensitive to item interdependency in showing UAB across surveys than others. In our study, we did not find any consistent response patterns across surveys. This means that item interdependency was unlikely to exert a structurally different influence on the various categories of respondents across surveys.

Our results seem to go beyond the absence of UAB consistency across surveys. As the more surveys were applicable to an UAB, the more contrasting outcomes were found; many categories were associated with relatively *more* of an UAB for some surveys, while relatively *less* of that UAB for other surveys. Most contrasting results were found for giving socially desirable responses. More evidence was found for contrasting UAB than for consistent UAB across surveys. This evidence is not compatible to our idea that specific groups will show consistency for at least some of the specific UABs across most or all surveys. *Overall, we conclude that the occurrence of UAB 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.*

Following this conclusion, we do not recommend survey-independent adaptive survey design for respondents based on their cognitive ability. The findings for age and educational level are not consistent

and clearly differ depending on both survey and UAB. In essence, this means that our outcomes confirm the different associations and their different directions of the existing literature. The added value of our study is the overarching overview for age and educational level, systematically examined across a set of ten different surveys for a range of eight different UABs. We conclude that age and educational level may be taken into account for adaptive survey design only for specific surveys and survey topics.

In our study, we did not focus on UAB 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 UAB to individual or groups of identified respondents, nor to compare them between surveys for the same category and UAB. 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 UAB across surveys. In other words, we did not attribute UAB to identified respondents, but to the specific category (for instance respondents aged 15-24) 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 and Smyth, 2015; Olson, Smyth and Ganshert, 2019) that takes into account repeated measurements of individual respondents. The focus of our study was placed on visualizing summaries of UAB and comparing subgroups that 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 UAB 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 UAB. 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, Buelens, Engelen, Van der Laan, Meertens and Schouten (2014); Campanelli et al. (2011); Gallhofer, Scherpenzeel and Saris (2007), and Saris and Gallhofer (2007) for overviews of item characteristics and their relation to

measurement error. Items can be coded on the presence or absence of characteristics like for instance question sensitivity. Hence, items that are coded as sensitive could be compared to items that are not coded as sensitive on the occurrence of UAB. In this way, the presence of item characteristics may be connected to UAB for the items of whole surveys specifically or across the items of multiple surveys more generally. Based on such associations, an overview of present item characteristics and their relation to UAB and measurement error may be obtained.

Acknowledgements

We would like to thank Joost van der Neut for contributing to the adaptation of Cliff's Delta. We would like to thank CentERdata for the availability of LISS Panel data.

Appendix A

Table A.1
The behaviour occurrence proportions for respondents aged 65-74 (65+) and respondents aged 75 or older (75+) for all behaviours*, in total and for all surveys**

	TO	AS	FA	HE	HO	IN	PE	PO	RE	WO	LF
SD 65+	0.66	0.95	0.61	0.66	***	0.79	0.77	0.59	0.27	0.77	
SD 75+	0.65	0.96	0.60	0.64		0.78	0.76	0.58	0.30	0.79	
PR 65+	0.33		0.49	0.65		0.36	0.25	0.18	0.68	0.17	
PR 75+	0.31		0.50	0.65		0.33	0.24	0.16	0.66	0.13	
DK 65+	0.06				0.07	0.16		0.06	0.00		
DK 75+	0.06				0.07	0.14		0.07	0.00		
ST 65+	0.10		0.05	0.36		0.32	0.02	0.07	0.24		
ST 75+	0.08		0.04	0.25		0.29	0.01	0.06	0.19		
WT 65+	0.05				0.02	0.04					0.03
WT 75+	0.04				0.01	0.03					0.03
AC 65+	0.47		0.44				0.50	0.45	0.19		
AC 75+	0.49		0.42				0.51	0.48	0.21		
NE 65+	0.22		0.28			0.25	0.21	0.22			
NE 75+	0.21		0.28			0.25	0.21	0.22			
EX 65+	0.19		0.37			0.11	0.23	0.11			
EX 75+	0.20		0.40			0.11	0.25	0.10			

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

References

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

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

Andersen, H., and 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., and Herzog, A.R. (1986). The quality of survey data as related to age of respondent. *Journal of the American Statistical Association*, 81(394), 403-410.

Antoni, M., Bela, D. and 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., and O'Malley, P.M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491-509.

Bais, F. (2021). *Constructing Behaviour Profiles for Answer Behaviour Across Surveys*. Dissertation, Utrecht University. https://doi.org/10.33540/538.

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M. and 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 and Research*, 48(2), 263-295. https://doi.org/10.1177/0049124117729692.

Bais, F., Schouten, B. and 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.

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

Beukenhorst, D., Buelens, B., Engelen, F.,Van der Laan, J., Meertens, V. and 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., and McClendon, J.M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, 7, 608-628. doi: http://dx.doi.org/10.1207/S15328007SEM0704_5.

Binswanger, J., Schunk, D. and Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly*, 77, 783-797.

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

Bradburn, N., Sudman, S., Blair, E. and Stocking, C. (1978). Question threat and response bias. *Public Opinion Quarterly*, 42, 221-234.

Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., Hope, S., Blake, M. and 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.

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

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

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

Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P. and 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.

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

De Leeuw, E.D., Hox, J.J. and Dillman, D. (2008). *International Handbook of Survey Methodology*. Taylor & Francis Group.

DeMaio, T.J. (1984). Social desirability and survey measurement: A review. In *Surveying Subjective Phenomena*, (Eds., C.F. Turner and E. Martin). New York: Russell Sage Foundation, 2, 257-281.

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

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

Galesic, M., Tourangeau, R., Couper, M.P. and Conrad, F.G. (2008). Eye-tracking data new insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72, 892-913.

Gallhofer, I.N., Scherpenzeel, A. and 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, 328-351. http://www.jstor.org/stable/2749156.

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

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

Heerwegh, D., and Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, 27, 49-63.

Hess, M.R., and Kromrey, J.D. (2004). *Robust Confidence Intervals for Effect Sizes: A Comparative Study of Cohen's d and Cliff's Delta under Non-Normality and Heterogeneous Variances*. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

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

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

Jann, B., Krumpal, I. and 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.

Johnson, T., and Van de Vijver, F.J.R. (2003). Social desirability in cross cultural research. In *Cross-Cultural Survey Methods*, (Eds., J. Harness, F.J.R. van de Vijver and P. Mohler.), New York: John Wiley & Sons, Inc., 193-202.

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

Kaminska, O., McCutcheon, A. and Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74, 880-906. doi: http://dx.doi.org/10.1093/poq/nfq062.

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

Kreuter, F., Presser, S. and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847-865. doi: http://dx.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, 213-236.

Krosnick, J.A. (1992). The impact of cognitive sophistication and attitude importance onresponse order effects and question order effects. In *Order Effects in Social and Psychological Research*, (Eds., N. Schwarz and S. Sudman), New York: Springer, 203-218.

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

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

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

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

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. and Conaway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66, 371-403.

Krosnick, J.A., Narayan, S. and Smith, W.R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 70, 29-44.

Leigh, J.H., and Martin, C.R. (1987). Don't know item nonresponse in a telephone survey: Effects of question form and respondent characteristics. *Journal of Marketing Research*, 24, 418-424.

Lensvelt-Mulders, G.J.L.M. (2008). Surveying sensitive topics. In *International Handbook of Survey Methodology*, (Eds., E.D. de Leeuw, J.J. Hox and D.A. Dillman). New York: Taylor and Francis, Psychology Press, EAM series, 461-478.

Marín, G., Gamba, R.J. and Marín, B.V. (1992). Extreme response style and acquiescence among hispanics. *Journal of Cross-Cultural Psychology*, 23, 498-509.

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

McClendon, M.J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, 20, 60-103.

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

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

Messick, S.J. (1966). The psychology of acquiescence: An interpretation of research evidence. In *Response Set in Personality Assessment*, (Ed., I.A. Berg), Chicago: Aldine, 115-145.

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

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

Olson, K., Smyth, J.D., and 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. and Helic, A. (2000). *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. Retrieved, October 1, 2009, from http://harrisschool.uchicago.edu/About/publications.

Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In *The Role of Constructs in Psychological and Educational Measurement*, (Eds., H.I. Braun, D.N. Jackson and D.E. Wiley), Mahwah, NJ: Erlbaum, 49-69.

Pickery, J., and 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.

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

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

Roßmann, J., Gummer, T. and Silber, H. (2017). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*.

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

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

Saris, W.E., and Gallhofer, I.N. (2007). 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. and Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, 4(1), 61-79. http://www.surveymethods.org.

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

Schonlau, M., and Toepoel, V. (2015). Straightlining in web survey panels over time. *Survey Research Methods*, 9, 125-137. doi: https://doi.org/10.18148/srm/2015.v9i2.6128.

Schouten, B., and Calinescu, M. (2013). Paradata as input to monitoring representativeness and measurement profiles: A case study of the Dutch Labour Force Survey. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, (Ed., F. Kreuter), Hoboken, NJ: Wiley, 231-258.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.

Shoemaker, P.J., Eichholz, M. and Skewes, E.A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, 14, 193-201.

Si, S.X., and Cullen, J.B. (1998). Response categories and potential cultural bias: Effects of an explicit middle point in cross-cultural surveys. *International Journal of Organizational Analysis*, 6, 218-230.

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

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

Tarnai, J., and Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail versus telephone surveys. In *Context Effects in Social and Psychological Research*, (Eds., N. Schwarz and S. Sudman), New York: Springer Verlag.

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859-883. doi: https://doi.org/10.1037/0033-2909.133.5.859.

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

Van Rosmalen, J., Van Herk, H. and Groenen, P.J.F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47(1), 157-172. doi: https://doi.org/10.1509/jmkr.47.1.157.

Vargha, A., and Delaney, H.D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational Behavioral Statistics*, 25(2), 101-132. doi: http://dx.doi.org/10.2307/1165329.

Vis-Visschers, R., Arends-Tóth, J., Giesen, D. and Meertens, V. (2008). *Het Aanbieden Van "Weet Niet" en Toelichtingen in Een Webvragenlijst*. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.

Ye, C., Fulton, J. and Tourangeau, R. (2011). More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, 75(2), 349-365. doi: https://doi.org/10.1093/poq/nfr009.

Zhang, C. (2013). *Satisficing in Web Surveys: Implications for Data Quality and Strategies for Reduction*, (Ph.D.) Ann Arbor, MI: University of Michigan. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97990.

Zhang, C., and Conrad, F.G. (2013). Speeding in Web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8, 127-135.