mda

methods, data, analyses

JOURNAL FOR QUANTITATIVE METHODS AND SURVEY METHODOLOGY

Volume 13, 2019 | 1

Social Desirability Bias in Surveys – Collecting and Analyzing Sensitive Data

Ben Jann, Ivar Krumpal & Felix Wolter (Editors)

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246282
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

Editorial: Social Desirability Bias in Surveys – Collecting and Analyzing Sensitive Data

Ben Jann[1], *Ivar Krumpal*[2] *& Felix Wolter*[3]

[1] University of Bern
[2] University of Leipzig
[3] Johannes Gutenberg University Mainz

Studying social phenomena and social problems often involves measuring and analyzing behaviors or attitudes that are sensitive in several ways. Topics such as delinquency, substance abuse, sexual issues, xenophobia or homophobia may oblige survey respondents to self-report information about very private issues or to report that they have acted against social or legal norms. Hence, survey participants could fear negative consequences of violating social desirability (SD) norms or of a disclosure of their private information to third parties (Tourangeau & Yan, 2007).

As cumulative empirical research has shown, this prompts respondents to engage in self-protective behavior when answering sensitive survey questions, namely by providing untruthful and biased answers (be it unconsciously or deliberately) or by refusing to answer at all (Krumpal, 2013; Lensvelt-Mulders, 2008; Wolter, 2012). This systematic misreporting or nonresponse leads to biased estimates and poor data quality. Statistical associations could be biased as well if the degree of misreporting varies systematically across subgroups or is related to other variables.

At the same time, research about sensitive topics and norm-violations is of particular interest for the social sciences and public discussions likewise: Public authorities, for instance, are interested in being informed about the prevalence of tax evasion, corruption, or illicit work. Media and political parties seek for accurate election forecasts. Researchers may want to study levels and determinants of deviant behaviors, political extremism, or health problems.

The demand for valid measurements of sensitive issues on the one hand and the well-confirmed difficulties due to SD bias on the other has occupied survey methodologists since the very beginning of modern survey research (Benson, 1941; Hyman, 1944). There are two main lines of research. The first one consists in theorizing about, identifying, and quantifying response biases and, if possible, in providing means for controlling such biases ex post, that is, after the data has been collected. One approach for instance concerns measuring and adjusting for socially desirable responding by using psychometric SD scales. The theoretical part of this research agenda seeks for explanations and clarifications of the mechanisms causing systematic misreporting or nonresponse. The second line of research aims at developing data collection techniques that alleviate or, at best, entirely avoid response biases. More conventional approaches in this regard encompass choosing a well-tailored (e.g., self-administered) survey mode or a mixed-mode design, using wording or filtering techniques, and reducing interviewer effects. Strategies that are more complex employ special questioning techniques that mostly pursue the goal of reducing misreporting by increasing the level of anonymity of the respondents' answers, for example via adding random statistical noise to the data. Randomized-response (RRT; Warner, 1965) and item count techniques (ICT; Droitcour et al., 1991) are probably the most prominent techniques in this regard.

Despite the long-standing research tradition in this field, one cannot allege that all problems have been solved. This holds for both theoretical and methodical questions on "best practices". For example, there is an ongoing theoretical discussion about the psychological mechanisms causing respondents to misreport on their true status (e.g., Holtgraves, 2004). Empirical findings regarding the performance of special questioning techniques such as RRT and ICT are mixed and often inconclusive (e.g., Holbrook & Krosnick, 2010). Hence, the objective of this special issue is to contribute to the ongoing debate about theoretical issues as well as about establishing best practices, survey designs, or measurement instruments for surveying sensitive topics.

The article by *Henrik Andersen* and *Jochen Mayerl* addresses the question whether socially desirable responding is more a deliberate, reflected editing of answers, or an automatic process occurring spontaneously. The authors find empirical evidence for both mechanisms depending on whether respondents report about positively connoted traits or about negatively connoted ones.

The paper by *Axel Franzen* and *Sebastian Mader* investigates whether "phantom questions", that is, questions on fictitious, non-existent issues, represent an opportunity to measure respondents' affinity for SD bias. The authors empirically compare classic SD scales (short versions of the Crowne-Marlowe SD scale) and phantom questions with respect to their internal and external consistency and validity.

The study by *Manfred Antoni*, *Daniel Bela*, and *Basha Vicari* deals with SD bias in reported earnings. Linking survey data to administrative validation data on an individual level, the authors investigate the degree of over- and underreporting depending on earnings levels, other individual characteristics, and interviewer effects.

Paula Fomby and *Narayan Sastry* discuss the use of interactive voice response technology (IVR) for collecting sensitive data among adolescents. The authors review questionnaire design, fieldwork protocols, data quality and completeness, and respondent burden of the IVR procedure employed in the Panel Study of Income Dynamics 2014 Child Development Supplement.

The paper by *Alessandra Gaia* and *Tarek Al Baghal* presents a new version of the ICT, namely the longitudinal ICT (L-ICT). While ICT is implemented in cross-sectional surveys with a random split into different sub-samples, L-ICT administers the long- and short-lists (one including the sensitive item, the other not) to the same respondents in different waves of a panel survey. The authors discuss general properties, pros, and cons of L-ICT and present empirical results from a first implementation in the *Understanding Society* Innovation Panel.

The article by *Anke Erdmann* presents empirical evidence on the performance of the triangular model (TM) for gathering sensitive survey data as compared to conventional direct questioning. The sensitive questions pertain to issues about mental stress among students. The author also addresses whether the TM has different effects for certain subgroups of respondents, such as for those scoring high on SD or depressiveness scales.

Finally, the study by *Felix Wolter* seizes a suggestion by Grant, Moon, and Gleason (2014) and introduces the person count technique (PCT), a new variant of ICT. PCT is empirically tested in an experimental survey against conventional direct questioning with respect to nonresponse and misreporting on attitude questions about asylum seekers.

Overall, we are confident that this special issue of mda provides various important contributions to both theoretical and practical challenges in the field of research on sensitive questions. We would like to thank all the authors for their valuable contributions and their patience during the review process. Our thanks also go to the editorial team of mda for their support, and the reviewers for their careful reading and commenting of the manuscripts.

References

Benson, L. E. (1941). Studies in Secret-Ballot Technique. *Public Opinion Quarterly, 5*(1), 79–82.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The Item Count Technique as a Method of Indirect Questioning: A Review of its Development and a Case Study Application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz., & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 185–210). New York: Wiley.

Grant, T., Moon, R., & Gleason, S. A. (2014). *Asking Many, Many Sensitive Questions: A Person-Count Method for Social Desirability Bias*: Unpublished Manuscript.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout by Using the Randomized Response Technique. Evidence Calling into Question the Method's Validity. *Public Opinion Quarterly, 74*(2), 328–343.

Holtgraves, T. (2004). Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding. *Personality and Social Psychology Bulletin, 30*(2), 161–172.

Hyman, H. (1944). Do They Tell the Truth? *Public Opinion Quarterly, 8*(4), 557–559.

Krumpal, I. (2013). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity, 47*(4), 2025–2047.

Lensvelt-Mulders, G. J. L. M. (2008). Surveying Sensitive Topics. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 461–478). New York: Lawrence Erlbaum.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859–883.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*(309), 63–69.

Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik.* Wiesbaden: Springer VS.

Responding to Socially Desirable and Undesirable Topics: Different Types of Response Behaviour?

Henrik Andersen & Jochen Mayerl
Chemnitz University of Technology

Abstract

Social desirability describes the tendency of respondents to present themselves in a more positive light than is accurate and is a serious concern in surveys. If researchers are better able to understand the underlying mechanisms responsible for social desirability bias, they may be able to devise ways to identify and correct for it. One possibility involves determining whether social desirability is more of a deliberate 'editing' of responses or an automatic, perhaps 'self-deceptive', act. Then researchers could potentially flag conspicuously fast or slow responses to improve data quality. We outline dual-process-related theoretical arguments for both scenarios and test their plausibility using data gathered in a tablet-based CASI survey of pre-service teachers in Germany that were asked to assess their suitability for their chosen profession. Our analysis involves the use of fixed-effects multilevel models that enable us to control for unobserved differences between respondent- and item-characteristics while also examining cross-level interactions between the predictors at various levels. Specifically, we examine the classic respondent- (i.e. need for social approval) and item-related characteristics (i.e. trait desirability) associated with social desirability bias, as well as the speed at which the respondents gave their answers. Doing so allows us to observe under what circumstances the respondents tended to overstate positive characteristics as well as understate negative ones. We find evidence for social desirability as an automatic as well as a deliberate response behaviour. However, the mechanism responsible for determining whether social desirability occurs automatically or deliberately seems to be whether the item content is desirable or undesirable. Desirable traits seem to elicit faster socially desirable responses whereas undesirable traits seem to elicit slower socially desirable responses.

Keywords: social desirability, sensitive questions, response latencies, paradata, response bias, survey research, multilevel models

Social desirability (SD) bias describes respondents systematically presenting themselves in a more positive light than is accurate in self-reported surveys. It is a serious concern in survey research and can impact prevalence estimates of behaviour and attitudes as well as observed relationships between variables (Stocké & Hunkler 2007). For decades, researchers have tried to better understand the underlying processes that result in SD bias. Doing so may make it possible to identify measurement error due to SD and improve data quality (Tourangeau & Yan 2007). Much of the research thus far has focused on the question of *whether SD is a deliberate or an automatic action*. The main goal of this article is to contribute to finding an answer to this question. If it is mostly deliberate and respondents carefully consider the desirability of their answer before giving it, then it may be possible to 'flag' answers that took the respondent particularly long to answer, for example. If SD is mostly automatic, the same could be true for unusually fast answers.

The measurement of response latencies (RLs) provides a promising method of indirectly assessing the underlying processes associated with SD bias. In psychological research, RLs have been used for decades as a common method of measuring cognitive processes (e.g. Fazio 1990b). In survey research, the development of computer assisted technology (e.g. CATI, CAPI, CASI) made it possible to include such measurements even in large-scale survey projects (e.g. Bassili & Fletcher 1991). One of the most prominent applications involves their use as a proxy measure for cognitive processing *modes* (e.g. Fazio 1990a; Mayerl 2009) with faster responses suggesting a more automatic-spontaneous mode; slower responses a deliberate-controlled one.

Regardless of the promise RL measurement shows, it has become clear that the solution to the problem of SD is not as simple as: "socially desirable responses are fast/slow". Rather, it seems a whole range of factors influence how respondents deal with survey questions. These include respondent-related personality traits, characteristics of the question content, the respondents' unknown 'true' answers and characteristics related to the survey situation (see Krumpal 2013; Tourangeau & Yan 2007 for a comprehensive overview).

This article looks to contribute to better understanding the factors that lead to SD responses and ways to sensibly incorporate RLs to improve data quality. We investigate the question as to whether SD is more the result of automatic or deliberate processes and outline theoretical arguments for both scenarios. We use data collected in a tablet-based CASI survey of pre-service teachers in Germany that were asked to assess their suitability for their chosen profession. To approach the

Direct correspondence to

Henrik Andersen, Technische Universität Chemnitz, Faculty of Behavioural and
Social Sciences, Institute of Sociology, Chair for Sociology with a Focus on
Empirical Social Research, Thüringer Weg 9, 09126, Chemnitz, Germany
E-mail: henrik.andersen@soziologie.tu-chemnitz.de

research question, we examine not only the classic respondent- (i.e. need for social approval) and item-related characteristics (i.e. trait desirability) associated with SD bias, but also the speed at which the respondents gave their answers. We observe under what circumstances the respondents tended to overstate positive characteristics as well as understate negative ones and tie the results back to the theoretical discussion.

In the next section, we outline a theory of SD responding that incorporates both automatic and deliberate viewpoints and allows us to generate logical expectations for the later analysis. After giving an overview of our data and variables, we outline the analytical strategy which involves the specification of successive multilevel models. We then present our empirical results and finally summarize and discuss the implications for future research.

Theoretical Background

In this section, we outline two typical ways to approach the topic of SD: as a deliberate utility maximizing- and an automatic norm-conforming behaviour. We focus on some influential works by researchers in the analytical-empirical tradition.

Dual Processes and the Determinants of Social Desirability

It is now well established that SD bias encompasses at least two distinct factors (Holtgraves 2004; Krumpal 2013; Paulhus 1984; Paulhus & Reid 1991; Tourangeau & Yan 2007; Wiggens 1964). What is referred to as *impression management* describes situations in which respondents deliberately misreport either to gain approval or avoid disapproval. *Self-deception*, on the other hand, describes self-reports that are inflated but sincere (Paulhus 1984). Two different underlying cognitive processes are implied: impression management is a rational, utility-maximizing action that is motivated by the goal of gaining approval or avoiding disapproval. Self-deception can be seen as an automatic reaction to highly accessible and internalized social norms (Esser 1990; Kroneberg 2006).

In order to properly examine SD bias, we thus need a theoretical framework that encompasses both utility-maximizing rational actions as well as automatic norm-guided ones. The Model of Frame Selection (MFS, Esser 1991b; Kroneberg 2006; Mayerl 2009) offers such a framework and has previously been applied to explain respondent behaviour by several researchers (Esser 1990; 1991b; Mayerl 2009, 2010; Skarbek-Kozietulska et al. 2012; Stocké 2004, 2007; Wolter 2012; Wolter & Junkermann 2018). The MFS extends the classical rational choice theory (RCT) by 1) accounting for ostensibly non-utility-maximizing behaviour based on actors' subjective experiences, i.e. the *framing* of the situation and 2) incorporating what is referred to as *variable rationality*; the idea that actors reduce complexity and effort with the help of symbols, norms, habits and emotions (Kroneberg 2006). Both of these extensions are important for the analysis of SD bias and will be discussed in turn.

The MFS assumes actors go through several implicit steps before acting. The actor must first interpret the situation (*frame selection*), then they must identify sets of appropriate behaviours for the situation (*script selection*), before then performing the action (*action selection*) (Kroneberg 2006). The extent to which actors go through these steps in a deliberate as opposed to a spontaneous fashion refers to the assumption of variable rationality. Frames, scripts and actions can thus be selected in either a deliberate *reflecting-calculating* (rc) or *automatic-spontaneous* (as) manner (Esser 1991b; Kroneberg 2006). The factors that are said to determine the *mode* of selection are opportunities, motivation, effort and accessibility (this is compatible with social psychological dual-process theory, e.g. Fazio 1990a; see for an overview Mayerl 2009). Opportunities refer to things like time or capabilities; motivation is often provided by fear of making a wrong decision; deliberate consideration requires effort (whereas automatic actions require little); accessibility refers to the ease of finding appropriate selections (Kroneberg 2006).

In terms of SD, two of the most prominent applications of the MFS, an article by Esser from 1990[1] and another by Stocké from 2004[2], present contradictory accounts with regards to the question of whether SD is an automatic or deliberate action. It is important that the reader is aware of the fact that we will first outline the arguments as they were originally presented, and that the discrepancies therein represent part of the puzzle we wish to contribute to solving.

SD as an Automatic Response Behaviour

Esser (1990) describes social desirability as an automatic action that is the result of the cognitive accessibility, or *match*, of the frame of SD. He sees SD as a response set; a temporary strategy employed by respondents with a strong internalized *need for social approval* (NSA) to simplify their choice of actions. He describes that in a low-cost situation[3] such as a survey, the default mode for respondents is one of cooperation (‘provide valid answers’). For the frame of SD to become activated

[1] Here it is important to note that when we refer to ‘Esser’s standpoint’, we are referring to the argument laid out in 1990. At various points, Esser has presented both perspectives: making the argument for social desirability as a utility-maximizing behaviour (1986; 1991b) as well as a spontaneous norm-conforming behaviour (1990).

[2] Stocké published a very similar article in English in 2007 that covers the same theoretical ground.

[3] ‘Low-cost’ describes situations with low direct costs, low absolute opportunity costs and a low utility differential (see Mayerl 2010 for a more detailed overview).

and override the cooperative survey frame, the normative expectations of the situation must be transparent. This means the respondent must be able to recognize the existence of a social norm and determine which response option best fulfils the expectation (see also Wolter 2012). This transparency is based on the so-called *trait desirability* (TD) of the item. Trait desirability describes the overall strength and direction of the desirability of the question's content. It can be operationalized in various ways and summarizes the individual-level desirability beliefs (e.g.: "I think smoking is an undesirable habit", "Having had many sex partners is desirable", "Is it desirable or undesirable to say negative things about refugees?"). Esser's conception of SD suggests an interaction between TD and the respondent's NSA. TD informs the respondent about the normative expectations of the situation, the salience of which is increased by the respondent's NSA.

Esser's outline of SD thus hinges on the respondent choosing the frame of SD (F_{sd}) out of the set of other possible frames ($S_F = \left\{F_{sd} \in \left\{F_1,\ldots,F_N\right\}\right.$ for all $j \in N,\ j \neq sd$), of which the assumed default frame of cooperation, F_c (lower case '*c*'), is part of *j*. This means the match of the frame of SD (m_{sd}) must be greater than the match of any other frame:

$$m_{sd} > m_j, \tag{1}$$

where, for him, $m_{sd} = TD \times NSA$. This conception of a match corresponds to the idea that there must be situational objects present relevant to the frame (TD) and that the respondent must connect these objects to the frame (NSA, see Kroneberg 2006). Furthermore, if the automatic mode is activated, that is, the match is strong enough to at least equal the effort relative the subjective expected utility of the rc-mode, then the respondent will act automatically based on the activated frame of SD:[4]

$$SEU(as) \geq SEU(rc)$$

[4] We use Kroneberg's (2005; 2006) formalization for the sake of simplicity for much of this paper although there are other variants (e.g. Esser (2001; 2003) and Mayerl (2009). For low-cost situations like the vast majority of surveys, all three of these variants come to the same conclusion that a perfect match (m = 1) will always block the rc-mode (see Mayerl (2009) for an in-depth discussion on this topic). In high-cost situations, the versions of Esser and Kroneberg differ from Mayerl's: his MFS$_E$ (with 'exit option') states that, especially when the costs of choosing wrongly are high, a person may deliberate before acting *even if the match is perfect*. This can be shown by his formalization of the conditions necessary for the switch from as- to rc-mode ($SEU(rc) > SEU(as) \rightarrow \left[(U_{rc} - C) - (m_i U_i - C_w) + U_{intrinsic\ motivation}\right] p > m_i U_i$, compare with Inequality (4) below). This means that the theory as outlined in this paper as well as the empirical findings applies to typical survey situations but may not be applicable for surveys dealing with extremely sensitive topics that present more high-cost situations (e.g. illegal behaviour or infidelity).

which derives (2)

$$m_{sd} \geq 1 - C/\left(p\left(U_{rc} + C_w\right)\right)$$

(Kroneberg 2005, 2006) where m_{sd} is the degree of match between the situation and the frame of SD, C represents the costs associated with a deliberate choice (i.e. effort), p is the opportunity for reflection and $U_{rc} + C_w$ summarizes the motivation; with U_{rc} as the utility of a deliberate choice and C_w the consequences of choosing wrongly (Fazio 1990a; Kroneberg 2005). This is at least the case in low-cost situations (e.g. surveys) where a sufficient match of a frame can directly influence action, thereby skipping the script- and action-selection phase (Esser 1990; Kroneberg 2005; 2006; Mayerl 2009).[5] Even if we cannot operationalize the right-hand side of Inequality (2), we can make the ceteris paribus assumption that the clearer the norm (TD) and the higher the salience (NSA), resulting in a high match of frame and situation, the more likely an automatic SD response.

SD as a Deliberate Response Behaviour

Stocké (2004; 2007) describes the opposite standpoint. He sees SD as a deliberate utility-maximizing action. While Esser assumes the cooperative frame (F_c) per default, Stocké expands on this assumption and states that the extent to which the respondent cooperates with the goals of the researcher is determined by the (strength of their) attitude towards surveys. The more positive and cognitively accessible their attitude towards surveys, the more likely they cooperate. Respondents stray from their cooperative role when the subjective utility of a SD response crosses a certain threshold. Specifically, the subjective expected utility (SEU) increases based on the presence of three components: 1) the respondent's approval motive $\left(U_{SD} \in [0,1]\right)$, 2) clear desirability beliefs $\left(\Delta w_{TD} \in [-1, +1]\right)$ and 3) an absence of privacy ($w_p \in [0,1]$, Stocké 2004; 2007):

$$SEU\left(SD\right) = U_{SD} \times \Delta w_{TD} \times w_p, \tag{3}$$

Being a multiplicative equation, each of these components must be given in order to expect an SD response and turn the respondent from a cooperator to a 'conformer'.

[5] When, for example, the frame clearly defines both the script and action: $a_j = 1$, $a_{j|i} = 1$ and $a_{k|j} = 1$, where $a_{j|i}$ is the accessibility of script *j* given frame *i*, a_j is the availability of script *j* and $a_{k|j}$ is the degree to which script *j* regulates action *k*. In such case, the *activation weight* of action *k* $\left(AW(A_k \mid S_j)\right)$ is governed solely by the match of the frame *i*: m_i; see Esser 1990; Kroneberg 2006.

Although not explicitly stated by Stocké, his argumentation seems to represent a truncated and somewhat altered version of the typical decision-theoretic specification of the conditions for the switch from an as- to an rc-mode:

$$SEU\left(rc\right) > SEU\left(as\right)$$

which derives (4)

$$p\left(1-m_c\right)\left(U_{rc}+C_w\right) > C,$$

(Kroneberg 2005; 2006; Kroneberg et al. 2010) where $\left(1-m_c\right)$ is the degree of mismatch between the situation and the default frame (in this case of cooperation).

We can assume that, for most respondents, the opportunity for reflection (*p*, i.e. ability) is given and thus equals one. If we can accept that $U_{SD} \times \Delta w_{TD} \times w_p$ represents the respondent's motivation to give an SD response,[6] we can state that the respondent may switch to a deliberate mode and consider the option of giving an SD answer if he or she identifies an alternative frame and has the motivation to do so:

$$\left(1-m_c\right)\left(U_{SD} \times \Delta w_{TD} \times w_p\right) > C. \tag{5}$$

Inequality (5) is our own interpretation of Stocké's (2004; 2007) argument brought together with the more general decision-theoretic specification of Kroneberg (2005; 2006; Kroneberg et al. 2010). While it is typically difficult to operationalize the degree of mismatch ($1-m_c$), we can state that, ceteris paribus, the likelihood of a deliberate SD response increases with motivation (the second bracketed parameter in Inequality (5)).

Stocké's (2004; 2007) assertion that privacy concerns are necessary to expect an SD response is problematic for several reasons. For one, findings on the effect of anonymity of SD bias are mixed. There is a great deal of empirical research finding anonymity has little or no effect on SD (e.g. Börger 2013; Dwight & Feigelson 2000; Hancock & Flowers 2001; Krysan 1998; Northover et al. 2017; Richman et

[6] It is not entirely clear, based on Stocké's (2004, 2007) argumentation, how the approval motive, trait desirability and privacy concerns should translate into the more general MFS framework. We could speculate that the respondent's approval motive multiplied by the desirability beliefs concerning a survey item could represent the utility of a deliberate choice $\left(U_{SD} \times \Delta w_{TD} = U_{rc}\right)$, and that privacy concerns represent the costs of a wrong choice $\left(w_p = C_w\right)$. This would change Stocké's assertion that a lack of privacy concerns should negate entirely the utility of an SD response (making the contribution of w_p additive rather than multiplicative) and brings it more in line with our belief that privacy concerns can increase the motivation to answer in an SD way, but are not necessary.

al. 1999; Weisband & Kiesler 1996).[7] From a theoretical standpoint, it can also be argued that intrinsic motivations can lead respondents to provide SD answers even in anonymous conditions. Wolter (2012), for example, points to the concept of cognitive dissonance which was introduced by Festinger (1957). Cognitive dissonance describes discomfort that results when conflicting attitudes exist at the same time or when one's attitude and behaviour does not match (Wolter 2012, p. 166). For example, cognitive dissonance could result when a pre-service teacher believes strongly that good teachers are funny, but realizes that they themselves are not funny. One way to deal with cognitive dissonance and relieve the feeling of discomfort (especially when other options – such as changing one's behaviour – are out of the question) is to trivialize or ignore the existence of dissonant attitudes, beliefs or behaviours. Thus, we can assume that non-conformity to social norms can create cognitive dissonance in respondents, and that this can occur even in anonymous conditions. In fact, as Wolter (2012) points out, it may be more accurate to say that intrinsically motivated desirable responses are the result of the frame of 'neutralizing cognitive dissonance' that is functionally equivalent to the frame of 'social desirability' as outlined above. Also, as early as 1986, Esser described this type of intrinsically motivated SD as 'cultural' and the more traditional type outlined by Stocké (2004; 2007) as 'situational' SD.

For these reasons, we expect SD responses to be the result of the respondent's need for social approval and their desirability beliefs – or, indirectly, the trait desirability of the item. A lack of anonymity, whether perceived or real, may increase the likelihood of a 'situationally' motivated SD response, but we do not expect that it is necessary. Rather, in accordance with the cognitive dissonance argument, the mere fact that the respondent realizes their behaviour or characteristics do not live up to either their own beliefs or attitudes, or the predominant views of society in general, should be enough to generate SD bias. The question remains whether the determinants of SD bias encourage an automatic norm-conforming- or a deliberate approval-maximizing response. This will be the focus of the next section.

Desirable vs. Undesirable Traits

The arguments for SD bias as an automatic and as a deliberate action both point to the same main determinants: the respondent's need for social approval and the trait desirability of the item. The argument for SD as an automatic action states that the presence of both determinants increases the likelihood that the frame of SD can be matched to the situation, leading to quick SD responses. The argument for SD as

[7] Although there are also examples of studies finding an effect (e.g. Bader et al. 2016; Booth-Kewley et al. 2007; Dodou & de Winter 2014; Joinson 1999; Kays et al. 2012; Kreuter et al. 2008; Krysan et al. 1994).

a deliberate action states that these same determinants increase the motivation to consider the option of providing an untruthful answer. This should lead to slower SD responses.

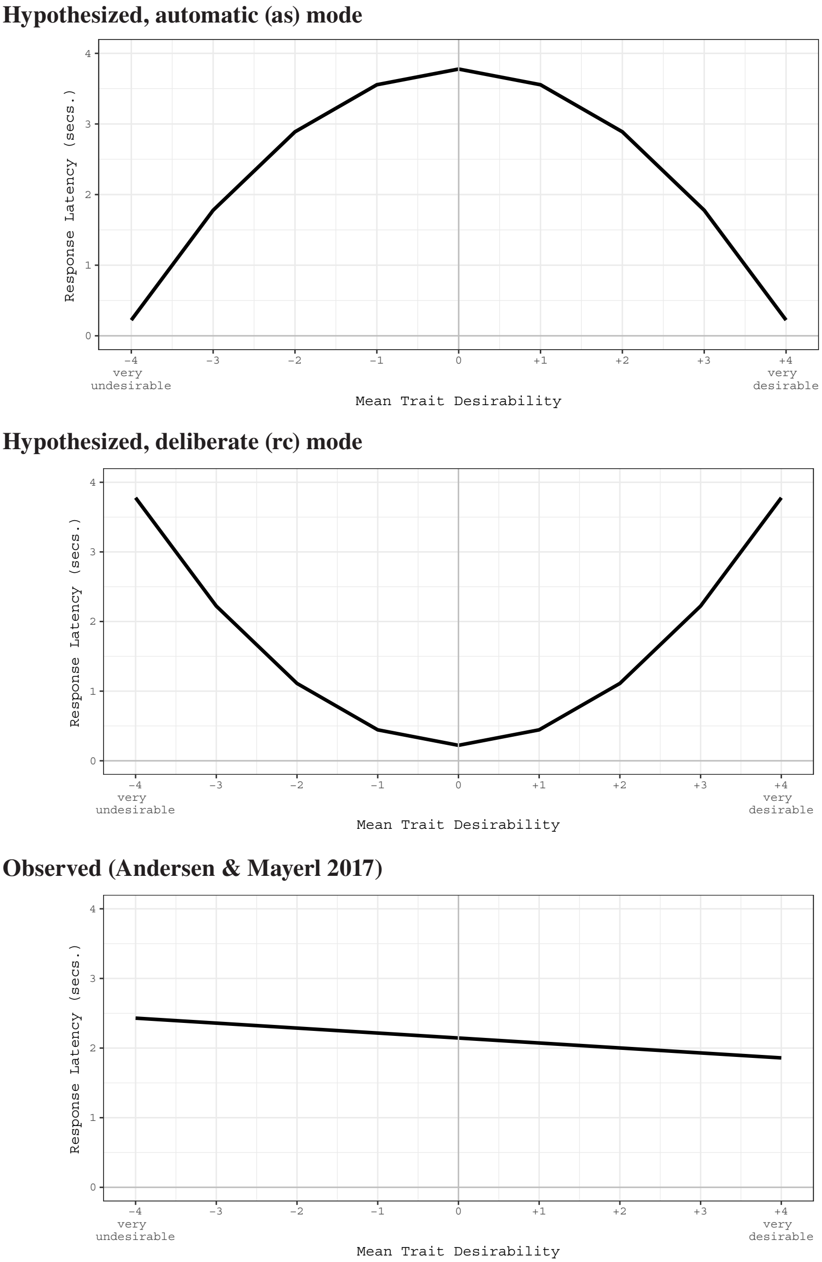
It is unlikely that the respondent's NSA on its own should govern the mode of responding. It does not seem plausible, for example, to assume that a respondent with a strong NSA will always answer faster or slower than a respondent with less of the characteristic. Rather, SD hinges on the *transparency* of the existence of normative expectations; i.e. the desirability beliefs of the respondents vis-à-vis the particular item content. NSA can be seen as heightening the salience of these subjective social norms (Esser 1990).

Thus, it would seem the mode selection, automatic or deliberate, is dependent primarily on the item content. If we imagine a graph with an item's TD on the x-axis ranging from very undesirable to very desirable (with neither undesirable nor desirable in the middle of the scale) and the response latencies on the y-axis, the automatic argument would suggest an inverted U-shape: the clearer the social norms are (increasing desirability and undesirability), the faster the responses should be. On that same graph, the deliberate argument would suggest the opposite: a U-shaped curve with responses becoming slower the clearer the social norms. The top two panels of Figure 1 summarize these theoretical expectations.

Results of a Previous Study

In a previous study, we examined the relationship between item- and respondent-related characteristics and response latencies (Andersen & Mayerl 2017). Whereas response latencies are the independent variable in this study, then they were the dependent variable. The aim of the study was to take a preliminary look at how the determinants of SD (particularly TD and NSA) affected the length of time the respondents took to answer the questions. In terms of TD, we did not find the expected U- or inverted U-shaped curve as outlined above, as the squared TD term had no significant effect. Rather, the main effect of TD was negative and significant. The bottom panel of Figure 1 shows the empirical results contrasted with the theoretical expectations outlined above.

On a bipolar scale, the negative effect means that response latencies become faster the more desirable the item content is. On the other hand, the more undesirable the item content, the slower the responses become. This effect remains when controlling for factors such as the respondent's baseline speed, the length of the question, its position in the questionnaire, etc. The effect is furthermore linear in nature; on a scale from -4 to +4, latencies become increasingly slower as the item content becomes more undesirable (meaning it does not seem to be merely a result of negative keying).



Hypothesized, automatic (as) mode

Hypothesized, deliberate (rc) mode

Observed (Andersen & Mayerl 2017)

Note. The scale of the y-axis as well as the exact shape of the curve in the hypothesized diagrams is arbitrary; see Andersen & Mayerl 2017

Figure 1 Hypothesized and observed relationship between trait desirability and response latency

We took this as evidence to suggest that not only the transparency of the social norm surrounding a survey item but also its *direction* is important for determining the mode of responding. Undesirable item content seems to trigger more deliberate responses while desirable item content seems to lead to more automatic ones. In fact, some research has dealt with the possibility that certain respondents react more strongly to desirable content and others to undesirable content. Paulhus has even suggested a four-factor typology of SD responding that differentiates between the *degree of awareness* (impression management vs. self-deception) as well as the *content* (Paulhus 2002). Along the content dimension, respondents are grouped according to their motivation for answering untruthfully. Respondents that are motivated by *egoistic* factors attempt to present themselves in an overly positive light, highlighting their social and intellectual traits such as dominance, fearlessness, emotional stability, intellect and creativity (Paulhus 2002, p. 63 f.). Respondents that are motivated by *moralistic* factors tend to deny socially-deviant characteristics and claim such social qualities as agreeableness, dutifulness and restraint (Paulhus 2002, p. 64). Uziel (2010) refers to a similar typology and uses labels previously coined in earlier work by Paulhus & Reid (1991): *adjustment* and *defensiveness*.[8] Defensiveness is characterized by the "avoidance of threatening situations" (Uziel 2010, p. 247), and that defensive respondents are motivated not by "social approval, but rather the avoidance of social disapproval" (Uziel 2010, p. 247). Adjustment describes respondents that tend to use the survey situation as a way to exaggerate positive characteristics like friendliness, stability and well-being (Uziel 2010, p. 248).

While research on the topic of a possible four-factor model of social desirability (impression management vs. self-deception and egoistic vs. moralistic) has not fully matured for various reasons,[9] we believe this line of reasoning may be promising in explaining why response latencies seem to react differently based on the content of the question. Without the possibility of operationalizing a fully differentiated NSA scale, our preliminary work nevertheless leads us to believe that not only the strength but also the direction of the TD should be of importance.

[8] There are other terms used to describe this difference, Damarin & Messick (1965), refer to 'propagandistic' and 'autistic' motives, for example.

[9] Personality-scales meant to assess those with a tendency towards a self-deceptive moralistic bias, or 'self-deceptive denial' have not been popularly implemented due to ethical concerns and factor analytic empirical evidence suggesting it is rather weakly pronounced (Paulhus & Reid 1991).

Research Questions and Hypotheses

This study is interested in examining two main questions. First, is SD the result of an automatic or a deliberate process? We examine this question by specifying a three-way interaction between the determinants of SD (trait desirability and need for social approval) and the response latencies, and observing the self-reported scores given by the respondents. If SD is an automatic behaviour as outlined by Esser (1990), then we should observe more biased scores when the match is sufficient ($m_{sd} = TD \times NSA$) *and the respondent answers quickly.* If SD is a deliberate behaviour as argued by Stocké (2004, 2007), then scores should be more biased when the motivation is sufficient ($U = TD \times NSA$) *and the respondent answers slowly.* Again, the inconsistency of the views of Esser and Stocké should be clear: how can the interaction between TD and NSA at once represent the match and the motivation? However, by looking at the three-way interaction $TD \times NSA \times RL$ and observing how respondents answered, we aim to identify which conceptualization is more plausible. It is entirely possible that any and all components of the interaction could fail to show significant effects on the scores of the respondents. It could be that the interaction $TD \times NSA$ affects scores but that speed at which the respondent answers plays no role, for example. We therefore proceed in a step-wise fashion, first looking at the main effects individually, then the two-way effects before finally moving on the suggested three-way interaction.

With our second research question we hope to contribute to finding a way to bridge the gap between the competing conceptualizations. It seems likely, based on an abundance of empirical evidence, that SD can be both an automatic as well as a deliberate behaviour. But what are the mechanisms responsible for determining the mode? Obviously, we cannot state that $TD \times NSA$ at once causes automatic and deliberate SD responses. However, based on the four-factor SD typology put forth by Damarin & Messick (1965), Paulhus & Reid (1991), Paulhus (2002), and Uziel (2010) and our observations from previous research, we have reason to believe that the *direction of the TD*, whether desirable or undesirable, may be an often-overlooked factor that influences how SD manifests.

We integrate the theoretical and empirical knowledge and formulate the following hypotheses:

Hypothesis 1: highly *desirable* item content and a *strong need for social approval* should mean that *faster responses* are associated with more positive scores.

Hypothesis 2: highly *undesirable* item content and a *strong need for social approval* should mean that *slower responses* are associated with more positive scores.

Data and Method

Data

The study uses data from a research project carried out at the Technische Universität Kaiseralutern called EVA3PLUS. The project is a longitudinal panel-study with computer assisted self-interview (CASI) tablet questionnaires with three survey waves taking place at intervals of around six months. The project attempted to conduct a complete sample of all biology and chemistry pre-service teachers at the Gymnasium-level (a university/college preparation-level secondary school form in Germany) in Rhineland-Palatinate from mid-2014 to mid-2017. In total, the overall sample size for the study is 631 with 416 individual respondents participating between one and three times. Substantively, the study looks at pre-service biology and chemistry teachers' attitudes and behaviours with regards to using experiments in the classroom. The methodological focus of the project is on the use of response latencies to improve the quality of survey data.

Variables

The dependent variable are scores on 30 items of teacher-related characteristics, each measured on a 7-point rating scale (Appendix 1 shows the descriptive statistics of the dependent variable and Appendix 3 reports the wording of the 30 items along with the mean trait desirability scores). The items asked respondents to self-assess their qualities as a teacher. They included statements such as "Spending time with teenagers is a lot of fun" and "I feel insecure when I have to speak in front of others". Normatively speaking, these are characteristics teachers should (or should not) possess: they should like spending time with teenagers and should not have problems speaking in front of others, for example. We assume, therefore, that they are principally sensitive topics for future teachers. Although the surveys were conducted anonymously and without the presence of an interviewer, we assume further that confronting the fact that one does not possess a desirable characteristic (or rather that one possesses an undesirable characteristic) will lead to uncomfortable cognitive dissonance ("I want to be a teacher, but I am not good at being a teacher", see Wolter 2012). Items suggesting undesirable characteristics were recoded so that higher values always indicate more desirable answers (agreeing to possessing positive characteristics and disagreeing to negative ones).

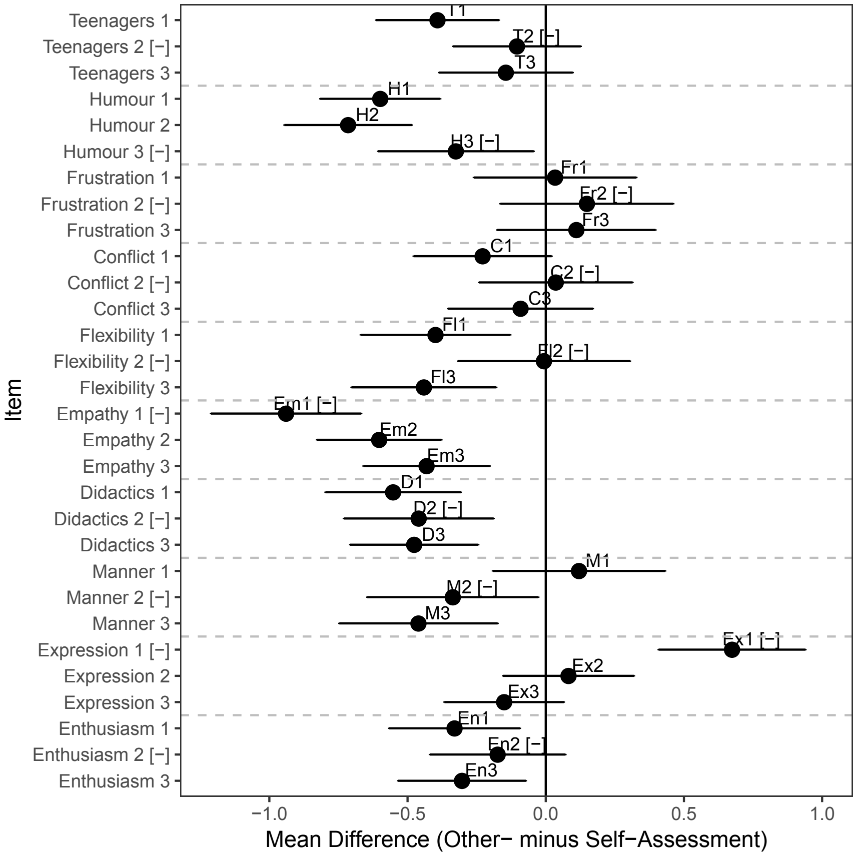
The method for collecting the response latencies is outlined in detail in Andersen & Mayerl (2017). Due to the large degree of non-normality of the distribution, and in order to eliminate outliers, the top and bottom 5% of the distribution was eliminated (see Mayerl & Urban 2008 for more on the preparation of raw RLs for analysis). This resulted in a mean response latency of 4.6 seconds (std. dev. = 2.0).

The response latency variable is continuous; it does not represent a dichotomous pair of options but rather illustrates a continuum with a deliberative-controlled mode on the pole of high elaboration and the automatic-spontaneous mode at the other extreme end of low elaboration (see for more on this Carlston & Skowronski 1986; Gibbons & Rammsayer 1999; Mayerl 2010; Schaffner & Roche 2016, Sheppard & Teale 2000; Shiv & Fedorikhin 2002).

Desirability beliefs can be assessed by either asking the respondent themselves whether a characteristic is desirable or undesirable in their opinion, or by asking the respondent to judge how the characteristic is viewed by society in general (Stocké 2004). In either case, the trait desirability of an item is generated by aggregating the individual desirability beliefs into an overall measure. In order to assess the trait desirability of the items, a small secondary pencil-and-paper survey of other students in biology and chemistry teachers' education programs at the Technische Universität Kaiseralutern was conducted (n = 77). The sample populations of the main study and the small trait desirability supplementary study can be seen as very homogenous groups. The students were asked to assess how desirable the various teacher characteristics were seen in society in general. The scale ranged from -4: "extremely undesirable" to +4: "extremely desirable" with 0 as the middle category: "neutrally seen". The mean scores can be found in Appendix 3.

The respondents' need for social approval was measuring using two items from the Crowne-Marlowe SD scale (Crowne & Marlowe 1960, p. 351). The index was created as an average of the two scores, displaying satisfactory characteristics ($\alpha = .65$). In cases in which the respondent answered the NSA scale in two different waves of data collection, the NSA score was averaged over the two occasions. If the respondent took part more than once in the overall survey but only provided valid NSA scores on one occasion, those values were copied over to the other wave(s). We feel comfortable in doing this as NSA is typically seen as a stable personality trait: z_i as opposed to z_{it} to put it in terms of a typical panel analysis, see the analytical strategy section (DeMaio 1984; Krumpal 2013; Tourangeau & Yan 2007). The descriptive statistics of the items are found in Appendix 2. In order to better interpret the three-way interaction between TD, NSA and RL, for the analysis we collapsed the scale into a dichotomous variable with 0 = weak to moderate NSA (< 6) and 1 = strong NSA (≥ 6).

We include other respondent- and item-characteristics into the models as fixed effects: the respondent's tendency to acquiescence (based on a count of the amount of times the respondent answered "completely agree" on 64 other survey items), sex (male = 1), year of birth, whether or not they had taken part in the survey before (repeat = 1) and the number of item syllables. As they are specified, the models allow us to include such variables and observe their effects but they are not strictly necessary. The use of respondent and item fixed-effects multilevel models through within-cluster centering allows us to control for unobserved differences between



Note. Error bars show confidence intervals (95%); [-] identifies undesirable item content; created with sjPlot package in R (Lüdecke 2017)

Figure 2 Mean differences between other- and self-assessments

respondents and items (more on that below, see Enders & Tofighi 2007; Rüttenauer 2018). Descriptive statistics of the predictors can be found in Appendix 1.

Self- vs. Other-Assessment

Without validation data, studies looking at SD bias are often forced to use the 'more (or less) is better' assumption (Wolter 2012). Here, we take higher item scores as an indication of more biased responses. Obviously, this assumption is problematic because it is not possible to disprove that high item scores are not just truthful answers by respondents that actually possess a desirable trait to a high degree. To some extent, this is not particularly troubling because we include explicit SD indicators as explanatory variables in the model. If high item scores are not at least partially the result of SD bias, then we should not expect any meaningful results from these predictors.

To further put concerns to rest, we collected a secondary sample in which we asked the instructors at the teachers' colleges (n = 175) to assess the study respondents' possession of the 30 characteristics. The 'other-assessment' questionnaires were sent out within a week or so of the respondents having completed the main survey. This other-assessment survey gives us an external criterion with which we can test the plausibility of the assumption that some scores are, indeed, biased by SD. Figure 2 summarizes the results of this secondary study. It shows the mean differences between the other- and the self-assessments (with 95% confidence intervals). Negative values indicate the respondents' instructor rated the person more poorly than the person rated themselves. We take mean values in the negative range as evidence that a substantial number of respondents answered in an SD fashion (i.e. presented themselves in a more positive light than the external criterion).

Unfortunately, due to the relatively small sample size and further item nonresponses, it was not practical to include this information in the following statistical models. However, the findings give us confidence in continuing on with the analysis under the assumption that more positive self-assessments are at least partially the result of SD bias.

Analytical Strategy

The data is structured as follows: respondents ($j = 1 \ldots J$) and items ($k = 1 \ldots K$) are crossed; each respondent answers each item and each item is answered by each respondent (at least ideally, given no item nonresponse). We refer to measurements at the lowest level ($i = 1 \ldots N$) as 'events' which are nested at once within both respondents and items. Events cover all variables that vary within respondents and items, including response latencies (which we can refer to as $x_{i(jk)}$) and our dependent variable, item scores ($y_{i(jk)}$). The respondents' need for social approval (NSA_j) and the item's trait desirability (TD_k) vary across respondents and items, respectively.

We use multilevel modeling to account for the hierarchical nature of the data (Hox et al. 2018). This allows us to account for the nested structure by including random effects for our grouping variables. Furthermore, we apply within-cluster centering to our level 1 predictor, response latencies. This has the effect of ensuring our level 1 predictor is uncorrelated with the higher level predictors, and makes the corresponding regression slopes based solely on within-cluster variation (see Enders & Tofighi 2007 for a comprehensive overview of within-cluster centering, see also Allison 2009). Thus, doing so allows us to control for unobserved differences between respondents and items. For this reason, such models are sometimes referred to as fixed-effects multilevel models (e.g. Rüttenauer 2018). In fact, for studies interested in the effect of level-1 predictors or cross-level interactions, Enders & Tofighi (2007) suggest always centering level-1 variables within-cluster. Variables at the higher level were centered around the grand mean except for the dummies for NSA, sex, and repeat respondents which retained their original metric.

We began our analysis by specifying an intercept-only model (Model 0, not shown in Table 1) that included random intercepts for respondents and items but no predictors at any level. The interclass correlations (ICCs) obtained from that model showed that 17% ($\rho_j = .17$) and 4% ($\rho_k = .04$) of the variance in item scores (y) is attributable to the respondents and the items, respectively (for more on this see Hox et al. 2018). In a second step, we tested whether the slopes of RL on scores varied systematically between respondents or items. The results showed a model that included by-respondent and by-item intercepts and by-item random slopes for RLs. We settled on this model specification based on a likelihood ratio test that showed significant by-item slope variation compared to one with only random intercepts ($\chi^2(1) = 29.566$, $p < 0.001$, see Baayen, Davidson & Bates 2008).

In order to gain a better understanding of the contribution of the various predictors, we proceed in a step-wise fashion, first introducing the main effects of all predictors at the various levels (Model 1), before then introducing two-way interactions between the predictors of interest (Model 2), and then finally introducing the three-way interaction between the determinants of SD (TD and NSA) and the RLs (Model 3). Doing so allows us to observe the effects in isolation before moving on to the interpretation of the more complicated ones. Model 1, which includes the main effects of all predictors at all levels can be written as

$$\begin{aligned} item\,score_{i(jk)} = \gamma_{0(00)} + \gamma_{1(00)}\left(RL_{i(jk)}\right) + \gamma_{0(10)}\left(NSA_j\right) + \\ \gamma_{0(01)}\left(TD_k\right) + \ldots + u_{0j} + u_{0k} + u_{1k}\left(RL_{i(jk)}\right) + e_{i(jk)}, \end{aligned} \tag{6}$$

where $\gamma_{0(00)}$ is the overall intercept, u_{0j}, u_{0k} and $e_{i(jk)}$ are the respondent-, item- and idiosyncratic deviations from the overall mean and u_{1k} the by-item random slope parameter. $\gamma_{1(00)}$, $\gamma_{0(10)}$ and $\gamma_{0(01)}$ are the coefficients for the variables RL, NSA and TD, respectively. For the sake of simplicity, the other control variables are not shown in the equation. The inclusion of cross-level interactions between predictors at various levels follows straight-forwardly from Equation (6).

Table 1 Fixed-effects multilevel models, dependent variable: item scores (recoded)

	Model 1		Model 2		Model 3	
	b	se	b	se	b	se
Intercept	4.306 ***	(.117)	4.316 ***	(.124)	4.316 ***	(.123)
Event-level variables						
Response latency (RL)	-.051 **	(.016)	-.049 **	(.018)	-.049 **	(.018)
Respondent-level variables						
Repeat respondent	-.039	(.088)	-.040	(.088)	-.042	(.088)
Need social approval (NSA)	.136	(.091)	.122	(.127)	.127	(.127)
Acquiescence	.407	(.496)	.409	(.495)	.409	(.495)
Male	.125	(.100)	.126	(.100)	.125	(.099)
Year of birth	-.010	(.020)	-.010	(.020)	-.010	(.020)
Item-level variables						
Syllables	-.020 *	(.009)	-.020 *	(.009)	-.020 *	(.009)
Trait desirability (TD)	.112 ***	(.022)	.026	(.040)	.104 *	(.048)
Cross-level interactions						
TD x NSA			-.014	(.017)	-.150 **	(.050)
TD x RL			-.022 **	(.007)	-.005	(.009)
NSA x RL			-.003	(.019)	-.002	(.019)
TD x NSA x RL					-.029 **	(.010)
Goodness of fit						
AIC	23,377.5		23,374.8		23,368.3	
BIC	23,472.8		23,490.6		23,490.9	
Log-Likelihood	-11,674.7		-11,670.4		-11,666.2	
Observations						
Total	6,693		6,693		6,693	
Groups						
Respondent	244		244		244	
Item	30		30		30	
Variance components						
Respondent $\left(\sigma^2_{j-\text{int}}\right)$	.381		.380		.379	
Item $\left(\sigma^2_{k-\text{int}}\right)$	.121		.087		.086	
Item $\left(\sigma^2_{k-slope}\right)$	.005		.003		.003	
Residual $\left(\sigma^2_{e}\right)$	1.765		1.765		1.763	

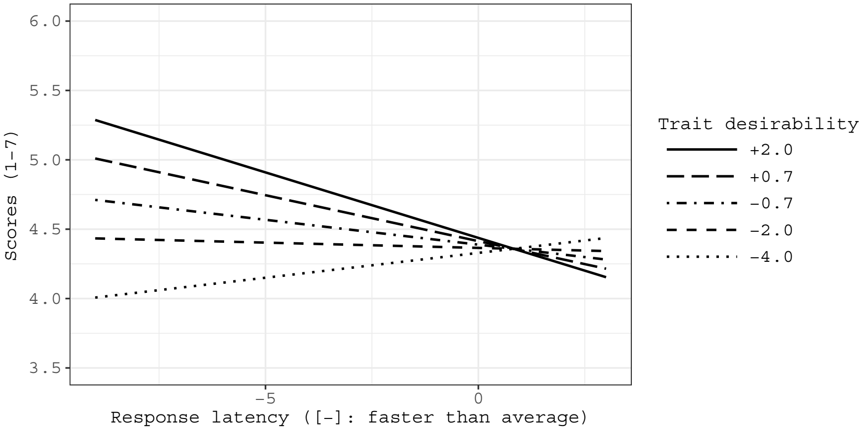
Note. Estimator: REML, for goodness of fit statistics model was re-ran with ML; event-level predictor RL was centered within-cluster, higher level variables centered around grand mean; unstandardized estimates; models estimated using lme4 package in R (Bates et al. 2015); ***p<.001, **p<.01, *p<.05, $^{+}$p<.10; two-sided test

Analysis

The results of the analysis can be found in Table 1. It shows the unstandardized coefficients (b) and standard errors (se). As for Model 1, which includes only the main effects of the predictors at all levels, we see that the TD of the item has a significant positive effect on scores (b = .112, p < .001), meaning the more desirable the trait, the more respondents tended to claim to possess it. Here it is important to note that while scores were recoded so that higher values always indicated more desirable responses, TD was measured on a bipolar scale (from -4 to +4 before centering).[10] This means that undesirable and desirable items were not treated equally by respondents. Desirable traits lead disproportionately to more positive answers than undesirable traits lead to less negative ones. Finally, we observe a significant negative effect of RLs (b = -.051, p < .01). The longer the respondent took to answer the question, the more negatively the respondents rated their qualities as a teacher. Looking just at the isolated effect of RLs on scores, however, does not tell us anything about SD responses. In order to better understand the extent to which RLs relate to SD, we must look at them in combination with the determinants of SD. This is shown in Model 2.

Model 2 introduces all two-way interactions that are implied by the three-way interaction in Model 3. Here, we see that the interaction between the TD and RL is significant (b = -.022, p < .01). Figure 3 shows the interaction graphically. The result suggests that *only fast responses seem to be influenced by the desirability of the item content.* This is evidenced by the intercepts, the ranking of which corresponds to the TD value. Amongst fast responses, the difference in scores between very desirable (solid line, +2) and very undesirable (lower dotted line, -4) items is fairly substantial, roughly one and a half scale-points. On the other hand, there is almost no difference in scores for slow responses based on TD. As mentioned earlier in reference to the previous study, here too the effect of TD does not seem to be simply due to the item keying. If it was, the regression lines would not fan out. If this was the case, the slopes for the items of above-average desirability would overlap; the same would go for the undesirable side. Also, as with the effect of TD in Model 1, scores are disproportionately affected by desirable item content. In fact, the regression slope for the most undesirable content (lower dotted line) is slightly positive, meaning respondents answering more slowly to these items rated their teaching characteristics more positively. However, it is difficult to interpret this as an 'editing' process (Tourangeau & Yan 2007) as the slowest of responses are not nearly as positive as the fast responses for desirable items.

[10] Also, for the sake of simplicity, we will often refer to 'desirable' vs. 'undesirable' traits – however, due to the centering of the variables, we are actually comparing items of 'above average desirability' with those of 'below average desirability'.

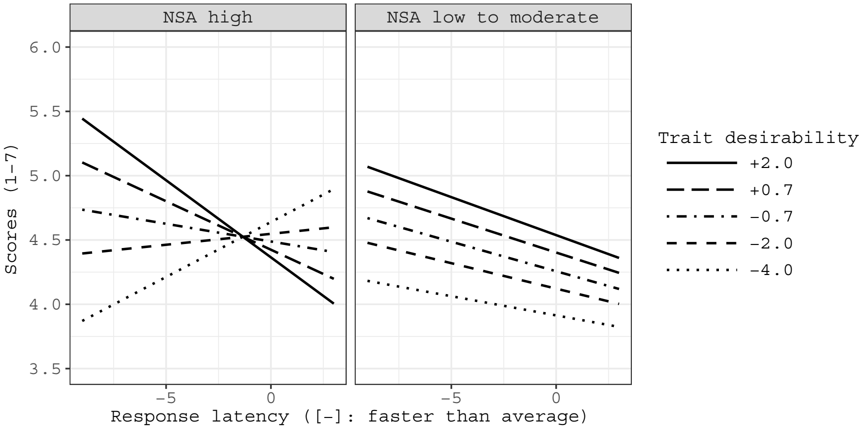


Note. Created with ggplot2 package in R (Wickham 2009)

Figure 3 Two-way interaction: trait desirability x response latency (Model 2)

In model 2, the main effect of TD falls out of significance. As is the case in all models, the effect of the number of syllables is significant (the effect stays constant throughout at b = -.020, p < .05). The longer the question, the more negatively the respondents rated their teaching qualities. On the other hand, the interactions of NSA with TD and RL are not significant. This means that the speed of responses does not moderate the effect of NSA on scores and that the central implied interaction $TD \times NSA$ does not systematically influence scores.

Although Model 2 shows the central interaction is not significant, we nevertheless test the three-way interaction $TD \times NSA \times RL$ in Model 3. This interaction is in fact significant (b = -.029, p < .01) and can be inspected graphically in Figure 4. Whereas Figure 3 suggests that *only fast responses are affected by TD*, Figure 4 shows us that this is not exactly the case. To illustrate this, we start by describing the right side of Figure 4 which shows the results for respondents with a weak to moderate NSA. Here, we see that TD has an effect on scores as evidenced by the spread of the intercepts. The more desirable the trait, the more respondents claimed to possess it (and vice versa). Curiously, for respondents with a weak NSA, longer responses are actually associated with more negative self-assessments.



Note. Created with ggplot2 package in R (Wickham 2009)

Figure 4 Three-way interaction: trait desirability x need for social approval x response latency (Model 3)

Now, if we compare this to the left side of Figure 4, the relevance of the result to the theoretical discussion above becomes clearer. As with Figure 3, we see a fairly pronounced effect of TD on scores for fast responses (see intercepts). However, the slopes for the extreme TD values (solid line and dotted line) are steeper amongst those with a strong NSA. For the most desirable traits, faster responses are substantially more positive than slower ones. For undesirable traits, it is the slower responses that are more positive. To summarize, we can state that the answers of respondents with a low to moderate NSA are influenced by the TD of the item, and that their answers are more consistent regardless of how long they take to answer the question. In fact, if anything they actually tend to become more reserved the longer they take to answer. For respondents with a strong NSA, the negative effect of desirable traits and the positive effect of undesirable ones are almost equally strong.

Discussion and Conclusion

The findings generally lend support to our hypotheses. If we can accept response latencies as an appropriate proxy for the degree of elaboration (with automatic and deliberate modes at each end of the spectrum), then social desirability seems to be the result of both automatic and deliberate actions. The mode of response seems to be in part dependent on the *desirability or undesirability* of the item content.

We take the results to indicate that respondents that answer quickly to desirable traits may be answering in a SD way, irrespective of their NSA. For undesirable traits, the longer the response, the more positive the self-reports become in the case of strong NSA. Thus, NSA seems have a strong moderating effect on the interaction between RL and TD. Taken together, we are left with the conclusion that both scenarios (automatic and deliberate) are as plausible now as when we started out. Our results suggest a strong need for social approval and a very desirable trait leads to more automatic SD answers as outlined by Esser's argument. On the other hand, Stocké's assertion that trait desirability and need for social approval lead to deliberate SD answers is supported if one looks only at the very undesirable traits. We suggest, therefore, that the content of the item may be an important factor that determines the *mode of response*. This has not been discussed by either Esser (1990, 1991a, 1991b) or Stocké (2004, 2007), but could be an overlooked factor that allows both views to exist simultaneously. In general, our results suggest it is unlikely that socially desirable responses are either simply fast or slow. However, at this point, the exact mechanism responsible for this observed relation can only be speculated on. More work is needed to investigate the interplay in greater detail and assess the generalizability of the results.

In fact, it could be that the results of this study are specific to our research/survey design: tablet-based CASI surveys of a relatively homogenous sample regarding a very specific, relatively low-cost topic. Other types of surveys (web-based, CATI, CAPI), samples and topics could yield different results. Also, the analytical framework does not make it possible to truly test whether, for example, *trait desirability leads to fast/slow responses* which has been taken for granted throughout this article. It is possible that the causal direction is actually the opposite: perhaps respondents that take their time with the survey tend to be more receptive to the TD of the item. Furthermore, our expectations in this study were strongly influenced by what we empirically observed in a previous study. While there is some research that supports the overall sentiment that respondents may react differently based on certain types of questions, we are still very much in the beginning stages of fleshing out our theoretical argument. More work is needed that brings together not only the psychological work on egoistic/moralistic bias but also the methods such as response latencies to measure cognitive processing modes.

We ultimately encourage a more systematic investigation and manipulation of the various components. Indeed, socially desirable responses seem dependent on a complex interplay between respondent-, item-, and survey-characteristics. We hope with this article to draw attention to this and contribute to a better understanding of the use of response latencies to identify and hopefully correct measurement bias due to social desirability.

References

Allison, P. (2009). *Fixed Effects Regression Models*. Thousand Oaks: Sage Publications.

Andersen, H., & Mayerl, J. (2017). Social Desirability and Undesirability Effects on Survey Response Latencies. *Bulletin of Sociological Methodology*, 135(1), 68-89.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subject and items. *Journal of Memory and Language*, 59, 390-412.

Bader, F., Bauer, J., Kroher, M., & Riordan, P. (2016). Privacy Concerns in Responses to Sensitive Questions. A Survey Experiment on the Influence of Numeric Codes on Unit Nonresponse, Item Nonresponse, and Misreporting. *methods, data, analyses (mda)*, 10(1), 47-72.

Bassili, J. (2003). The minority slowness effect: Subtle inhibitions in the expression of views not shared by others. *Journal of Personality and Social Psychology*, 84(2), 261-276.

Bassili, J., & Fletcher, J. (1991). Response-Time Measurement in Survey Research a Method for CATI and a New Look at Nonattitudes. *Public Opinion Quarterly*, 55(3), 331-346.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects model using lme4. *Journal of Statistics Software*, 67(1), 1-48.

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior*, 23(2007), 463-477.

Börger, T. (2013). Keeping up appearances: Motivations for socially desirable responding in contingent valuation interviews. *Ecological Economics*, 87(2013), 155-165.

Carlston, D., & Skowronski, J. (1986). Trait Memory and Behavior Memory: The Effects of Alternative Pathways on Impression Judgment Response Times. *Journal of Personality and Social Psychology*, 50(1), 5-13.

Crowne, D., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349-354.

Damarin, F., & Messick, S. (1965). Response Styles as Personality Variables: A Theoretical Integration of Multivariate Research. *ETS Research Bulletin Series*, 1(1965), i-116.

DeMaio, T. (1984). Social Desirability and Survey Measurement: A Review. In: Turner, C. & Martin, E. (Eds.), *Surveying Subjective Phenomena, Vol. 2*. New York: Russel Sage Foundation.

Dodou, D., & de Winter, J. C. F. (2014). Social desirability is the same in offline, online and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36(2014), 487-495.

Dwight, S., & Feigelson, M. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, 60(3), 340–360.

Enders, C., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, 12(2), 121-138.

Esser, H. (1986). *Können Befragte lügen? Zum Konzept des „wahren Wertes“ im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung.* ZUMA-Arbeitsbericht 1986/02. Mannheim: Zentrum für Umfragen, Methoden und Analysen – ZUMA.

Esser, H. (1990). “Habits”, “Frames” und “Rational Choice”. Die Reichweite von Theorien der rationalen Wahl (am Beispiel der Erklärung des Befragtenverhaltens). *Zeitschrift für Soziologie*, 19(4), 231-247.

Esser, H. (1991). Die Erklärung systematischer Fehler in Interviews: Befragtenverhalten als "rational choice". In: Wittenberg, R. (Ed.), *Person – Situation – Institution – Kultur. Günter Büschges zum 65. Geburtstag.* Berlin: Duncker & Humblot.

Esser, H. (2001). *Soziologie. Spezielle Grundlagen: Sinn, und Kultur. Band 6.* Frankfurt a.M.: Campus.

Esser, H. (2003). Der Sinn der Modelle: Antwort auf Götz Rohwer. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 55, 395-368.

Fazio, R. (1990a). Multiple Processes by which Attitudes Guide Behavior: the MODE Model as an Integrative Framework. *Advances in Experimental Social Psychology*, 23, 75-109.

Fazio, R. (1990b). A practical guide to the use of response latency in social psychological research. In: Hendrick, C., & Clark, M. S. (Eds.), *Review of Personality and Social Psychology. Vol. 11. Research Methods in Personality and Social Psychology.* Newbury Park: Sage Publications.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Stanford: Stanford University Press.

Gibbons, H., & Rammsayer, T. (1999). Auswirkung der Vertrautheit mit einer Reizdimension auf Entscheidungsprozesse: Der modulierende Einfluss kontrollierter vs. automatischer Informationsverarbeitung. In Wachsmuth & Jung (Eds.), K*ogWis99, Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft.* Bielefeld/ St. Augustin.

Hancock, D., & Flowers, C. (2001). Comparing social desirability responding on world wide web and paper-administered surveys. *Educational Technology Research and Development*, 49(1), 5–13.

Holtgraves, T. (2004). Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding. *Personality and Social Psychology Bulletin*, 30(2), 161-172.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel Analysis. Techniques and Applications. Third Edition.* New York: Routledge.

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments and Computers*, 31(3), 433-438.

Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior*, 28, 251-256.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in cati, ivr, and web surveys. *Public Opinion Quarterly*, 72(5), 847-865.

Kroneberg, C. (2005). Die Definition der Situation und die variable Rationalität der Akteure. Ein allgemeines Modell des Handelns. *Zeitschrift für Soziologie*, 34(5), 344-363.

Kroneberg, C., Yaish, M., & Stocké, V. (2010). Norms and Rationality in Electoral Participation and in the Rescue of Jews in WWII: An Application of the Model of Frame Selection. *Rationality and Society*, 22(1), 3-36.

Kroneberg, Clemens (2006). The Definition of the Situation and Variable Rationality: The Model of Frame Selection as a General Theory of Action. *Sonderforschungsbereich 504. Rationalitätskonzepte, Entscheidungsverhalten und ökonomische Modellierung*, 06-05.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality and Quantity*, 47, 2025-2047.

Krysan, M. (1998). Privacy and the expression of white racial attitudes. a comparison across three contexts. *Public Opinion Quarterly*, 62, 506–544.

Krysan, M., Schuman, H., Scott, L.J., & Beatty, P. (1994). Response rates and response content in mail versus face-to-face surveys. *Public Opinion Quarterly*, 58, 381-399.

Lüdecke, D. (2017). sjplot: *Data visualization for statistics in social science* [Compute software manual]. Retrieved from https://CRAN.R-project.org/package=sjPlot.

Mayerl, J. (2009). *Kognitive Grundlagen sozialen Verhaltens. Framing, Einstellungen und Rationalität*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Mayerl, J. (2010). Die Low-Cost-Hypothese ist nicht genug. *Zeitschrift für Soziologie*, 39(1), 38-59.

Mayerl, J., & Urban, D. (2008). *Antwortreaktionszeiten in Survey-Analysen. Messung, Auswertung und Anwendungen*. Wiesbaden: VS Verlag.

Northover, S., Pedersen, W., Cohen, A., & Andrews, P. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Computers in Human Behavior*, 38, 144-153.

Paulhus, D. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609.

Paulhus, D. (2002). Socially Desirable Responding: The Evolution of a Construct. In: Braun, H., Jackson, D., & Wiley, D. (Eds.), *The role of constructs in psychological and educational measurement*. Mahwah: Erlbaum.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology*, 60(2), 307-317.

Richman, W., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A metaanalytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires and interviews. *Journal of Applied Psychology*, 84(5), 754–775.

Rüttenauer, T. (2018). Neighbors matter: A nation-wide small-area assessment of environmental inequality in Germany. *Social Science Research*, 70, 198-211.

Schaffner, B., & Roche, C. (2016). Misinformation and Motivated Reasoning Responses to Economic News in a Politicized Environment. *Public Opinion Quarterly*, 81(1), 86-110.

Sheppard, L., & Teasdale, J. (2000). Dysfunctional thinking in major depressive disorder: A deficit in metacognitive monitoring? *Journal of Abnormal Psychology*, 109(4), 768-776.

Shiv, B., & Fedorikhin, A. (2002). Spontaneous versus controlled influences of stimulus-based affect on choice behavior. *Organizational Behavior and Human Decision Processes*, 87(2), 342-370.

Skarbek-Kozietulska, A., Preisendörfer, P., & Wolter, F. (2012). Leugnen oder Gestehen? Bestimmungsfaktoren wahrer Antworten in Befragungen. *Zeitschrift für Soziologie*, 41(1), 5-23.

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Prognosen der Rational-Choice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie*, 33(4), 303-320.

Stocké, V. (2007). The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys. *Journal of Official Statistics*, 23(4), 493-514.

Stocké, V., & Hunkler, C. (2007). Measures of Desirability Beliefs and Their Validity as Indicators for Socially Desirable Responding. *Field Methods*, 19(3), 313-336.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin*, 133(5), 859-883.

Uziel, L. (2010). Rethinking Social Desirability Scales: From Impression Management to Interpersonally Oriented Self-Control. *Perspectives on Psychological Science*, 5(3), 243-262.

Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. *Proceedings of the SIGHI Conference on Human Factors in Computer Systems*, 3–10.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wiggens, J. (1964). Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement*, 24, 551-562.

Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer VS.

Wolter, F., & Junkermann, J. (2018). Antwortvalidität in Survey-Interviews: Meinungsäußerungen zu fiktiven Dingen. In Wolbring, T., & Menold, N. (Eds.) (forthcoming*), Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente (ASI-Schriftenreihe)*. Wiesbaden: Springer VS.

Wolter, F., & Preisendörfer, P. (2013). Asking Sensitive Questions: An Evaluation of the Randomized Response Technique Versus Direct Questioning Using Individual Validation Data. *Sociological Methods and Research*, 42(3), 321-353.

Appendix

Appendix 1 Descriptive Statistics

	mean	sd	min	max	n	missing
Event-level variables						
Item scores	4.60	1.54	1.00	7.00	9720	0
Response latency	4.65	2.00	1.88	11.86	8769	951
Respondent-level variables						
Repeat respondent	.39	.49	.00	1.00	9720	0
Need for social approval	5.78	.83	3.00	7.00	7306	2414
Acquiescence	.11	.10	.00	.63	9660	60
Sex (male)	.29	.45	0.00	1.00	9660	60
Year of birth	1987.42	2.49	1979.00	1991.00	9660	60
Item-level variables						
Syllables	17.26	4.72	7.00	29.00	9720	0
Trait desirability	.75	1.92	-2.77	2.83	9720	0

Note. Original metrics before centering; constant of 1,900 subtracted from Year of birth

Appendix 2 Descriptive Statistics of Need for Social Approval Items

	mean	sd	min	max	n	missing	reliability
No matter who I'm talking to, I'm always a good listener.	5.75	0.95	2.00	7.00	242	389	.65
I am always courteous, even to people who are disagreeable.	5.83	1.08	2.00	7.00	242	389	

Note. Cronbach's Alpha reliability; statistics based on the untransformed wide-format data (one row per respondent) rather than the long-format data used for the rest of the analysis (with one row per 'event'); 1: does not apply to me at all … 7: applies fully and completely to me

Appendix 3 Mean Trait Desirability Score Per Item and Standard Deviation (sd)

Item		mean	sd
Interaction with younger people (Teenagers)			
1	Spending time with teenagers is a lot of fun.	2.82	1.10
2 [-]	Teenagers tend to annoy me quickly.	-2.77	1.40
3	I always get along with teenagers.	2.38	1.35
Humour			
4	I find it easy to make others laugh.	1.29	1.57
5	My friends and acquaintances appreciate my friendly disposition.	1.74	1.59
6 [-]	I sometimes have trouble being funny at the right moment.	-.81	1.40
Tolerance for frustration (Frustration)			
7	I take being insulted well.	1.64	1.67
8 [-]	I am very sensitive to personal accusations and attacks.	-2.01	1.50
9	I can cope with disappointment better than many other people.	.74	1.82
Ability to assert oneself (Conflict)			
10	I am able to stick by my opinions in conflicts.	1.73	1.42
11 [-]	When I am challenged I sometimes find it difficult to argue my point convincingly.	-1.70	1.73
12	I am good at winning arguments.	1.69	1.41
Flexibility			
13	I deal well with unforeseen situations.	2.08	1.49
14 [-]	I need things to go as planned.	-.91	1.61
15	I can adapt myself to new situations without any problems.	1.90	1.18
Social sensibility (Empathy)			
16 [-]	I find it difficult to put myself in someone else's shoes.	-2.32	1.82
17	I have good feeling for how to deal with people.	2.55	1.32
18	I am aware of problems other people may be having.	2.22	1.12
Didactic abilities (Didactics)			
19	I am good at explaining complex situations.	2.82	1.33
20 [-]	Sometimes I am not able to communicate complex topics so that other people are able to understand.	-1.91	2.09
21	I find it easy to teach others.	2.83	1.31
Comfort speaking in front of others (Manner)			
22	I don't mind talking in front of a group unprepared.	1.60	2.02
23	When I have to speak or present in front of a group, I am able to overcome my nervousness.	2.17	1.27
24 [-]	I feel insecure when I have to speak in front of others.	-2.45	1.47

Item		mean	sd
Ability to express oneself (Expression)			
25 [-]	My ability to express myself in discussions is sometimes limited.	-1.66	1.77
26	I am able to express complicated things clearly and concisely.	2.09	1.36
27	I can adjust the way I express myself depending on who I am talking to.	1.94	1.30
Ability to awake interest (Enthusiasm)			
28	I am good at getting people excited about things.	2.45	1.29
29 [-]	I find it difficult to convince others of things.	-1.94	1.50
30	I am good at getting people interested in things.	2.34	1.40

Note. [-] denotes undesirable item content; -4: strongly undesirable… 0: neutrally seen… +4: strongly desirable

methods, data, analyses | Vol. 13(1), 2019, pp. 37-57 DOI: 10.12758/mda.2019.01

Do Phantom Questions Measure Social Desirability?

Axel Franzen & Sebastian Mader
University of Bern

Abstract

Social desirability is a major problem in survey research. One way of handling the problem is to measure social desirability and to incorporate it into the statistical analysis. There are different techniques of measuring social desirability. We investigate and compare the performance of the well-known Crowne-Marlowe scale with the less common use of phantom questions. Up to now, there is only one study, which tests the comparative performance of both instruments (Randall & Fernandes 1991). In this paper we replicate the test and introduce a few innovations. In difference to the former study, we compare two short versions of the Crowne-Marlowe scale, the 10-items version as suggested by Clancy and Gove (1974) and a 10-items version suggested by Stocké (2014). First, we test both scales with respect to their internal consistency. Second, we investigate which of the two versions has the strongest impact on different sensitive behaviors (alcohol consumption, shoplifting, law compliance, and reported life satisfaction). Third, we construct 20 phantom questions, 10 with fictitious answering categories that can hardly be confused with existing things, and 10 where the fictitious categories resemble existing persons or sites. We then investigate whether the phantom questions pick up social desirability better than the Crowne-Marlowe scale. The study was conducted online with 365 student subjects. Our results indicate that the short version of the Crowne-Marlowe scale suggested by Clancy and Gove (1974) performs best. But none of our phantom questions or any combination of them is able to pick up social desirability. Instead over-claiming is associated with a lack of knowledge.

Keywords: social desirability, phantom questions, overclaiming, sensitive behavior, Crowne-Marlowe Scale

Social desirability is a major problem in survey research. Respondents usually have more or less the desire to report their true attitudes and behavior. However, when questions relate to sensitive topics they are also ashamed of reporting the true values and adapt their response towards what they believe is socially accepted or expected. This social desirability bias is well known and there are many examples of it in the literature (e.g. Tourangeau & Yan 2007; Wolter 2012). Very prominent examples stem from research about voting behavior or sexual behavior. For instance, the General Social Survey (GSS) asks men and women in the US for the number of sexual partners during their lifetime. Men report an average of 12.3 and women of 3.3 (Smith 1992). Similar results are obtained for Great Britain, France or New Zealand (Wiederman 1997). Assuming that both groups have roughly the same size and that sex involves usually one man and one woman the average must be the same. Hence, either men vastly exaggerate the true number or women reduce it or both. Also surveys about the participation in the last election or referendum usually generate much larger numbers than the known voting participation (Belli et al. 2001). There are many other examples that relate to tax evasion (Korndörfer et al. 2014) and other types of deviant behavior (e.g. Preisendörfer & Wolter 2014).

Basically, there are three ways of dealing with the social desirability bias. First, one possibility is obviously to not use surveys in sensitive research areas or at least to complement survey data with other observational or process generated data. A second strategy is to increase the anonymity of respondents. Besides using closed envelopes or question wording (which is actually not increasing anonymity but downplaying the sensitivity of the questions) anonymity can be increased by using self-administered interviews, or implementing special techniques like the randomized response technique (RRT) or related approaches like the crosswise model, or the item count model (ICT). The existing evidence suggests that self-administered interviews are less prone to socially desirable response behavior than personal interviews (Tourangeau & Yan 2007). Recent research on using RRT, the crosswise model or ICT suggests that they do not perform very well in surveys (e.g. Coutts & Jann 2011; Holbrook & Krosnick 2010; Höglinger et al. 2016; Wolter & Preisendörfer 2013). Often the sensitive behavior under investigation (e.g. plagiarism, or shoplifting) is lower when using these techniques as compared to direct questioning. Furthermore, a paper by Höglinger and Diekmann (2017) suggests that the "more is better" assumption does not always hold. In their study the number of participants who reported to have a very rare disease was higher using the crosswise model technique and hence further away from the true value than without

Direct correspondence to
Axel Franzen, University of Bern, Institute of Sociology, Fabrikstrasse 8, 3012 Bern, Switzerland
E-mail: franzen@soz.unibe.ch

this technique. A study by Höglinger and Jann (2018) compares different versions of RRT (forced-response and unrelated-question) and the crosswise model with respect to direct questioning and respondents' known behavior. They also report false positives using the crosswise model, and none of the RRT implementation outperforms direct questioning. One problem with indirect question techniques is that respondents do not understand the mechanism and react with high suspicion or increased random answering behavior.

A third strategy to deal with social desirability is to measure it. This was already suggested by Crowne and Marlowe in 1960. The original Crowne-Marlowe scale consists of 33 items that describe extreme behaviors or attitudes that hardly always apply to a respondent. An example is the item "I have never intensely disliked anyone". Respondents are then asked whether this statement describes their behavior or attitude as "true" or "false". The more "true" answers are given to socially desirable behaviors (as the example) or "false" answers to undesirable behaviors the higher is a respondent's score on the social desirability scale. The Crowne-Marlowe scale is the most applied measure of social desirability in survey research. Already Phillips and Clancy (1972) found that respondents scoring high on the Crowne-Marlowe (CM) scale also report higher overall life happiness (for similar results see Kozma and Stones 1987, Carstensen and Cone 1983) or report to have more friends as compared to respondents with low CM values. However, there is also some counterevidence. For example Johnson et al. (2012) find no association between the CM scale and cocaine use underreporting or with actual cocaine use as assessed by respondents' hair, saliva or urine samples. One problem with measures of social desirability like the CM scale is that it lacks "true" scores. Hence, it could be the case that the scale does not measure over- or underreporting but respondents' true behavior or attitude, or at least a mixture of both (Tourangeau & Yan 2007). One way to circumvent this problem is to use phantom questions. Such questions were already used by Phillips and Clancy (1972) in order to validate the social desirability scale of Crowne and Marlowe (1960, 1964). Phantom questions ask respondents whether they are familiar with certain people, books, movies or sites that do not exist. Hence, as opposed to the items used in the CM scale the true values of phantom questions are known and respondents are clearly overclaiming when responding to be familiar with non-existing people or sites.

So far there is only one study which compares the performance of the CM scale with the performance of phantom questions (Randall & Fernandes 1991). Randall and Fernandes (1991) use the full 33-items Crowne-Marlowe scale and five phantom questions that relate to consumer goods (movies, products, music albums, TV programs, and designer labels). The sensitive behavior under study referred to ten different acts of self-reported student misconduct (e.g. having plagiarized a term paper, turning in the same paper for two classes, cheating in exams). They find a negative correlation (r = - 0.24) between the Crowne-Marlowe scale and students' misconduct and no statistically significant correlation for the phantom questions. However, none of both measures were significant in the final multiple OLS regression analysis in which the authors included also a measure for the self-rated desirability of the ten sensitive behaviors in question. Consequently, the authors conclude "that further use of the M-C scale is not advisable" (ibid. 814). Similar conclusions apply to the phantom questions.

However, these conclusions are disputable. Randall and Fernandes (1991) measure trait desirability by asking respondents how desirable they believe each behavior under investigation is. These item-specific ratings are correlated with the CM scale (ibid. 811). From a theoretical perspective it is reasonable to assume that respondents' general measure of social desirability (CM scale) affects the desirability of specific behaviors (and not the other way round). For example, respondents' rating of the desirability of shoplifting could be influenced by the general tendency to answer in a socially desirable way. Under this assumption, trait desirability is a mediator variable and should not be included in one multiple regression model investigating the relation of the CM scale on sensitive behavior. Doing so wipes out (over-controlling) the correlation between the two (Morgan & Winship 2008, p. 65). Hence, the study might not be a reliable test of the performance of the CM scale.

Our study differs in a number of respects from the former study by Randall and Fernandes (1991): First, we refrain from including trait desirability for the reason already outlined above. Second, Randall and Fernandes (1991) use the full 33-items CM scale. However, this is a very long instrument and impractical for general population surveys. Therefore, we use two short versions of the CM scale, which are often found in the literature (Clancy 1971; see also Clancy & Gove 1974; Stocké 2014), and compare them with respect to their dimensionality, internal consistency, and their performance. Third, Randall and Fernandes (1991) use ten specific sensitive questions that only relate to typical student behavior like cheating in exams. In difference, our study includes questions from different areas such as respondents' level of norm compliance, alcohol consumption (Welte & Russell 1993; Embree & Whitehead 1993), shoplifting, and life satisfaction (Kozma & Stones 1987). Fourth, one reason why respondents might claim familiarity with non-existent objects or people in phantom questions might be the confusion with existing things. To study the impact of the confusion potential of phantom questions we designed one version having little confusion potential and one with a larger potential, and split the sample in such a way that every group received five phantom questions of each type.

The remainder of the article is structured as follows: Section two describes the two short versions of the CM scale and discusses their measurement characteristics. Section three presents the 20 phantom questions and contrasts the ones with and without the risk of confusion. Section four compares the criterion-related validity of the CM scales with the performance of the phantom questions. The final section concludes and discusses the results.

The Crowne-Marlowe Scale

To study the characteristics and performance of the CM-scale in comparison to phantom questions we conducted an online survey among the student population of the University of Bern. For this purpose, we randomly selected 2000 email addresses from the student email register and sent them an email including a link leading to the online survey in the beginning of March 2017. Overall, 463 students participated in the survey, which constitutes a response rate of 23.2%. The questionnaire contained about 70 questions, including 18 items of the Marlowe-Crowne scale, 10 phantom questions, and various questions on sensitive topics such as attitudes towards norm compliance, shoplifting, alcohol consumption, and life satisfaction. The median completion time of the survey was about 14 minutes. We excluded 70 participants from further analyses since their completion time was below 50% or above 200% of the median completion time. The rationale for doing so is that answering 70 questions in 7 minutes properly is probably not possible. Also, using 28 minutes for a 14 or 15-minute survey seems suspicious and might be due to respondents' attempt to search or google for true answers. We excluded an additional 28 respondents since they answered a test item instructing them not to provide any answer. The item reads "In this question we show you four answer categories. Please do not check any of the provided answer categories." Dropping cases with either a very short or a very long completion time and those with an invalid answer to the test item left us with 365 valid cases. However, the exclusion of these cases did not change any of the results substantially.

The original CM-scale consists of 33 items. Since this is a rather large instrument for a general survey most authors have used a reduced version of the CM scale. A prominent example is the 10-item short version suggested by Clancy 1971 (see also Clancy & Gove 1974; Phillips & Clancy 1972). Another short version was suggested by Stocké (2014). First, we discuss some measurement qualities of both scales separately. Second, we investigate whether the measurement qualities can be improved by some combination of both scales. The 10 items suggested by Clancy (1971) are shown in Table 1.

First, the distribution of CM1 is very close to normal and only slightly skewed to the left (skewness = -.13). Second, an exploratory principle component analysis (PCA) extracts four factors consisting of one or three items each. Third, Cronbach's alpha is .39 suggesting that the short version has low internal consistency. Both latter characteristics suggest that the items of Clancy's short version are rather heterogeneous. Next, we compare the short version suggested by Clancy (CM1) to a

Table 1 The short CM-Scale of Clancy 1971 (CM1)

		Polarity	CM1
I	(1) I never hesitate to go out of my way to help someone in trouble.	T	.73
	(2) On occasion I have had doubts about my ability to succeed in life.	F	.75
	(3) I sometimes try to get even, rather than forgive and forget.	F	.52
II	(4) No matter who I'm talking to, I'm always a good listener.	T	.63
	(5) At times I have really insisted on having things my own way.	F	.65
	(6) I have never been irked when people expressed ideas very different from my own.	T	.67
III	(7) If I could get into a movie without paying and be sure I was not seen, I would probably do it.	F	.69
	(8) There have been times when I felt like rebelling against people in authority even though I knew they were right.	F	.76
	(9) I never resent being asked to return a favour.	T	.24
IV	(10) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.62

Note: N = 365, min = 0, max = 10, mean = 5.4, median = 5, modus = 5, sd = 1.92, Cronbach's alpha = 0.39. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, and IV, respectively. A component is identified if eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering T to the first question or answering F to the second question.

different version suggested by Stocké (2014) (hereafter CM2). Also, Stocké (2014) picked 10 items from the original list (Table 2). Eight items differ from the CM1 version but two items appear in both short versions. These are the item number 5 of the Stocké version ("No matter who I'm talking to, I'm always a good listener.") and item number 7 ("Before voting I thoroughly investigate the qualifications of all the candidates."). Also, CM2 is almost normally distributed (skewness = -.20) and an exploratory factor analysis extracts also four components. Cronbach's alpha of CM2 is 0.53 and, hence, slightly better than the internal consistency of CM1 but still unsatisfactory.

Since both short versions have undesirable measurement qualities, e.g. multidimensionality and low consistency, we combined both scales to a 16-items ver-

Table 2 The short CM-Scale of Stocké 2014 (CM2)

		Polarity	CM2
I	(1) I sometimes feel resentful when I don't get my way.	F	.70
	(2) I'm always willing to admit it when I make a mistake.	T	.71
	(3) I am sometimes irritated by people who ask favors of me.	F	.46
	(4) I have never deliberately said something that hurt someone's feelings.	T	.64
II	(5) No matter who I'm talking to, I'm always a good listener.	T	.64
	(6) I am always courteous, even to people who are disagreeable.	T	.91
III	(7) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.79
	(8) I keep getting myself on principles whose observance I expect from others.	T	.58
IV	(9) I can remember "playing sick" to get out of something.	F	.81
	(10) There have been occasions when I took advantage of someone.	F	.64

Note: N = 365, min = 0, max = 10, mean = 5.35, median = 5, modus = 6, sd = 2.09, Cronbach's α = 0.53. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, and IV, respectively. A component is identified if the eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering F to the first question or answering T to the second question.

sion (CM3).[1] This 16-items version is depicted in Table 3. The CM3 version of the social desirability scale has a Cronbach's alpha of 0.62 and therefore, outperforms the CM1 and CM2 versions. However, like the other two short versions the scale is not one-dimensional but consists of five components as indicated by a principal component analysis (PCA).

[1] Two items were dropped since their inclusion resulted in lower Cronbach's alpha values.

Table 3 A composite scale of social desirability (CM3)

		Polarity	CM3
I	(1) There have been times when I felt like rebelling against people in authority even though I knew they were right.	F	.72
	(2) I sometimes try to get even, rather than forgive and forget.	F	.64
	(3) I have never deliberately said something that hurt someone's feelings.	T	.69
II	(4) If I could get into a movie without paying and be sure I was not seen, I would probably do it.	F	.59
	(5) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.56
	(6) There have been occasions when I took advantage of someone.	F	.54
	(7) I keep getting myself on principles whose observance I expect from others.	T	.69
III	(8) At times I have really insisted on having things my own way.	F	.71
	(9) I sometimes feel resentful when I don't get my way.	F	.69
	(10) I am always willing to admit it when I make a mistake.	T	.37
	(11) I am sometimes irritated by people who ask favors of me.	F	.57
IV	(12) No matter who I'm talking to, I'm always a good listener.	T	.60
	(13) I have never been irked when people expressed ideas very different from my own.	T	.68
	(14) I am always courteous, even to people who are disagreeable.	T	.76
V	(15) I never hesitate to go out of my way to help someone in trouble.	T	.70
	(16) On occasion I have had doubts about my ability to succeed in life.	F	.65

Note: N = 365, min = 0, max = 16, mean = 8.50, median = 9, modus = 7, sd = 2.97, Cronbach's α = 0.62. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, IV, and V respectively. A component is identified if the eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering F to the first question or answering T to the thrid question.

Phantom Questions

The CM-scale has the disadvantage that it lacks true values. Hence, it might not only pick up social desirability but also some true personality differences between respondents. This confusion might be one reason why the scale is multidimensional without any clear evidence why some items fall into one and others into another component. One alternative to measure social desirability or the need for social approval are phantom questions. Such questions ask respondents whether they are familiar with some objects, places or personalities that do not exist. The idea is that subjects with a strong need for social approval have a higher chance to claim that they are familiar with the person or object even if it does not exist since admitting not knowing something might create social disapproval. An example of a phantom question would be: "In the following we list four important international organizations. Which of these organizations do you know?" which is then followed by four answer categories "UNO", "OECD", "WIO", and "NATO". Obviously, "WIO" does not exist. But the answer has one problem. It is very close to "WHO" and thus, respondents might claim familiarity with WIO because they confuse it with WHO. Because of this risk of confusion, and because there is generally very little experience with phantom questions we generated 10 different phantom questions from various areas such as politics, geography, literature, architecture, science, movies, or generally concerning publicly known personalities. Additionally, we created two versions of every phantom question, one version of which we thought that the risk of confusion is low and one in which it is higher. Generally, the risk of confusion is higher if the fictitious issue sounds similar to an existing issue in contrast to an issue that is distinct from an existing site or person. All twenty questions are listed in Table 4. Because it would be cumbersome for respondents to answer twenty such knowledge questions in one survey we randomly split the questionnaire in two versions. Version one contained the first five phantom questions without the risk of confusion and the last five with the risk of confusion. The other version was the other way around and contained the first five phantom questions with the risk of confusion and the last five without risk of confusion. This way we had two groups of respondents who answered each ten phantom questions. This design enables us to study the effect of low or high risk of confusion on the answering behavior of phantom questions.

After a short introduction, the online questionnaire started with the 10-items CM scale of Clancy and Gove (1974), followed by five phantom questions, continued with 10-items of the CM scale of Stocké (2014), and was again followed by the remaining five phantom questions. Questions on more or less sensitive opinions and behaviors followed in the middle. The questionnaire concluded with sociodemographic information. Each block of phantom questions was split into two parts showing three phantom questions on the first screen and two on the following screen.

Table 4 displays the proportion of respondents who answered "yes" to the four categories of the 10 phantom questions. We are interested here in the proportions of "yes" answers to fake categories. Table 4 shows that the proportion of yes-answers to fake items without risk of confusion varies between 0% and 9% and is therefore relatively low. None of the respondents said that they are familiar with an Oscar – winning movie called "sense of delight" and 9% thought that Peter Dickens was an American President. The proportions of yes answers are considerably higher when the fake answer is formulated in such a way that the risk of confusing it with existing places or people is higher. Proportions vary between 5% and 49% with the risk of confusion. 5% of respondents said that they are familiar with a Nobel Prize Winner called Jassir Peres, and 49% claimed that they are familiar with an architectural style called "futurism". The proportions of yes-answers are consistently higher when we purposely tried to increase the risk of confusion. Hence, this intended manipulation worked quite well. However, surprisingly phantom questions do not correlate very high among each other. This is true for the first five phantom questions without risk of confusion and the last five one with risk of confusion in group one (highest $r = .50$) as well as for those phantom questions in group 2 (highest $r = .28$). Practically none of our respondents consistently claimed familiarity with all fictitious items.

This already points into the direction that phantom questions are very context specific but do not pick up consistently a personality trait such as the need for social approval. Furthermore, there are also no obvious sequence effects. Phantom questions were presented in the order displayed in Table 4. Only 1% answered that they are familiar with EBO (first item), 2% with the author Jean-François Le Gouguec, 6% with Sevenstone Cave, 7% with Modular Style, and 3% with the Fun Loving Animals. Hence, there is no indication of learning effects, such that respondents improved their performance with the number of phantom questions. Similar observations apply to the sequence of the other phantom questions.

Comparing the Criterion-Related Validity of the CM-Scale with the Performance of Phantom Questions

The questionnaire contains a number of questions on sensitive topics, such as whether respondents ever took something from a store without paying for it (shoplifting), how many glasses of alcohol they consume during a week, whether they believe that laws should always be adhered, and on their general life satisfaction.

Table 4 Description of the Phantom Questions by risk of confusion (ROC)

		Without ROC		With ROC	
I	International Organizations: In the following, we list four important international organizations. Which of these organizations do you know?	UNO	(98)	UNO	(99)
		OECD	(70)	OECD	(68)
		EBO	(1)	WIO	(8)
		NATO	(99)	NATO	(100)
II	Authors of Universal Literature: In the following, we list four authors of Universal Literature. Which of these authors do you know?	Johann Wolfgang von Goethe	(99)	Johann Wolfgang von Goethe	(99)
		William Shakespeare	(100)	William Shakespeare	(100)
		Mark Twain	(91)	Mark Twain	(88)
		Jean-François Le Gouguec	(2)	Niki de Saint Phalle	(24)
III	UNESCO World Heritage Sites: Now we list four UNESCO World Heritage Sites. Which of these sites do you know?	Venice	(99)	Venice	(100)
		Sevenstone Cave	(6)	The Mexican Wall	(19)
		Pyramids of Gizeh	(90)	Pyramids of Gizeh	(80)
		Taj Mahal	(92)	Taj Mahal	(92)
IV	Architectural Styles: Now we list four architectural styles. Which of these epochs do you know?	Bauhaus	(52)	Bauhaus	(46)
		Jugendstil	(85)	Jugendstil	(86)
		Gothic Style	(98)	Gothic Style	(97)
		Modular Style	(7)	Futurism	(49)
V	Environmental Protection Organizations: In the following, we list four important international environmental protection organizations. Which of these organizations do you know?	Fun Loving Animals	(3)	World Climate Protection Trust	(9)
		United Nations Environmental Programme	(20)	United Nations Environmental Programme	(20)
		World Wildlife Fund	(64)	World Wildlife Fund	(67)
		Greenpeace	(100)	Greenpeace	(100)

Table 4 continued

		Without ROC		With ROC	
VI	Famous Musicians: Which of the following four musicians do you know?	Kurt Cobain	(91)	Kurt Cobain	(95)
		Paul McCartney	(95)	Paul McCartney	(94)
		Sandy Lawn	(2)	Bob Cohen	(21)
		Amy Winehouse	(100)	Amy Winehouse	(99)
VII	US Presidents: In the following, we list four former US Presidents. Which of these politicians do you know?	Barack Obama	(100)	Barack Obama	(100)
		Peter Dickens	(9)	George Adam	(10)
		John F. Kennedy	(100)	John F. Kennedy	(100)
		Abraham Lincoln	(98)	Abraham Lincoln	(98)
VIII	Charitable Celebrities: In the following, we list four famous personalities that take a stand for charity. Which of these persons do you know?	George Clooney	(99)	George Clooney	(100)
		Angelina Jolie	(99)	Angelina Jolie	(100)
		Gabriele Goldau	(2)	Jenifer Cruz	(20)
		Bill Gates	(99)	Bill Gates	(99)
IX	Oscar-winning Movies: In the following, we list four Oscar-winning movies. Do you know these films?	Sense of Delight	(0)	A Beautiful Girl	(8)
		Gladiator	(91)	Gladiator	(89)
		Titanic	(99)	Titanic	(98)
		Forrest Gump	(97)	Forrest Gump	(97)
X	Nobel Peace Prize Winners: Which of the four following Nobel Peace Prize Winners do you know?	Dalai Lama	(100)	Dalai Lama	(100)
		Michail Gorbatschow	(73)	Michail Gorbatschow	(73)
		Martin Luther King	(97)	Martin Luther King	(100)
		Aleksander Islic	(0)	Jassir Peres	(5)
Dummy for ≥ 1 'yes' answer (I)-(X) (n=365)			(12)		(56)

Note: Arabic numbers in parentheses indicate the percentage of 'yes' answers. Fake items are underlined. For questions (I)-(V) n = 174 for the fake items without ROC and n=191 with ROC. For questions (VI)-(X) n = 191 for the fake items without ROC and n = 174 with ROC.

All these questions were taken in the exact same formulation as they usually appear in large general population surveys. Agreeing to law compliance and life satisfaction are socially desirable matters. Respondents who are identified of having a high need for social approval should therefore more strongly over-report those behaviors as compared to individuals who care less about social approval. Results of multiple OLS regression analyses are displayed in Table 5. Every line of the table represents the results of an independent multiple regression model in which we control for all available socio- demographic variables (age, sex, subject of study, nationality, main language, household size, designated study degree). As can be seen, all CM scales are positively related to agreement with law compliance and life satisfaction as expected. Hence, the CM scale does pick up over-reporting. The effects of the CM2 and CM3 scales are a little weaker than the effects of the CM1 scale, and are statistically insignificant with respect to life satisfaction and norm compliance. Alcohol consumption and shoplifting should be underreported by respondents with a high need of social approval and this is what can be observed from the results of Table 5. Here, all three CM versions perform equally well. Respondents with a high need for social approval report to drink less alcohol and report less often that they have shoplifted before (logit model).

Next, we investigate how the phantom questions perform. For this purpose, we constructed two different scales. Respondents are coded as being sensitive towards social desirability if they have claimed familiarity with at least one or more fake sites, objects, organizations, or persons when the risk of confusion was low (without ROC), and when the risk of confusion was high (with ROC). The results of both versions are displayed in lines 4 and 5 of Table 5. As can be seen from the results neither version is statistically significantly associated with any of the four dependent variables. These results are robust if we ran 20 models including each time a different phantom question or if the index is composed of both versions of phantom questions (with and without the risk of confusion), or if the index is constructed continuously by summing up the number of wrongly answered phantom questions.

Hence, phantom questions do obviously not measure social desirability. This raises the question of what phantom questions measure instead. One obvious answer is that they simply measure knowledge. We therefore conducted a second study trying to find evidence for this explanation. The second study was conducted in May 2017 at the University of Bern with N = 318 respondents. The original purpose of the second study was to investigate the relation of IQ test scores (see Liepmann et al. 2012) with emotional intelligence and empathy. However, the online questionnaire which respondents had to answer in the laboratory contained also some of the same phantom questions used in Study 1 (questions I, II, III, VI and IX without risk of confusion). The relevant results of Study 2 are displayed in Table 6. The dependent variable is the dichotomous characteristic of whether respondents answered

Table 5 Regressions of various traits on CM and overclaiming

Model		(1)	(2)	(3)	(4)
		OLS	OLS	OLS	Logit
Dependent Variable (z-stand.)		Law compl.	Happiness	Alcohol	Shoplifting
CM1 (z-stand.)		0.11*	0.22***	-0.11*	-0.50***
		(0.06)	(0.05)	(0.05)	(0.12)
	adjusted R^2	0.03	0.05	0.08	
	pseudo R^2				0.08
CM2 (z-stand.)		0.08	0.12	-0.12*	-0.38**
		(0.05)	(0.06)	(0.05)	(0.12)
	adjusted R^2	0.02	0.02	0.08	
	pseudo R^2				0.07
CM3 (z-stand.)		0.10	0.18**	-0.10*	-0.45***
		(0.05)	(0.06)	(0.05)	(0.12)
	adjusted R^2	0.02	0.04	0.08	
	pseudo R^2				0.07
Overclaiming (without ROC)		-0.00	-0.21	-0.12	0.31
		(0.16)	(0.17)	(0.11)	(0.34)
	adjusted R^2	0.01	0.01	0.07	
	pseudo R^2				0.04
Overclaiming (with ROC)		-0.07	0.07	0.00	-0.40
		(0.11)	(0.11)	(0.11)	(0.24)
	adjusted R^2	0.01	0.00	0.07	
	pseudo R^2				0.05
n		348	348	348	347

Note: Displayed are the standardized regression coefficients. * = p<0.05, ** = p<0.01, *** = p<0.001. All standard errors (in parentheses) are robust with respect to heteroscedasticity. All models control for sex, age, German mother tongue, Swiss nationality, designated study degree, household size, and study subject. Table A1 summarizes the descriptive statistics of all variables in the models. Note that all results remain robust even if respondents with a very low or very high completion time remain in the sample.

Table 6 Logistic Regression of Overclaiming on IQ

Model Dependent Variable (z-stand.)	(1) Overclaiming
IQ (z-stand.)	-0.65*** (0.19)
Female	-0.07 (0.39)
Age	-0.03 (0.11)
Mother Tongue: German	0.40 (0.68)
Swiss	-0.03 (0.81)
Household Size	-0.04 (0.17)
Designated degree: Master	0.05 (0.66)
University of Bern	-0.03 (0.44)
Constant	-1.59 (2.91)
n	297
pseudo R^2	0.05
Loglikelihood	-98.53

Note: Displayed are logit coefficients. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. All standard errors (in parentheses) are robust with respect to heteroscedasticity.

one or more phantom questions wrongly.[2] The logistic regression results of Table 6 indicate that subjects with high IQ test scores claimed statistically significantly less often of being familiar with non-existing things, objects, or people. None of the other control variables (sex, age, subject of study and so on) were found to be associated with overclaiming of phantom questions. This is also true for the same

[2] The dependent variable has values ranging from 0 (no claim of familiarity with any fictitious item) to 5 (claiming familiarity with each fictitious item). Of all respondents 37 (12%) stated to be familiar with at least one fictitious item, and eleven respondents with more than one. Giving this skewed distribution, we dichotomized the dependent variable and used a logistic regression. However, using a negative binomial model gives the same results.

analysis of the data from the first study. Taken together, our results suggest that phantom questions measure knowledge but not the need for social approval.

Summary and Discussion

A comparison of the performance of three different short versions of the CM scale with respect to self-reports on law conformity, shoplifting, alcohol consumption, and life satisfaction suggests that the CM scale picks up social desirability. As expected, higher values on the CM scale are positively associated with opinions on law compliance and life satisfaction. The standardized coefficients show that the effect sizes are small. Furthermore, all three versions detect also underreporting of shoplifting and alcohol consumption as expected. Moreover, our study shows that it basically does not matter whether we use the short version suggested by Clancy (1971), or a combined version of the Stocké and Clancy scale with 16 items. The combined version has a higher Cronbach's alpha value but the associations with sensitive behavior are almost the same as with the CM1 scale. Hence, our study confirms the finding of other studies suggesting that the CM scale works.

However, we did not find a single association with one or any combination of phantom questions with sensitive behavior (shoplifting, alcohol consumption, norm compliance, life satisfaction). Also, phantom questions have small correlations among each other and no correlation with any short version of the CM scale (see Table A2). These results suggest that phantom questions measure knowledge but not the need for social approval. Of course, our study results are obtained from a student sample which raises questions on the generalizability. However, limitations of generalizability mainly apply to descriptive results but less to associational findings. Theoretically, it is possible that phantom questions pick up social desirability in a general population sample but not in a student sample. However, practically this is very unlikely.

In contrast to phantom questions, we find that all three versions of the CM scale are associated with the sensitive behaviors studied, and that the CM1 version outperforms the other two versions slightly. This finding might suggest, that the CM scale measures social desirability. However, the finding is also compatible with the interpretation that the CM scale as well as the sensitive behavior(s) are both caused by true but unobserved personality differences. In that case, the correlation between the CM scale and the desirable behavior in question would be spurious. This omitted variable bias can only be avoided in validation studies in which the true behavior of respondents is known. Such studies are rare. One recent study by Preisendörfer and Wolter (2014) does not find a statistically significant relation between the CM scale and truthful answering whether respondents have been convicted of a crime. However, also Preisendörfer and Wolter (2014) included trait desirability in their analysis together with the CM scale and, therefore, might have introduced an over-control bias into their study. Hence, further research on the validity of the CM scale and improvements on measuring social desirability are still in need.

References

Belli, R. F., Traugott, M. W., & Beckmann, M. N. (2001). What leads to voting overreports? Contrasts of overreporters to validated voters and admitted nonvoters in the American National Election Studies. *Journal of Official Statistics*, 17(4), 479-498.

Carstensen, L. L., & Cone, J. D. (1983). Social desirability and the measurement of psychological well-being in elderly persons. *Journal of Gerontology*, 38(6), 713-715.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research*, 40(1), 169-193.

Clancy, K. (1971). Systematic bias in field studies of mental illness. Ph.D. dissertation, New York University.

Clancy, K., & Gove, W. (1974). Sex Differences in Mental Illness: An Analysis of Response Bias in Self-Reports. *American Journal of Sociology*, 80(1), 205-216.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology,* 24(4), 349-354.

Crowne, D., & Marlowe, D. (1964). The Approval Motive. Studies in Evaluative Dependence. New York: John Wiley & Sons.

Embree, B. G., & Whitehead, P. C. (1993). Validity and reliability of self-reported drinking behavior: Dealing with the problem of response bias. *Journal of Studies on Alcohol*, 54(3), 334-344.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods,* 10(3), 171-187.

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis,* 25(1), 131-137.

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLoS ONE* ,13(8), e0201770.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly,* 74(2), 328-343.

Johnson, T. P., Fendrich, M., & Mackesy-Amiti, M. E. (2012). An evaluation of the validity of the Crowne-Marlowe need for approval scale. *Quality and Quantity,* 46(6), 1883-1896.

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*, 45, 18-32.

Kozma, A., & Stones, M. J. (1987). Social desirability in measures of subjective well-being: A systematic evaluation. *Journal of Gerontology,* 42(1), 56-59.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). Intelligenz-Struktur-Test. Screening. Göttingen: Hogrefe.

Morgan, S. L., & Winship, C. (2008). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press.

Phillips, D. J., & Clancy, K. J. (1972). Some effects of „social desirability" in survey studies. *American Journal of Sociology,* 77(5), 921-940.

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly,* 78(1), 126-146.

Randall, D. M., & Fernandes, M. F. (1991). The Social Desirability Response Bias in Ethics Research. *Journal of Business Ethics,* 10(11), 805-817.

Smith, T. (1992). Discrepancies Between Men and Women in Reporting Number of Sexual Partners: A Summary from four Countries. *Social Biology,* 39(3-4), 203-211.

Stocké, V. (2014). Deutsche Kurzskala zur Erfassung des Bedürfnisses nach sozialer Anerkennung. Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS, doi:10.6102/zis159.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin,* 133(5), 859-883.

Welte, J. W., & Russell, M. (1993). Influence of socially desirable responding in a study of stress and substance abuse. *Alcoholism – Clinical and Experimental Research,* 17(4), 758-761.

Wiederman, M. W. (1997). The Truth Must Be in Here Somewhere: Examining the Gender Discrepancy in Self-Reported Lifetime Number of Sex Partners. *The Journal of Sex Research*, 34(4), 375-386.

Wolter, F. (2012). Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik. Wiesbaden: Springer VS.

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, 42(3), 321-353.

Appendix

Table A1 Descriptive statistics of all dependent and control variables of Table 5

	Variable	mean	sd	min.	max.	n	Question wording
Dependent Variables	Norm compliance	3.64	1.07	1	5	361	"Laws should always be complied to, no matter how agreeable they are." Five answering categories from 1 'disagree strongly' to 5 'agree strongly'.
	Happiness	7.42	1.63	0	10	361	"All in all how satisfied are you with your life?" 11-point Likert scale ranging from 0 'very unsatisfied' to 10 'very satisfied'.
	Alcohol Consumption	2.98	3.58	0	25	369	"How many glasses of wine, beer, or other alcoholic beverages do you drink in a usual week?" Number of weekly glasses of wine, beer, or other alcoholic beverages.
	Shoplifting	.37		0	1	366	"Have you ever in your life taken something deliberately from a store without paying for it?" Answer categories 1 'yes' and 0 'no'.

Table A1 continued

	Variable	mean	sd	min.	max.	n	Question wording
Control Variables	Sex: Female	.60		0	1	360	Dummy, 1 if female
	Age	24.73	4.79	19	59	360	in years
	German Mother Tongue	.92		0	1	360	Dummy, 1 if yes
	Swiss Nationality	.91		0	1	360	Dummy, 1 if yes
	Household Size	3.19	1.18	1	5	360	Number of people living in the household
	Bachelor	.62		0	1	358	Dummy, 1 if designated study degree is Bachelor (reference category)
	Master	.33		0	1	358	Dummy, 1 if designated study degree is Master.
	Ph.D.	.05		0	1	358	Dummy, 1 if designated study degree is Ph.D.
	Study Subject:						
	Economics and Social Sciences	.34		0	1	369	Dummy, 1 if yes (reference category)
	Law	.13		0	1	369	Dummy, 1 if yes
	Natural Sciences	.15		0	1	369	Dummy, 1 if yes
	Medicine	.14		0	1	369	Dummy, 1 if yes
	Humanities	.24		0	1	369	Dummy, 1 if yes

Table A2 Correlation matrix of the different CM scales and overclaiming scales (phantom questions)

	CM1	CM2	CM3	Overclaiming without ROC	Overclaiming with ROC
CM1					
CM2	0.59***				
CM3	0.86***	0.87***			
Overclaiming without ROC	-0.00	-0.02	-0.00		
Overclaiming with ROC	0.00	0.05	0.00	0.20***	

Note: Displayed are Pearson's correlation coefficients. *** = p < 0.001.

Validating Earnings in the German National Educational Panel Study. Determinants of Measurement Accuracy of Survey Questions on Earnings

Manfred Antoni[1], *Daniel Bela*[2] & *Basha Vicari*[1]

[1] *Institute for Employment Research (IAB), Nuremberg*

[2] *LIfBi – Leibniz Institute for Educational Trajectories, Bamberg*

Abstract

Questions on earnings are counted among sensitive topics that often produce high rates of item nonresponse or measurement error. Both types of bias are well documented in the literature and are found to concentrate in the tails of the earnings distribution. In this paper, we explore whether measurement error on earnings could be explained by socially desirable reporting and whether the error is impacted by interviewer characteristics.

Using the linked dataset NEPS-SC6-ADIAB, which contains survey data from the German National Educational Panel Study, Starting Cohort "Adults", linked to administrative earnings records from the German Federal Employment Agency, we analyze the extents of over- and underreporting and the influence of respondent and interviewer characteristics on these behaviors for different quartiles of the earnings distribution.

Our results show that the average level of misreporting is relatively low (approximately 6% of median earnings). Our main logistic model reveals that female and more highly educated respondents report significantly more accurately while those with higher earnings misreport to a significantly greater extent. Regarding the impact of personality traits on reporting accuracy, we find significant positive effects for more agreeable respondents and significant negative effects for extraverted respondents. When differentiating by the direction of misreporting, we find, for instance, that women are less likely to overreport across all earnings quartiles. However, the influence of interviewer characteristics is negligible.

Keywords: Measurement error, earnings, social desirability, interviewer effects, NEPS-SC6-ADIAB

Information on earnings is among the statistics that are most pervasively collected in population surveys. It provides a basis for a wide array of research conclusions and policy decisions related to topics such as a country's overall wealth distribution and social inequality trends (Moore et al. 2000; Bound & Krueger 1991). From an individual perspective, it is often used to approximate a person's socioeconomic status in order to explain decisional or behavioral patterns. However, any survey data are prone to some kind of nonresponse or measurement error. This is especially true in regard to sensitive questions, such as questions on respondents' earnings, collected in interviewer-administered surveys (Moore et al. 2000; Groves et al. 2009).

Questions on sensitive topics often produce relatively high rates of item nonresponse and measurement error because such questions can be perceived as threatening to disclose private information or deviant behavior (Jann 2014; West & Blom 2017). Tourangeau and Yan (2007) expect high rates of item nonresponse for questions on personal income because these questions are perceived to be intrusive. They do not necessarily expect high rates of misreporting, however, because such questions are not associated with a disclosure of violation of social norms. This expectation is supported by the findings of Krumpal (2013) which show that in German population surveys the earnings question has the highest nonresponse rate among all items. Moreover, missing earnings reports are not randomly distributed; instead, the missing values are concentrated in the tails of the earnings distribution (Riphahn & Serfling 2005; Bollinger et al. 2018).

The statement of Tourangeau and Yan (2007) is, however, contrasted by a wealth of studies revealing the prevalence of misreporting in response to survey questions on earnings (see, e.g., Duncan & Hill 1985; Bound & Krueger 1991; Bound et al. 1994; Bollinger 1998; Moore et al. 2000; Pedace & Bates 2000; Gottschalk & Huynh 2005; Kapteyn & Ypma 2007; Bricker & Engelhardt 2008; Gottschalk & Huynh 2010; Kim & Tamborini 2014; Angel et al. 2017). All these studies assess the quality of earnings reports by linking survey information to auxiliary data sources, most commonly administrative records that offer more reliable measures of earnings, which are thus treated as the "true" values. Regarding the nature

Acknowledgments

We thank two anonymous reviewers and the editors whose valuable comments helped to improve the quality of our manuscript.

Direct correspondence to

Manfred Antoni, Research Data Centre (FDZ) of the Federal Employment Agency at
the Institute for Employment Research (IAB), Regensburger Str. 100,
D-90478 Nuremberg
E-mail: manfred.antoni@iab.de

of earnings measurement error, these studies find a U-shaped pattern similar to that of item nonresponse: there is a negative correlation between the measurement error and the assumed true earnings value, indicating that low earners tend to overreport their earnings, while high earners tend to underreport (Bound & Krueger 1991; Bollinger 1998; Bricker & Engelhardt 2008). Nevertheless, we still know little about why respondents edit their answers depending on their positions in the earnings distribution.

We are not the first to examine the misreporting of individual earnings using a validation study, but most previous studies were conducted in the Anglo-American context, used small or restricted samples (e.g., male workers), or used a cross-sectional design. We use the linked data product called NEPS-SC6-ADIAB, which contains survey data from the German National Educational Panel Study, Starting Cohort "Adults", (NEPS SC6) – a panel survey representative of the German adult population and covering a rich set of respondent characteristics – linked to administrative earnings records from the German Federal Employment Agency. Because interviewers either can have a positive influence on participation and data quality or can cause interviewer effects (Essig & Winter 2009; Landrock 2017), we also include interviewer characteristics and paradata on the interview situation in our analysis.

Thus, we contribute to the literature in three ways: First, we provide evidence on a cultural context of money taboo, where talking about financial issues causes feelings of uneasiness (Trachtman 1999). Germany is counted among such cultural contexts (see, e.g., Kirkcaldy et al. 1992). As responding to sensitive questions is, in general, a highly context-specific behavior (Jann 2014), we assume that the cultural context of money taboo changes a merely intrusive question into one that might create embarrassment or shame. These factors make it particularly unpleasant for respondents to report very low or very high earnings (see, e.g., Bound & Krueger 1991), especially when it comes to admitting to living in poverty or in luxury in the presence of an interviewer.

Our second contribution directly derives from this fact because we combine our earnings validation study with the concept of socially desirable reporting. On the one hand, respondents might edit their reports towards some subjectively estimated norm of individual wealth. On the other hand, a competent interviewer might be able to create a trustful interviewing atmosphere and hence minimize the social desirability bias. Using a rich set of respondent and interviewer characteristics as well as variables reflecting the interview situation allows us to examine this aspect closely.

Third, we conduct several analyses that allow us to identify the direction of misreporting (over- vs. underreporting). Because the literature already documents the phenomenon of "mean-reverting measurement error" as manifested in increased misreporting in the tails of the earnings distribution (Kim & Tamborini

2014; Angel et al. 2017), we further analyze the tendencies to under- or overreport in different quartiles of the earnings distribution. This strategy also yields deeper insight into the impact of socially desirable reporting on the determinants of such tendencies.

Theoretical Background and Hypotheses

There are various reasons why collecting information on earnings is difficult. First, we should differentiate between unintentional and deliberate misreporting. Answering the question "What are your monthly gross earnings?" requires a cognitive process that passes through several stages, including interpretation of the question, retrieval of the exact amount, estimation and judgment, and response production (see, e.g., Tourangeau 1984; Moore et al. 2000; Groves et al. 2009; Kim & Tamborini 2014). Problems in interpretation/understanding, recall and response production result in unintentional misreporting. These, however, should generate an approximately randomly distributed measurement error or heaping[1] at round numbers.

In the case that respondents perceive answering an earnings question as uncomfortable or a violation of privacy, they could either refuse to answer or deliberately misreport values. This is consistent with findings that earnings questions have the highest rates of item nonresponse in general population surveys (see, e.g., Tourangeau & Yan 2007; Krumpal 2013) and that there is a substantial level of misreporting mainly in the left and right tails of the earnings distribution (Pedace & Bates 2000; Riphahn & Serfling 2005; Essig & Winter 2009; Bollinger et al. 2018). This "mean-reverting measurement error" (Bound & Krueger 1991; Bricker & Engelhardt 2008) gives rise to our assumption that such response behavior is caused by socially desirable reporting rather than by problems of understanding or recall.

According to social desirability theory, respondents reflect on an expected mainstream view in their cultural context and then edit their answers to comply with this view (see, e.g., DeMaio 1984; Krumpal 2013; Lipps & Lutz 2017). In other words, they are more likely to report desirable attributes than undesirable ones to present themselves in a positive light (Stocké & Hunkler 2007). The presence of an interviewer might either increase the tendency to edit answers, especially when the social distance between the respondent and interviewer is perceived as high (Diekmann 2008), or decrease misreporting, particularly when the interviewer is able to create a trustful atmosphere or help the respondent to interpret a question correctly (see, e.g., Landrock 2017).

[1] Heaping refers to reporting numbers in increments (Zinn & Würbach 2016).

A rich literature on the influence of respondent characteristics exists, even more so for interviewer effects on item nonresponse and measurement error in the case of sensitive questions (for an overview, see West & Blom 2017). We consider both types of influences to explain the misreporting of earnings in the survey data by accounting for socially desirable reporting. Thus, we derive our three main research questions and assign several hypotheses to each of them:

1. What is the extent of earnings misreporting in the survey data?
2. How do respondent characteristics influence measurement error on earnings questions?
3. How do interviewer characteristics influence measurement error on earnings questions?

First, we are interested in the overall extent of misreporting, measured as the deviation between the two data sources. Some evidence exists that the measurement error on earnings questions is modest in panel studies (Bound & Krueger 1991; Kühne 2018). The results are inconsistent, however, with regard to whether earnings are underreported mainly by high earners (Paulus 2015; Angel et al. 2017) or low earners (Meyer & Mittag 2017), are overreported by low earners (Bollinger 1998), or are both over- and underreported depending on the characteristics of different subgroups (Pedace & Bates 2000; Kim & Tamborini 2014). Taken together, these findings lead us to expect a mean-reverting measurement error with more pronounced rates of misreporting in both tails of the earnings distribution (*hypothesis 1*).

Concerning the impact of the sociodemographic characteristics of the respondents on responses to sensitive questions, Preisendörfer and Wolter (2014, p. 126) find that female, older, and better-educated respondents are more likely to underreport delinquent behavior than male, younger, and less-educated respondents are. Regarding income questions, however, Bound and Krueger (1991) show that the average measurement error is larger for men than for women (confirmed by, e.g., Bricker & Engelhardt 2008). Bollinger (1998) finds that low-income men are most likely to overreport their earnings. Following these findings, we assume that female respondents report more accurately in general (*hypothesis 2*). The effects of age and education are less clear. Some studies find no evidence for correlations of age and education with misreporting (Bound & Krueger 1991; Gottschalk & Huynh 2005), whereas others find that the measurement error rises with reported education level (Bricker & Engelhardt 2008) or decreases with better education at higher earnings levels (Kim & Tamborini 2014). As these findings are rather ambiguous, we follow the more general study of Preisendörfer and Wolter (2014) and hypothesize that younger (*hypothesis 3*) and less-educated (*hypothesis 4*) respondents report their earnings more accurately than other groups do.

In general, respondent characteristics that are associated with the level of earnings are often considered to affect misreporting on earnings questions. Therefore, we include personality traits of the respondents in our models. Several studies confirm an effect of personality traits, as measured in terms of the "Big Five Inventory" dimensions, on earnings (e.g., Mueller & Plug 2006; Heineck & Anger 2010; Spurk & Abele 2011), although the effects differ depending on whether these traits are considered independently or in combination with sociodemographic characteristics. For our analysis, we choose two personality traits out of five that we assume to exert a significant influence on the ability to cope with the interview situation. These traits are the dimensions of "extraversion" and "agreeableness" in the Big Five Inventory.[2] In NEPS SC6, each personality trait was measured by two items on a scale from 1 through 5, as recommended by Rammstedt and John (2007), who developed this short version of the Big Five Inventory. We assume that in an interview situation, a distinctly extraverted respondent will tend to exaggerate his or her earnings and thus be more likely to overreport (*hypothesis 5*), whereas a respondent with a high score in agreeableness will tend to stay close to the true value of his or her earnings and hence report more accurately (*hypothesis 6*).

To explore how interviewers influence the measurement error on earnings, we also formulate hypotheses on the sociodemographic characteristics of the interviewer and on the interview situation. West and Blom (2017, p. 187) give an overview of the effects of the interviewer's gender, age and experience on misreporting in response to sensitive questions. The majority of studies they consider find female interviewers to elicit more accurate responses than male interviewers do. However, the interaction between the genders of the interviewer and respondent also seems to play a role (Lipps & Lutz 2017). Regarding the age of the interviewer, West and Blom (2017) find greater evidence for a positive relationship between response quality and interviewer age, although this relationship is again moderated by the interaction between the interview partners. Because the similarity between interviewer and respondent seems to be an important source of influence (cp. Diekmann 2008), we hypothesize that interviewers of the same gender (*hypothesis 7*) and of similar age (*hypothesis 8*) and educational level (*hypothesis 9*) to those of the respondent induce less misreporting.

Furthermore, West and Blom (2017) consider the effect of the interviewer's experience on response quality. These authors distinguish between overall experience as an interviewer and survey-specific experience. Although they find ambiguous evidence for both experience measures, we nevertheless hypothesize that more experienced interviewers should, in general, elicit more accurate reports (*hypothesis 10*) and that accuracy should also be positively correlated with the interviewer's

[2] All Big Five Inventory dimensions seem to have some effects on earnings, however, we do not assume a significant influence of the other three dimensions (openness to experience, conscientiousness, and neuroticism) on respondent's answering behavior.

experience with the current survey. Following Preisendörfer and Wolter (2014), we thus hypothesize that reporting accuracy should increase with the number of interviews conducted within any given survey wave (*hypothesis 11*). Finally, interviewer effects inducing socially desirable reporting are stronger in face-to-face interviews, during which the interviewers' characteristics are directly observable by the respondents, than they are in telephone interviews, during which the interviewers' characteristics can only be estimated by the respondents (West et al. 2013). We therefore expect to find smaller measurement errors in computer-assisted telephone interviewing (CATI) responses than in computer-assisted personal interviewing (CAPI) responses (*hypothesis 12*).

Data and Research Strategy

Data

Survey data: NEPS SC6

For our analyses, we use survey data from the German National Educational Panel Study, Starting Cohort "Adults" (NEPS SC6, https://doi.org/10.5157/NEPS:SC6:8.0.0). An overview of the content and theoretical basis of this survey can be found in Allmendinger et al. (2011). From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The survey data were first collected as part of the survey "Working and Learning in a Changing World" (IAB-ALWA, cp. Antoni et al. 2010). This survey consisted of a start-up survey wave, conducted in winter 2007/2008, and was followed by seven additional annual survey waves, starting in winter 2009/2010, conducted in the NEPS framework. In each of these eight waves, the employment histories of the respondents were recorded. Naturally, information for all employment episodes before each respondent's first interview had to be collected retrospectively (i.e., all past employment experiences were queried in the first interview). In each follow-up wave, all current employment episodes of each person were surveyed in real time. For episodes of the latter type, net and gross income information was queried in all survey waves except the first one; thus, our estimation sample is limited to waves 2 through 8. Additionally, more than 93% of the respondents consented for their survey information to be linked to administrative data from the German

Federal Employment Agency. The phrasing of the survey questions on earnings and linkage consent is available in the online documentation of the NEPS surveys.[3]

Linked data: NEPS-SC6-ADIAB

The longitudinal administrative data available at the Institute for Employment Research (IAB), the research unit of the German Federal Employment Agency, originate from a number of different sources within the German social security system. On the one hand, these data contain information on various aspects of unemployment insurance and assistance. On the other hand, and more importantly for our analyses, the IAB data also contain information provided by employers about all of their dependent employees.[4] These employment history data include data on every person who has been dependently employed at least once since 1975 for West Germany and since 1992 for East Germany. Information on employees is reported by employers via mandatory notifications to the social security system. For NEPS SC6 consenting respondents, the two data sources are linked. The joint data product we use for our analyses, NEPS-SC6-ADIAB (doi:10.5164/IAB.NEPS-SC6-ADIAB7515.de.en.v1), has been documented by Antoni et al. (2018).

While employers' notifications to the social security system also contain a small number of sociodemographic and job characteristics, the information most crucial to our analyses is the sum of gross earnings during each reported job episode. These earnings directly determine the contributions to social insurance. Employers who fail to issue correct notifications on earnings and to directly transfer the proper amounts to the social security system are subject to considerable sanctions, ranging from financial penalties to imprisonment of up to ten years.[5] For these reasons, the information on gross earnings contained in the administrative data at hand is considered to be highly reliable. We therefore treat the resulting measure of earnings as the true value and any deviation from that value by the respondent during an interview as measurement error. There are, however, some caveats with regard to these administrative data, some of which require us to adapt our analyses.

Sample restrictions

Our goals are to measure the accuracy of earnings reports as precisely as possible and to distinguish the different factors contributing to any deviations we find. To do so, we take several steps during data preparation to restrict our sample to rule out

[3] https://www.neps-data.de/en-us/datacenter/dataanddocumentation/startingcohortadults/documentation

[4] See Antoni, Ganzer, and vom Berge (2016) for more details on these administrative data and https://fdz.iab.de/en.aspx for how to access them via the Research Data Centre of the Federal Employment Agency at the IAB.

[5] As of § 266a of the German Criminal Code, "StGB".

any factors in the two data-generating processes that might contribute to deviations that respondents cannot understand or influence.

Because the administrative employment histories include only dependent employees whose earnings are subject to mandatory social security contributions, these data do not contain information on civil servants (the German "Beamte") or the self-employed. Any such employment episodes are therefore not considered in our analyses, even when they are reported in the survey data. Moreover, employer notifications do not include the working hours corresponding to employment episodes. This prevents us from calculating hourly earnings, which is particularly problematic for part-time workers. For this reason, we restrict all analyses to full-time employment episodes.

Another reason for this restriction to full-time employment episodes is that the record linkage procedure merely identifies the administrative data corresponding to a given person from the survey data. The linkage process does not extend to the assignment of every single employment episode from one dataset to its exact counterpart in the other dataset. We therefore restrict our estimation sample to employment episodes that were either ongoing on the date of the interview or had ended shortly before the interview. In this way, we can ensure that we are actually comparing earnings measures related to the same employment episode. If we did not remove part-time employment episodes from the analyses, our sample could include respondents with two parallel part-time jobs at the time of the interview. This would strongly increase the risk of assigning two unrelated job episodes to each other and, thus, of comparing the wrong earnings measures.

For observations with administrative earnings beyond the social security contribution ceiling ("Beitragsbemessungsgrenze"), the measure is truncated at this threshold value. Because it would be impossible to determine a valid administrative earnings measure in these cases, we eliminate them from the estimation sample in accordance with the procedures recommended by Drews, Groll, and Jacobebbinghaus (2007, p. 32).

Finally, some special payments made to employees (e.g., end-of-year bonuses) may be reported in separate but parallel notifications, usually with a much shorter duration than that of the main employment episode. We do not add the wage sums reported in such notifications to our administrative earnings measures because it would be impossible to determine whether a given respondent considered such a payment when reporting on his gross earnings. Because such special payments may introduce natural deviations between the earnings measures that are unrelated to the response behavior during the interview, we include a number of variables in our estimations that at least allow us to control for the existence of such factors.

In addition to these deliberate exclusions of cases, we also drop observations for which any of the dependent or independent variables are missing. The greatest loss in observations for our complete case analyses results from the fact that the

Big Five personality traits were surveyed only in waves 5 and 8 of the NEPS SC6 survey. However, the stability of the Big Five instruments over time is documented in the literature, especially for the adult population (cp. Cobb-Clark & Schurer 2012; Rantanen et al. 2007). This allows us to transfer the reported data for a given person to all corresponding interviews from other waves without measurements of these traits. Ultimately, because we exclude all respondents without any personality trait measurements, our sample is reduced to all respondents who answered the relevant questions in at least one of waves 5 and 8.

Due to omitting such a large part of the original NEPS SC6 sample, and especially due to restricting the estimation sample to full-time dependently employed persons, we have to acknowledge that our remaining estimation sample is no longer representative. In Table A1[6] in the Appendix, we compare the subsamples (estimation sample vs. non-estimation sample) for the main respondent characteristics used in our analyses. Unsurprisingly, nearly all characteristics show significant biases between the two groups. As a result, we cannot claim representativeness for our estimations.

Dependent Variables

Our main focus is on (deliberate) deviations in reported earnings relative to the administrative measure. As our main dependent variable, we chose an indicator that reflects deviation of reported earnings from the administrative measure by more than 20%.[7]

Additionally, we extended the dependent variable to a multinomial indicator reflecting the direction of deviation (“underreporting”, “no deliberate deviation”, or “overreporting”). Again, we chose a threshold of deviation by more than 20% in each direction. Because the three categories are mutually exclusive by nature, this generated variable can be used as a dependent variable in multinomial logit models without violating the assumption of the independence of irrelevant alternatives. This indicator separates our estimation sample into 1464 instances of underreporting and 760 instances of overreporting, corresponding to 10% and 5%, respectively, of the total number of observations.

[6] All tables in this paper were generated using the user-written Stata routine estout (Jann 2005).

[7] Three alternative variations of all models, using a threshold of 10%, a full standard deviation of the earnings distribution, and one-half standard deviation of the earnings distribution, have been calculated. The results are available by request.

Respondent and Interviewer Characteristics

The estimation sample comprises 14065 observations from 4087 respondents, i.e., the average number of observations per respondent is approximately 3.4 (see Table 1 for a tabular overview). A descriptive analysis of the respondents' characteristics reveals that 70% of the observations correspond to male respondents. This overrepresentation is attributed to the fact that we consider only full-time employment episodes, which are still more common among men than among women in Germany. Most respondents were aged between 30 and 49 years (52%); only 9% were younger.[8] A small minority of 4% of the respondents reported no vocational degree after schooling; the education level of the majority (38%) corresponded to intermediate schooling with vocational training. Approximately one-fifth of the observations were collected from respondents who had completed lower secondary education and vocational training (20%), another one-fifth to respondents who had completed

Table 1 Respondent characteristics

	Mean	SD	Min	Max
Resp. gender				
Male	0.70	0.46	0	1
Female	0.30	0.46	0	1
Resp. age				
Aged 29 and lower	0.09	0.28	0	1
Aged 30-49	0.52	0.50	0	1
Aged 50 or older	0.39	0.49	0	1
Resp. education				
Schooling, no training	0.04	0.20	0	1
Lower secondary, voc. train.	0.20	0.40	0	1
Intermediate, voc. training	0.38	0.49	0	1
Upper secondary, voc. train.	0.18	0.38	0	1
Higher education degree	0.21	0.41	0	1
Personality traits				
Big 5: Extraversion	3.32	0.92	1	5
Big 5: Agreeableness	3.54	0.59	1.3	5
Survey mode				
CAPI	0.49	0.50	0	1
CATI	0.51	0.50	0	1

Source: NEPS-SC6-ADIAB, own calculations.
Notes: Number of observations: 14065 of 4087 respondents.

[8] We classified respondent age to brackets similar to those available for interviewers.

upper secondary education and vocational training (18%), and a similar number are associated with respondents holding a higher education degree (21%). The two personality traits considered in our analyses, “extraversion” and “agreeableness”, show means of 3.32 and 3.54, respectively. Approximately one half of the interviews included in our estimation sample were conducted via CAPI; the other half were conducted via CATI.

Table 2 presents a comparison of the interviewer characteristics for each survey mode. Most of the available interviewer attributes show significant differences between the CATI and CAPI modes, as shown by t-tests of the differences between the means for the two groups. The most striking findings are that the interviewers in the CAPI group were significantly older and more experienced than those in the CATI group. On the other hand, the CAPI interviewers performed significantly fewer interviews on average than the CATI interviewers did. This finding is not surprising, considering that CAPI interviewers must travel to their respondents' locations before conducting interviews, while a CATI interviewer may be assigned

Table 2 Interviewer characteristics, t-test by interview mode

	CAPI	CATI	Difference	t
Interviewer's gender				
I: male	0.569	0.528	0.041***	4.907
I: female	0.431	0.472	-0.041***	-4.907
Interviewer's age				
I: aged 29 and lower	0.008	0.311	-0.303***	-53.148
I: aged 30-49	0.151	0.372	-0.221***	-30.604
I: aged 50-65	0.607	0.277	0.330***	41.765
I: aged older than 65	0.234	0.041	0.194***	35.054
Interviewer's education				
I: lower secondary	0.165	0.082	0.082***	15.001
I: intermediate	0.246	0.185	0.061***	8.790
I: upper secondary	0.589	0.732	-0.143***	-18.142
Interviewer's experience				
I: exp. less than 2 years	0.144	0.285	-0.141***	-20.653
I: exp. 2-3 years	0.294	0.291	0.003	0.454
I: exp. 4-5 years	0.200	0.235	-0.034***	-4.948
I: exp. 6 or more years	0.362	0.189	0.172***	23.361
I: no. of interviews conducted so far	28.574	47.179	-18.605***	-28.102

Source: NEPS-SC6-ADIAB, own calculations.
Notes: Number of observations: 14065. Number of interviewers: 800. *** indicates significance at the 0.1% level.

another interview immediately after the previous one without leaving the telephone studio. We expect the interviewer's knowledge of this specific NEPS SC6 questionnaire to be a possible factor in reducing the risk of eliciting socially desirable answers to the earnings question. Thus, we include a variable reflecting the interviewer's individual familiarity with the specific survey instrument to test our *hypothesis 11.* This variable counts the number of interviews an interviewer has conducted in each wave up to and including the current one.

Control Variables for Multivariate Analyses

As mentioned earlier, our aim is to reduce all deviations between the two earnings measures to only those that can be considered deliberate misreporting, to the greatest possible extent. Thus, all regression analyses are performed on a set of control variables that can potentially support this distinction. Most importantly, we introduce four dummy variables based on the survey data that may influence the accuracy of the earnings measures. First is an indicator of paid overtime, complemented by a second indicator of other special payments. Third is a dummy that indicates whether a child benefit ("Kindergeld") is integrated into the earnings report. These variables act as approximations of factors that make deviations more likely because they represent monetary benefits that may be counted as earnings in one data source but not the other. A fourth dummy variable indicates whether the person is working for a public or private employer. This may influence how accurately respondents recall their gross earnings because public employees are assigned to highly standardized wage schemes, whereas employees of private companies have considerably more individual bargaining power over their earnings. We also control for the region of birth (West Germany, East Germany, or outside of Germany) as a proxy to reduce potential cultural differences in reporting. Finally, we include indicators of the panel wave in which an interview was performed.

Results

Extent and Determinants of Item Nonresponse of the Earnings Question

Before beginning a detailed analysis of the measurement error on reported earnings, we substantiate one of our central assumptions, namely, that information on earnings is considered sensitive, at least for respondents from a cultural context in which money is a taboo topic, such as in Germany. This is corroborated by the following findings: First, we encounter a substantially higher share of item nonresponse on the gross earnings question compared to questions on more generic information, e.g., a respondent's job. The number of answers in the categories "don't know" and "refuse to answer" together represent more than 11% of the responses to the earnings question, a much higher share than those for questions on, for instance, part-time vs. full-time employment (0.7%) or attendance of training courses during a given job (less than 0.5%). Only one-half of the item nonresponse on the gross earnings question correspond to recall problems (i.e., answering "don't know"), either claimed or true.

We derive our second confirmation of question sensitivity by estimating the effects of the respondents' and interviewers' characteristics on the propensity to validly answer the open-ended question[9] about gross earnings. The results of this standard logit model are presented in Table A2 in the Appendix. The lack of a face-to-face presence of the interviewer during a telephone interview (in contrast to the CAPI mode) reduces the risk of item nonresponse, which is consistent with our assumption that the mere presence of an interviewer might lead respondents to avoid answering the earnings question. On the other hand, we see a significant influence of interviewer experience, with field personnel with at least two years of experience successfully reducing the risk of item nonresponse on the gross earnings question. Using the administrative data source, we are able to classify the survey participants by their "true" earnings, even if they did not respond to the earnings question. To do so, we include the variable that reflects the appropriate quartile of the earnings distribution. The results show that respondents in the lowest earnings quartile are the most likely to provide a valid answer to the open gross earnings question, although only the marginal effects of the second and fourth quartile are significant. These results are consistent with the findings in the existing literature, which indicate that respondents with lower earnings levels are more likely to answer the earnings question, whereas persons with higher earnings levels are more likely to refuse to answer. Overall, these findings are well consistent with our assumption regarding the sensitive nature of the earnings question.

Descriptive Comparison of Earnings Measures

A comparison of the survey gross earnings measure and the more reliable administrative gross earnings measure illustrated in Table 3 shows considerable deviation between these two central attributes. When inspecting the two earnings measures separately, we find both of their means to be slightly above 3000 Euros (rows 1 and 2). From row 3 onward, Table 3 shows the deviation of the survey earnings measure from the administrative earnings measure based on the difference computed for

[9] For each relevant job, the question on gross earnings was first asked using an open-ended question to elicit the exact value. Only if the respondent was unable or unwilling to answer that question would he or she be asked to at least classify his or her earnings relative to a list of earnings brackets.

Table 3 Descriptive statistics of survey and administrative gross earnings measures and individual deviations (in Euro)

	Mean	Median	Min	Max	N
Gross earnings measures:					
Administrative data	3353.1	3247.3	1216.8	6050.5	14065
Survey data	3162.5	3000.0	980.0	18000.0	14065
Deviation of survey measure from administrative measure:					
Overall	-190.6	-194.8	-3707.8	12532.3	14065
>20% underreporting	-1041.6	-980.6	-3707.8	-297.6	1464
>20% overreporting	1268.9	955.9	257.3	12532.3	760

Source: NEPS-SC6-ADIAB, own calculations.

Notes: All values in the three bottom rows are calculated based on individual observations, not by calculating the difference between the first two table rows. In the two bottom rows, a difference is counted as under- or overreporting ifthe survey measure is more than 20% below or above the administrative measure, respectively.

each individual observation. Both mean and median of the overall deviation are roughly 200 Euros (190.6 Euros and 194.8 Euros, respectively). The absolute value of the mean deviation represents underreporting by almost 6% relative to the mean of administrative earnings. The two bottom rows of Table 3 present additional details on observations with under- or overreporting by more than 20% compared to the given administrative earnings measure. Among these observations, the mean absolute deviations are more than 1000 Euros in each direction.

By comparing the whole distribution of each of the two variables, we find typical heaping structures in the survey data. In addition, the distribution of the survey measure is slightly but visibly shifted towards the lower side of the distribution (see Figure 1).

A closer look at the differences between the two earnings measures shows an interesting pattern: While the deviation is balanced in the lower earnings groups, it becomes broader with higher earnings. For illustration, Figure 2 visualizes this pattern across the four quartiles of the earnings distribution, as drawn from the administrative data. While the standard deviation of the difference might be expected to increase for higher earnings groups because respondents with higher earnings have a broader range of possible answers to which they could deviate, the positions of the quartiles are emphasized here. The higher the earnings quartile is, the more likely is underreporting compared to the more reliable administrative earnings. This result, although only initial descriptive evidence, supports parts of *hypothesis 1.*

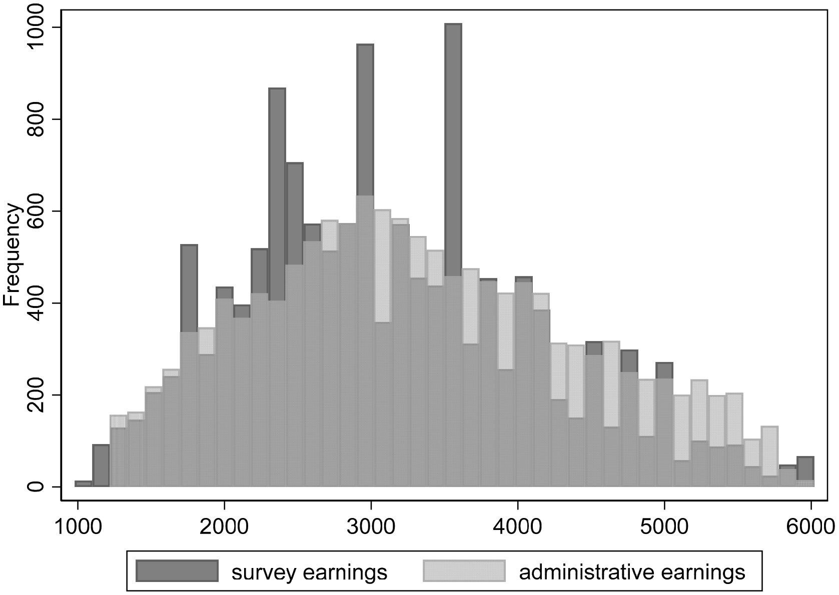


Figure 1 Histograms of reported and administrative monthly gross earnings (excluding outliers; in Euro)

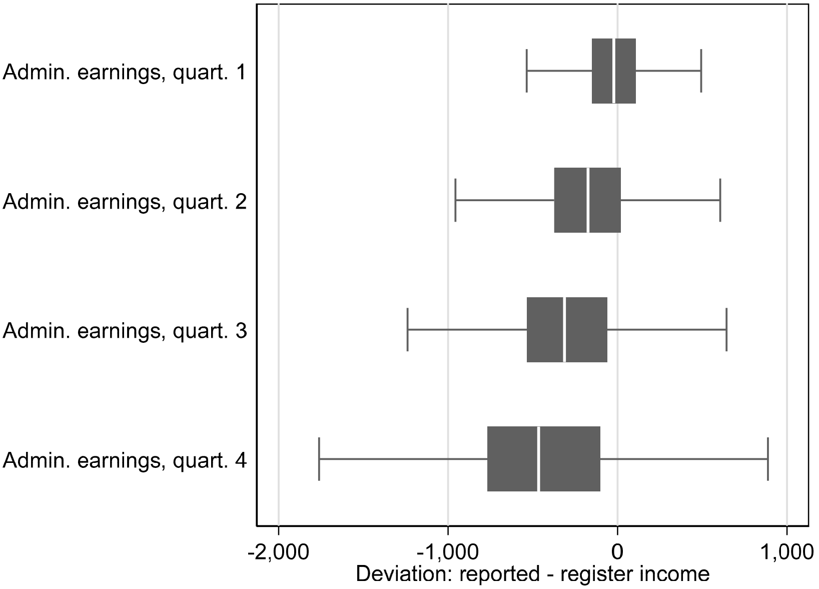


Figure 2 Box plots of deviations between the reported and administrative monthly gross earnings for each quartile of the administrative earnings (excluding outliers; in Euro)

Bias in Reported Earnings

We next examine the extent of bias in case that the open-ended earnings question is answered. Therefore, we calculate two basic model specifications with the dependent binary variable “reported monthly gross earnings differ from administrative earnings by more than 20%”. The first specification (“restricted model”, left column of Table 4) displays the average marginal effects for respondent and interviewer characteristics without controlling for quartiles of the earnings distribution. As existing literature reflects the relevance of the earnings quartiles (e.g., Angel et al. 2017; Meyer & Mittag 2017), we add respective dummies to the specification “basic model” (right column of Table 4). The results of both standard logit models are presented and discussed in the following.

In the restricted model we find an effect of gender that confirms our *hypothesis 2*: female respondents have a lower likelihood of deviating from their “true” earnings when answering the survey question. We do not find any effect regarding age (*hypothesis 3*). However, there is a clear educational effect. Highly educated respondents are less likely to report earnings that differ from their administrative earnings. This result contradicts our *hypothesis 4*. Considering the influence of personality traits, we find that extroverted persons show a higher tendency to inaccurately report their earnings, while the opposite is true for persons with a high score in agreeableness. These findings are in accordance with *hypotheses 5* and *6*. We also find significantly higher accuracy for interviews conducted via telephone, which supports our *hypothesis 12*. However, we do not find any significant effects for interviewer characteristics (*hypotheses 7* to *11*).

In the basic model most of these effects persist, except the effect of respondents’ gender. Additionally, we find a clear tendency for respondents with the highest level of earnings to misreport, which is partly consistent with *hypothesis 1*. Both calculated measures of model fit (Pseudo R^2 and AIC) indicate that the basic model that controls for earnings quartiles is more appropriate. Thus, all subsequent analyses are based on this model specification.

Direction of Bias in Reported Earnings

To gain deeper insight into the topic, we analyze the direction of misreporting as a categorical dependent variable (underreporting vs. no deviation vs. overreporting) in a multinomial logit model. Table 5 presents the results of this model, with non-deviating respondents being the baseline category for the multinomial calculation.

Table 4 Logit regressions, basic model and restricted model without monthly gross earnings quartiles as control variables, results displayed as average marginal effects

	Restricted model		Basic model	
Resp. gender (ref.: male)				
Female	-0.024**	(-3.17)	-0.015	(-1.94)
Resp. age (ref.: aged 29 and lower)				
Aged 30-49	-0.006	(-0.51)	-0.021	(-1.67)
Aged 50 or older	0.005	(0.38)	-0.014	(-1.07)
Resp. education (ref.: schooling, no training)				
Lower secondary, voc. training	0.005	(0.25)	0.007	(0.38)
Intermediate, voc. training	-0.010	(-0.57)	-0.016	(-0.85)
Upper secondary, voc. training	-0.012	(-0.70)	-0.026	(-1.43)
Higher education degree	-0.036*	(-2.13)	-0.064***	(-3.46)
Admin. earnings (ref.: quart. 1)				
Admin. earnings, quart. 2			0.010	(1.03)
Admin. earnings, quart. 3			0.022	(1.91)
Admin. earnings, quart. 4			0.085***	(6.75)
Personality traits				
Big 5: Extraversion	0.010**	(2.72)	0.010**	(2.68)
Big 5: Agreeableness	-0.015*	(-2.42)	-0.013*	(-2.18)
Survey mode (ref.: CAPI)				
CATI	-0.030*	(-2.01)	-0.031*	(-2.13)
I: gender (ref.: male)				
I: female	0.004	(0.52)	0.004	(0.46)
I: age (ref.: aged 29 and lower)				
I: aged 30-49	-0.018	(-1.62)	-0.018	(-1.69)
I: aged 50-65	0.003	(0.24)	0.004	(0.32)
I: aged older than 65	-0.012	(-0.69)	-0.012	(-0.69)
I: education (ref.: lower secondary)				
I: intermediate	0.019	(1.17)	0.019	(1.20)
I: upper secondary	-0.001	(-0.09)	-0.001	(-0.06)
I: experience (ref.: exp. below 2 years)				
I: exp. 2-3 years	0.002	(0.19)	0.001	(0.14)
I: exp. 4-5 years	-0.008	(-0.70)	-0.008	(-0.67)
I: exp. 6 or more years	0.001	(0.07)	0.001	(0.06)
I: no. of interviews cond. so far	0.000	(1.84)	0.000	(1.77)

	Restricted model	Basic model
Pseudo R^2	0.013	0.020
AIC	12179	12105
Observations	14065	14065

Source: NEPS-SC6-ADIAB, own calculations.
*Note*s: Indicator for absolute deviation by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

As expected, we find strongly diverging effects for underreporters and overreporters.

For instance, female respondents show a significantly higher tendency to underreport their earnings than their male peers, whilst they simultaneously show a clear tendency to overreport less often. By differentiating the direction of bias in reported earnings, we can see that the former support of *hypothesis 2* was driven mainly by females' non-overreporting behavior.

In contradiction to the results from Table 4, we find significant effects for the respondent's age: Compared to persons below the age of 30 years, older respondents are less likely to underreport their earnings. Simultaneously, the older age groups show a higher tendency to overreport their earnings. This finding now partly supports *hypothesis 3*, while the basic model did not reveal such an effect. The same is true for the impact of education on misreporting. We see that persons with upper secondary or higher educational degrees are less likely to underreport but more likely to overreport their earnings. This result partly supports *hypothesis 4.*

The enhanced model also shows that the effects of the respondents' personality traits, as found earlier, are driven only by the overreporting respondents. Extraverted persons show a significantly higher likelihood to overstate their earnings, and agreeable respondents show a reduced likelihood to do so. This supports our *hypotheses 5* and *6*. Nevertheless, the former effect of a lower likelihood of misreporting in CATI mode vanishes (*hypothesis 12*). Again, we do not see any effects of interviewer characteristics (*hypotheses 7* to *11*).

Table 5 Multinomial logit regressions to differentiate between over- and underreporting, results displayed as average marginal effects

	Underreporting		Overreporting	
Resp. gender (ref.: male)				
Female	0.024**	(3.18)	-0.038***	(-10.44)
Resp. age (ref.: aged 29 and lower)				
Aged 30-49	-0.049***	(-3.97)	0.016**	(3.02)
Aged 50 or older	-0.054***	(-4.31)	0.027***	(4.44)
Resp. education (ref.: schooling, no training)				
Lower secondary, voc. training	-0.000	(-0.02)	0.006	(0.70)
Intermediate, voc. training	-0.024	(-1.26)	0.005	(0.56)
Upper secondary, voc. training	-0.052**	(-2.92)	0.024*	(2.42)
Higher education degree	-0.087***	(-4.75)	0.029**	(2.89)
Admin. earnings (ref.: quart. 1)				
Admin. earnings, quart. 2	0.050***	(8.14)	-0.056***	(-6.75)
Admin. earnings, quart. 3	0.090***	(9.81)	-0.082***	(-10.27)
Admin. earnings, quart. 4	0.166***	(15.88)	-0.081***	(-9.71)
Personality traits				
Big 5: Extraversion	-0.005	(-1.56)	0.015***	(5.88)
Big 5: Agreeableness	-0.005	(-0.87)	-0.008**	(-2.61)
Survey mode (ref.: CAPI)				
CATI	-0.022	(-1.66)	-0.009	(-1.39)
I: gender (ref.: male)				
I: female	-0.001	(-0.15)	0.005	(1.41)
I: age (ref.: aged 29 and lower)				
I: aged 30-49	-0.015	(-1.75)	-0.003	(-0.47)
I: aged 50-65	0.003	(0.29)	0.001	(0.08)
I: aged older than 65	-0.007	(-0.50)	-0.004	(-0.47)
I: education (ref.: lower secondary)				
I: intermediate	0.021	(1.41)	-0.004	(-0.53)
I: upper secondary	-0.001	(-0.12)	-0.000	(-0.03)
I: experience (ref.: exp. below 2 years)				
I: exp. 2-3 years	-0.002	(-0.24)	0.003	(0.59)
I: exp. 4-5 years	0.001	(0.08)	-0.008	(-1.42)
I: exp. 6 or more years	0.008	(0.84)	-0.007	(-1.27)
I: no. of interviews conducted so far	0.000	(1.43)	0.000	(1.60)

	Underreporting	Overreporting
Pseudo R^2	0.065	
AIC	14291	
Observations	14065	

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for over-/underreporting by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

Finally, we find substantive effects of the earnings quartiles. Compared to persons from the lowest quartile (the reference category), belonging to a higher quartile of the earnings distribution is correlated with a higher tendency of underreporting and a lower tendency of overreporting. This finding again only partly supports our *hypothesis 1.* Moreover, it suggests that there may be different mechanisms of report bias, and different directions of bias, in different parts of the earnings distribution. As the bias does not appear to be randomly distributed, we interpret it as evidence for deliberate misreporting.

Bias in Reported Earnings Across the Earnings Distribution

To account for the strong impact of the earnings quartiles in the former models and to follow an approach similar to Kim and Tamborini (2014), we recalculate our multinomial model separately for each of the four quartiles of the earnings distribution. These models are presented in Table 6. We find a consistent gender effect, indicating that females are less likely to overreport across all earnings quartiles, supporting *hypothesis 2*. Yet, in the two earnings groups below the median, the model reveals an increased likelihood for females to underreport their earnings.

The effects of the respondent's age are somehow contradictory. While older persons are less likely than respondents below the age of 30 years to underreport in earnings quartiles 2 and 4, this result does not hold for the other quartiles. Older respondents also show a higher tendency to overreport but only in the tail quartiles. Thus, *hypothesis 3* again appears to be only partly supported. However, these effects foster the idea of a U-shaped pattern of misreporting across the earnings distribution, dependent on the age groups, which indirectly supports *hypothesis 1.*

The effect of higher education making respondents less likely to underreport but more likely to overreport is clearly driven only by the second quartile of the

Table 6 Multinomial logit regressions to differentiate between over- and underreporting, estimated separately by quartiles of administrative earnings, results displayed as average marginal effects

	Earnings quart. 1		Earnings quart. 2		Earnings quart. 3		Earnings quart. 4	
	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.
Resp. gender (ref.: male)								
Female	0.023 **	-0.058***	0.037**	-0.037***	0.004	-0.027***	0.021	-0.032***
	(2.902)	(-5.801)	(2.774)	(-4.677)	(0.263)	(-5.358)	(1.208)	(-5.423)
Resp. age (ref.: aged 29 and lower)								
Aged 30-49	-0.003	0.026*	-0.070***	0.018	-0.050	-0.003	-0.082*	0.029**
	(-0.252)	(2.043)	(-3.525)	(1.753)	(-1.808)	(-0.305)	(-2.000)	(2.997)
Aged 50 or older	-0.019	0.045**	-0.049*	0.029*	-0.044	0.017	-0.113**	0.030**
	(-1.483)	(2.984)	(-2.249)	(2.466)	(-1.588)	(1.485)	(-2.691)	(2.936)
Resp. education (ref.: schooling, no training)								
Lower secondary, voc. training	-0.008	-0.003	-0.047	0.036**	0.009	-0.001	0.053	-0.007
	(-0.519)	(-0.162)	(-1.266)	(2.841)	(0.192)	(-0.042)	(1.150)	(-0.333)
Intermediate, voc. training	-0.005	0.006	-0.084*	0.032**	-0.010	-0.007	0.007	-0.001
	(-0.279)	(0.277)	(-2.396)	(2.716)	(-0.239)	(-0.413)	(0.159)	(-0.044)
Upper secondary, voc. training	-0.010	0.027	-0.092**	0.049***	-0.037	0.009	-0.064	0.019
	(-0.570)	(1.062)	(-2.688)	(3.334)	(-0.839)	(0.473)	(-1.534)	(0.945)
Higher education degree	-0.008	0.052	-0.121***	0.041*	-0.091*	-0.005	-0.103*	0.026
	(-0.413)	(1.784)	(-3.671)	(2.559)	(-2.106)	(-0.298)	(-2.495)	(1.259)

	Earnings quart. 1		Earnings quart. 2		Earnings quart. 3		Earnings quart. 4	
	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.	Underrep.	Overrep.
Personality traits								
Big 5: Extraversion	-0.002	0.024***	-0.000	0.012**	-0.017**	0.011***	-0.002	0.011*
	(-0.689)	(3.597)	(-0.002)	(2.582)	(-2.960)	(3.330)	(-0.254)	(2.574)
Big 5: Agreeableness	-0.002	-0.024**	0.001	0.009	-0.025*	-0.016**	0.001	-0.003
	(-0.278)	(-3.082)	(0.059)	(1.161)	(-2.021)	(-3.023)	(0.057)	(-0.457)
Pseudo R^2	0.053		0.066		0.071		0.057	
AIC	3402		3514		3396		4029	
Observations	3517		3516		3516		3516	

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for over-/underreporting by >20% of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. The constant, all interviewer variables from the basic model in table 4 and the following control variables are omitted from the table: region of birth, public employer, paid overtime, special payments, child benefits, survey mode and panel wave. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

earnings distribution. The model shows no significant effects for the corresponding variables in the lowest earnings group and only a slight similar effect for highly educated respondents in the two upper quartiles. This result still supports our *hypothesis 4* but only for one earnings group. Also, the results contradict the findings of Kim and Tamborini (2014).

Regarding the personality trait of "extraversion", the analyses show results consistent with those of the joint estimation. The tendency of extraverted persons to be more likely to overreport their earnings can be seen across all earnings quartiles, aided by a small decrease in the likelihood of underreporting in the third quartile, which supports *hypothesis 5*. More "agreeable" persons, however, do not show a consistently smaller likelihood to overreport across all earnings groups. By contrast, only persons in the first and third earnings quartiles show an effect of this kind, conveyed by the lower tendency to underreport in the third quartile. This result only weakly supports *hypothesis 6*.

Interviewer Effects

In the literature on interviewer effects, several studies suggest estimating multilevel models to calculate the extent of the interviewer's impact (see, e.g., O'Muircheartaigh & Campanelli 1999; Lipps & Pollien 2011; Korbmacher & Schroeder 2013). In the next step, we follow this approach to validate our previous findings of any interviewer effects. Given that we have thus far found little evidence for interviewer effects on reporting accuracy, it is not surprising that the intraclass correlation coefficient (ICC) is very low (0.0234). This result indicates that very little of the variance in our misreport indicator is explained by interviewer characteristics.

Moreover, the literature suggests possible influences not only of the interviewer characteristics themselves but also of the similarity in socio-demography between the interviewer and the respondent, as we have stated in *hypotheses 7* to *9* (e.g., Diekmann 2008; Lipps & Lutz 2017; West & Blom 2017). Thus, we recalculate both the basic and multinomial models after introducing dummy variables representing similarity in gender, age and education between both interview counterparts. This newly introduced indicators presented in Table 7 suggest an effect of the educational similarity between the interviewer and respondent, showing that interviewers who are more or less educated than their respective respondents are more likely to elicit underreported answers to the earnings question. This notable effect corroborates *hypothesis 9*. For the similarity in gender and age, we however do not find any support (*hypotheses 7* and *8*). Additionally, we do not see evidence to confirm our assumptions that the overall interviewing experience (*hypothesis 10*) or the interviewers' familiarity with the NEPS SC6 survey instrument (*hypothesis 11*) reduce the tendency to misreport.

Table 7 Logit and multinomial logit regressions, basic model estimated with additional indicators for difference between respondent and interviewer, results displayed as average marginal effects

	Logit		Underreporting		Overreporting	
I: experience (ref.: exp. below 2 years)						
I: exp. 2-3 years	0.002	(0.17)	-0.002	(-0.18)	0.003	(0.59)
I: exp. 4-5 years	-0.005	(-0.46)	0.003	(0.27)	-0.008	(-1.34)
I: exp. 6 or more years	0.001	(0.12)	0.009	(0.87)	-0.007	(-1.21)
I: no. of interviews conducted so far	0.000*	(1.98)	0.000	(1.70)	0.000	(1.69)
I: gender disparity (ref.: same gender as resp.)						
I: different gender than respondent	0.001	(0.16)	-0.003	(-0.58)	0.005	(1.41)
I: age disparity (ref.: same age class as resp.)						
I: higher age class than respondent	-0.005	(-0.56)	-0.005	(-0.59)	-0.001	(-0.14)
I: lower age class than respondent	-0.004	(-0.41)	-0.007	(-0.99)	0.003	(0.48)
I: educational disparity (ref.: same schooling as resp.)						
I: higher schooling than respondent	0.023*	(2.31)	0.023**	(2.59)	-0.000	(-0.01)
I: lower schooling than respondent	0.022*	(2.27)	0.027**	(2.91)	-0.005	(-0.94)
Survey mode (ref.: CAPI)						
CATI	-0.038*	(-2.44)	-0.027	(-1.92)	-0.011	(-1.65)
Pseudo R^2	0.019		0.065			
AIC	12110		14290			
Observations	14065		14065			

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for absolute deviation by >20% or for over-/underreporting by >20%, respectively, of administrative monthly gross earnings as dependent variable, z-statistics in parentheses. Interviewer characteristics sex, age and schooling are removed from the model. The constant and the remaining independent variables from the basic model in Table 4 are omitted from the table. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 800 interviewers.

Summary and Conclusions

We used linked survey data from the German National Education Panel Study's Starting Cohort "Adults" (NEPS SC6) and administrative data from the German Federal Employment Agency to estimate and analyze the drivers giving rise to measurement error in monthly gross earnings based on a sequence of logistic and multinomial logistic models. Constraints in comparability between the earnings measures in both data sources lessen the generalizability of our results. Following the latest validation studies (see, e.g., Kim & Tamborini 2014; Angel et al. 2017), we are able to classify inaccurate responses as either over- or underreporting. Gaining insight into the different mechanisms driving these two kinds of misreporting, we show that the higher response accuracy of female respondents is driven by a reduced tendency to overreport, while the inaccuracy effects for older and better-educated respondents are primarily driven by a reduced likelihood to underreport earnings. Moreover, the reporting inaccuracy of extraverted persons results from a higher tendency to overreport, whereas agreeable respondents are less likely to follow this pattern.

In regressions separated by earnings groups, we find mainly consistent effects of gender, age and personality traits. The education level effect persists only in the second quartile of the earnings distribution.

None of our calculations indicate important direct effects of interviewer characteristics on either reducing or amplifying the tendency to misreport. This may be an indication of highly competent field personnel and, if so, is good news in general for users of NEPS survey data. However, we find evidence that interviewers with education levels similar to those of their respective respondents may elicit more accurate results and, especially, reduce the risk of underreporting. This can be interpreted as the result of respondents' tendency to provide socially desirable answers.

In addition to all aspects covered by this article, cognitive factors may also affect the reporting of income. To validly answer a question about earnings, the respondent must at least pass through the cognitive stages of interpretation or understanding, retrieval, and response production (cp. Tourangeau 1984; Groves et al. 2009). Cognitive effects, if present, might be misinterpreted as an influence of socially desirable behavior. Thus, further analyses should aim to make use of competency assessment data to approximate these cognitive aspects and to narrow down the subset of misreporting that is truly due to social desirability.

References

Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R. & Ruland, M. (2011). Adult education and lifelong learning. *Zeitschrift für Erziehungswissenschaft*, 2 (Special Issue 14), 283–299. https://doi.org/10.1007/s11618-011-0197-0

Angel, S., Heuberger, R., & Lamei, N. (2017). Differences between household income from surveys and registers and how these affect the poverty headcount: evidence from the Austrian SILC. *Social Indicators Research*. https://doi.org/10.1007/s11205-017-1672-7

Antoni, M., Bachbauer, N., Eberle, J., & Vicari, B. (2018). *NEPS-SC6 survey data linked to administrative data of the IAB (NEPS-SC6-ADIAB 7515)* (FDZ-Datenreport No. 02/2018 EN). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg. https://doi.org/10.5164/IAB.FDZD.1802.en.v1

Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2010). *Working and learning in a changing world. Part I: overview of the study* (FDZ Methodenreport No. 05/2010). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Antoni, M., Ganzer, A., & vom Berge, P. (2016). *Sample of integrated labour market biographies (SIAB) 1975-2014* (FDZ-Datenreport No. 04/2016(en)). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Education as a Lifelong Process: The German National Educational Panel Study (NEPS). (2011). In H.-P. Blossfeld, H. G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft (Vol. 2, Special Issue 14)*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Bollinger, C. R. (1998). Measurement error in the current population survey: a nonparametric look. *Journal of Labor Economics*, 16(3), 576–594. https://doi.org/10.1086/209899

Bollinger, C. R., Hirsch, B., Hokayem, C. M., & Ziliak, J. P. (2018). *Trouble in the tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch* (IZA Discussion Paper No. 11710). IZA – Institute of Labor Economics. Bonn.

Bound, J., Brown, C., Duncan, G. J., & Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3), 345–368.

Bound, J. & Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: do two wrongs make a right? *Journal of Labor Economics*, 9 (1), 1–24.

Bricker, J. & Engelhardt, G. V. (2008). Measurement error in earnings data in the health and retirement study. *Journal of Economic and Social Measurement*, 33(1), 39–61.

Cobb-Clark, D. A. & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1), 11–15. https://doi.org/10.1016/j.econlet.2011.11.015

DeMaio, T. J. (1984). Social desirability and survey measurement: a review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 257–282). Russell Sage Foundation.

Deutscher Bundestag. (1998). Criminal Code in the version promulgated on 13 November 1998, Federal Law Gazette [Bundesgesetzblatt] I p. 3322, last amended by Article 1 of the law of 24 September 2013, Federal Law Gazette I p. 3671 and with the text of Article 6(18) of the law of 10 October 2013, Federal Law Gazette I p 3799.

Diekmann, A. (2008). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (19. Aufl.). Reinbek bei Hamburg: Rowohlt.

Drews, N., Groll, D., & Jacobebbinghaus, P. (2007). *Programmierbeispiele zur Aufbereitung von FDZ Personendaten in Stata* (FDZ Methodenreport Nr. 06/2007). Institut für Arbeitsmarkt- und Berufsforschung (IAB). Nürnberg.

Duncan, G. J. & Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(4), 508–532.

Engel, U., Jann, B., Lynn, P., & Scherpenzeel, A. (Eds.). (2014). *Improving survey methods: Lessons from recent research*. European Association of Methodology Series. London: Routledge.

Essig, L. & Winter, J. (2009). Item non-response to financial questions in household surveys: an experimental study of interviewer and mode effects. *Fiscal Studies*, 30(4), 367–390.

Gottschalk, P. & Huynh, M. (2005). Validation study of earnings in the SIPP – do older workers have larger measurement error? *Center for Retirement Research Working Papers*, (2005-07).

Gottschalk, P. & Huynh, M. (2010). Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *Review of Economics and Statistics*, 92(2), 302–315. https://doi.org/10.1162/rest.2010.11232

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, New Jersey: Wiley.

Heineck, G. & Anger, S. (2010). The returns to cognitive abilities and personality traits in Germany. *Labour Economics*, 17 (3), 535–546. https://doi.org/10.1016/j.labeco.2009.06.001

Jann, B. (2005). Making regression tables from stored estimates. *Stata Journal*, 5(3), 288–308.

Jann, B. (2014). Asking sensitive questions: overview and introduction. In U. Engel, B. Jann, P. Lynn, & A. Scherpenzeel (Eds.), Improving survey methods: Lessons from recent research (pp. 101–105). European Association of Methodology Series. London: Routledge.

Kapteyn, A. & Ypma, J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. *Journal of Labor Economics*, 25(3), 513–551.

Kim, C. & Tamborini, C. R. (2014). Response error in earnings: an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research*, 43(1), 39–72. https://doi.org/10.1177/0049124112460371

Kirkcaldy, B. D., Furnham, A. F., & Lynn, R. A. (1992). National differences in work attitudes between the UK and Germany. *European Work and Organizational Psychologist*, 2(2), 81–102. https://doi.org/10.1080/09602009208408537

Korbmacher, J. M. & Schroeder, M. (2013). Consent when linking survey data with administrative records: the role of the interviewer. *Survey Research Methods*, 7(2), 115–131. https://doi.org/10.18148/srm/2013.v7i2.5067

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, 47 (4), 2025–2047.

Kühne, S. (2018). From strangers to acquaintances? Interviewer continuity and socially desirable responses in panel surveys. *Survey Research Methods*, 12(2), 121–146. https://doi.org/10.18148/srm/2018.v12i2.7299

Landrock, U. (2017). How interviewer effects differ in real and falsified survey data: using multilevel analysis to identify interviewer falsifications. *methods, data, analyses*, 11(2), 163–188. https://doi.org/10.12758/mda.2017.03

Lipps, O. & Lutz, G. (2017). Gender of interviewer effects in a multi-topic centralized CATI panel survey. *methods, data, analyses*, 11(1), 67–86. https://doi.org/10.12758/mda.2016.009

Lipps, O. & Pollien, A. (2011). Effects of interviewer experience on components of nonresponse in the European Social Survey. *Field Methods*, 23(2), 156–172. https://doi.org/10.1177/1525822x10387770

Meyer, B. D. & Mittag, N. (2017, August). Using linked survey and administrative data to better measure income: implications for poverty, program effectiveness and holes in the safety net (IZA Discussion Paper No. 10943). IZA Institute of Labor Economics. Bonn.

Moore, J. C., Stinson, L. L., & Welniak, Jr., E. J. (2000). Income measurement error in surveys: a review. *Journal of Official Statistics*, 16(4), 331–361.

Mueller, G. & Plug, E. (2006). Estimating the effect of personality on male and female earnings. *ILR Review*, 60(1), 3–22. https://doi.org/10.1177/001979390606000101

National Research Council. (1984). *Cognitive aspects of survey methodology* (T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau, Eds.). Washington, D. C.: National Academies Press. https://doi.org/10.17226/930

O'Muircheartaigh, C. & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(3), 437–446. https://doi.org/10.1111/1467-985X.00147

Paulus, A. (2015, June). *Income underreporting based on income-expenditure gaps: survey vs tax records* (ISER Working Paper Series No. 2015-15). Institute for Social and Economic Research, University of Essex. Essex.

Pedace, R. & Bates, N. (2000). Using administrative records to assess earnings reporting error in the survey of income and program participation. *Journal of Economic and Social Measurement*, 26(3,4), 173–192.

Preisendörfer, P. & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, 78(1), 126–146. https://doi.org/10.1093/poq/nft079

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: a 10-item short version of the Big Five inventory in english and german. *Journal of Research in Personality*, 41(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rantanen, J., Metsäpelto, R.-L., Feldt, T., Pulkkinen, L., & Kokko, K. (2007). Long-term stability in the big five personality traits in adulthood. *Scandinavian Journal of Psychology*, 48(6), 511–518. https://doi.org/10.1111/j.1467-9450.2007.00609.x

Riphahn, R. & Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30(2), 521–538.

Spurk, D. & Abele, A. E. (2011). Who earns more and why? A multiple mediation model from personality to salary. *Journal of Business and Psychology*, 26(1), 87–103. https://doi.org/10.1007/s10869-010-9184-3

Stocké, V. & Hunkler, C. (2007). Measures of desirability beliefs and their validity as indicators for socially desirable responding. *Field Methods*, 19 (3), 313–336. https://doi.org/10.1177/1525822X07302102

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology*. Washington, D. C.: National Academies Press. https://doi.org/10.17226/930

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Trachtman, R. (1999). The money taboo: its effects in everyday life and in the practice of psychotherapy. *Clinical Social Work Journal*, 27(3), 275–288. https://doi.org/10.1023/A:1022842303387

Turner, C. F. & Martin, E. (Eds.). (1984). *Surveying subjective phenomena*. Russell Sage Foundation.

West, B. T. & Blom, A. G. (2017). Explaining interviewer effects: a research synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175–211. https://doi.org/10.1093/jssam/smw024

West, B. T., Kreuter, F., & Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: a function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29 (2), 277–297. https://doi.org/10.2478/jos-2013-0023

Zinn, S. & Würbach, A. (2016). A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, 43(4), 682–703. https://doi.org/10.1080/02664763.2015.1077372

Appendix

Table A1 T-test of characteristics of respondents within and outside the estimation sample

	Not in est. sample	In est. sample	Difference	t
Resp. gender				
Male	0.444	0.675	-0.231***	-26.215
Female	0.556	0.325	0.231***	26.215
Resp. age				
Aged 29 and lower	0.061	0.046	0.015***	3.592
Aged 30-49	0.388	0.460	-0.072***	-8.177
Aged 50 or older	0.551	0.494	0.057***	6.370
Region of birth				
West Germany	0.708	0.660	0.048***	5.772
East Germany	0.180	0.250	-0.070***	-9.759
Abroad	0.112	0.090	0.022***	3.954
Resp. education				
Schooling, no training	0.095	0.046	0.049***	9.868
Lower secondary, voc. training	0.204	0.190	0.014	1.910
Intermediate, voc. training	0.294	0.360	-0.066***	-7.966
Upper secondary, voc. training	0.142	0.174	-0.032***	-4.912
Higher education degree	0.265	0.230	0.035***	4.475
Personality traits				
Big 5: Extraversion	3.411	3.343	0.068***	3.899
Big 5: Agreeableness	3.597	3.534	0.063***	5.532
Survey mode				
CAPI	0.314	0.250	0.064***	7.791
CATI	0.686	0.750	-0.064***	-7.791

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Number of respondents: 16873. Unlike all other tables, this analysis only includes the most recent observation for each respondent instead of including all valid observations of respondents within the observation period. *** indicates significance at the 0.1% level.

Table A2 Logit regression on the availability of an open gross earnings report, results displayed as average marginal effects

	Dep. var.: open earnings	
Resp. gender (ref.: male)		
Female	-0.018**	(-3.05)
Resp. age (ref.: aged 29 and lower)		
Aged 30-49	-0.021**	(-2.80)
Aged 50 or older	-0.033***	(-4.49)
Resp. education (ref.: schooling, no training)		
Lower secondary, voc. training	-0.018	(-1.51)
Intermediate, voc. training	-0.005	(-0.43)
Upper secondary, voc. training	0.000	(0.04)
Higher education degree	0.009	(0.75)
Admin. earnings (ref.: quart. 1)		
Admin. earnings, quart. 2	-0.017**	(-2.63)
Admin. earnings, quart. 3	-0.010	(-1.67)
Admin. earnings, quart. 4	-0.035***	(-4.94)
Personality traits		
Big 5: Extraversion	-0.000	(-0.12)
Big 5: Agreeableness	0.006	(1.50)
Survey mode (ref.: CAPI)		
CATI	0.048***	(4.08)
I: gender (ref.: male)		
I: female	0.008	(1.16)
I: age (ref.: aged 29 and lower)		
I: aged 30-49	-0.012	(-1.33)
I: aged 50-65	-0.010	(-0.99)
I: aged older than 65	-0.003	(-0.21)
I: education (ref.: lower secondary)		
I: intermediate	-0.019	(-1.55)
I: upper secondary	-0.013	(-1.42)
I: experience (ref.: exp. below 2 years)		
I: exp. 2-3 years	0.020*	(2.11)
I: exp. 4-5 years	0.020	(1.91)
I: exp. 6 or more years	0.023*	(2.13)
I: no. of interviews conducted so far	-0.000	(-1.87)

	Dep. var.: open earnings
Pseudo R^2	0.035
AIC	7659
Observations	15162

Source: NEPS-SC6-ADIAB, own calculations.

Notes: Indicator for the availability of a valid response to the open question on gross earnings as dependent variable, z-statistics in parentheses. The constant and the following control variables are omitted from the table: region of birth, panel wave, public employer, paid overtime, special payments and child benefits. *, **, *** indicate significance at the 5%, 1% and 0.1% level, respectively. Standard errors clustered for 808 interviewers. Contrary to all other regressions, this regression also includes observations without valid responses to the open-ended question on gross earnings. The number of observations and interviewers is therefore higher than in all other tables. The administrative monthly gross earnings quartiles as control variables are recalculated to accommodate the different sample size.

methods, data, analyses | Vol. 13(1), 2019, pp. 91-110 DOI: 0.12758/mda.2018.05

Data Collection on Sensitive Topics with Adolescents Using Interactive Voice Response Technology

Paula Fomby & Narayan Sastry

University of Michigan

Abstract

We describe the development and implementation of a survey administered using interactive voice response (IVR) technology to collect information on sensitive topics in a US national sample of adolescents age 12-17. Respondents were participants in the Panel Study of Income Dynamics 2014 Child Development Supplement (N=1,098). We review questionnaire design, fieldwork protocols, data quality and completeness, and respondent burden. We find that in the context of research on sensitive topics with adolescents, IVR is a cost-efficient and flexible method of data collection that yields high survey response rates and low item nonresponse rates with distributions on key variables that are comparable to other national studies.

Keywords: interactive voice response; adolescents; sensitive topics; social desirability bias

Modes of data collection that allow greater anonymity, such as the internet, text messages, or interactive voice response, generally lead to more reporting of sensitive behaviors compared to standard telephone interviewing (Kreuter et al. 2008; Midanik & Greenfield 2010; Schober et al. 2015). Telephone interviewing is more likely to elicit accurate reports of sensitive behaviors when respondents are able to find a private setting in which to complete the interview or questions are worded or a response booklet is used so as to not require respondents to provide sensitive responses aloud. These conditions may be harder to achieve in telephone interviews with adolescents for three reasons: first, adolescents may have less control over the presence or interference of others during a telephone interview compared to adults, thus increasing the risk that sensitive information will be disclosed to a parent or sibling; second, the consequences of such disclosure may be uniquely consequential and detrimental for adolescents; and third, adolescents' greater tendency to provide socially desirable responses in survey settings compared to adults potentially compromises the quality of information on sensitive topics collected during an interviewer-administered telephone interview (Paulhus 1991; Reynolds & Richmond 1978).

Interactive voice response (IVR) technology provides an attractive method to overcome these concerns (Corkrey & Parkinson 2002; Stritzke et al. 2005; Tourangeau et al. 2002). In the survey context, IVR technology uses a pre-recorded or computer-generated voice to deliver questionnaire content to respondents and allows respondents to use their telephone keypads to input responses. This method allows participants to respond to sensitive questionnaire content without disclosing their answers directly to an interviewer and without the risk of inadvertent or intentional verbal disclosure to others. Responses are recorded in an electronic database without personally identifying information and the database is delivered securely from the IVR vendor to the survey operations team.

Acknowledgments

An earlier version of this publication was presented at the 2017 conference of the European Survey Research Association in Lisbon, Portugal. The authors thank Jennifer Arrieta, Gregg Peterson, and Ryan Yoder for technical guidance.

Funding

Research reported in this publication was supported by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development of the National Institutes of Health under award number R01HD052646 (PI: Sastry). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Direct correspondence to

Paula Fomby, University of Michigan, Institute for Social Research,
426 Thompson St. 1248, Ann Arbor, MI 48106-1248, USA
E-mail: pfomby@umich.edu

We describe the development and implementation of an IVR-administered questionnaire as one part of the Panel Study of Income Dynamics (PSID) 2014 Child Development Supplement (CDS-2014), a large-scale national study of children aged 0-17 years in U.S. households. While telephone audio computer-assisted self-interview (A-CASI) and IVR data collection methods have been used with small regional samples of adults (Beach et al. 2010; Cooley et al. 2000) and youth (Stritzke et al. 2005), we are aware of no other national study that has used IVR technology to collect information on sensitive topics from adolescents. Below we describe the design, protocols, and implementation of IVR data collection in this context, discuss participant cooperation rates and data quality, and offer lessons learned and recommendations for future data collection using this mode.

Context

The U.S. Panel Study of Income Dynamics (PSID) began in 1968 with a nationally-representative sample of 4,800 U.S. families. As the world's longest-running household panel study, it is a cornerstone for empirical social science research on socioeconomic mobility, health, and status attainment. It includes data collected over 40 waves (annually until 1997 and biennially since then) from up to five generations of family members descended from original PSID householders. Immigrant sample refreshers in 1997 and 2017 combined with low rates of attrition from wave to wave have kept the sample broadly representative of the U.S. population. PSID has been directed by a research investigator team at the Institute for Social Research at the University of Michigan since its inception.

CDS-2014 is a multidisciplinary study of child development and well-being embedded in PSID. The sample includes all children aged 0-17 years who resided in a household that completed the 2013 PSID Core interview and their primary caregivers, usually a child's mother (N=4,333 children in 2,517 households, 88% response rate). Study content includes information on children's family, neighborhood, and school contexts and on their cognitive, emotional, behavioral, and social development. Data were collected primarily through telephone interviews with primary caregivers and adolescents aged 12-17 years. In addition, a random 50 percent of households were selected to receive an in-home visit to collect information that could not be obtained reliably by telephone. The in-home component included cognitive achievement assessments for children and primary caregivers, children's time diaries for a randomly-assigned weekday and weekend day, interviewer observations, and interviews with children aged 8-11 years. Data collection occurred between November 2014 and April 2015 and between November 2015 and February 2016. CDS-2014 builds upon the original PSID Child Development Supplement, which began in 1997 to collect information on up to two children aged

0-12 years per household. Where CDS-2014 included home visits with a random half of participating families, the original CDS included home visits with all families. During these visits children completed computer-assisted personal interviews (CAPI) and A-CASI interviews on sensitive topics.

The Choice to Use IVR Technology

The CDS survey interview with adolescents includes sensitive questions on bullying, physical development, sexual activity, drug and alcohol use, and delinquent behavior. Data collected on these topics via A-CASI in the original CDS have been used widely in research spanning a variety of disciplines including economics, criminology, psychology, and epidemiology (Agnew et al. 2008; deBlois & Kubzansky 2016; Neymotin & Downing-Matibag 2013; Wen & Shenassa 2012).

Given broad public interest in these topics and the demonstrated value of related CDS data to the research and policy communities, the study's investigator team was committed to retaining the related questionnaire content in CDS-2014. However, the shift to telephone interviewing with adolescents required converting the A-CASI instrument used in the original CDS to a different mode of data collection. Criteria for selecting another mode included protection of respondent privacy; minimizing disclosure risk, social desirability bias, and respondent burden; and consistency with the A-CASI mode of administration in order to minimize mode effects. Options including a mail-out/mail-back questionnaire and a web-based instrument were discarded because no mechanisms were available to ensure respondent privacy or confidentiality or to authenticate a respondent's identity prior to administration. For example, another person in an adolescent's household could intercept a paper questionnaire or observe questionnaire content on a computer screen during a web-based interview.

In contrast, IVR technology minimizes the potential for interference or intervention. Survey questions are read by a pre-recorded or computer-generated voice and respondents enter responses on their telephone keypad, thus limiting the potential for others to hear or read interview content. Because no interviewer involvement is required to record responses or to transmit data to the IVR service provider, the risk of social desirability bias is also substantially reduced compared to an interviewer-administered questionnaire. Further, the IVR instrument may be programmed to require login credentials provided only to respondents, thereby reducing opportunity for another household member to intervene and complete the interview in place of the targeted respondent. Beyond these gains, IVR was a relatively inexpensive mode of data collection compared to the costs of paper questionnaire production, postal service, web programming, or field interviewer time.

IVR Questionnaire Development

The IVR instrument was adapted from the A-CASI instrument used in previous rounds of CDS. In the A-CASI administration, respondents listened to question wording and response categories through headphones and were able to read the questionnaire content on a laptop computer screen at the same time. Because interviews were done during a household visit, an interviewer was always present to ensure that the respondent completed the A-CASI task in private without interference. In contrast, the IVR administration was prepared with the expectation that respondents would only hear questionnaire content and would have no visual cues to prompt their progress through the instrument. (In advance of the interview, respondents received a printed booklet that contained response categories to each item in the questionnaire, but the booklet did not include question wording and respondents were not required to have the response booklet on hand to complete the IVR interview.) The shift to a new mode of administration required modifications to the presentation of content, strategies to allow respondents to skip items they did not wish to answer, and methods to train respondents in how to use the instrument. We review these modifications here.

IVR technology allows response entry using the keypad on a conventional landline telephone or on a cellular telephone or smartphone. In advance of the interview, CDS-2014 respondents received an inexpensive set of earbuds so that those using a cellular telephone could hear the interview questions and see the keypad at the same time. To ensure that respondents understood question intention, the programmed voice stressed the most salient words in each item, and at the outset of the interview, respondents were instructed to use their keypad to have any question repeated. For standalone questions and the first in any series of questions that used the same response set, the entire question and all response categories were presented before the respondent could enter a response on their telephone keypad. Higher-order items in a series required the respondent to hear the complete question wording but permitted response entry before the complete response set was presented. For all items, the response set repeated after three seconds if no response or an out of range response was entered. Respondents were permitted to skip over any item after the question and response set were repeated once, and the instrument automatically skipped to the next item if no response was entered after the response set was presented three times. As in the earlier A-CASI administration, “do not know” was not permitted as a valid response.

To train respondents to interact with the IVR instrument, three practice questions were included at the beginning of the interview. These items asked respondents to report their gender and age and whether their response booklet was available. Three questions at the end of the interview assessed the respondent’s perception of task difficulty. Questionnaire content is available at https://psidonline.isr.umich.edu/cds/questionnaires/cds-14/child.pdf.

All questionnaire content and protocols were developed by the research investigator team. A commercial service provider programmed the instrument, hosted the toll-free telephone line and secure server for data collection, and transferred content data files to Survey Research Operations at University of Michigan twice each week during the fieldwork period. The cost per eligible case for these services was approximately $9. The service provider had no identifying information about or means to contact respondents.

Protocol

Adolescents' eligibility to participate in the CDS-2014 interview required informed consent and assent from, respectively, an adolescent respondent's primary caregiver and the adolescent. Eligibility for the IVR interview further required that the adolescent first complete the interviewer-administered portion of the telephone interview (N=1,098). Three protocols to connect eligible respondents to the IVR interview were used in the course of fieldwork. At the outset, technical limitations prevented interviewers from being able to transfer respondents directly to the IVR interview.[1] Instead, interviewers provided each eligible adolescent with the toll-free telephone number to access the IVR instrument and a randomly-generated unique identifier to use as a login credential. If the respondent had not called in to connect to the IVR instrument within three days, the interviewer made a follow-up call to the adolescent's household. The interviewer provided the telephone number and unique identifier again only if speaking directly to the adolescent. Approximately 46 percent of eligible respondents (N=509) initiated the telephone call to access the interview within the first 16 weeks of fieldwork under this protocol (November 2014 to mid-February 2015).

Ten weeks before the end of the initial fieldwork period, an endgame strategy was introduced. Letters were mailed to eligible adolescent respondents who had not yet initiated the IVR interview with instructions on three ways to connect: Those who still had the telephone number and login credential could call in directly; those who no longer had the contact information could either call a centralized survey lab at the University of Michigan to be transferred directly to the IVR interview; or the respondent could await a call from the survey lab to connect them. Respondents were offered a $10 conditional incentive for their participation. This incentive was

[1] Decentralized interview staff conducted telephone interviews from their own homes on personal telephone lines that were not equipped to accommodate call transfers. The cost to transition to dedicated business lines for the purpose of enabling call transfers was prohibitive.

offered in addition to the $25 incentive already provided upon completion of the interviewer-administered portion of the telephone interview. Approximately 29% of adolescents who had not responded prior to the endgame initiated the IVR interview before the initial fieldwork period ended under this protocol (N=172, 15.7% of all eligible adolescents, mid-February-April 2015).

A four-month fieldwork extension began in November 2015 to contact IVR nonrespondents in order to increase response rates, sample size, and population representativeness. The protocol for the IVR interview included a letter mailed to the adolescent respondent that contained the toll-free telephone number and login credential, a $5 cash pre-payment for participation, and the offer of a $20 conditional incentive. Field staff followed up with reminder calls to households where adolescents did not initiate a telephone call within one week of the mailing. Approximately 40% of eligible adolescents who had not previously initiated the IVR interview did so during this fieldwork extension period (N=191, 17.4% overall). We investigated whether interviews completed during the fieldwork extension period (among respondents with a higher nonresponse propensity) displayed worse data quality (Fricker & Tourangeau 2010). No statistically significant differences were found in item nonresponse rates or in perceived burden between late responders and participants who completed the IVR interview during the main data collection period, although average administration time was about two minutes shorter during the fieldwork extension (p<.05).

In total, 872 respondents (79.4%) connected to the IVR system to begin the interview. (See Table 1) A slightly smaller fraction provided complete or partial data, as we describe below. The endgame strategy and fieldwork extension period increased sample size and improved the race/ethnic representativeness of the sample. Latino adolescents were more likely than their non-Latino white and black peers to participate in the endgame. This may be due in part to a later fieldwork start date for families with Spanish-speaking caregivers, which meant Latino adolescents were more likely to complete the interviewer-administered portion of the interview during the endgame period. Non-Latino black and Latino adolescents were also somewhat more likely to initiate the IVR interview during the fieldwork extension period rather than during the initial period compared to non-Latino white youth. The gender and age distribution of respondents was similar across the three periods.

Cooperation Rates

Of the 872 adolescents who initiated the IVR interview, 802 completed it, 30 provided partial data, and 40 broke off during the interview introduction or practice questions. The overall cooperation rate combining partial and complete data was

Table 1 Interactive voice response (IVR) interview, PSID 2014 Child Development Supplement, adolescents aged 12-17 years

Final status	N	Percentage
Began interview	872	79.4%
Completed interview	*802*	*73.0%*
Partial interview	*30*	*2.7%*
Breakoff	*40*	*3.6%*
No contact	226	20.6%
Total	1098	

75.8% ([802+30]/1098). Table 2 characterizes the eligible sample overall and by IVR interview outcome, comparing the subset of adolescents who did not call in or who broke off early in the interview to those who provided complete or partial data. Characteristics are weighted to be representative of U.S. adolescents aged 12 and older who were born in 1997 or later and whose families had resided in the U.S. at least since that year. IVR participants were similar to the full sample on child and caregiver age, child gender, family size, and educational attainment of the household head. Non-Latino black and Latino adolescents and youth in households with lower family income were over-represented among nonparticipants in the IVR interview.

Table 3 summarizes results from a random effects logistic regression model estimating the log-odds of adolescent non-cooperation in the IVR interview as a function of the characteristics presented in Table 2. The random effects model is clustered on the household identifier in order to estimate the share of variance in the probability of non-cooperation that is attributable to differences between compared to within households. Adjusting for other sociodemographic characteristics, the log-odds of non-cooperation was similar by adolescent age and gender, sample origin, geographical region, and household composition. Adolescents in households where the head had some college or a Bachelor's degree or higher were more likely to participate compared to those in households where the head had a high school education. Latino adolescents and those living in households with family income in the bottom quartile were more likely not to participate compared, respectively, to non-Latino white adolescents and peers with family income in the top quartile. Although these individual coefficients were statistically significant, the full sets of coefficients associated with categorical variables were not jointly statistically significant for any of the multi-category covariates. Coresident siblings' log-odds of

Table 2 Descriptive statistics, adolescents aged 12-17 years eligible to complete the PSID 2014 Child Development Supplement IVR interview overall and by interview outcome

			IVR interview outcome				
	Eligible sample		Partial or complete data		No contact or breakoff		
	Mean	SD	Mean	SD	Mean	SD	
Child characteristics							
Age in years	14.423	1.648	14.476	1.669	14.231	1.558	
Male	0.506	0.500	0.493	0.500	0.556	0.498	
Race/ethnicity							
Non-Latino white	0.580	0.494	0.634	0.482	0.386	0.488	*
Non-Latino black	0.157	0.364	0.145	0.352	0.202	0.402	*
Latino any race	0.223	0.416	0.186	0.389	0.356	0.480	*
Other race	0.036	0.187	0.034	0.182	0.044	0.205	*
Race/ethnicity unknown	0.003	0.057	0.001	0.028	0.012	0.110	
Family characteristics							
Sample source							
1968 general population	0.747	0.435	0.785	0.411	0.611	0.489	*
1968 low-income oversample	0.077	0.267	0.068	0.251	0.112	0.316	*
1997 immigrant refresher	0.175	0.381	0.147	0.355	0.277	0.448	*
Region of the United States							
Northeast	0.135	0.342	0.142	0.349	0.110	0.314	
North Central	0.260	0.439	0.281	0.450	0.182	0.386	*
South	0.375	0.484	0.366	0.482	0.407	0.492	
West	0.231	0.421	0.211	0.408	0.301	0.459	*
Metropolitan area	0.743	0.437	0.736	0.441	0.769	0.422	
Family income in 2012							
Bottom quartile	0.153	0.361	0.125	0.331	0.254	0.436	*
2nd quartile	0.241	0.428	0.215	0.411	0.336	0.473	*
Third quartile	0.253	0.435	0.271	0.445	0.188	0.392	*
Top quartile	0.352	0.478	0.388	0.488	0.221	0.416	*
No. of children in household (topcoded at 5)	2.404	1.141	2.384	1.137	2.476	1.153	
Two parents in household (biological, adoptive, or step)	0.608	0.488	0.636	0.481	0.505	0.501	*
Household head employed	0.822	0.383	0.851	0.356	0.717	0.451	*

Table 2 continued			IVR interview outcome				
	Eligible sample		Partial or complete data		No contact or breakoff		
	Mean	SD	Mean	SD	Mean	SD	
Household head age							
29 years or younger	0.031	0.174	0.027	0.163	0.046	0.209	
30-45 years	0.599	0.490	0.585	0.493	0.648	0.478	
46 years or older	0.370	0.483	0.388	0.487	0.306	0.462	
Household head education							
<12 years	0.144	0.351	0.130	0.337	0.191	0.394	
High school graduate	0.292	0.455	0.274	0.446	0.357	0.480	
Some college	0.255	0.436	0.254	0.436	0.258	0.438	
Bachelor‘s degree or higher	0.302	0.459	0.339	0.474	0.171	0.377	*
Unknown	0.007	0.082	0.002	0.046	0.024	0.152	
N	1098		832		266		

*p<.05

Table 3 Random effects logistic regression estimates of the log-odds of IVR interview nonparticipation, PSID 2014 Child Development Supplement

	B	SE	
Child characteristics			
Age in years	-0.192	0.121	
Male (vs. female)	0.365	0.395	
Race/ethnicity (vs. non-Latino white)			
Non-Latino black	1.262	0.841	
Latino any race	2.264	0.971	*
Other race	1.292	1.355	
Race/ethnicity unknown	5.346	3.534	
Family characteristics			
Sample source (vs. 1968 general population)			
1968 low-income oversample	0.592	0.808	
1997 immigrant refresher	0.437	1.143	
Region of the United States (vs. West)			
Northeast	-0.180	0.999	
North Central	-1.203	0.799	
South	-0.574	0.747	
Metropolitan area	0.195	0.595	

Table 3 continued	B	SE	
Family income in 2012 (vs. top quartile)			
Bottom quartile	1.984	0.907	*
2nd quartile	1.454	0.799	
Third quartile	0.471	0.742	
No. of children in household (topcoded at 5)	0.142	0.235	
Two parents in household (biological, adoptive, or step)	-0.211	0.554	
Household head employed	-0.137	0.626	
Household head age (vs. 30-45 years)			
29 years or younger	0.485	1.179	
46 years or older	-0.177	0.585	
Household head education (vs. high school graduate)			
<12 years	-1.031	0.827	
Some college	-1.285	0.595	*
Bachelor‘s degree or higher	-1.959	0.729	*
Unknown	1.574	2.853	
Constant	-2.906	2.300	
Rho	0.903	0.016	
N=1,098			
k=880 (observations clustered on household identifier)			
Wald chi-square (df=24)	46.14 *		

*p<.05

participation were correlated at .90 (rho=.903), meaning that most variation in the likelihood of participation was due to differences across rather than within households. A weighted logistic regression clustered on the household identifier produced substantively similar associations.

We cannot establish definitively why Latino and lower-income adolescents were less likely to participate in the IVR interview compared to their peers, but a few explanations are plausible. Among Latinos, a higher probability of nonresponse may have resulted from the later fieldwork start for primary caregiver interviews conducted in Spanish. Adolescents in lower-income families who did not immediately complete the IVR interview might have been more difficult to reach in follow-up compared to those in higher-income families if they changed residence (Desmond et al. 2015) or contact telephone numbers more often during the study period. Variation in concerns about intrusiveness, disclosure, or social desirability associated with sensitive topics also may have contributed to sociodemographic patterning of nonresponse (Tourangeau & Yan 2007). However, we note that eligibility for the IVR interview was conditional on completion of several other study components, suggesting that the eligible sample overall might have been more open to an interview on sensitive topics compared to samples selected unconditionally.

Partial Data

Cases with partial data are those that advanced beyond the first three practice questions but did not reach the end of the interview. Among those who provided partial data, breakoff points varied; that is, it did not appear that interview length or any single questionnaire item disproportionately increased the risk of breakoff. The share of cases providing only partial data (4 percent of those with any data) is low despite two circumstances. First, because of a programming limitation, respondents who terminated the IVR interview early were required to start from the beginning when they called in to resume the interview, thus increasing respondent burden. Second, the case management system flagged respondents who had not yet initiated their IVR interviews for interviewer follow-up but did not flag IVR interviews that contained only partial data, so interviewers did not recontact adolescents who terminated the IVR interview early. Nevertheless, approximately 14% of respondents who eventually provided a complete interview terminated the interview early at least once and re-entered the system to complete the interview from the beginning. Adolescents might have done so in order to overwrite their initial responses to sensitive questions, for example, to change their reported history of sexual activity or substance use. However, a review of the partial and completed interview records demonstrated that responses were consistent across administrations.

A number of issues potentially contributed to breakoffs among the 40 respondents (3.6% of total) who terminated the interview before advancing beyond the practice questions. First, the interview script for the second practice item prompted respondents to enter the *pound* (#) sign after entering their age in years. Adolescents who recognized this symbol as a *hash* sign and who were unfamiliar with the term “pound sign” might have been uncertain about how to proceed. Second, the third practice item asked whether respondents had their response booklet on hand for the interview but did not state that the booklet was not required to proceed. Respondents who had disposed of or misplaced the response booklet might have interpreted this question to mean that they would be unable to complete the interview. Third, respondents who found the IVR interview experience cumbersome or dull might have decided to terminate the interview near the outset, particularly if they were aware that they had already qualified to receive the incentive for participation.

Respondent Burden

The IVR instrument included a total of 94 items with some path-dependent content. Adolescents who completed the interview responded to 51.3 questionnaire items on average, and the average administration time was 18.7 minutes. The oldest adolescents required three minutes more to complete the interview compared to the youngest (20.9 minutes for 17-year-old respondents vs. 17.8 minutes for 12-year-old respondents) and were presented with 7.4 more items (55.5 items vs. 48.1 items respectively) on average.

Rules for response entry on the telephone keypad introduced a source of respondent burden beyond interview length. Where items required a single-digit response, participants advanced to the next questionnaire item upon keying in a response value. Where items allowed or required a response with two digits or more (e.g., age or year), respondents were asked to use the pound (hash) sign as a delimiter to indicate when the entry was complete. The opportunity to enter multiple digits introduced more room for error in any response compared to single-digit coding schemes, and the requirement to enter the pound sign added the potential for confusion about how to advance to the next item. Nevertheless, items with multiple-digit response options did not have higher rates of nonresponse or subsequent breakoff compared to questionnaire items with single-digit responses. When asked about perceived burden at the end of the interview, 94 percent of respondents with complete data reported that they had answered questions carefully and accurately, and 93 percent reported that the IVR questionnaire was either easy or “neither difficult nor easy” to complete.

Data Quality and Social Desirability Bias

Non-random variation in three study participant behaviors potentially threatens data quality: survey nonresponse, item nonresponse, and inaccurate reporting. To the extent that social desirability bias increases the risk that respondents evade or provide misleading responses on questions pertaining to sensitive topics, the CDS-2014 IVR interview may be particularly susceptible to compromised data quality.

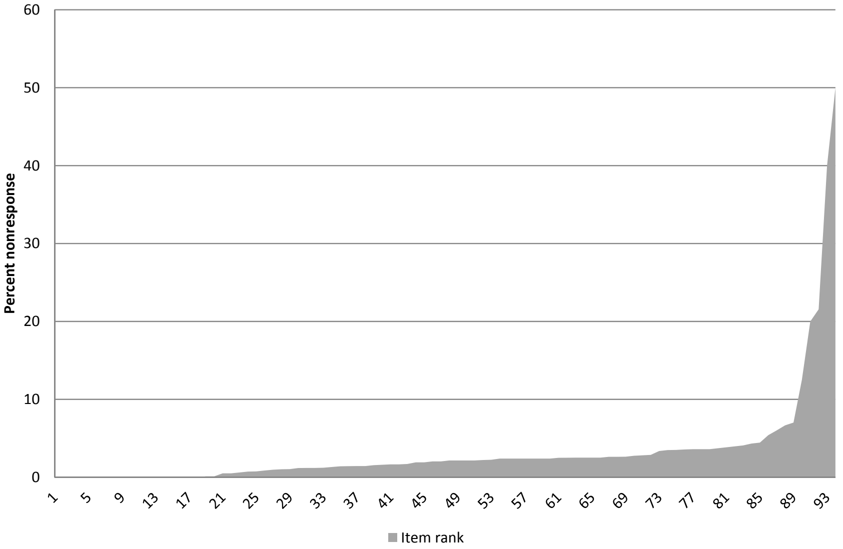
With regard to survey nonresponse, approximately one-quarter of eligible adolescents either did not initiate the IVR interview or did not advance beyond the practice questions. We do not know the reason some adolescents declined to participate. Certainly, the study’s protocol requiring the adolescent to call in to initiate the interview likely reduced willingness to participate regardless of interview content. Beyond that, a subset of adolescents or their parents may have been discouraged from participating after learning about the sensitive content during the informed consent process. As Table 2 demonstrates, nonresponse was not random:

racial and ethnic minority youth, younger adolescents, and those from socioeconomically disadvantaged families were less likely to respond than their peers. To the extent that nonrespondents differ from participants on the attributes measured by the IVR questionnaire, survey results are not fully representative of the target population, but sample selectivity could be reduced by constructing and applying non-response weights to the IVR component of the study based on characteristics obtained in other parts of the study.

Figure 1 presents the distribution of item nonresponse across the 94 items included in the IVR questionnaire ranked from the lowest to the highest nonresponse rate. Item nonresponse is measured as the share of respondents to whom a questionnaire item was administered who did not provide a valid response. It excludes individuals who were skipped out of the item or who terminated the interview before reaching it. Three-quarters of items had nonresponse rates below three percent, and all but five items had nonresponse rates below 10 percent. Items with the highest nonresponse rates (40 to 50 percent) pertained to the circumstances surrounding a live birth (birth complications, placement for adoption) that were reported by the small set of adolescents (10 or fewer) who had this experience. Nonresponse rates were also high in response to questions about the calendar month and year of menarche and first sexual intercourse, but most respondents subsequently reported their age at these events instead. For example, 23 percent of girls who had reached menarche did not report the calendar date of the event (N=83), but 94 percent of those respondents (N=78) reported age at onset in the follow-up question.

Lastly, we compare weighted distributions on key variables to two other data sources. There is no gold standard for prevalence of sensitive behaviors, and distributions vary across studies as a function of sample design, mode of data collection, and change over historical time. Nevertheless, to the extent that results are roughly consistent across studies, we may conclude that CDS-2014 captured reports on similar constructs. We compare reported age at sexual initiation for CDS-2014 respondents aged 15 to 17 to reports from the 2013-15 U.S. National Survey of Family Growth (NSFG) (National Center for Health Statistics 2016) and reports on lifetime smoking behavior for all adolescents compared to the 2015 wave of Monitoring the Future (MTF) (Miech et al. 2017). These studies are frequently cited as high-quality surveys of U.S. adolescent sexual behavior and substance use respectively. Information on sexual activity is collected in NSFG via an in-person interview.

Figure 1 Distribution of item nonresponse, PSID 2014 Child Development Supplement interactive voice response interview, complete and partial cases, N=832



Information on substance use is collected in MTF via an in-school administration of a paper-and-pencil interview.[2]

Table 4 summarizes these comparisons. In all cases, the confidence intervals from the CDS-2014 sample include the population estimates reported from external data sources. The prevalence of sexual initiation and average age at first sexual intercourse reported in CDS-2014 among adolescents aged 15-17 years is comparable to estimates provided by NSFG respondents in the same age group. The prevalence of reported sexual initiation among boys is 4.3 percentage points lower in CDS-2014 compared to NSFG, while among girls it is 1.6 percent points higher. Average reported age at sexual initiation is 0.14 years higher for boys and 0.28 years lower for girls in CDS-2014 compared to NSFG. Lifetime prevalence of reported

[2] Beyond differences in mode of administration, the multiconditional nature of unit nonresponse probability in CDS-2014 (i.e., the requirement that participants and their families complete various interview components in order to reach the IVR interview) may yield a sample selected on characteristics that are more difficult to adjust for in probability weights compared to the unconditional cross-sectional MTF and NSFG samples, potentially contributing to divergent weighted population estimates. Further, those studies differ from CDS-2014 in their broad questionnaire content and in their sampling frames, which include foreign-born adolescents and adolescents with foreign-born parents who entered the United States since 1997.

Table 4 Prevalence of sexual activity and substance use, CDS-2014 and national comparison surveys (weighted estimates with 95% confidence intervals in italics)

	CDS-2014	National Survey of Family Growth (2013-15), ages 15-17	Monitoring the Future (2015)
R ever had sexual intercourse			
Male (15-17 years) (N=201)	24.2% *(16.8%-31.6%)*	28.5%	
Female (15-17 years) (N=176)	27.9% *(19.7%-36.1%)*	26.3%	
Age in years at first intercourse			
Male (15-17 years) (N=68)	14.67 *(14.02-15.33)*	14.53	
Female (15-17 years) (N=60)	14.62 *(14.26-14.99)*	14.90	
R ever tried smoking a cigarette			
9th grade or lower (N=495)	10.4% *(6.9%-13.8%)*		13.3% (8th grade)
10th-11th grade (N=258)	20.8% *(14.7%-26.9%)*		19.9% (10th grade)
12th grade or higher (N=58)	29.4% *(15.1%-43.8%)*		31.1% (12th grade)

cigarette smoking in CDS-2014 roughly aligns with estimates from Monitoring the Future for students in grades 8, 10, and 12, although estimates are somewhat lower for the youngest and oldest adolescents in CDS-2014. We conclude that population estimates based on data collected in the IVR interview are comparable to estimates generated from similar samples interviewed using other modes of data collection.

Research Ethics

Protection of respondent privacy and confidentiality and strategies to minimize the risk of deductive disclosure are paramount in any study of children, who are considered a vulnerable population in human subjects research. In a supplemental study derived from a genealogical sample design like CDS-2014 in the context of the Panel Study of Income Dynamics, these concerns are further heightened because family members are likely aware of children's participation and may seek to find their responses to sensitive items once the data are publicly released. We adopted a variety of strategies to address these concerns.

Protection of privacy and confidentiality, especially from parents and siblings, drove the choice to adopt IVR technology to administer sensitive questionnaire content in the context of a telephone interview with adolescents. Further, login credentials provided directly to the adolescent were developed to preserve respondent fidelity and prevent any tampering or intervention. While parents were allowed to inquire about the content of the questionnaire (an option few actually exercised), no one was permitted to access the child's survey responses.

Three strategies protect respondent confidentiality after data collection. First, all data transfer and storage policies comply with standards developed by Panel Study of Income Dynamics staff and approved by the University of Michigan Institutional Review Board. Second, a Certificate of Confidentiality issued by the U.S. Department of Health and Human Services prior to the start of fieldwork protects the study investigators from being compelled through a legal proceeding to provide individually-identifying information about a respondent. Third, data on sensitive topics collected through the IVR interview are made available to researchers to use only in a secure data enclave under terms of a restricted-use data agreement. (Details available at https://simba.isr.umich.edu/restricted/ChildReportSensitive.aspx.)

Lessons Learned and Recommendations

CDS-2014 is the first large-scale national study to collect information on sensitive topics from adolescents using interactive voice response technology. The preceding review demonstrates that IVR is a cost-efficient and flexible method of data collection that yields high survey response rates and low item nonresponse rates with distributions on key variables that are comparable to other national studies. We conclude with an assessment of lessons learned and recommendations based on the CDS-2014 fieldwork experience.

IVR provides an interview context that is expected to reduce measurement error arising from social desirability bias and to increase item response rates compared to data collection methods that are perceived to be less anonymous (Sakshaug et al. 2010). To the extent that such gains were achieved in CDS-2014, the tradeoff was a decline in survey response rates compared to the CATI interview that immediately preceded the IVR interview which occurred at least in part because of technical limitations in the capacity to transfer respondents to the IVR telephone line directly. Substantial field staff time and resources were invested in a variety of strategies to follow up with and engage respondents to complete the interview. Ultimately, this additional effort paid off, as the weighted samples from the main child interview and the IVR interview are substantively similar on key sociodemographic characteristics. Nevertheless, a primary recommendation for future CATI-based data collection efforts supplemented by IVR technology is to have in place a mechanism to transfer respondents directly from one interview mode to another. For respondent protection, this mechanism should require the interviewer or the participant to provide unique login credentials in order to launch the interview. Even under optimal transfer conditions, some respondents will choose to break off or will be lost during the transfer. Depending on the design of IVR implementation, such costs can be weighed against the gains from achieved data quality in subsequent analysis and evaluation.

Other recommendations pertain to the IVR instrument itself. First, instructions should be developed with the assumption that the respondent will have no written material on hand as an additional learning support (even if such materials are provided in advance of the interview), and instructions should be evaluated for clarity prior to fieldwork. Second, vocabulary used in instructions should be familiar to respondents. For example, in the case of CDS-2014, adolescents recognized the symbol # as a "hash" sign rather than as a "pound" sign. Third, to balance consistency in the administration of questionnaire items against respondent burden, the programmed instrument should require the respondent to hear the complete question and set of response options on the first administration of an item or at the beginning of a set of related items, and then allow flexibility in the presentation of response options so that respondents may key over the repeated full set when they know how they wish to respond. Finally, minimize the number of keypad strokes required by the respondent. In the case of CDS-2014, single-digit response categories worked best.

To summarize, IVR interviewing carries some tradeoffs compared to other modes of data collection and requires substantial forethought and planning to maximize survey response rates and minimize respondent burden and error. To the extent that these costs are counterbalanced by complete data and diminished social desirability bias among respondents, IVR interviewing can provide an effective method to collect high-quality data on sensitive topics with adolescents.

References

Agnew, R., Matthews, S. K., Bucher, J., Welcher, A. N., & Keyes, C. (2008). Socioeconomic Status, Economic Problems, and Delinquency. *Youth & Society, 40*(2), 159-181. doi: 10.1177/0044118X08318119

Beach, S. R., Schulz, R., Degenholtz, H. B., Castle, N. G., Rosen, J., Fox, A. R., & Morycz, R. K. (2010). Using audio computer-assisted self-interviewing and interactive voice response to measure elder mistreatment in older adults: Feasibility and effects on prevalence estimates. [Article]. *Journal of Official Statistics, 26*(3), 507-533.

Cooley, P. C., Miller, H. G., Gribble, J. N., & Turner, C. F. (2000). Automating telephone surveys: using T-ACASI to obtain data on sensitive topics. *Computers in Human Behavior, 16*(1), 1-11. doi: http://dx.doi.org/10.1016/S0747-5632(99)00048-5

Corkrey, R., & Parkinson, L. (2002). Interactive voice response: Review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers, 34*(3), 342-353. doi: 10.3758/BF03195462

deBlois, M. E., & Kubzansky, L. D. (2016). Childhood self-regulatory skills predict adolescent smoking behavior. *Psychology, Health & Medicine, 21*(2), 138-151. doi: 10.1080/13548506.2015.1077261

Desmond, M., Gershenson, C., & Kiviat, B. (2015). Forced Relocation and Residential Instability among Urban Renters. *Social Service Review, 89*(2), 227-262. doi: 10.1086/681091

Fricker, S., & Tourangeau, R. (2010). Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys. *Public Opinion Quarterly, 74*(5), 934-955. doi: 10.1093/poq/nfq064

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly, 72*(5), 847-865. doi: 10.1093/poq/nfn063

Midanik, L. T., & Greenfield, T. K. (2010). Reports of alcohol-related problems and alcohol dependence for demographic subgroups using interactive voice response versus telephone surveys: The 2005 US National Alcohol Survey. *Drug and Alcohol Review, 29*(4), 392-398. doi: 10.1111/j.1465-3362.2009.00161.x

Miech, R. A., Schulenberg, J. E., Johnston, L. D., Bachman, J. G., O'Malley, P. M., & Patrick, M. E. (2017). National Adolescent Drug Trends in 2017: Findings Released Retrieved February 27, 2018, from http://www.monitoringthefuture.org/data/17data/17drtbl1.pdf

National Center for Health Statistics. (2016). 2013-2015 National Survey of Family Growth Public Use Data and Documentation. Hyattsville, MD: CDC National Center for Health Statistics.

Neymotin, F., & Downing-Matibag, T. M. (2013). Religiosity and adolescents' involvement with both drugs and sex. *Journal of religion and health, 52*(2), 550-569.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Reynolds, C. R., & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6*(2), 271-280. doi: 10.1007/BF00919131

Sakshaug, J. W., Yan, T., & Tourangeau, R. (2010). Nonresponse Error, Measurement Error, And Mode Of Data Collection: Tradeoffs in a Multi-mode Survey of Sensitive and Nonsensitive Items. *Public Opinion Quarterly, 74*(5), 907-933. doi: 10.1093/poq/nfq057

Schober, M. F., Conrad, F. G., Antoun, C., Ehlen, P., Fail, S., Hupp, A. L., . . . Zhang, C. (2015). Precision and Disclosure in Text and Voice Interviews on Smartphones. *PloS one, 10*(6), e0128337. Retrieved from http://europepmc.org/abstract/MED/26060991http://europepmc.org/articles/PMC4465184?pdf=renderhttp://europepmc.org/articles/PMC4465184https://doi.org/10.1371/journal.pone.0128337 doi:10.1371/journal.pone.0128337

Stritzke, W. G. K., Dandy, J., Durkin, K., & Houghton, S. (2005). Use of interactive voice response (IVR) technology in health research with children. *Behavior Research Methods, 37*(1), 119-126. doi: 10.3758/BF03206405

Tourangeau, R., Steiger, D. M., & Wilson, D. (2002). Self-Administered Questions by Telephone: Evaluating Interactive Voice Response. *The Public Opinion Quarterly, 66*(2), 265-278.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883.

Wen, X., & Shenassa, E. D. (2012). Interaction Between Parenting and Neighborhood Quality on the Risk of Adolescent Regular Smoking. *Nicotine & Tobacco Research, 14*(3), 313-322. doi: 10.1093/ntr/ntr215

methods, data, analyses | Vol. 13(1), 2019, pp. 111-137 DOI: 10.12758/mda.2018.09

The Longitudinal Item Count Technique: A New Technique for Asking Sensitive Questions in Surveys

Alessandra Gaia [1,2] & *Tarek Al Baghal* [2]

[1] *University of Milan-Bicocca*

[2] *Institute for Social and Economic Research, University of Essex*

Abstract

Asking respondents sensitive questions directly may lead to socially desirable responding. As alternative, some have proposed using the Item Count Technique (ICT). The problem with ICT methods is that these can have low statistical efficiency, but also do not provide an indicator of the behavior at the respondent level. We propose a new variant of the ICT to overcome these issues: the Longitudinal Item Count Technique (LICT). Instead of administering different lists (one including the sensitive item and one without) to two random groups in a single survey, the LICT administers both lists to each respondent, but at different survey waves. The sensitive attribute can be estimated as the difference within individuals across waves. Like the ICT, the LICT can be extended to a two-list version. In this paper we discuss the assumptions, implementation, limitations, and ethical implications of this novel technique, and present application of the method in the *Understanding Society* Innovation Panel, estimating the prevalence of the gay, lesbian, and bisexual population in the United Kingdom. In this first application, the LICT in some ways appeared to provide better estimates than the traditional ICT, but also provided some inconsistency in estimates. We discuss the implications of these results and point to routes for further research.

Keywords: Item Count Technique; sensitive questions; social desirability; longitudinal data; LGBT research

The Item Count Technique (ICT) – also called “Unmatched Count Technique” or “List Experiments” – is used to improve the measurement of sensitive topics, reducing social desirability bias. This promising technique protects respondents’ privacy when it works as planned, with no “ceiling” and “floor” effects (i.e. every or no item in the list applying). The ICT, introduced by Smith, Federer and Raghavarao (1974), is an indirect questioning technique to ask sensitive questions in surveys. Instead of inferring the population prevalence of a sensitive behavior by asking respondents directly whether they engaged in that behavior, using the ICT the researcher can extrapolate this information experimentally.

Specifically, in the ICT sample members are randomly divided into two groups; respondents in each group are presented with a list of items and asked to count how many items apply to them. Each group’s list is identical but for the sensitive item appearing only in one of them. Items should be selected such that it is reasonable for respondents to select some but not all items. While ICT methods produce estimates that can be useful in estimating prevalence of sensitive behaviors within subgroups and for regression analyses (see Corstange 2009; Holbrook & Krosnick 2010; Blair & Imai 2010; Imai 2011; Glynn 2013), such results from the ICT are typically imprecise due to low statistical efficiency. The lack of indicators at the respondent-level is also problematic as ICT methods do not allow analyses at the individual-level, but rather at the aggregate-level only.

To overcome these issues, we propose a variation to the ICT: the Longitudinal Item Count Technique (LICT). Instead of splitting the sample in two groups, all respondents are presented with the list which includes the sensitive item in one survey wave and the list that does not include the sensitive item in another survey wave. Since the entire sample is used, there are less concerns of statistical efficiency, as with standard ICT. Importantly, LICT methods also provide an individual-level indicator of the behavior of interest under certain circumstances, since both lists (with and without the sensitive item) are administered to each respondent. In these cases, analyses can be made directly at the individual-level, including multivari-

Acknowledgments

We are grateful to Noah Uhrig for contributing to the study design and to the development of the Longitudinal Item Count Technique. We are also thankful for comments from Peter Lynn and Annette Jäckle on the preliminary draft of this paper. Finally, we thank two anonymous reviewers and the guest editors of the special issue for their helpful comments on this manuscript.

The second author is funded by a research award from the UK Economic and Social Research Council (award no. ES/K005146/1) for “Understanding Society: the UK Household Longitudinal Study, phase 3.”

Direct correspondence to

Alessandra Gaia, Department of Sociology and Social Research,
University of Milan-Bicocca, Milan, Italy
E-mail: alessandra.gaia@unimib.it

ate methods such as regression models without the need for multiple steps such as those proposed by Imai (2011).

The circumstances where these meaningful individual-level indicators are met mostly likely occur when the items are time invariant, e.g. items that refer to past events, like where the respondents grew up ("I have grown-up in the country-side"), dates in the past which are significant to the respondents, like birthdays of significant others ("My father's birthday is in October"), *etc*. If the selected items are not time invariant (e.g. "I have travelled to Spain"), the event may occur between data collection waves. If that is the case, respondents answering the survey question accurately would report a higher number of items in the second wave compared to the first survey wave.

Time invariance in LICT methods is not always necessary, except that the LICT also rests on the assumption that there is no trend in the list items (upward or downward, across waves). If there is a trend in the list items (including the sensitive behavior) measuring differences will be confounded with change over time. As long as there is not a trend, individual respondent time variability is acceptable, although it will increase the variance of estimates. In some ways, individual time variability may be desirous in the LICT. Generally, the LICT allows researchers to identify whether the trait of interest applies to the respondent, although it provides less privacy than the ICT. In particular, if a respondent remembers the lists across waves, time invariant items may lead respondents to realize they will be reporting on the sensitive behavior by reporting higher or lower counts (depending on which list is presented at which wave).

Conversely, time variant items may allow respondents to maintain a sense of anonymity intended by ICT methods. For example, travelling to Spain may occur between waves, or a tattoo may be removed – in either case, changes to the counts are therefore not directly related to the indication of the sensitive behavior. Further, in the LICT design proposed here, respondents are divided where half of the sample receives the list with the sensitive item in an earlier wave and the list without in a later wave, while the reverse ordering occurs for the other group. To the extent that time variant behaviors do not trend and are distributed equivalently across groups over time, the averaging of estimates will tend to eliminate any bias introduced by time invariant items.

Further, for both the ICT and LICT to work properly, items should be selected to avoid "ceiling" and "floor" effects. If lists contain the non-sensitive items where all items are likely to be selected among respondents ("ceiling" effect), those with the sensitive item list would self-identify by counting all the items. There is also some concern that respondents may view themselves as self-identifying in the case where the list has items where the respondent is likely to select none ("floor" effect). However, this "floor" effect is less problematic, as it requires the assumption that the interviewer can infer that respondents with the sensitive-item list are indicating the sensitive behavior applies to them when reporting a count of one (Kuha & Jackson 2014).

While LICT methods have not yet been explored in research previously, ICT has been used to estimate sensitive behaviors across a number of disciplines. For example, disciplines like development economics or political studies often adopt the ICT (at times referring to it as "list experiments") to elicit very sensitive behaviors – e.g. vote buying in Turkey (Çarkoğlu & Aytaç 2015), voter intimidation in Russia (Frye et al. 2018) attitudes toward Female Genital Mutilation in Ethiopia (De Cao & Luz 2015) the presence of drug trafficking organizations in Mexico (Magaloni et al. 2012); and in conflict settings such as contemporary Afghanistan (Blair et al. 2014). These studies can be extended to explore these phenomena across time using LICT. The implementation of the technique in fields such as development economics is facilitated by the fact that often researchers implement small scale experiments which require observation before and after treatment, where the measures across time allow for implementation of the LICT. Given the frequent need for indicators of sensitive behaviors in many disciplines, LICT may be of particular use.

We motivate the usage of the technique in the next section through the description of a sensitive topic asked in surveys: sexual orientation. Then we describe the features of this innovative technique and the underlying assumptions, provide guidance on its implementation, discuss its limitations, and the ethical implications associated with it. We then present an empirical application of the method on the sensitive topic of sexuality. The implementation of the method is conducted using experimental data from a large scale nationally representative survey of the UK population, the Innovation Panel of *Understanding Society*: the UK Household Longitudinal Study. Three sets of estimates are compared using this experimental data: first, standard direct questions frequently asked in surveys to measure sexual identity; second, we explore ICT and LICT indicators measured at two consecutive waves of the longitudinal study; third, we examine extensions of the ICT and LICT using two lists to generate estimates. We conclude with a discussion of our findings, and implications for further research.

Measuring Sensitive Questions in Surveys: Sexual Orientation

This substantive topic of analysis for the current research, sexual orientation, is chosen for both the importance and the complexity of obtaining reliable estimates in this area. Indeed, providing sound statistical information on the gay, lesbian, or bisexual populations (also called "sexual minorities") is needed to inform policy makers on disadvantage and discrimination. However, obtaining good quality data is methodologically challenging, as sexuality is one of the most sensitive topics when asked about directly in social surveys.

An additional complication is that classification of people's sexuality is complex as "sexual orientation" is a multidimensional construct involving three different dimensions: sexual attraction, sexual behavior, and self-identification (Laumann et al. 1994). "Heterosexual/homosexual/bisexual attraction" indicates whether a person is sexually attracted by someone of the same sex, of the opposite sex, or of both sexes, whereas "heterosexual/homosexual/bisexual behavior" indicates whether someone has had sexual experiences with someone of the same sex, opposite sex, or of both sexes. And sexual identity indicates self-identification into "heterosexual", "homosexual", "bisexual", or "other" sexual identities. Classification of the population could occur along any of these three dimensions (sexual attraction, behavior, and identity) or amongst any combination of them, and it is not clear which are most relevant for population estimation much less monitoring of equality (Aspinal 2009). Until now, large scale multi-purpose UK studies have measured sexual identity as self-identification into "heterosexual", "homosexual", "bisexual", or "other" sexual identities, rather than these various dimensions.

In addition to being a sensitive behavior "non-heterosexual" sexual identity, homosexual attraction and homoerotic behavior are also rare in the general population. Indeed, nationally representative surveys suggest a low prevalence of "non-heterosexual" sexual identity, homosexual attraction and homoerotic behavior in the UK. Results from the UK National Survey of Sexual Attitudes and Lifestyles III show that 3.3% of respondents identified as gay, lesbian, bisexual or other, 3.2% in the UKHLS and 1.9% self-identify as gay, lesbian or bisexual (the option "other" was not provided) in the 2013 British Social Attitudes Survey. In terms of same-sex sexual attraction and homoerotic behavior, data from the National Survey of Sexual Attitudes and Lifestyles III (2010) show that 10.6% of respondents declare being attracted by a person of the same sex and 10.5% declare having had sexual experiences with a person of the same sex.

Overall, there appears to be a low true prevalence of the behaviors of interest, which may have consequences for using methods such as the ICT and LICT. Although Ahlquist (2017) finds that the ICT does not perform well with rare behaviors, Kiewiet de Jonge and Nickerson (2013) find empirical evidence that the ICT is more effective in estimating low prevalence behaviors than high prevalence. In particular, they find that while low prevalence items do not show evidence of artificial inflation (more reports than expected), high prevalence items show a tendency toward artificial deflation (less reports than expected). Given the possibility that the ICT (and by extension LICT) may bolster the measurement of low prevalence behaviors, the sensitive nature of sexual behaviors and the complexity of measuring sexual orientation, we consider the estimation of the all three dimensions of sexual orientation (attraction, behavior, and identity), as an interesting case study for the first implementation of the LICT.

Methodology of the ICT and LICT

In the ICT, survey sample members are divided randomly into two groups, with each being provided a list for which to provide a count of items that apply to them. One list has an additional item, the sensitive behavior of interest. The mean difference in list counts across the two groups theoretically should range between 0 and 1. The result is the estimated prevalence of the sensitive behavior in the population. Formally, the estimated prevalence of the sensitive item using ICT is calculated as following:

$$\hat{p}_{ICT} = \overline{x}_{a+s} - \overline{x}_a \tag{1}$$

where:

$\overline{x}_{a+s}$ is the average number of items counted in list *a* plus the sensitive item;

$\overline{x}_a$ is the average number of items counted in list *a*.

As long as the two samples are independent the variance is the sum of the variances of each of these means, that is $Var\left(\overline{x}_{\mathrm{a+s}}\right) + Var\left(\overline{x}_{\mathrm{a}}\right)$. Since the ICT only uses half of the sample for each mean estimate, there is a loss in precision in the estimate, and the variance is larger than if the entire sample was used for each mean.

As outlined above, one alternative to solve this problem of precision, as well as provide individual-level estimates, is the LICT. Each respondent is given both lists, one with the sensitive item and one without, and asked for counts of relevant items. These lists are given in different waves, although which lists goes in the earlier wave and which list goes in the later can vary. In particular, it is recommended that the sample is divided randomly such that half gets the list without the sensitive item and half the list with the sensitive item in the earlier wave, with each group getting the other list in the later wave. This balancing allows for effects from time invariant items to potentially average out, assuming events are equally likely to occur for groups over time. The LICT then takes the differences in lists within individuals, opposed to the mean group differences of the ICT. The prevalence of the sensitive behavior is estimated as the mean of the within individuals differences for the entire sample, formally:

$$\hat{p}_{LICT} = \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}\left(x_{i,\ a+s,\ w(s)} - x_{i,\ a,\ w}\right)\right] \tag{2}$$

where:

n is the total number of respondents

$x_{i,\ a+s,\ w(s)}$ is the number of items counted in list *a* plus the sensitive item for respondent *i* at the wave with the sensitive item in the list;

$x_{i,\ a,\ w}$ is the number of items counted in list *a* for respondent *i* at the wave without the sensitive item in the list.

The variance of this estimate is based on the difference of dependent observations, hence can be expressed as

$$\begin{aligned} Var\left(\hat{p}_{LICT}\right) &= Var\left(x_{i,\ a+s,\ w(s)}\right) + Var\left(x_{i,\ a,\ w}\right) \\ &\quad -2Cov\left(x_{i,\ a+s,\ w(s)}, x_{i,\ a,\ w}\right) \end{aligned} \tag{3}$$

Where the covariance term accounts for the dependency in measures. This expression can be simplified as $Var\left(\bar{d}\right)$ where $d_i = \left(x_{i,\ a+s,\ w(s)} - x_{i,\ a,\ w}\right)$, and there is no need to compute the separate variances and the covariance. It is then possible to take the difference at the individual level and apply the standard variance estimator to the mean of these individual differences.

Both the ICT and the LICT can also be extended using two lists. The Two-List ICT has been proposed to take advantage of the full sample in a cross-sectional setting, to overcome efficiency problems (Droitcour et al. 1991, Biemer & Brown 2005). Each subsample receives one list with the extra item of interest and one short list without the item of interest (list sets *a* and *b*). As such the estimated prevalence of the sensitive item in the Two List ICT can be formalized as:

$$\hat{p}_{2ICT} = \left(\hat{p}_{s1} + \hat{p}_{s2}\right)/2 \tag{4}$$

where:

$$\hat{p}_{s1} = \bar{x}_{a+s} - \bar{x}_{a}$$

$$\hat{p}_{s2} = \bar{x}_{b+s} - \bar{x}_{b}$$

Each list sets *a* and *b* lead to an ICT estimate in the same way as in (1), but then these are averaged to take the overall sample mean. The estimated variance for the Two Lists ICT is as follows:

$$Var\left(\hat{p}_{2ICT}\right) = \left(\frac{1}{4}\right)\left(Var\left(\hat{p}_{s1}\right) + Var\left(\hat{p}_{s2}\right) + 2\rho_{s1s2}\sqrt{Var\left(\hat{p}_{s1}\right)Var\left(\hat{p}_{s2}\right)}\right) \tag{5}$$

Where ρ_{s1s2} is the correlation between the estimators of $\hat{p}_{s1}$ and $\hat{p}_{s2}$, with the expectation that this correlation is negative (Biemer & Brown 2005). The variance can also be estimated (as it is done here) using just the first two terms, i.e. $\left(\frac{1}{4}\right)\left(Var\left(\hat{p}_{s2}\right) + Var\left(\hat{p}_{s2}\right)\right)$, given the complications in estimated ρ_{s1s2} (see Biemer & Brown 2005). However, using this form of the variance will likely overestimate the true variance, as the last term in (5) is likely negative. This overestimate means a reduction in precision and wider confidence intervals, but conversely means there will be greater conservativism in significance testing.

While Two-List ICT methods improve efficiency in estimates, there is still a lack of individual indicators. Since the LICT already uses the full sample, the benefit of having Two-List LICT is that there are multiple indicators of the sensitive behavior, rather than one, which may solidify conclusions by relying on multiple rather than single data points. Like the LICT, the Two-List LICT is estimated within individuals, as all respondents receive both lists with and without the sensitive item. In one wave, respondents receive list *a* with the addition of the sensitive item, and list *b* without the additional sensitive item, and in the other wave (again the order of wave can vary), the other version of each list *a* and *b* is given. Like the Two-List ICT, the Two-List LICT prevalence can be estimated via averaging the estimated prevalence of each of the two list sets *a* and *b,*

$$\hat{p}_{2LICT} = \left(\hat{p}_{LICT(a)} + \hat{p}_{LICT(b)}\right)/2 \tag{6}$$

Where $\hat{p}_{LICT(a)}$ and $\hat{p}_{LICT(b)}$ are estimated separately via (2). The variance of the Two-List LICT then takes the form of the Two-List ICT reported in Biemer and Brown (2005)

$$\begin{aligned} Var\left(\hat{p}_{2LICT}\right) = \left(\frac{1}{4}\right)\Big(&Var\left(\hat{p}_{LICT(a)}\right) + Var\left(\hat{p}_{LICT(b)}\right) \\ &+2\rho_{LICT(a),LICT(b)}\sqrt{Var\left(\hat{p}_{LICT(a)}\right)Var\left(\hat{p}_{(LICT(b))}\right)}\Big) \end{aligned} \tag{7}$$

Where $\rho_{LICT(a),LICT(b)}$ is the correlation between the estimators of $\hat{p}_{(LICT(a))}$ and $\hat{p}_{(LICT(b))}$. Given both list sets are used for each individual, the correlation estimate is more direct, and this is the variance estimator used in the following empirical example.

Data and Methods

Data come from an experiment implemented in the *Understanding Society* Innovation Panel waves 8 and 9 (IP8 and IP9) (University of Essex 2018). *Understanding Society: the UK Household Longitudinal Study* (UKHLS) is a multidisciplinary study that focuses on a wide range of topics such as living arrangements, fertility, housing, economic activity, income, health, and political attitudes. *Understanding Society* includes an Innovation Panel (IP), a separate sample used to test methodological innovations in longitudinal surveys, in general, and *Understanding Society*, in particular. The Innovation Panel target population is adults (aged 16+) living in Great Britain. The study aim is to interview each adult member of the household and individuals are followed when they move to other parts of Great Britain. Sample members are interviewed every 12 months. The Innovation Panel mirrors *Understanding Society* in its design and it is a stratified, clustered, probability sample. Prior to the fifth wave (IP5), all interviews were conducted by interviewers, but moved to sequential mixed-mode web and CAPI design at IP5. Two-thirds of households were allocated to the mixed-mode design, while the other third were administered the standard single-mode CAPI design. In the mixed-mode treatment, if any household member did not respond to the web survey within three weeks, an interviewer was sent to attempt a face-to-face interview. A mop-up period allows respondents to complete in either web or telephone interviews, although no respondents in the sample completed via telephone. All experimental allocations used in the current study are made independent of the mixed-mode experiment (described in detail in Jäckle et al. 2017).

To ensure that results of the various measures explored are comparable, and because the analysis of interest is across lists across waves, the analytic sample is defined as those who answered all lists given across both waves. Respondents who did not answer all of the lists, including those not responding to any list within a wave or those only responding at one wave are not included in this analysis. Overall, refusal to list questions across both waves was low, ranging from 3.4% of respondents in IP8 on a question on sexual behavior to 0.5% of respondents in IP9 on a question on sexual identity. Also "don't know" answers were rare, to levels lower than 0.7% in all items and waves. Further, due to the possibility that respondents could change waves in the mixed-mode allocation, the data are further restricted to respondents answering in the same mode across wave. This restriction removes any effect that the change of mode could have on responses across waves within respondents. This analytic sample has 1370 respondents.

Experimental Design

Experimental design

The LICT in the IP was designed to measure all three dimensions of sexual orientation (attraction, behavior, and identity), using two lists for each dimension, six in total. The lists are then repeated at the subsequent survey wave to derive the longitudinal element of the ICT. Respondents were randomly allocated at IP8 to one of two conditions. Each of the two conditions received three lists without a key and three with a key item, with the two groups differing on which set of lists were received. At IP9, each group received the reverse set of lists; i.e. if the respondent received a list with the key item at IP8, that list with the same non-key items was presented at IP9 minus the key item or *vice versa*. Given two lists were used for each dimension, Two-List ICT and LICT estimates can also be made. Table 1 shows the experimental design of the LICT.

Table 1 LICT implemented at IP8 and IP9

	IP8	IP9
Group 1	List A	List A + S1
	List B + S1	List B
	List C	List C + S2
	List D + S2	List D
	List E	List E + S3
	List F + S3	List F
Group 2	List A + S1	List A
	List B	List B + S1
	List C + S2	List C
	List D	List D + S2
	List E + S3	List E
	List F	List F + S3

Note: S1 refers to being sexual attracted from someone of the same sex, S2 refers to having had homoerotic sexual experiences (sexual experiences with someone of the same sex), and S3 refers to self-identifying as gay, lesbian, or bisexual.

A basic check of whether the randomization worked tested differences across groups on age (in 7 categories), sex (male, female), marital status (single, formerly married, married), education (university/professional degree, A-level/GSCE, less education) and urbanicity (urban, rural). Generally, the randomization appears to have worked, with all comparisons across conditions not significantly different at p<0.05.

Before the sexual identity ICT questions, the respondent was presented with a brief preamble which explained what was needed for each question; that is, only the counts of behaviors relevant to them. The wording of the introduction (as well as the full question wording for each ICT question) is presented in Appendix 1. As examples, three item lists are presented below, one on sexual attraction, one on sexual behavior and one on sexual identity, each including the sensitive item of interest. After each list on the same screen, respondents were presented with the question: “How many statements are true for you?” with the options “None are true”, “One statement”, “Two statements” “Three statements”, “Four statements”, “Five statements”. Questions without the key item did not have the “Five statements” response option.

Example of item count on sexual attraction:

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

Example of item count on sexual experience:

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

Example of item count on sexual identity:

I would describe myself as **being** …

- gay, lesbian or bisexual
- stylish and fashionable
- disabled
- patient
- British

At each wave, the ordering of item counts (i.e. the different lists) was randomized across respondents, and the statements within lists were also randomized.

The wording of the ICT questions was designed with the aim of mixing non-sensitive items that were expected to be high prevalence with non-sensitive items that were expected to be low prevalence; this is consistent with the indication of the literature (see Glynn 2013). Indeed, if all items in the list are of a high prevalence, gay, lesbian, and bisexual respondents may count all items in the list, and thus self-identify themselves as gay, lesbians, and bisexuals, i.e. a "ceiling effect"; conversely, if all "non-sensitive" items are very rare (and perceived by respondents as being more rare than belonging to the gay, lesbian, and bisexual population), a "floor" effect may occur.

Therefore we combined items that we expected to be low prevalence (e.g. "I would describe myself as being disabled"), with items that we expected to be high prevalence (e.g. "I would describe myself as being British"). When items were designed, in early 2014, items: “I consider myself as being British” (list E) and “I consider myself as being European” (list F) were considered non-sensitive high prevalence items. However, the debate on the United Kingdom European Union membership (which developed in conjunction with the referendum, held on 26th June 2016) pervaded public opinion during the fieldwork for IP9 (summer 2016). This parallel timing may have increased the sensitive nature of these two items, and altered the estimating prevalence of the two items at IP9. Finally, the questions were designed so that the list of items would fit together and make sense to respondents – as suggested by Droitcour et al. (1991).

To explore the possibility of “ceiling” and “floor” effects, Figures 1 and 2 present the distribution of the items reported as true for each list which does not include the sensitive item. We focus on the extremes of the distribution (i.e. 0 and 4 true statements). In the dimensions of attraction (lists A and B) and behavior (lists C and D), the large majority (29.2% - 44.0%) of respondents, in both waves, reported that none of the items presented applied to them; conversely, in the identity questions (lists E and F) the “floor” effect was not problematic, as “none of the statements are true” was selected by only a small percentage of respondents (2.2% - 3.8%).

The evidence for “ceiling” effects is mixed. While lists A (attraction), list C (behavior) and E (identity) resulted with only a small proportion of respondents selecting that all “four statements are true”, ranging between 1.1% and 5.1%, lists B (attraction) and F (identity) respondents reporting that all four behaviors range between 16% and 20%. Similarly, while not quite as high, list D had 7.4% of respondents (in IP8) and 10.4% (in IP9) selecting four statements are true. The more limited evidence for “ceiling” effects is reassuring, as “ceiling” effects are more problematic to ICT than “floor” effects (Kuha & Jackson 2014).

In addition to the ICT, respondents were also asked a direct question on sexual identity; sample members were randomly allocated to two different protocols, which vary in question wording and in mode of administration. These two protocols are currently adopted in two large scale studies in the United Kingdom, i.e. *Understanding Society*: the UK Household Longitudinal Study (UKHLS) and the Integrated Household Survey (IHS). The protocols for the two studies are as follows:

Protocol 1 – UKHLS:

The question is asked in self-completion either by Computer Assisted Self-Interview (CASI) or by Web.

Protocol 2 – IHS:

The question is asked Face-to-Face (in Computer Assisted Personal Interview, CAPI) with the aid of a showcard

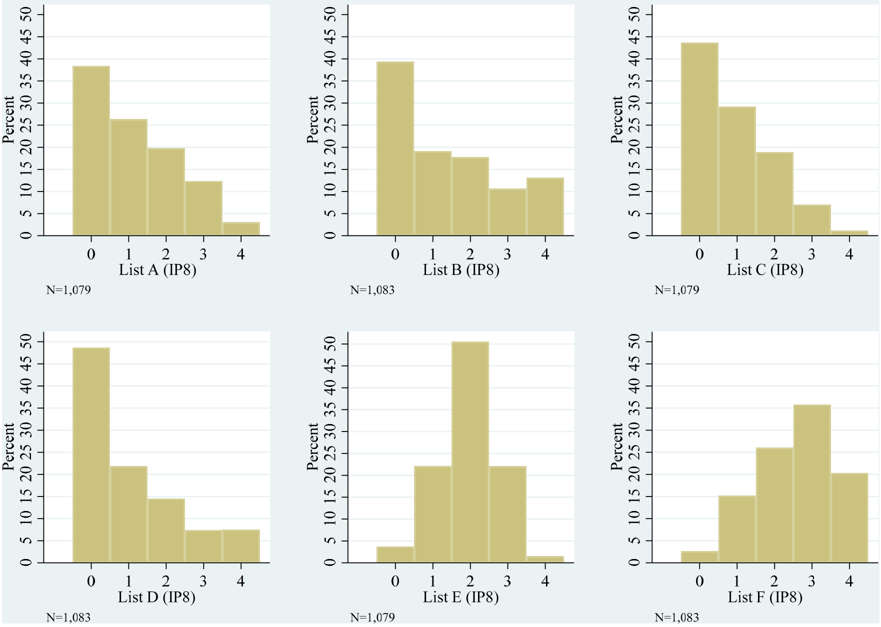


Figure 1 Distribution of Reported Items Excluding Sensitive Items, IP8

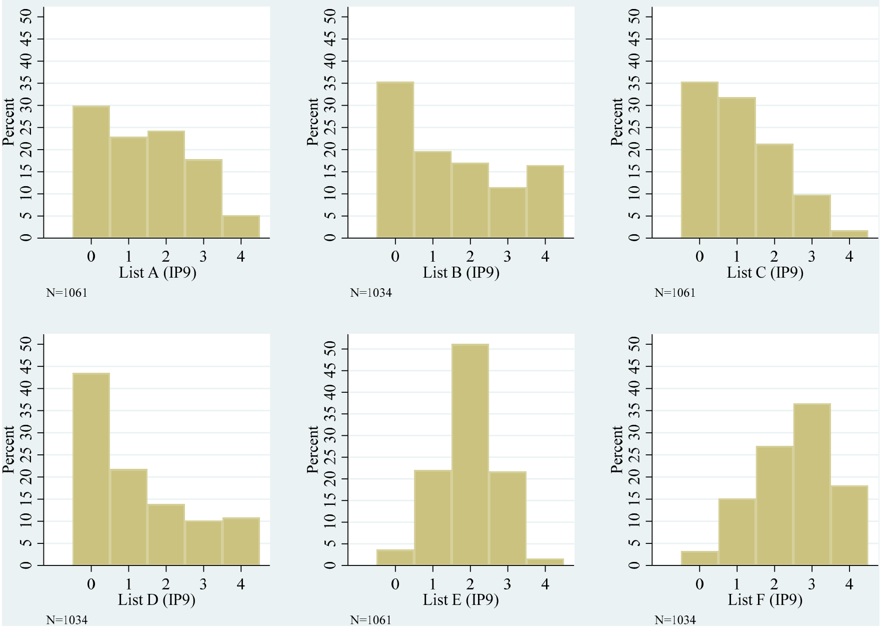


Figure 2 Distribution of Reported Items Excluding Sensitive Items, IP9

The visual design was identical in the Web and CASI versions of the UKHLS question. The question wording for the two protocols, the showcard, and the interviewer instructions are presented in Appendix 1. The ICT questions were separated from the direct sexual identity question in the questionnaire in order to avoid carry-over effects between these survey tasks.

Sample members were randomly allocated to receive either the UKHLS or IHS protocol. The experimental allocation was fully crossed with the allocation to the two lists ICT groups. Respondents were given the same protocol/question in both waves. Deviations to the experimental allocations were implemented to accommodate the mixed-mode nature of the survey design (Jäckle et al. 2017). Specifically, respondents completing the survey by Web answered the question according to the self-completion UKHLS protocol, regardless of their original allocation.

Results

Most surveys have attempted to directly measure sexual identity in questionnaires using a single question (or a small set of questions). These standard forms of questions are the basis of comparison for the Two-List LICT proposed here. Table 2 below presents the self-reported sexual identity using the three different protocols: the UKHLS Web protocol; the UKHLS face-to-face protocol using CASI; and the IHS protocol directly asked by an interviewer using a showcard. While most respondents provided a response at both waves, the UKHLS protocols, which offers an explicit "Prefer Not to Say" option and are self-administered, has more respondents refusing to respond than the IHS protocol.

In all instances, the large majority of responses indicated a heterosexual identification, with more than ninety-percent identifying so in all cases. Slightly more respondents identified as heterosexual in the IHS protocol in both waves, which was asked directly by an interviewer with a showcard. The small cell sizes for non-heterosexual responses make significance testing of the entire response distributions unreliable. However, tests of heterosexual/non-heterosexual responses (binary) show that the UKHLS CASI protocol elicited significantly less (at $p<0.05$) heterosexual responses than either the UKHLS Web ($t(1362)=-2.76$, $p<0.01$) or IHS protocol at IP8 ($t(1362)=-2.04$, $p<0.05$). At IP9, the UKHLS CASI protocol received significantly less heterosexual responses than the IHS protocol ($t(1356)=-2.22$, $p<0.05$), but is not significantly different from the UKHLS Web protocol. While not conclusive, these results are suggestive that, as expected, interviewer-administered questions may lead to more responses seen as socially desirable.

Although the above results suggest mode may reduce socially desirable reporting, it is unlikely to have entirely removed these pressures. As such, item count

Table 2 Self-reported sexual identity using direct questioning

	IP8			IP9		
	UKHLS-Web	UKHLS-CASI	IHS	UKHLS-Web	UKHLS-CASI	IHS
Heterosexual	94.9% (n=590)	91.6% (n=348)	95.1% (n=350)	93.4% (n=581)	91.6% (n=348)	94.6% (n=348)
Gay or Lesbian	1.6% (n=10)	1.6% (n=6)	0.8% (n=3)	1.9% (n=12)	1.8% (n=7)	0.8% (n=3)
Bisexual	1.9% (n=12)	1.1% (n=4)	2.2% (n=8)	1.1% (n=7)	2.4% (n=9)	1.6% (n=6)
Other	NA	1.3% (n=5)	1.1% (n=4)	1.0% (n=6)	1.3% (n=5)	0.3% (n=1)
Prefer Not to Say/ Refused	1.6% (n=10)	4.5% (n=17)	0.5% (n=2)	2.6% (n=16)	2.9% (n=11)	1.4% (n=5)
Don't Know	NA	NA	0.3% (n=1)	NA	NA	1.4% (n=5)
n	622	380	368	622	380	368

techniques may improve reporting and estimates. Table 3 presents the estimates from the IP8 and IP9 ICT, as well as the LICT using data from both waves. Standard errors for each estimate are also presented. These standard errors show that as expected, given the use of the full sample in the LICT versus half in each ICT estimate, the LICT improves efficiency over the ICT estimators. In every comparison between LICT and ICT estimates, LICT estimates have smaller standard errors.

Beyond that result, it is difficult to make other substantive conclusions. This difficulty is largely due to negative values that occur throughout the estimates. If ICT and LICT methods work, negative values should not occur, as respondents with longer lists (i.e. with the sensitive item) are expected on average to provide higher counts. This negative value indicates a negative prevalence of a sensitive behavior, and so is not interpretable. There is some evidence presented in Table 3 to suggest how this may occur.

For example, the IP8 ICT estimate for List B is negative, while at IP9 the List B estimate is positive. This result may occur if those assigned to the List B without the sensitive item at IP8 truly had more non-sensitive items to report on average than those assigned to List B + S (with the sensitive item) at IP8, particularly given the expected low prevalence of the behavior. Respondents with the higher true average without the sensitive behavior in the list could report a higher mean at one wave

Table 3 ICT and LICT estimates

Dimension	IP8 ICT	IP9 ICT	LICT
Attraction			
List A	0.12	-0.05	0.04
(S.E.)	(0.06)	(0.07)	(0.03)
List B	-0.08	0.21	0.07
(S.E.)	(0.08)	(0.08)	(0.03)
Experience			
List C	0.15	0.05	0.09
(S.E.)	(0.06)	(0.06)	(0.03)
List D	0.07	0.09	0.09
(S.E.)	(0.07)	(0.08)	(0.03)
Identity			
List E	-0.01	-0.04	-0.03
(S.E.)	(0.04)	(0.04)	(0.02)
List F	-0.20	0.02	-0.09
(S.E.)	(0.06)	(0.06)	(0.03)

(in this case List B at IP8) than those given the list with the sensitive behavior. Since these same respondents with the higher average are asked the same list with the sensitive item and the group with the lower average asked the list with only non-sensitive items, the expected difference would now be positive. Also, since the higher average respondents would also add in reports of the sensitive behavior, this average could be even larger than the negative value identified. This pattern is what occurs for List B in IP8 and List A in IP9.

This explanation may not actually be what is occurring, and does not clearly explain all of the negative values in Table 3. There are negative values for List E and List F estimates at IP8. At IP9, while the List F ICT estimate is now positive, which could fit with the above explanation, the List E estimate is still negative, and somewhat larger in absolute value. Other explanations may also explain these negative values in ICT estimates, for example various forms of measurement error, such as counting and reporting error of relevant items.

The LICT also leads to negative estimates for List E and List F, and group differences cannot explain these values in the same way, given estimates are within individuals for the entire sample. One explanation is that the items used in these lists are not necessarily time invariant as these can change within respondents. For example, a respondent could count they were healthy (in List F) in one wave, but could be feel unhealthy in the other wave. However, to the extent that changes occur equally over groups assigned to different lists at each wave, these changes should balance out and negative estimates avoided.

While these time invariant items are very much a possible explanation for these negative values in the LICT, as well as other measurement errors (e.g. counting), it should be pointed out that List E contains the item being "British" and List F has the item being "European". As noted above, the lead-up and vote for the UK to leave the European Union occurred during the IP9 fielding period, which may have affected respondents' counts of these items in a differential way than from IP8. If this was the case, which seems possible, the need to avoid a trend (i.e. an event affecting one wave differentially) in the LICT is violated. If this explanation is the case, it underscores the need to avoid items that may trend (although in this case, the possible trend was unforeseen at the design stage).

This trend explanation does not obviously explain the ICT estimates seen for Lists E and F at IP8 and IP9, as these are both cross-sectional estimates. To the extent that the trend explanation holds, at least LICT results are understandable. The LICT also appears to provide better estimates elsewhere, as there are no other negative estimates, unlike for the ICT. Further, the estimates across lists within a dimension (which are estimating the same sensitive item) vary less for LICT estimates than for ICT estimates. The similarity in LICT estimates across lists within dimension suggests the possibility (although not certainly) that the LICT estimates do not depend on list, whereas with ICT the larger variation across lists does not suggest this possibility.

Although the direct questions asked only about identity, which can be a very different construct to attraction and experience, it is also potentially useful to compare ICT and LICT estimates to these direct questions. Using the results originally presented in Table 2 as a baseline is also suggestive about the usefulness of estimates of list methods. For example, while the standard of assessing methods to improve reporting of sensitive behaviors is "more is better" (e.g. Tourangeau & Yan 2007), ICT estimates in Table 3 are at times very much more than those of the direct questions. For example, the UKHLS and IHS protocols provide estimates ranging from 2.7% to 3.5% identifying as being homosexual or bisexual. Comparatively, based on List A at IP8, the ICT estimates 12% of respondents have homosexual attraction and using List B the ICT provides an estimate 21% for the same (these may be due to the differences in non-sensitive items across groups, explained above). Conversely, for the LICT estimates for homosexual attraction is 4% based on List A and 7% on List B, so more than the direct questions, but not as drastically as the ICT estimates. The ICT estimate for homosexual experience based on List A is also 15%; however, the remainder of ICT estimates is relatively smaller or negative.

A suggested improvement to the ICT which may improve estimates is the Two-List ICT (Biemer & Brown 2005). In this case, Two-List ICT averages esti-

Table 4 Two-List ICT and Two-List LICT Estimates

Dimension	IP8 Two-List ICT	IP9 Two-List ICT	Two-List LICT
Attraction	0.02 (0.05)	0.08 (0.05)	0.06 (0.04)
Experience	0.11 (0.05)	0.07 (0.05)	0.09 (0.03)
Identity	-0.11 (0.04)	-0.01 (0.04)	-0.06 (0.03)

mates from the two lists within each dimension presented in Table 3, within waves. The LICT can also be extended to the Two-List LICT using the same averaging of estimates from lists within dimension. The estimates of Two-List ICT and Two-List LICT and the standard errors for these are presented in Table 4.

Both methods lead to negative estimates for Identity (Lists E and F), continuing to suggest problems with the method, noting the potential issues with these specific lists. However, there are no other negative values identified for any other estimate, which is an improvement over single-list ICT estimates, but consistently the same for LICT estimated. The Two-List ICT estimates are relatively smaller due to the averaging effect, and the drastically larger values are generally gone. The Two-List ICT estimate standard errors are also smaller than the single-list ICT estimates, demonstrating the benefit of Two-List ICT over the single-list version (even with the possibly conservative estimate of variance). Comparatively, the Two-List LICT estimates and standard errors are largely the same, given the small variation in individual list estimates. This consistency is reassuring in that lack of consistency (as in the ICT) is suggestive of possible problems. While there is still problematic evidence, and it does not prove the success of the LICT, lack of consistency is not a problem in the current application.

Discussion and Conclusions

This paper describes a new technique for collecting data on sensitive topics in surveys, extending on Item Count Technique methods: the Longitudinal Item Count Technique. Unlike the traditional ICT, this method uses the full sample and provides individual-level data. While results suggest some problems, the LICT results also provide evidence of the method's potential usefulness. The main problem identified is negative LICT estimates in two instances. Certainly negative estimates are problematic in any item count method; a negative prevalence is obviously not a true outcome. However, it is suggested that in this instance, the failure of the LICT to produce realistic estimates are due to the violation of the assumption that there is no trend in the data over time. The two lists that led to negative LICT estimates contained non-sensitive items regarding being British and European; the second administration of these lists occurred during the lead up-to and aftermath of the UK referendum to leave the European Union. Although problematic, if these negative values are due to items that trended, then future implementations of LICT may be able to avoid this problem with careful selection of items. Still, this explanation is not the only one which may explain the problems identified. In particular, the LICT lists used time variant items, which may have caused instability in responses; however, the balancing of lists across waves with a two-group design hopefully countered much of this impact.

Evidence suggesting the potential usefulness of the LICT exists in that it outperformed traditional ICT methods in a number of ways: it had lower standard errors, varied less on lists measuring the same dimension, and provided estimates that were greater, but not drastically so, than differing direct questions on sexual identity, the sensitive behavior of interest. While these results do not prove that the LICT is reliable or accurate, it is suggestive and at least does not prove that the method definitively does not work.

To ensure that the LICT method is useful, further research is needed. In particular, more applications of the LICT are needed using differing sensitive behaviors, especially where true values are known (if possible). The LICT methods here were all completed using self-completion data collection (CASI and Web). Research using face-to-face interviewing is also needed, as self-completion may have a differential impact on response and respondents, as some respondents may not be able to self-complete the questions.

From a design perspective, the downside of the LICT is that it requires multiple waves of data collection, which increases costs, while ICT or direct questions can be handled in a cross-sectional study. It should also be noted that other guidelines for the design of the traditional ICT are relevant also for the LICT (see Glynn (2013) for a recent summary of guidelines). Among the important design issues, in the application of the LICT, researchers need to consider whether an ethical approval is needed for data collection. Indeed, the LICT poses more challenges than the ICT from an ethical point of view, as respondents are revealing their sensitive behaviors by answering both, and they may not be aware of revealing them.

Furthermore, if respondents do realize they are being asked to reveal their sensitive behavior without being asked explicitly may lead to survey drop-out, or, in the context of a longitudinal study, panel attrition. The impact of asking the sensitive behavior to all respondents in the LICT may vary on which list (with or without the sensitive item) is presented at the earlier and later waves. For example, respondents may remember having answered already the short list in an earlier wave, the additional item in the later wave may make the realization of revealing the sensitive behavior more likely. Additionally, the length between waves may impact the method; longer lags between waves may increase the chance respondents do not remember whether they answered a similar question before. Shorter lengths could have the opposite effect.

References

Ahlquist, J. (2018). List experiment design, non-strategic respondent error, and item count technique estimators. *Political Analysis, 26*(1), 34–53. doi:10.1017/pan.2017.31

Aspinal, P. J. (2009). *Estimating the size and composition of the lesbian, gay, and bisexual population in Britain* (Equalities and Human Rights Commission Research Report 37). Retrieved November 12, 2018, from the Equalities and Human Rights Commission website: https://www.equalityhumanrights.com/sites/default/files/research-report-37-estimating-lesbian-gay-and-bisexual-population-in-britain.pdf

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics*, *21*(2), 287–308.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, *20*(1), 47–77. doi: 10.1093/pan/mpr048

Blair, G., Imai, K., & Lyall, J. (2014). Comparing and combining list and endorsement experiments: Evidence from Afghanistan. *American Journal of Political Science*, *58*(4), 1043–1063. doi: 10.1111/ajps.12086

Corstange, D. (2009). Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. *Political Analysis, 17*(1), 45–63. doi: 10.1093/pan/mpn013

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research, 40*(1), 169–93. doi: 10.1177/0049124110390768

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base-rates for sensitive behavior. *Personnel Psychology, 47*(4), 817–828. doi: 10.1111/j.1744-6570.1994.tb01578.x

De Cao, E., & Lutz, C. (2015). *Measuring attitudes regarding female genital mutilation through a list experiment* (CSAE Working Paper Series No. 2015-20), Retrieved November 12, 2018, from the Centre for the Study of African Economies, University of Oxford website: https://www.csae.ox.ac.uk/materials/papers/csae-wps-2015-20.pdf

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The Item Count Technique as a method of indirect questioning: A review of its development and a case study application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, S. Sudman, (eds) *Measurement Errors in Surveys*, pp. 185–210. Hoboken, New Jersey: John Wiley & Sons.

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly, 77*(S1), 159–172. doi: 10.1093/poq/nfs070

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly, 74*(1), 37–67. doi:10.1093/poq/nfp065

Jäckle, A., Gaia, A., Al Baghal, T., Burton, J. & Lynn, P. (eds) (2017). *Understanding Society the UK household longitudinal study Innovation Panel, waves 1-9, user manual*. Colchester: University of Essex.

Jensen, N. M., Mukherjee, B. & Bernhard, W. T. (2014). Introduction: survey and experimental research in international political economy. *International Interactions*, *40*(3), 287–304. doi: 10.1080/03050629.2014.899222

Johnson, A., London School of Hygiene and Tropical Medicine. Centre for Sexual and Reproductive Health Research, NatCen Social Research, & Mercer, C. (2017). *National Survey of Sexual Attitudes and Lifestyles, 2010-2012*. [data collection]. *2nd Edition*. UK Data Service. SN: 7799, doi:10.5255/UKDA-SN-7799-2

Kiewiet de Jonge, C. P. and Nickerson, D. W. (2014). Artificial inflation or deflation? assessing the Item Count Technique in comparative surveys. *Political Behavior 36*(3), 659–682. doi:10.1007/s11109-013-9249-x

Kuha, J. & Jackson, J. (2014). The item count method for sensitive survey questions: modelling criminal behavior. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(2), 321–341. doi: 10.1111/rssc.12018

Laumann, E. O., Gagnon, J. H., Michael, R. T. & Michaels, S. (1994). *The social organization of sexuality: sexual practices in the United States*. Chicago: University of Chicago Press.

Magaloni, B., Diaz-Cayeros, A., Romero, V. & Matanock, A. M. (2012). The enemy at home: exploring the social roots of criminal organizations in Mexico. Available at: https://ssrn.com/abstract=2122950

NatCen Social Research. (2014). *British Social Attitudes Survey, 2013*. [data collection]. UK Data Service. SN: 7500. doi: 10.5255/UKDA-SN-7500-1

Smith, L. L., Federer, W. T. & Raghavarao, D. (1974). A comparison of three techniques for eliciting answers to sensitive questions. *American statistical association. Proceedings of the social statistics section* pp. 447–452. Washington D.C.: American Statistical Association.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. doi: 10.1037/0033-2909.133.5.859

University of Essex. Institute for Social and Economic Research, NatCen Social Research, & Kantar Public. (2017). *Understanding Society: Waves 1-7, 2009-2016 and Harmonised BHPS: Waves 1-18, 1991-2009*. [data collection]. *9th Edition*. UK Data Service. SN: 6614.

University of Essex. Institute for Social and Economic Research. (2018). Understanding Society: Innovation Panel, Waves 1-10, 2008-2017. [data collection]. 9th Edition. UK Data Service. SN: 6849. doi: 10.5255/UKDA-SN-6849-10

Appendix 1: Question wording

Item Count Technique (CASI & WEB)

Introduction

“The next set of questions will ask you to count the number of statements that are true for you. Please only count the number of statements. We are not interested in knowing which statements are relevant for you.”

Group 1

Item count list A

I have at least once been sexually **attracted** to someone who …

- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Item count list B + sensitive item

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list C

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Item count list D + sensitive item

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list E

I would describe myself as **being** …

- stylish and fashionable
- disabled
- patient
- British

How many statements are true for you?

None are true
One statement

Two statements
Three statements
Four statements

Sexuality item count list F + sensitive item

I would describe myself as **being** …

- gay, lesbian or bisexual
- healthy
- tolerant
- European
- working class

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Group 2

Sexuality item count list A + sensitive item

I have at least once been sexually **attracted** to someone who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list B

I have at least once been sexually **attracted** to someone who …

- wears the latest trends and fashions

- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements

Sexuality item count list C + sensitive item

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- is the same sex as me
- has a disability
- is fit and muscular
- grew up with me in my local area
- is ten or more years older than me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list D

I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …

- wears the latest trends and fashions
- has a tattoo or body piercing
- is of a different ethnicity to me
- is from a different class background to me

How many statements are true for you?
None are true
One statement
Two statements
Three statements
Four statements

Sexuality item count list E + sensitive item

I would describe myself as **being** …

- gay, lesbian or bisexual
- stylish and fashionable
- disabled
- patient
- British

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements
Five statements

Sexuality item count list F

I would describe myself as **being** …

- healthy
- tolerant
- European
- working class

How many statements are true for you?

None are true
One statement
Two statements
Three statements
Four statements

Direct questions:

Protocol 1 – IHS

Mode: Face-to-Face with showcard

Question wording: “Which of the options on this card best describes how you think of yourself? Please just read out the number next to the description.”

SHOWCARD
27. Heterosexual / Straight
21. Gay / Lesbian
24. Bisexual
29. Other

Note: “Don’t Know” and “Refuse” were not displayed in the showcard. Interviewers recorded “Don’t Know” and “Refuse” if those where spontaneous answers of the respondent.

Mode: Telephone

Question wording: “I will now read out a list of terms people sometimes use to describe how they think of themselves: “Heterosexual or Straight”, “Gay or Lesbian”, “Bisexual”, or “Other”. As I read the List Again please say ‘yes’ when you hear the option that best describes how you think of yourself.

Heterosexual or Straight
Gay or lesbian
Bisexual
Other”

Interviewer Instruction: on first reading, read list to end without pausing. Note that “heterosexual or straight” is one option “gay or lesbian” is one option. On second reading, please pause briefly after each option.

Protocol 2 – UKHLS

Mode: WEB or CASI

“Which of the following options best describes how you think of yourself?

Heterosexual or Straight
Gay or Lesbian
Bisexual
Other
Prefer not to say”

Non-Randomized Response Models: An Experimental Application of the Triangular Model as an Indirect Questioning Method for Sensitive Topics

Anke Erdmann
Bielefeld University

Abstract

When it comes to sensitive questions, data is often affected by bias due to non-response or effects of social desirability. Several methods have been introduced to eliminate answer bias by using randomization processes and probabilistic theory to obscure the respondent's answer and create anonymity, thus facilitating honest answers. The probably most traditional method is the Randomized Response Technique by Warner (1965). However, this method is loaded with certain disadvantages. Therefore, in the last decade, newer methods were introduced that aim at balancing the disadvantages and weaknesses of previous methods, for instance, the non-randomized models Crosswise Model and Triangular Model (Yu et al. 2008) as well as the Parallel Model (Tian 2014). Although especially the Triangular Model is easy to implement in a study, there is only little empirical evidence on its application in different survey modes and populations. Further, it is to assume that certain questions are not equally sensitive for everybody due to specific personal characteristics. Thus, indirect questioning might not be effective in general but only for certain populations. The present study extends prior work on the Triangular Model by evaluating it for different subgroups. The conducted experiment asks for sensitive characteristics in the context of mental stress among students. The Triangular Model achieves significantly higher percentages than conventional direct questioning for illegal drug use among persons that answer socially desirable according to the characteristic of Self-Deception. For the other analyzed subgroups (Impression Management, gender, and depressiveness), the Triangular Model could not achieve higher prevalence rates compared to direct questioning on a sufficient probability level. But still, hard evidence on the effectiveness of indirect questioning models is thin and further critical discussion is needed.

Keywords: Triangular Model, Social Desirability, Indirect Questioning, Survey Methodology, Non-Randomized Response

Collecting data is substantial for empirical research. Yet, the reliability and validity of data gathered in surveys is at risk of being limited due to non-response, effects of social desirability or other bias. For that reason, continuous research in survey methodology is essential to further improve modes of data collection and analysis. Especially social desirability has concerned scholars for some time now. It means that a respondent – deliberately or not – adjusts his or her answer according to what he or she thinks is socially accepted. Several scales have been developed to measure this construct and new interrogation techniques have been constantly introduced to take into account systematic bias in surveys. A promising possibility to collect data on sensitive topics is indirect questioning. Such techniques anonymize the respondent's answer using probability theory and try to facilitate honest answers by protecting the respondent's information. Probably the most up-to-date techniques are so-called non-randomized response models. However, to this day, only few studies examine the performance and the viability of these methods. For some of those models, to the best of my knowledge, there is even no empirical testing at all. For this reason, this research article presents an evaluation of one selected non-randomized response model – the Triangular Model – that compares its estimated prevalence rates with the ones obtained with direct questioning.

The present study is mainly inspired by previous work by Jerke & Krumpal (2013) and aims at extending it by evaluating the Triangular Model in different subgroups. To test this assumption, an online survey was conducted in which the method was applied in the context of mental stress and psychological problems.

This research paper starts with a brief overview on social desirability. Second, non-randomized response models are presented in detail to give an overview on these indirect questioning models. After that, the conducted study is described and the results are presented and discussed.

Acknowledgments

This research article is based on the author's master thesis at the Faculty of Sociology at Bielefeld University. The author would like to thank Andreas Pöge (Faculty of Sociology, Bielefeld University) for the main supervision, Fridtjof Nußbeck (Faculty of Psychology and Sport Science, Bielefeld University) for the secondary supervision as well as the editors and the anonymous reviewers for their helpful and constructive comments on previous versions of this paper.

Direct correspondence to

Anke Erdmann, M.A., Faculty of Sociology, Bielefeld University,
Universitätsstraße 25, 33615 Bielefeld
E-mail: anke.erdmann@uni-bielefeld.de

The Concept of Social Desirability

When conducting an empirical investigation, it is advisable to pay attention to effects of social desirability. A traditional scale to measure this answering behavior is the M-C SDS (Marlowe-Crowne Social Desirability Scale) by Crowne & Marlowe (1960). Redesigns for German studies are, for example, the SDS-CM (Social Desirability Scale by Crowne & Marlowe; Lück & Timaeus 1969, 1997b), the SDS-E (Social Desirability Scale by Edwards, Lück & Timaeus 1997a) and the SES-17 (Soziale Erwünschtheitsskala-17; Stöber 1999, 2001). These scales are easy to handle by using a summed score but there is criticism that they assume a one-dimensionality of the construct. In 1984, Paulhus argued that social desirability consists of two dimensions: Impression Management (IM) and Self-Deception (SD). Whereas IM means a deliberate deception to create a positive image towards others to gain social acknowledgment, SD describes the unconscious deception of one's own to maintain an optimistic and positive self-image (Krumpal & Näher 2012; Paulhus 1984; Winkler et al. 2006). To measure those two dimensions, Paulhus (1984) developed the Balanced Inventory of Desirable Responding (BIDR). Yet, this scale contains 40 items, which makes it inappropriate for most surveys. To overcome this, Winkler et al. (2006) developed a short scale that measures both dimensions of social desirability while containing only six items. The scale fulfills the criteria for reliability, internal and external validity and complies with the theoretical and empirical assumptions of the BIDR-scale by Paulhus (1984). The scale's formulation is described in the measurement section.

How strongly a question is affected by social desirability bias depends on the question's content. A strong vulnerability to social desirability is given when a question is about sensitive, illegal or embarrassing content that is a potential danger for the respondent to reveal his or her true answer (e.g., sexuality, drug consumption, political opinions, violation of social norms). However, there is no exact definition of what a sensitive question is. Tourangeau & Yan (2007) define it as follows:

> "A question is sensitive when it asks for a socially undesirable answer, when it asks, in effect, that the respondent admits he or she has violated a social norm" (Tourangeau & Yan 2007, p. 860).

So in fact, the sensitivity of a question is not objective but depends on many factors (Wolter 2012). For instance, whether a question is sensitive or not might depend on who is asked. For example, Tourangeau & Yan (2007) mention political elections where the question whether someone voted or not is only sensitive for the ones who did not. Further, questions about political topics are more sensitive among higher educated people (Tourangeau & Yan 2007).

Further, it is possible that a question is equally sensitive for everybody, but different answers are the socially desirable ones. For example, when regarding information about drug or alcohol consumption, in general, "no" seems to be the desirable answer, but it is possible that within certain groups (e.g., among peers), "yes" is the more accepted answer. Additionally, whether a question is sensitive or not might depend on "who is asking." For instance, being asked by a friend about sexuality or drug consumption is probably not as sensitive as being asked by a teacher, the parents or a research interviewer. Furthermore, it is possible that a question is differently biased in different subgroups. For example, questions about sexuality (e.g., number of sexual partners) might be equally sensitive for men and women but in opposite ways: While for one group, a high number is socially desirable, it is a low number for the other group. This extension that a question's sensitivity depends on many circumstances is part of a definition by Porst (2009):

> "A question is sensitive when the person answering it expects any negative responses of any kind as consequence of his or her answer in general or as consequence of a specific answer – this is independent from the content of the question" (Porst 2009, p. 124, own translation).

Therefore, a question is not sensitive per se but becomes sensitive through the situation, the involved persons, and their expectations.

Indirect Questioning Models

There are several methods to avoid or at least soften bias caused by social desirability. Mostly, they function by anonymizing answers or giving the respondent a feeling of confidentiality by adjusting the interview circumstances. Also, questions could be asked in a way to "de-dramatize the deviation of a social norm" (Häder 2015, pp. 213) by using special ways of wording and framing (Barton 1958; Porst 2009; Preisendörfer 2008). Other methods take a further step and use probability theory to anonymize answers and to estimate the prevalence rate of a critical question. For example, so-called Randomized Response and Non-Randomized Response Models belong to this category of indirect questioning. The Randomized Response Technique (RRT) was introduced by Warner (1965). The RRT links a randomization process to a sensitive question which serves the anonymization of the respondent's answer. A randomization device is needed that has two possible outcomes with known probabilities. Depending on the outcome, the respondent answers one of two statements where a sensitive characteristic is formulated in exactly opposite ways. Fox & Tracy (1986) illustrate an example where one out of ten balls of two different colors is drawn from a ballot box. When drawing a blue ball, the statement "I have used heroin" had to be answered, otherwise "I have never used heroin" when drawing a green ball. By knowing how many blue and green balls the box contains, the probabilities of receiving one of the statements are known. Hence it is obscured whether the sensitive characteristic applies. In this way, the general willingness to answer at all as well as the motivation to answer truthfully is expected to rise (Droitcour Miller 1981).

The RRT is well-researched and the body of literature offers many applications and methodological evaluations on different sensitive topics (e.g., Coutts & Jann 2011; Kirchner et al. 2013; Abernathy et al. 1970; Pitsch et al. 2012). But, although many studies justify using the RRT by attesting its success (e.g., Lara et al. 2016; van der Heijden et al. 2016), there are also several investigations that provide evidence that the RRT fails to yield more valid estimates as compared to DQ (e.g., Beldt et al. 2016; Buchman & Tracy 1982; Wolter & Preisendörfer 2013). Some empirical studies further discuss a general failure of the technique due to incorrect following of the instructions and cheating. For example, Holbrook & Krosnick find that the RRT failed in reducing response bias because "respondents were either unable or unwilling to implement the randomized response technique properly" (2010, p. 328). This raises concerns about the viability of the RRT – especially in interview situations like online or telephone surveys that lack control whether the interviewees really use the randomization device. To investigate the effects of determinants of misreporting by question mode, Wolter & Preisendörfer (2013) conducted an experimental study with criminal convicts to compare direct questioning (henceforth: DQ) with RRT. Their findings include that "the success of the RRT varies systematically depending on the interview situation and the actors involved" (Wolter & Preisendörfer 2013, p. 344), which challenges the assumption of a general usefulness of the RRT. Further, the factors that determine response behavior vary by question mode. This finding might explain the mixed results on the performance of the RRT: If response behavior varies by certain characteristics, different compositions of analyzed samples lead to diverging results in spite of using the same technique. Additionally, besides mixed evidence, a key disadvantage of RR-models is their complexity. The respondents have to understand the instructions and trust the procedure (Jann et al. 2012). Thus, cognitive overload, misunderstanding, and suspiciousness might result in answering errors (Jerke & Krumpal 2013). This and other weaknesses of RR-models shall be overcome by so-called non-randomized response models (NRR-models). The three techniques Crosswise Model, Triangular Model and Parallel Model are introduced in the following section.

Crosswise Model

In 2008, the Crosswise Model (CM) was introduced by Yu et al. (2008) alongside the Triangular Model. This technique combines a sensitive question to a non-sensitive one and asks for a combined answer on both questions simultaneously. The respondents choose between "both answers are equal" and "both answers are

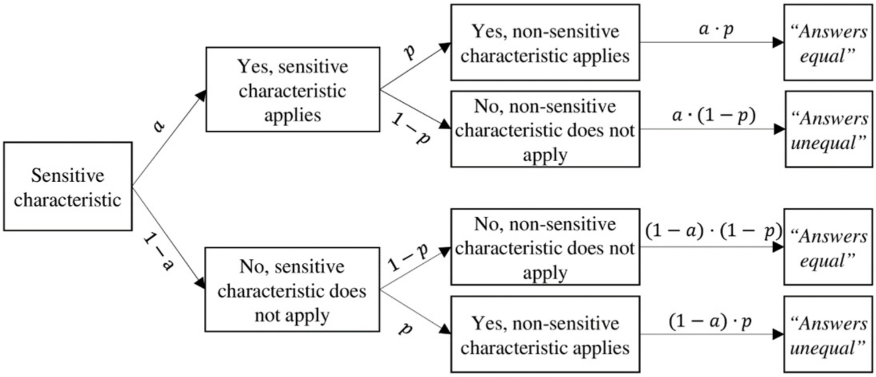

Sensitive characteristic
a
1 − a
Yes, sensitive characteristic applies
No, sensitive characteristic does not apply
p
1 − p
1 − p
p
Yes, non-sensitive characteristic applies
No, non-sensitive characteristic does not apply
No, non-sensitive characteristic does not apply
Yes, non-sensitive characteristic applies
a · p
a · (1 − p)
(1 − a) · (1 − p)
(1 − a) · p
"Answers equal"
"Answers unequal"
"Answers equal"
"Answers unequal"

Figure 1 Design of the Crosswise Model

unequal." The decisive element is that the probability distribution of the non-sensitive item is known (e.g., birth dates or random numbers like the last digit of a phone number). The model's theoretical construction is shown in Figure 1. The parameter *a* contains the unknown prevalence rate of the sensitive item and *p* is the probability to answer "yes" on the non-sensitive question.

The term $s(\text{"}equal\text{"})$ describes the share of "both answers are equal"-answers and is gathered from the sample. Thus, the estimator for the prevalence rate *a* – which is called $\hat{a}_C$ for the CM in this paper – is the following (Jann et al. 2012; Yu et al. 2008):

$$\hat{a}_C = \frac{s(\text{"}equal\text{"}) + p - 1}{2 \cdot p - 1}, \quad p \neq 0.5 \tag{1}$$

$\hat{a}_C$ = *Estimated proportion of "yes"-answers on the sensitive item*
s = *Proportion of "both equal"-answers in the sample*
p = *Probability of the non-sensitive item*

The variance of the estimator can be obtained through the following formula (Jerke & Krumpal 2013; Tang et al. 2013; Yu et al. 2008; Liu & Tian 2014):

$$Var(\hat{a}_C) = \frac{a \cdot (1 - a)}{n} + \frac{p \cdot (1 - p)}{n \cdot (2p - 1)^2}, \ p \neq 0.5 \tag{2}$$

The CM is a non-randomized version of Warner's RRT (Tian 2014). It is characterized by the same estimator, the same variance and is affected by the same mathematical restrictions. The CM does also have the same qualities regarding the best possible choice for *p* and the same calculations of optimal sample size (Ulrich et al. 2012). The first empirical evaluation is by Jann et al. (2012), who use the method for analyzing plagiarism and they compare the CM to DQ. Other methodological applications can be found in, for example, Kundt, Misch, & Nerré (2013) and Hoffmann & Musch (2016).

Triangular Model

The Triangular Model (TM) is similar to the CM but the essential distinction lies in the answering options. The sensitive question is once again linked to a non-sensitive characteristic with a known probability. But instead of choosing if either both answers are equal or not, the interviewee provides information whether his or her answers are both “no” or he or she affirms at least one of the two questions. Considering these answering options, a disadvantage in comparison with the CM becomes evident: The TM has an “option for protection.” Choosing “no on both questions” will definitely reveal that the respondent does not have the sensitive characteristic (Jann et al. 2012). So it can be criticized that the TM does not have a sufficient concealment of the answer “no” thus still being vulnerable to underreporting and the TM might not deliver adequate anonymization under certain circumstances (Tian 2014). Despite this drawback, the TM is worth testing because it surpasses other models regarding efficiency, revealment of the “yes”-answer, and is simple to implement in a survey (Wu & Tang 2016). Additionally, empirical evidence is rather scarce and it is still to be tested how this limitation really affects the model’s effectiveness.

An outline of the model can be seen in Figure 2. The proportion of “both no”-answers in the sample is the product of *p*’s inverse probability and the inverse proportion of the amount of persons carrying the sensitive item:

$$s\left("both\,no"\right) = \left(1 - a\right) \cdot \left(1 - p\right) \tag{3}$$

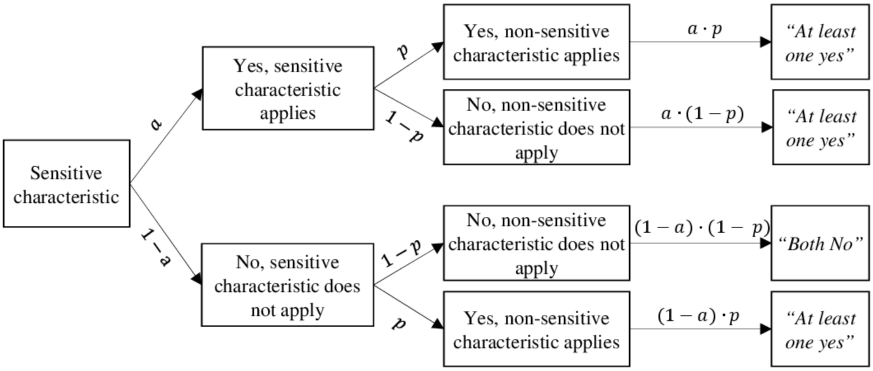

Sensitive characteristic
a
1 − a
Yes, sensitive characteristic applies
No, sensitive characteristic does not apply
p
1 − p
1 − p
p
Yes, non-sensitive characteristic applies
No, non-sensitive characteristic does not apply
No, non-sensitive characteristic does not apply
Yes, non-sensitive characteristic applies
a · p
a · (1 − p)
(1 − a) · (1 − p)
(1 − a) · p
“At least one yes”
“At least one yes”
“Both No”
“At least one yes”

Figure 2 Design of the Triangular Model

Rearranging the term (3) provides the estimator $\hat{a}_T$ for the TM (Jerke & Krumpal, 2013; Tang et al., 2013; Yu et al., 2008):

$$\hat{a}_T = 1 - \frac{s\left("both\,no"\right)}{1-p} \tag{4}$$

$\hat{a}_T$ = *Estimated proportion of "yes"-answers on the sensitive item*
s = *Proportion of "both no"-answers in the sample*
p = *Probability of the non-sensitive item*

The estimator's variance is described by the following formula (Jerke & Krumpal 2013; Tang et al. 2013; Yu et al. 2008):

$$Var\left(\hat{a}_T\right) = \frac{a\cdot(1-a)}{n} + \frac{p\cdot(1-a)}{n\cdot(1-p)} \tag{5}$$

These formulae reveal that the CM's restriction of choosing a p other than 0.5 is eliminated for the TM. However, although Yu et al. (2008) do not exclude any probabilities mathematically[1], a probability of 1 is not reasonable from a contentual perspective. If the probability of the non-sensitive item is 1 (i.e., the respondent's answer is definitely "yes"), the answer "both no" is not possible. Thus, all respondents have to answer with "at least one yes" so an estimation of the prevalence rate is impossible since the proportion of "both no"-answers is always 0 independently from the true prevalence rate a. In this case, total anonymity is given but also no result.

The opposite case of p=0 is not advisable as well: If the answer on the non-sensitive item is definitely "no," then it is clear that "at least one yes" means a "yes" on the sensitive question. Regarding the estimator and its variance, this means that the parts containing p are cancelled. So in fact, a TM with p=0 is basically just direct questioning resulting in total revelation of the answers but no anonymity. In conclusion, it is advisable to choose a probability that balances the relation between anonymity and efficient estimation.

To my best knowledge, the only application of the model is by Jerke & Krumpal. (2013) on student plagiarism at a German university. The study reveals higher prevalence rates for partial as well as for full plagiarism. In comparison to the CM, the authors find a smaller standard error for the TM and thus a more efficient estimation. However, the differences achieved with the TM are not significantly higher than in DQ.

[1] But it is evident from the formulae that a p of 1 would result in a denominator of 0.

Parallel Model

Despite the advantages of the CM and the TM, they both have a certain limitation: one category (usually the “no”-answer) has to be non-sensitive (Tian 2014, p. 293). To eliminate this restriction, Tian (2014) introduces another NRR-model: the Parallel Model (PM). This technique uses *two* non-sensitive items with a known probability (named as W and U). The respondents belong to two groups (W=1 and W=0, i.e., the first non-sensitive characteristic applies or not). Then, the answer on this first non-sensitive question (W) decides whether the respondent answers the second non-sensitive (U) or the sensitive question (Y) (for an example, see Tian 2014, p. 300). Since the answer on the first question is unknown, the interviewer does not know which question is answered. Figure 3 shows an outline of the PM and how the amount of “yes” and “no” answers in the sample is composed. From this Figure, the following estimator can be derived (Tian 2014, p. 301):

$$\hat{a}_P = \frac{s("yes") - q \cdot (1 - p)}{p} \tag{6}$$

$\hat{a}_P$ *= Estimated proportion of “yes”-answers on the sensitive item*
s *= Proportion of “yes”-answers in the sample*
p, q *= Probabilities of the non-sensitive items*

Again, the estimator’s variance consists of the usual sampling variance and additionally a part that is induced by the randomization process.

$$Var\left(\hat{a}_P\right) = \frac{a \cdot (1 - a)}{n} + \frac{(1 - p) \cdot \varphi}{n \cdot p^2}$$

$$\text{where } \varphi = (p - 1) \cdot q^2 + (1 - 2 \cdot a \cdot p) \cdot q + a \cdot p \tag{7}$$

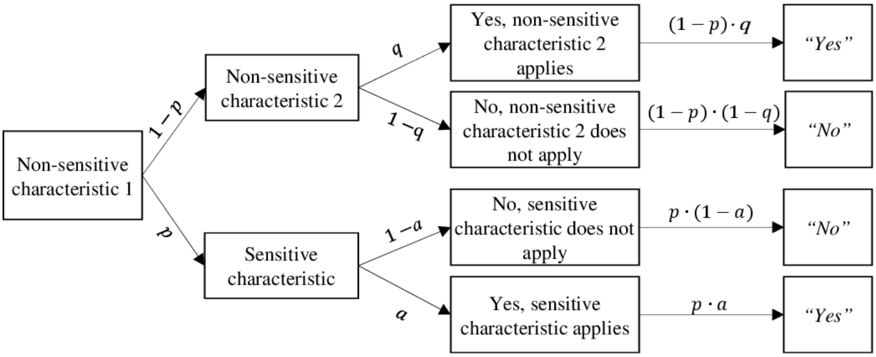

Non-sensitive characteristic 1
1 – p
p
Non-sensitive characteristic 2
Sensitive characteristic
q
1 – q
1 – a
a
Yes, non-sensitive characteristic 2 applies
No, non-sensitive characteristic 2 does not apply
No, sensitive characteristic does not apply
Yes, sensitive characteristic applies
(1 – p) · q
(1 – p) · (1 – q)
p · (1 – a)
p · a
“Yes”
“No”
“No”
“Yes”

Figure 3 Design of the Parallel Model

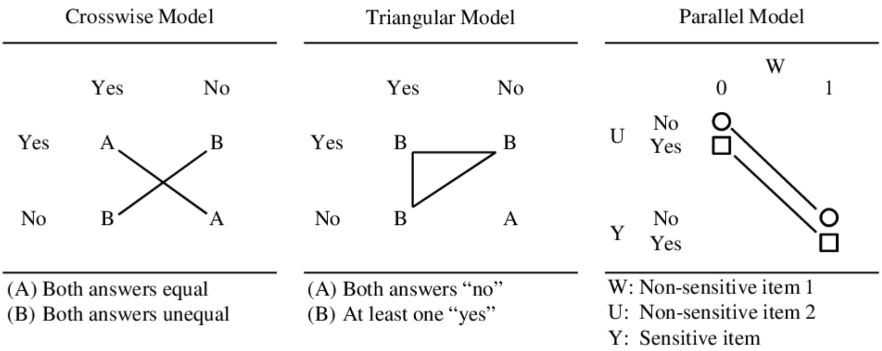


Figure 4 Answering options for the Crosswise, Triangular and Parallel Model

The combination of answers leads to a parallelism (for more details, see Tian 2014, p. 300) which is displayed in Figure 4 alongside the answering options for the CM and TM.

The logic of the PM is comparable to the Unrelated Question Model by Horvitz et al. (1967). Thus, the PM combines the advantages of this specific RR-model with the strengths of an NRR-model: The design is a device-free technique but has – compared to the CM and the TM – a better anonymization of answers. The information whether the sensitive characteristic applies or not are both protected. So far, to the best of my knowledge, there are no experimental applications that evaluate the PM in comparison to DQ.

The Present Study

Inspired by the work of Jerke & Krumpal (2013), the present study examines the TM by comparing its estimated prevalence rates to the ones that are achieved using DQ. It is assumed that anonymized questioning "cancels out the costs that make respondents misreport in DQ mode" (Wolter & Preisendörfer 2013, p. 329). This includes persons that strive for social acknowledgment, thus answering socially desirable. Further, several authors point out that misreporting in surveys is most likely for the persons who "have the most to lose" when reporting truthfully (Bernstein et al. 2001; Wu & Tang 2016), i.e., the persons that have the sensitive characteristic. Thus, this study puts the focus on the assumption that a question might have different levels of sensitivity for different persons or groups, so the TM might prove to be efficient only in certain subgroups in the sample. For this purpose, an online survey on the topic "Mental stress among students" was conducted (field time from 13th to 27th July 2015). First, the TM is compared to DQ in general.

Since the answers are anonymized when using the TM, a higher prevalence rate can be expected in comparison to DQ because a more honest answering behavior is assumed (*Hypothesis 1*).

Second, the TM will be analyzed separately for gender (*Hypothesis 2*), social desirability (with the two dimensions IM and SD; *Hypothesis 3*), and depressiveness (*Hypothesis 4*). Regarding social desirability, a stronger effect of the indirect questioning method is assumed for persons who have the characteristic of answering socially desirable. But, it is expected that the anonymization is only effective for IM. Deceiving according to IM is a conscious act to create a more positive image of oneself, for instance, in an interview situation. Self-deceptive behavior, however, is subconscious. Thus, anonymization of an interview situation should not affect the bias created by this characteristic. Regarding depressiveness, in the present study, the TM is supposed to be more efficient for persons with a high level of depressiveness because the questions in this specific questionnaire are assumed to be more sensitive for this group than for persons who are not depressive. The effect that is postulated for gender is assumed to be indirect. Prior research indicates that social desirability varies by gender. Females are more prone to answer socially desirable (Becker & Cherny 1994; Dalton & Ortegren 2011) – especially regarding IM. Further, studies suggest that female students are more strongly strained by depressiveness than their male colleagues (Burger & Scholz 2018; Margitics & Pauwlik 2009). Thus, it is assumed that the TM works better for females.

The questionnaire was conducted as an online survey because this method offers advantages considering the possibility to contact many people and to randomly sort the respondents into the two survey conditions.

Measurements

Sensitive Questions

According to the topic of mental stress amongst students, the respondents were asked whether they ever did the following acts during their studies:

> *Did you ever make use of a psychological consultancy?*
>
> *Did you ever use prescriptive medication for enhancing mental performance?*[2]
>
> *Did you ever use illegal drugs for enhancing mental performance?*[3]

[2] Additional explanation: "*For example, to learn more fastly and efficiently, to manage a workload or to be more focused during an exam.*"

[3] Additional explanation: "*This means, for example, substances like amphetamine ("speed"), cocaine, methamphetamine, etc.*"

Respondents in the DQ condition received the questions as they are and were asked to answer with “yes” or “no.” For the TM, the questions were combined with the following non-sensitive questions:

Is your mother’s birthday in January, February, or March?

Is your birthday in May?

Is your birthday in January?

The two possible answering options were:

*The answer is “no” on **both** questions.*

*The answer is “yes” on **at least one** of the questions.*

Independent Variables

The concept of social desirability was measured using the scale by Winkler et al. (2006). The scale contains six items that represent both dimensions of social desirability, Impression Management (IM) and Self-Deception (SD). Table 1 shows the wording of the items and which dimension is measured. The notes + and – depict whether a high or a low value represents the tendency to answer socially desirable.

To check for the scale’s dimensionality, a *principal component factor analysis* (PCA; Bortz 1989) was performed. The PCA confirms two factors and also the polarity assumed by Winkler et al. (2006). The results are in line with the findings by the authors and reflect the scale’s theoretical assumptions.

In consideration of the items’ polarity, two mean indices are designed for IM and SD by summing up the values of the items and dividing by their number. The correlation between the two dimensions is rather low (r=0.13, p=0.000), which confirms that these are two distinct concepts which are only slightly correlated. According to Paulhus, only extreme answers can be interpreted as socially desirable answering behavior. Thus, for each dimension, two subgroups are constructed using the same method as Winkler et al. (2006) by generating a dichotomous variable where values of 6 and higher are marked as 1 and all other values below this line are marked as 0.

Depressiveness is operationalized using a scale from Mohr & Müller (2014) which contains eight items that measure depressiveness in a non-clinical context (Table 2).

Applying a PCA confirms the one-dimensionality of the scale. The latent factor has an explained variance of 49.2 percent, which is in line with the data structure found by Mohr & Müller (2014). Thus, the items are condensed into a mean index by adding the values of the items and dividing by their number. Further, two subgroups are constructed based on this index. Since Mohr & Müller (2014) do

Table 1 Operationalized BIDR short scale by Winkler et al. (2006)

Instruction: Please take position to the following behaviors. What would you say: To what extent does the sentence apply to you?	
My first impression of people usually turns out to be right.	SD +
I am often insecure in my judgment.	SD –
I always know why I like things.	SD +
I have received too much change from a salesperson without telling him or her.	IM –
I am always honest to other people.	IM +
There have been occasions when I have taken advantage of someone.	IM –

Note: Answers on a 7-point-Likert scale from 1= "does not apply at all" to 7= "fully applies".

Table 2 Depressiveness scale by Mohr & Müller (2014)

Instruction: Use the following answering options to state whether resp. how often the following statements apply to you. There is no right or wrong answer. Please do not leave out any questions!
I have to push myself to do things.
Many things seem pointless to me.
I am oppressed by feelings of guilt.
I feel lonely even when I am around other people.
I have sad moods.
It is hard for me to make decisions.
At the beginning of the day, I feel worst.
I look into the future without hope.

Note: Own translation, answering options: 1=never, 2=very rarely, 3=rarely, 4=occasionally, 5=often, 6=very often, 7=almost always.

not define a cut point that marks depressiveness, the values 5, 6, and 7 (*often*, *very often*, and *almost always*) are coded to indicate a high level of depressiveness.

The collected demographic information are age and gender. For gender, the respondents could choose between *male*, *female*, and *other*. The information on age is used to refine the probability of the non-sensitive questions in the TM (see below). The questions were placed at the end of the questionnaire. No further demographic information were retrieved to keep the survey short and parsimonious.

Sampling and Data Collection

As apparent from the previous description, the variance for the estimators of indirect questioning models is always inflated due to an additional variance induced by the randomization process. So there is a need for a preferably large sample size to oppose the inaccuracies accompanied by the increased standard errors. Therefore, a main objective was to reach a large number of participants. The call to participate in the survey was sent to students via diverse mail distribution systems at different universities in Germany. First, ten public universities were chosen non-randomly. Then, e-mails were sent out to persons in charge (e.g., secretaries at the dean offices) at all faculties, resp. institutes at these selected universities. Thus, there is no specialization and all kinds of study programs are included. This way, a total sample size of n=1,546 was achieved for this study.[4]

Table 3 shows the sample size by the two survey conditions DQ and TM as well as for gender.[5] It is obvious that there is a bias regarding the distribution by males and females: Around 70% of respondents are female. The reason for this discrepancy is unclear. It is unlikely that this relation reflects the true gender distribution in the general population or distribution at the universities since a broad variety of study programs was selected. Instead, it is possible that this is the result of a higher willingness for females to participate in studies as well as a greater interest in surveys about psychological problems. This bias is considered to be irrelevant for the present experimental study, thus the data will be analyzed as it is.

Table 3 Sample size by survey condition and gender

	Total	Gender			
		Female	Male	Other	N.A.
Direct Questioning	688	478	196	13	1
Triangular Model	628	448	163	15	2
Total	1,316	926	359	28	3

Note: N.A.=no answer.

[4] All in all, 230 persons aborted the online survey before reaching the experimental part of the questionnaire where the random sorting into DQ and TM condition takes place. Thus, the following analyses are based on a sample of 1,316 persons.

[5] The group of persons that report *other* as their gender will not be considered as a separate group in the following gendered analysis due to very low sample size.

Analytical Strategy

The TM will be evaluated by estimating the prevalence rates using the formulae presented above. Additionally, the differences between the prevalence rates achieved with TM and DQ will be examined. These differences will be tested for statistical significance using the following formula (Jerke & Krumpal 2013, p. 364):

$$t = \frac{\hat{a}_T - \hat{a}_D}{\sqrt{Var\left(\hat{a}_T\right) + \frac{\hat{a}_D \cdot \left(1 - \hat{a}_D\right)}{n_D}}} \tag{8}$$

The parameters $\hat{a}_T$ and $Var\left(\hat{a}_T\right)$ have been described before. The abbreviation $\hat{a}_D$ marks the prevalence rate estimated with direct questioning (with n_D as belonging sample size). The distribution is the Student t-Distribution with $n_D + n_T - 2$ degrees of freedom.

The probabilities of the non-sensitive questions in the TM were determined based on data from the German Federal Statistical Office using age, resp. the birth year of the respondents. For this, the individual probability for each person was estimated by considering the birth rates of males and females for each month within a certain year. Then, the average was calculated for the whole sample. The probability for the mother's birth month was determined in the same way. Prior to this, however, the mother's birth year was estimated based on the respondent's birth year and the average age a mother gave birth to a child. Thus, the probabilities for the non-sensitive characteristics in this specific sample are the following:

"Is your mother's birthday in January, February, or March?"	*p=0.258*
"Is your birthday in May?"	*p=0.084*
"Is your birthday in January?"	*p=0.085*

Additionally, to analyze whether the TM works differently in certain groups of respondents, the *differences-in-differences (DID)* are considered. Analyzing DID is a technique to identify causal relationships by examining the influence of a certain treatment (Bertrand et al. 2003). Usually, it analyzes two groups – one group receives a treatment and the other group does not – that are measured at two time points. Then, the difference between the two time points of measurement *within* each group is determined followed by analyzing the difference *between* these two differences. Transferred to the present study, the "treatment" is belonging to a certain subgroup. The survey conditions represent two measurements. So first, the differences between DQ and TM that occur in the subgroups are considered. Second, the difference between these is determined. Therefore, the DID is calculated as follows:

$$(TM - DQ)_{Subgroup\,1} - (TM - DQ)_{Subgroup\,2} \tag{9}$$

If this difference-in-differences turns out to be non-random, this would suggest that the difference can be traced back to the subgroup, i.e., the TM works differently in the compared subgroups.

Results

Table 4 shows the descriptive results of the three main independent variables by gender. According to this dichotomization, 15.2 percent of the persons in the sample feature the characteristic of SD. Regarding IM, the proportion of persons classified as having this characteristic amounts to 20.4 percent.

Table 4 Proportions and means for Self-Deception, Impression Management and depressiveness by gender

	Total	Gender		Diff.	
		female	male		
Self-Deception	*n=1419*				
SD=1 (in %)	15.2 (1.0)	15.0 (1.2)	16.7 (2.0)	-1.7 (2.3)	p=0.454
95% CI for SD=1	[13.4 , 17.1]	[12.7 , 17.3]	[12.8 , 20.6]	[-6.1 , 2.7]	
Ø Mean Index	4.8 (1.0)	4.8 (1.0)	4.9 (1.0)	-0.1 (0.1)	p=0.107
95% CI for Mean Index	[4.7 , 4.9]	[4.7 , 4.9]	[4.8 , 5.0]	[-0.2 , 0.0]	
Impression Management	*n=1419*				
IM=1 (in %)	20.4 (1.1)	22.6 (1.4)	15.3 (1.9)	7.3 (2.5)	p=0.004
95% CI for IM=1	[18.3 , 22.5]	[19.9 , 25.3]	[11.6 , 19.1]	[2.4 , 12.2]	
Ø Mean Index	4.7 (1.2)	4.8 (1.2)	4.5 (1.3)	0.3 (0.1)	p=0.000
95% CI for Mean Index	[4.6 , 4.7]	[4.7 , 4.8]	[4.3 , 4.6]	[0.2 , 0.5]	
Depressiveness	*n=1366*				
Depr=1 (in %)	12.0 (0.9)	12.7 (1.1)	9.2 (1.5)	3.5 (2.0)	p=0.084
95% CI for Depr=1	[10.3 , 13.7]	[10.5 , 14.8]	[6.2 , 12.2]	[-0.5 , 7.4]	
Ø Mean Index	3.6 (1.0)	3.7 (1.0)	3.5 (1.0)	0.2 (0.1)	p=0.007
95% CI for Mean Index	[3.6 , 3.7]	[3.6 , 3.7]	[3.4 , 3.6]	[0.1 , 0.3]	
n		925	359		

Note: Category “other” and “no answer” on gender not displayed, “Total” for full sample incl. “other” and “no answer” on gender, mean index on a scale of 1 to 7, standard error (for proportions) and standard deviation (for mean indices) in parentheses.

Men feature a slightly higher proportion of SD than women, but this difference is not statistically significant on a 5%-level. For IM, however, there is a considerably and significantly (p=0.004) higher share for female persons. Similar results about gender differences for these two dimensions were found by other authors as well (Becker & Cherny 1994; Winkler et al. 2006).

Regarding the average depressiveness by gender, it becomes evident that female students feature a rather slightly but significantly higher level of depressiveness compared to male students. The dichotomized variable shows that the proportion of persons classified as depressive is more than three percentage points higher, but not significantly, among females.

Indirect Questioning – Full Sample Analysis

Table 5 shows the prevalence rate for the sensitive questions when asking directly as well as the rates that were estimated using the TM. The results show that the indirect questioning model reveals slightly higher percentages for the sensitive

Table 5 Prevalence rates of the sensitive questions

	DQ	TM	Diff.
Use of psychological consultancy			
Prop. (in %)	21.9	22.1	0.2 (p=0.951)
Std. Err.	1.6	2.7	3.0
95% CI	[18.8 , 25.0]	[16.9 , 27.3]	[-5.8 , 6.1]
Misuse of prescriptive medication			
Prop. (in %)	4.2	5.6	1.4 (p=0.403)
Std. Err.	0.8	1.5	1.6
95% CI	[2.7 , 5.7]	[2.7 , 8.5]	[-1.8 , 4.6]
Use of illegal drugs			
Prop. (in %)	3.6	4.7	1.1 (p=0.513)
Std. Err.	0.7	1.5	1.6
95% CI	[2.2 , 5.0]	[1.8 , 7.5]	[-2.0 , 4.1]
n	688	≥ 627	

Note: n for TM: 628, 628, 627; DQ=Direct Questioning, TM=Triangular Model, Prop. (in %)=(Estimated) proportion of "yes"-answers, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

questions. However, none of these differences turn out to be statistically significant. Hence, the TM does not achieve higher estimates when analyzing the total sample of students.

Indirect Questioning – Subgroup Analysis

According to the assumption that a question might only be sensitive for a certain group of people, the TM's effectiveness is checked within subgroups. As stated in the hypotheses section, the analysis is conducted for gender, the two dimensions of social desirability, and depressiveness.

Gender

The results with respect to gender are displayed in Table 6. The TM reveals slightly higher estimates for females but the differences between the survey conditions are small and not statistically significant. Although the differences between TM and DQ are larger for males, the effect is not significant as well. Thus, for these two subgroups, the indirect questioning model could not achieve non-randomly higher

Table 6 Prevalence rates of the sensitive questions by gender

	Female			Male		
	DQ	TM	Diff.	DQ	TM	Diff.
Use of psychological consultancy						
Prop. (in %)	23.4	25.7	2.3 (p=0.535)	16.8	10.8	-6.1 (p=0.283)
Std. Err.	1.9	3.2	3.7	2.7	5.0	5.4
95% CI	[19.6 , 27.2]	[19.5 , 31.9]	[-4.9 , 9.5]	[11.6 , 22.1]	[0.9 , 20.6]	[-16.8 , 4.6]
Misuse of prescriptive medication						
Prop. (in %)	4.8	5.0	0.2 (p=0.932)	2.6	6.3	3.7 (p=0.248)
Std. Err.	1.0	1.7	2.0	1.1	3.0	3.0
95% CI	[2.9 , 6.7]	[1.6 , 8.4]	[-3.7 , 4.0]	[0.3 , 4.8]	[0.4 , 12.1]	[-2.2 , 9.6]
Use of illegal drugs						
Prop. (in %)	2.5	2.7	0.2 (p=0.913)	5.6	10.9	5.3 (p=0.159)
Std. Err.	0.7	1.6	1.7	1.6	3.3	3.5
95% CI	[1.1 , 3.9]	[-0.5 , 5.9]	[-3.2 , 3.6]	[2.4 , 8.9]	[4.3 , 17.4]	[-1.7 , 12.2]
n	478	448		196	163	

Note: DQ=Direct Questioning, TM=Triangular Model, Prop. (in %)=(Estimated) proportion of "yes"-answers, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

Table 7 Differences-in-differences for gender

	DID	p	Std. Err.	95% CI
Use of psychological consultancy	8.4	p=0.227	6.9	[-5.2 , 22.0]
Misuse of prescriptive medication	-3.5	p=0.355	3.8	[-10.9 , 3.9]
Use of illegal drugs	-5.1	p=0.148	3.5	[-12.0 , 1.8]

Note: DID=Differences-in-differences, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

percentages for the sensitive questions. As opposed to the theoretical assumptions, the TM even yielded a lower prevalence rate than DQ among men for the question of psychological consultancy.

Since the TM achieves a higher prevalence rate than DQ for females while yielding a lower rate for males, the DID between females and males amounts to 8.4 percentage points for the first question. As to be seen in Table 7, the discrepancies of the survey conditions' differences between the subgroups are lower for the other two questions and also reversed (the TM achieves higher prevalence rates for men). However, none of these DID reach a sufficient level of statistical significance. Therefore, a systematic influence of gender on the TM's performance cannot be supported.

Social Desirability

Further, the analysis is conducted for the two dimensions of social desirability of which the results are displayed in Table 8 and Table 10. For persons that answer socially desirable according to IM, it becomes evident that the TM achieves higher percentages of persons having the sensitive characteristics. For example, the prevalence rate of using a psychological consultancy is seven percentage points higher when asking the question indirectly using the TM. However, this difference fails to achieve statistical significance. A similar difference can be found for the use of illegal drugs: When asking directly, only 0.8 percent of the persons admit to having used drugs during their studies. When asked using the TM, 6.1 percent in this subgroup state having used illegal drugs to enhance mental performance. However, none of these differences turn out to be statistically significant on a $p \leq 0.05$ level. Regarding the subsample of persons not having the characteristic of IM, no relevant or significant effect of the indirect questioning model can be found.

Although the TM yields higher estimates for the IM=1 group for the first and third question, the DID, as portrayed in Table 9, show no statistical significance. Thus, considering the DID is also in line with the finding that the TM's estimates

Table 8 Prevalence rates of the sensitive questions by social desirability: Impression Management

	IM=1			IM=0		
	DQ	TM	Diff.	DQ	TM	Diff.
Use of psychological consultancy						
Prop. (in %)	20.9	27.9	7.0 (p=0.298)	22.2	20.6	-1.6 (p=0.643)
Std. Err.	3.6	5.7	6.8	1.8	3.0	3.4
95% CI	[13.8 , 28.0]	[16.8 , 39.0]	[-6.5 , 20.5]	[18.7 , 25.6]	[14.7 , 26.5]	[-8.2 , 5.0]
Misuse of prescriptive medication						
Prop. (in %)	2.3	2.4	0.1 (p=0.990)	4.7	6.6	1.9 (p=0.321)
Std. Err.	1.3	2.8	3.2	0.9	1.7	1.9
95% CI	[-0.3 , 5.0]	[-3.1 , 7.9]	[-6.3 , 6.4]	[2.9 , 6.4]	[3.2 , 10.0]	[-1.8 , 5.6]
Use of illegal drugs						
Prop. (in %)	0.8	6,1	5.3 (p=0.105)	4.3	4.3	0.0 (p= 0.999)
Std. Err.	0.8	3.2	3.4	0.9	1.6	1.8
95% CI	[-0,8 , 2.3]	[-0.1 , 12.4]	[-1.4 , 12.1]	[2.6 , 6.0]	[1.1 , 7.5]	[-3.5 , 3.5]
n	129	142		559	≥ 484	

Note: In case of differences, the least number of observations is displayed; n for TM and IM=0: 485, 485, 484; DQ=Direct Questioning, TM=Triangular Model, Prop. (in %)=(Estimated) proportion of "yes"-answers, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

Table 9 Differences-in-differences for social desirability: Impression Management

	DID	p	Std. Err.	95% CI
Use of psychological consultancy	8.6	p=0.238	7.3	[-5.7 , 22.9]
Misuse of prescriptive medication	-1.8	p=0.646	3.9	[-9.5 , 5.9]
Use of illegal drugs	5.3	p=0.164	3.8	[-2.2 , 12.8]

Note: DID=Differences-in-differences, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

Table 10 Prevalence rates of the sensitive questions by social desirability: Self-Deception

	SD=1			SD=0		
	DQ	TM	Diff.	DQ	TM	Diff.
Use of psychological consultancy						
Prop. (in %)	16.8	12.4	-4.4 (p=0.557)	22.8	24.1	1.3 (p=0.712)
Std. Err.	3.7	6.5	7.4	1.7	2.9	3.3
95% CI	[9.4 , 24.3]	[-0.2 , 25.1]	[-19.0 , 10.3]	[19.4 , 26.2]	[18.4 , 29.8]	[-5.2 , 7.7]
Misuse of prescriptive medication						
Prop. (in %)	5.0	1.8	-3,2 (p= 0.419)	4.1	6.4	2.3 (p=0.217)
Std. Err.	2.2	3.3	3.9	0.8	1.7	1.8
95% CI	[0.6 , 9.3]	[-4.7 , 8.2]	[-10.9 , 4.6]	[2.5 , 5.7]	[3.1 , 9.6]	[-1.2 , 5.8]
Use of illegal drugs						
Prop. (in %)	3.0	12.6	9.6 (p=0.042)	3.7	3.2	-0.5 (p= 0.755)
Std. Err.	1.7	4.4	4.7	0.8	1.5	1.7
95% CI	[-0.4 , 6.3]	[4.0 , 21.2]	[0.4 , 18.9]	[2.2 , 5.3]	[0.2 , 6.2]	[-3.8 , 2.7]
n	101	100		587	≥ 526	

Note: In case of differences, the least number of observations is displayed; n for TM and SD=0: 527, 527, 526; DQ=Direct Questioning, TM=Triangular Model, Prop. (in %)=(Estimated) proportion of "yes"-answers, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

do not systematically differ from DQ and there is also no effect that could be traced back to socially desirable answering behavior according to IM.

As stated earlier, it is assumed that effects of the TM could only be found for IM but not for SD, since SD is not a deliberate form of deception. The estimated percentages show that no significant effects can be found for persons that do not feature the characteristic of SD (Table 10) and the differences between the survey conditions are small. For persons in subgroup SD=1, the TM yields lower percentages as DQ for the first two questions but also not on a statistically significant level.

However, there is a considerably and statistically significant higher prevalence rate for use of illegal drugs when using the TM (12.6 percent as compared to 3.0 percent using DQ). In fact, the SD=1 group even shows the highest percentage of drug consumption compared to all other subgroups when asking indirectly. These results are reasonable on the assumption of the personality that is ascribed to per-

Table 11 Differences-in-differences for social desirability: Self-Deception

	DID	p	Std. Err.	95% CI
Use of psychological consultancy	-5.7	p=0.489	8.2	[-21.9 , 10.5]
Misuse of prescriptive medication	-5.5	p=0.219	4.5	[-14.3 , 3.3]
Use of illegal drugs	10.1	p=0.022	4.4	[1.5 , 18.7]

Note: DID=Differences-in-differences, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

sons with a high level of SD: First of all, a certain level of SD characterizes a psychologically stable person and a positive self-image (Winkler et al. 2006, p. 3). This is also reflected in the amount of persons that used a psychological consultancy, which is rather low among persons with SD=1 (16.8 percent). Also, this is supported by a negative correlation between the mean indices for Self-Deception and depressiveness in this sample (r= –0.31, p=0.000). It is conceivable that persons with a high level of SD are also very outgoing and adventurous, thus having a higher tendency toward behavior like drug consumption. Therefore, this question might be especially sensitive to *these* persons because they are the ones that tend to misuse drugs. This could explain why there is a significant effect of the TM for this subgroup although it is not theoretically assumed according to social desirability.

Regarding the discrepancies between the differences in the survey conditions, it is evident for the first and second question that the TM mostly achieves only slightly higher estimates or even lower percentages which is also reflected in the DID (Table 11). As a consequence, for questions 1 and 2, there is no evidence for an influence of SD on the survey conditions' estimates. However, for the question about use of illegal drugs, also the DID shows to be statistically significant on the conventional 5%-level. Therefore, it can be concluded that the TM achieves a higher prevalence rate for persons with a high level of self-deceptive attitudes and there is evidence that the model works differently for these two SD groups.

Depressiveness

As compared to the other subgroups, the prevalence rate of using a psychological consultancy is highest among students that are classified as depressive (35.7 percent). The TM increases this percentage by nearly seven percentage points. Further, the percentage for misuse of prescriptive medication is nearly nine percentage points higher when asking indirectly instead of directly (Table 12). But these differences between the survey conditions are not statistically significant. For the use of illegal drugs, the TM cannot achieve a higher prevalence rate for this subgroup. In fact, the estimation is even slightly lower. Further, there are only marginal and no

Table 12 Prevalence rates of the sensitive questions by depressiveness

	Depr=1			Depr=0		
	DQ	TM	Diff.	DQ	TM	Diff.
Use of psychological consultancy						
Prop. (in %)	35.7	42.5	6.8 (p=0.467)	20.0	19.5	-0.5 (p=0.864)
Std. Err.	5.3	7.7	9.2	1.6	2.8	3.2
95% CI	[25.3 , 46.2]	[27.3 , 57.7]	[-11.3 , 24.9]	[16.8 , 23.2]	[14.0 , 25.0]	[-6.8 , 5.7]
Misuse of prescriptive medication						
Prop. (in %)	7.1	15.6	8.5 (p= 0.162)	3.8	4.3	0.5 (p=0.779)
Std. Err.	2.8	5.3	5.8	0.8	1.5	1.7
95% CI	[1.5 , 12.8]	[5.2 , 26.0]	[-3.1 , 20.0]	[2.3 , 5.3]	[1.3 , 7.3]	[-2.8 , 3.8]
Use of illegal drugs						
Prop. (in %)	7.1	3.8	-3.3 (p=0.512)	3.1	4.8	1.7 (p= 0.327)
Std. Err.	2.8	4.1	4.9	0.7	1.6	1.7
95% CI	1.5 , 12.8]	[-4.1 , 11.9]	[-13.0 , 6.4]	[1.7 , 4.5]	[1.8 , 7.9]	[-1.6 , 5.0]
n	84	75		604	≥ 551	

Note: In case of differences, the least number of observations is displayed; n for TM und Depr=0: 552, 552, 551; Prop. (in %)=(Estimated) proportion of "yes"-answers, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

significant differences between direct and indirect questioning for the subsample of persons that are not depressive.

So although the TM generates higher prevalence rates for the first and second sensitive question, there is no effect of depressiveness on the model's performance as suggested by the DID in Table 13. None of the discrepancies is significant on the conventional level. Therefore, it cannot be concluded that the indirect questioning technique might work differently for persons that are classified as depressed when asking sensitive questions about mental stress.

Table 13 Differences-in-differences for depressiveness

	DID	p	Std. Err.	95% CI
Use of psychological consultancy	7.3	p=0.434	9.3	[-11.0 , 25.6]
Misuse of prescriptive medication	8.0	p=0.116	5.1	[-2.0 , 18.0]
Use of illegal drugs	-5.0	p=0.313	4.9	[-14.7 , 4.7]

Note: DID=Differences-in-differences, Std. Err.=Standard Error, 95% CI=95% Confidence Interval.

Conclusion and Discussion

Regarding the full sample, the analysis revealed that there is no significant difference in the percentages achieved by the TM as compared to DQ. The same results can be found for gender: Although differences were expected for females, no significant higher prevalence rate could be achieved by the TM. Thus, there is no evidence for hypotheses 1 and 2.

Regarding social desirability, the TM could achieve higher percentages in the IM=1 group, but not in a statistically significant way. Although not expected, there is a significant higher prevalence rate for drug use within the group with the characteristic of SD. Testing the DID reveals that this performance of the TM differs significantly in this subgroup. Therefore, hypothesis 3 can be partially supported: An effect can be found for *one* of the dimensions of social desirability but not for the one that was theoretically assumed. Further, the effect can only be found for one of the three questions.

Within the group that is classified as depressed, higher prevalence rates can be found for usage of psychological consultancy and misuse of prescriptive medication, but again not on a sufficient probability level. Thus, no empirical valid support for hypothesis 4 can be found.

In conclusion, the evidence for the postulated assumptions and hypotheses is rather thin. Further, there are some limitations regarding the methodological perspective. First, it has to be stated that the results are not representative and the numbers of observations in the subgroups are small. A sample of university students was used and the mode of data collection was an online survey. Hence, the sample's representativeness is affected by selection through the mail distribution system, through online access, resp. internet affinity, and through self-selection (e.g., willingness to participate in a survey). Therefore, it should be kept in mind that the results are not transferrable to a general population but only to this very specific sample. Hence, there is still the need to evaluate the technique in other, more general samples and with other modes of data collection.

Another criticism – not only in this study but also in general – is that we cannot know whether the participants follow the instructions of the TM. Although it is unknown as well in DQ mode whether the respondents lie or tell the truth, indirect questioning methods might be especially vulnerable to deliberate cheating due to distrust. Very recently, Wu & Tang (2016) discussed noncompliance in NRR-models. They argue that especially the persons that "have the most to lose" (Wu & Tang 2016, p. 2828), i.e., the persons that carry the sensitive characteristic, tend to answer falsely due to distrust in the technique. As mentioned earlier in this paper, the TM has a clear protective answer ("both no") so it might be especially sensitive to cheating that would result in underreporting thus concealing the model's effectiveness. For that reason, the authors introduce the *dual non-randomized response triangular model* (DNRRTM) and the *alternating non-randomized response triangular model* (ANRRTM). In the DNRRTM, the respondents are randomly assigned to two groups where each group gets a different non-sensitive question combined with the sensitive question of interest. Thus, two non-sensitive characteristics with known probabilities are needed. The ANRRTM, however, functions with only one non-sensitive question where the two categories are alternated in the two groups. In a test of their models, Wu & Tang (2016) find that the DNRRTM as well as the ANRRTM provide higher prevalence rates compared to the TM. The authors recommend the ANRRTM since it is easier to implement by using only one innocuous question.

These results are useful regarding the results of the present study. Wu & Tang (2016) argue that the TM underestimates the true prevalence rate due to deliberate cheating especially by those who have the sensitive characteristic. In this study, the main assumption was that the TM is especially efficient for subgroups that are somehow related with the sensitive question or social desirability (e.g., depressed persons and questions about psychological consulting). In conclusion, it would be a possible perspective for future research to combine these two findings and to test the improved ANRRTM with regard to relevant subgroups.

However, indirect questioning models should not be thoughtlessly praised as the indisputable solution for underreporting in studies about sensitive characteristics. Instead, there is also fundamental criticism of such techniques. As already mentioned, empirical evidence on, for example, the RRT is mixed and there is no clear proof for its effectiveness. Actually, Holbrook & Krosnick even question "whether this technique has ever worked properly to achieve its goals" (Holbrook & Krosnick 2010). Further, the effectiveness of indirect questioning methods is mostly judged by the fact whether they can achieve higher estimates than direct questioning. But very recently, Höglinger & Diekmann (2017) as well as Höglinger & Jann (2018) drew attention to false positives (i.e., respondents falsely admitting to having the sensitive characteristic). In their validation studies, they show that the CM produces "false positives to a nonignorable extent" (Höglinger & Diekmann 2017, p. 135) which challenges the assumption that higher estimates are more valid. Even further, it calls into question the CM's good performance that has been suggested in previous studies. It is possible that these studies are biased by these false positives that inflate the model's estimates. Overall et al. (2018, p. 1) summarize that, in their study, none of the three tested indirect questioning models subtantially outperform direct questioning.

In conclusion, the authors speak against relying blindly on the more-is-better-assumption (Höglinger & Diekmann 2017, p. 136) which has been most prominent when examining (non-) randomized response models. Instead, validation strategies should be considered to evaluate indirect questioning models more accurately. In this paper, the validity and performance of the TM was also mainly judged in comparison to direct questioning. Therefore, future studies that evaluate this NRR-model might surely benefit from using validation data as it is suggested in current studies. In summary, the present study cannot deliver evidence for the hypothesis that indirect questioning models might be more effective in certain subgroups but it provides hints that a more precise analysis might be fruitful. We should improve future research on that topic and encourage further theoretical and empirical discussion on randomized and non-randomized response models.

References

Abernathy, J. R., Greenberg, B. G., & Horvitz, D. G. (1970). Estimates of Induced Abortion in Urban North Carolina. *Demography, 7(1)*, 19–29.

Barton, A. H. (1958). Asking the Embarrassing Question. *Public Opinion Quarterly, 22(1)*, 67–68.

Becker, G., & Cherny, S. S. (1994). Gender-controlled measures of socially desirable responding. *Journal of Clinical Psychology, 50(5)*, 746–752. https://doi.org/10.1002/1097-4679(199409)50:5<746::AID-JCLP2270500512>3.0.CO;2-V

Beldt, S. F., Daniel, W. W., & Garcha, B. S. (2016). The Takahasi-Sakasegawa Randomized Response Technique. *Sociological Methods & Research, 11(1)*, 101–111. https://doi.org/10.1177/004912418201100 1006

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting Voting: Why It Happens And Why It Matters. *Public Opinion Quarterly, 65*, 22–44.

Bertrand, M., Duflo, E., & Mullainathan, S. (2003). How Much Should We Trust Differences-In-Differences Estimates? NBER Working Paper Series, 8841. Cambridge, Massachusetts: National Bureau of Economic Research.

Bortz, J. (1989). *Statistik für Sozialwissenschaftler*. Heidelberg: Springer.

Buchman, T. A., & Tracy, J. A. (1982). Obtaining Responses to Sensitive Questions: Conventional Questionnaire versus Randomized Response Technique. *Journal of Accounting Research, 20(1)*, 263–271.

Burger, P. H. M., & Scholz, M. (2018). Gender as an underestimated factor in mental health of medical students. Annals of Anatomy = Anatomischer Anzeiger : *Official Organ of the Anatomische Gesellschaft, 218*, 1–6. https://doi.org/10.1016/j.aanat.2018.02.005

Coutts, E., & Jann, B. (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research, 40(1)*, 169–193.

Crowne, D. P., & Marlowe, D. (1960). A new Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology, 24(4)*, 349–354.

Dalton, D., & Ortegren, M. (2011). Gender Differences in Ethics Research: The Importance of Controlling for the Social Desirability Response Bias. *Journal of Business Ethics, 103(1)*, 73–93. https://doi.org/10.1007/s10551-011-0843-8

Droitcour Miller, J. (1981). Complexities of the Randomized Response Solution. *American Sociological Review, 46(6)*, 928–930.

Fox, J. A., & Tracy, P. E. (1986). Randomized Response. A Method for Sensitive Surveys. Beverly Hills, California: Sage Publications (A Sage University Papers Series. Quantitative Applications in the Social Sciences, No. 07-058).

Häder, M. (2015). *Empirische Sozialforschung: Eine Einführung* (3. Aufl.). Wiesbaden: Springer VS. Retrieved from http://dx.doi.org/10.1007/978-3-531-19675-6

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods, 48(3)*, 1032–1046. https://doi.org/10.3758/s13428-015-0628-6

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis, 25(01)*, 131–137. https://doi.org/10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PloS One, 13(8)*, e0201770. https://doi.org/10.1371/journal.pone.0201770

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring Voter Turnout By Using The Randomized Response Technique: Evidence Calling Into Question The Method's Validity. *Public Opinion Quarterly, 74(2)*, 328–343. https://doi.org/10.1093/poq/nfq012

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The Unrelated Question Randomized Response Model, 65–72.

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism. *Public Opinion Quarterly, 76(1)*, 32–49. https://doi.org/10.1093/poq/nfr036

Jerke, J., & Krumpal, I. (2013). Plagiarism in Student Papers: An Empirical Study Using the Triangular Model. *Methoden, Daten, Analysen, 7(3)*, 347–368. https://doi.org/10.12758/mda.2013.017

Kirchner, A., Krumpal, I., Trappmann, M., & Hermanni, H. von. (2013). Messung und Erklärung von Schwarzarbeit in Deutschland: Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit. *Zeitschrift Für Soziologie, 42(4)*, 291–314.

Krumpal, I., & Näher, A.-F. (2012). Entstehungsbedingungen sozial erwünschten Antwortverhaltens. Eine experimentelle Onlinestudie zum Einfluss des Wording und des Kontexts bei unangenehmen Fragen. *Soziale Welt, 63*, 65–89.

Kundt, T. C., Misch, F., & Nerré, B. (2013). Re-assessing the merits of measuring tax evasions through surveys: Evidence from Serbian firms. *ZEW Discussion Papers*, No. 13-047.

Lara, D., Strickler, J., Olavarrieta, C. D., & Ellertson, C. (2016). Measuring Induced Abortion in Mexico. *Sociological Methods & Research, 32(4)*, 529–558. https://doi.org/10.1177/0049124103262685

Liu, Y., & Tian, G.-L. (2014). Sample size determination for the parallel model in a survey with sensitive questions. *Journal of the Korean Statistical Society, 43(2)*, 235–249. https://doi.org/10.1016/j.jkss.2013.08.002

Lück, H., & Timaeus, E. (1997a). Soziale Erwünschtheit (SDS-E).

Lück, H., & Timaeus, E. (1997b). Soziale Erwünschtheit SDS-CM.

Lück, H. E., & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM). Diagnostica, 15, 134–141.

Margitics, F., & Pauwlik, Z. (2009). *Depression, subjective well-being, and individual aspirations of college students*. New York: Nova Science Publishers. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=281161

Mohr, G., & Müller, A. (2014). Depressivität im nichtklinischen Kontext. In D. Danner & A. Glöckner-Rist (Eds.), *Zusammenstellung sozialwissenschaftlicher Items und Skalen*.

Paulhus, D. L. (1984). Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology, 46(3)*, 598–609.

Pitsch, W., Emrich, E., & Pierdzioch, C. (2012). Match Fixing im deutschen Fussball: Eine empirische Analyse mittels der Randomized-Response-Technik. Diskussions-Papier. Helmut-Schmidt-Universität. Fächergruppe Volkswirtschaftslehre. Nummer 120.

Porst, R. (2009). *Fragebogen. Ein Arbeitsbuch* (1.th ed.). Wiesbaden: VS Verlag für Sozialwissenschaften (Studienskripten zur Soziologie).

Preisendörfer, P. (2008). Heikle Fragen in mündlichen Interviews: Ergebnisse einer Methodenstudie im studentischen Milieu.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität. Diagnostica, 45(4), 173–177.

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent Validity, Dis-criminant Validity, and Relationship with Age. *European Journal of Psychological Assessment, 17(3)*, 222–232.

Tang, M.-L., Wu, Q., Tian, G.-L., & Guo, J.-H. (2013). Two-sample Non Randomized Response Techniques for Sensitive Questions. *Communications in Statistics - Theory and Methods, 43(2)*, 408–425. https://doi.org/10.1080/03610926.2012.657323

Tian, G.-L. (2014). A new non-randomized response model: The parallel model. *Statistica Neerlandica, 68(4)*, 293–323. https://doi.org/10.1111/stan.12034

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133(5)*, 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods, 17(4)*, 623–641. https://doi.org/10.1037/a0029314

Van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2016). A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning. *Sociological Methods & Research, 28(4)*, 505–537. https://doi.org/10.1177/0049124100028004005

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60(309)*, 63–69.

Winkler, N., Kroh, M., & Spiess, M. (2006). Entwicklung einer deutschen Kurzskala zur zweidimensionalen Messung von sozialer Erwünschtheit. DIW Discussion Papers (579).

Wolter, F. (2012). *Heikle Fragen in Interviews*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Wolter, F., & Preisendörfer, P. (2013). Asking Sensitive Questions. *Sociological Methods & Research, 42(3)*, 321–353. https://doi.org/10.1177/0049124113500474

Wu, Q., & Tang, M.-L. (2016). Non-Randomized Response Model for Sensitive Survey with Noncompliance. *Statistical Methods in Medical Research, 25(6)*, 2827–2839. https://doi.org/10.1177/0962280214533022

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika, 67(3)*, 251–263. https://doi.org/10.1007/s00184-007-0131-x

A New Version of the Item Count Technique for Asking Sensitive Questions: Testing the Performance of the Person Count Technique

Felix Wolter

Johannes Gutenberg University Mainz, Institute of Sociology

Abstract

This paper presents empirical evidence on a recent advancement of the item count technique (ICT, a survey technique for asking sensitive questions), namely, the person count technique (PCT; Grant, Moon, & Gleason, 2014). PCT utilizes person lists instead of lists of filler questions, as is the case in the classic ICT design. This simplifies the questioning procedure, but leads to some methodological challenges such as floor and ceiling effects. The main part of this paper presents empirical evidence stemming from an experimental postal survey in Germany (N = 580) investigating how well PCT performs as compared to standard direct questioning (DQ) with regard to alleviating misreporting for questions on attitudes towards refugees.

PCT prevalence estimates for hostile attitudes towards refugees are significantly higher than DQ estimates for one item, and non-significantly higher for three items. Although not consistently significant, the differences are substantial, amounting to a threefold increase of the proportion of respondents expressing negative attitudes towards refugees. Even though the findings are not unequivocally in favor of PCT, this new ICT variant still deserves consideration in the future and warrants further development. Specifically, more knowledge is required with respect to its statistical properties and the best practices of its implementation.

Keywords: Sensitive questions, response bias, misreporting, item count technique, person count technique, refugees, xenophobia.

Background and Research Question

The issue of so-called sensitive questions has occupied survey methodology for several decades (Barton 1958; Hyman 1944; Krumpal 2013; Tourangeau & Yan 2007). It is a well-established fact that respondents, when answering survey questions on socially undesirable or desirable behaviors or attitudes, tend to tailor their answers in a socially desirable manner rather than answering truthfully. This pertains to questions on socially loaded behaviors (e.g., self-reported delinquency, voting behavior, or substance abuse), attitudes (e.g., xenophobia or homophobia), as well as other personal traits (e.g., health issues or personality characteristics). Generally speaking and according to Tourangeau and Yan (2007, p. 860), sensitive questions in surveys can be defined as questions which are private or intrusive, which pose a threat of disclosure for the respondent, and/or touch upon socially undesirable or desirable topics. The primary problem of misreporting on such questions by respondents in standard survey settings is that prevalence estimates of sensitive behaviors or attitudes will be biased. For example, Bradburn and Sudman (1979, p. 24) compare survey estimates of self-reported alcohol consumption with official sales figures, finding that "reported beer, wine, and liquor consumption [...] reaches only 51, 67, and 36 percent of the taxed sales figures, respectively". Furthermore, correlations between the sensitive issue under investigation and its determinants are also biased if the likelihood of misreporting is related to the determinants (Ganster, Hennessey, & Luthans 1983). Yet another issue when asking sensitive questions in surveys is item-nonresponse, which occurs if respondents refuse to answer the respective question at all. While this is a well-known phenomenon concerning questions on income (Moore, Stinson, & Welniak 2000; Yan, Curtin, & Jans 2010), empirical evidence is less consistent for sensitive questions on other topics

Editorial note

The author of this article is one of the guest editors of this MDA special issue. To avoid conflict of interest, the article has been handled by the other two guest editors and the author has not been involved in any of the editorial work related to this article (such as selecting and inviting reviewers, deciding on revisions, and accepting the manuscript for publication). The identities of the reviewers have not been revealed to the author.

Acknowledgments

For helpful comments, I would like to thank Ben Jann, Peter Preisendörfer, Justus Junkermann, Bastian Laier, who came up with the idea of "fixed person counts", and two anonymous reviewers of mda. I would also like to thank Jürgen Schiener and our students of the teaching project in which the data were collected.

Direct correspondence to

Dr. Felix Wolter, Institut für Soziologie, Johannes Gutenberg-Universität Mainz,
Jakob-Welder-Weg 12, D-55128 Mainz, Tel.: +49 (0) 6131 39 20831
felix.wolter@uni-mainz.de

(Tourangeau & Yan 2007, p. 862). This could be because respondents may interpret an answer refusal as an "admission of guilt".

In order to tackle the problem of misreporting (and item-nonresponse), survey methodologists have come up with a number of special questioning techniques. Conventional approaches encompass, for instance, anonymity assurances, "forgiving wording", or the sealed envelope technique (Benson 1941; Perry 1979). A more elaborate procedure is the randomized response technique (RRT; Fox & Tracy 1986; Warner 1965), which has probably gained the most attention in the methodological literature on sensitive questions in surveys. However, RRT procedures in surveys are usually complicated both for respondents and for interviewers. Moreover, doubts have been raised regarding the efficacy of RRT in avoiding response biases (Wolter & Preisendörfer 2013). An alternative to RRT is the item count technique (ICT, also referred to as list experiment or unmatched count technique; Droitcour et al. 1991; Kuklinski, Cobb, & Gilens 1997), which has attracted increased interest within the research community in recent years.

As with RRT, the idea behind ICT is the anonymization of the interview situation by adding noise to the data concealing the respondents' answers. This is achieved by randomly splitting the sample into (at least) two groups. One group, the "short-list group", receives a list of binary yes-no questions which are "harmless" and function as filler items (i.e., they are not important with regard to their content). The other group, the "long-list group", receives the same list of non-key items, but this time, the list additionally contains the (binary) sensitive item of interest. Respondents in both groups are asked not to answer each item individually, but rather to merely report the number of "yes" answers to the whole list. Therefore, the individual answer to the sensitive item is not disclosed to anyone, not even the interviewer (unless ceiling or floor effects occur, see below). For the whole sample, however, it is possible to calculate an estimate of the prevalence of the sensitive item by simply subtracting the mean of the short list from the mean of the long list. This classic ICT design for binary yes-no questions has recently been expanded upon with a version called item sum technique (IST; Trappmann et al. 2014; Wolter & Herold 2018), designed for quantitative sensitive items (such as the frequency of drug usage).

The person count technique (PCT) is another new variant of the classic ICT approach, originally proposed by Grant et al. (2014). PCT also applies to binary sensitive items, but instead of using lists of filler questions, it utilizes lists of persons. The short list is a number of people, and respondents are asked to report the number of persons for whom something applies. The long list corresponds to a list of persons as well, but also contains the respondent himself or herself.

This study presents empirical evidence on the performance of PCT as compared to standard direct questioning (DQ) with regard to alleviating misreporting on sensitive questions. To my knowledge, apart from the original (unpublished) study by Grant et al. (2014), there exists, as yet, no published research investigating the performance of the only just recently proposed PCT. The empirical data presented here were gathered in a postal survey of N = 580 respondents in the City of Mainz, Germany. The PCT-DQ comparison is investigated for four questions on attitudes towards refugees/asylum seekers in Germany. According to the literature (Krumpal 2012; Stocké 2007), expressing negative or hostile attitudes towards immigrants is prone to underreporting. Therefore, due to the enhanced anonymity in PCT mode as compared to DQ, self-reports on hostile attitudes towards refugees should be higher in PCT mode as compared to DQ mode (and, if item-nonresponse is a problem, it should be lower in PCT mode than in DQ mode).

The structure of this article is as follows: The next section will give a brief overview of methodological research on response biases pertaining to attitude questions about immigrants. Afterwards, I will first present the principles of ICT and PCT in more detail, followed by a discussion of methodological aspects and some general pros and cons of PCT vis-à-vis ICT. The "Study Design and Methods" section is devoted to the description of the survey design and some issues of the statistical analyses. The "Results" section depicts the results regarding the PCT-DQ comparison, which are subsequently discussed within a broader framework in the final "Discussion" section.

Social Desirability Bias in Research about Xenophobia

There is a long tradition of research on anti-immigrant or xenophobic attitudes in the social sciences (Allport 1954; Czymara & Schmidt-Catran 2016; Quillian 1995; Weins 2011, to cite but a few). One of the motivations driving this literature is the public and scientific concern regarding political extremism, or, more specifically, regarding voting for (right wing) extremist parties in elections, for which anti-immigrant attitudes are seen as a major influencing factor (Arzheimer 2008). Studying the causes and consequences of xenophobia, however, requires a valid measurement of these attitudes. Several authors have argued that survey estimates from questions on anti-immigrant attitudes are prone to social desirability bias (An 2015; Cea D'Ancona 2014; Janus 2010; Krumpal 2012; Stocké 2007). Since there are social norms inhibiting the public utterance of such attitudes or opinions, some respondents may seek to avoid expressing them in survey interviews. This leads to the underreporting and underestimation of xenophobic attitudes.

In contrast to other (behavioral) sensitive issues, studying misreporting on attitude questions such as on xenophobia is not straightforward with respect to the level of response bias, because a "true value" cannot be observed (by using external validation records, for instance). Hence, in order to assess the amount of social desirability bias, existing studies concentrate on comparing varying estimates according to different questioning techniques or survey modes. The ensuing evaluation is then carried out relying on the “more is better” assumption, which means that for socially undesirable traits like anti-immigrant attitudes, higher estimates are considered to be more valid than lower ones.

There are three studies comparing DQ and RRT estimates. Krumpal (2012) finds a significant improvement due to RRT for one out of three items on xenophobia, the prevalence of respondents expressing a xenophobic attitude amounting to 27 percent in DQ mode and to 35 percent in RRT mode. The estimates for the remaining two items are virtually the same in both question formats and amount to about 40 and 30 percent, respectively. Ostapczuk, Musch, and Moshagen (2009) observe a non-significant difference between a DQ and an RRT question on xenophobia. Depending on the education level of the respondents, the prevalence estimates of expressing a xenophobic attitude range from 25 to 45 percent in DQ mode and from 47 to 76 percent in RRT mode. Finally, Hoffmann and Musch (2016) compare the crosswise-RRT, an adjusted version of RRT (Yu, Tian, & Tang 2008), with DQ for one item on xenophobia and one on islamophobia. They observe significantly higher estimates (49 versus 27 percent) using the crosswise model for the first item, but not for the second (52 versus 43 percent).

Studies investigating the effect of ICT on self-reports of anti-immigrant attitudes have also been conducted. An (2015) finds that, when asked directly, around 59 percent (depending on education) of the respondents are against “cutting off immigration to the United States”. When asked using ICT, this fraction shrinks significantly to around 33 percent. Significant differences between DQ and ICT have also been reported by Janus (2010) for the same item (58 vs. 39 percent), and by Cappelen and Midtbø (2016) for an item on welfare benefits for immigrants in Norway. The study by Creighton and Jamal (2015), in contrast, yields mixed results with respect to the DQ-ICT comparison. While there is no difference for an item on “granting citizenship to a legal immigrant who is Muslim”, a significant difference (28 vs. 11 percent) was observed for “granting citizenship to a legal immigrant who is Christian”.

In sum, empirical research clearly shows that survey questions on anti-immigrant or xenophobic attitudes suffer from social desirability bias. The evidence regarding the performance of special survey techniques such as RRT or ICT to alleviate this problem, however, is mixed. The remainder of this article will present evidence on the performance of PCT in this regard.

Person Count: A Recent Advancement of the Item Count Technique

As explained above, the basic idea of ICT and PCT lies in concealing respondents' answers to binary sensitive survey questions by overlaying the data with noise. This noise is created by adding information about respondents' answers to other filler items (ICT) or third persons (PCT) to the individual answer to the sensitive item. Both ICT and PCT require a random split of the sample into a short-list group and a long-list group. When using ICT, respondents in the short-list group receive a list of harmless yes-no items, for example (Wolter & Laier 2014): "Below you see a list of four questions. Please indicate only the number of questions you answer with 'yes', thus, a number between zero and four. 1. Have you ever been abroad? 2. Have you ever used a taxi? 3. Have you used a plane this week? 4. Did you wash your car this week?". Respondents in the long-list group receive a list containing the same four non-key items plus the sensitive item of interest, for example "Have you ever driven a car while drunk?". Again, respondents are asked to only report the number of items they answer with "yes". In doing so, the individual answer to the sensitive item is not disclosed. Of course, this is true only if no ceiling or floor effects occur, i.e., if the respondent does not negate all items in the list or reports that all items apply. In order to avoid ceiling and floor effects, the non-key items should contain both low-prevalence and high-prevalence questions which ideally are negatively correlated among each other (Droitcour et al. 1991).

The PCT replaces the list of filler questions with a list of persons, and respondents are asked to report the number of persons for whom (they think that) something (sensitive) applies. In the short-list group, the list only consists of other uninvolved people; in the long-list group the respondent himself is added to the list; respondents report the number of persons for whom something applies including themselves. In the original proposition by Grant et al. (2014, p. 11–12) respondents were asked the following question: "We want to know what type of candidates people would support for President of the United States. Because this is a sensitive topic, we are not going to single you out. Instead, please think about three people you see or talk to often and we're going to ask you how many of these three people might be willing to vote for each type of candidate. We're going to ask about five candidates: a Republican, a Democrat, a Tea Party candidate, a Mormon, and a woman. It's ok to guess if you are not sure how many of the three people would vote for each candidate. […]" In the short-list group the introduction subsequently read "Thinking of these three people, how many would be willing to vote for [a republican, a democrat, a woman etc.]", while in the long-list group, it read "Thinking of you and these three people […]".

For both the basic ICT and the PCT, a prevalence estimate of the sensitive item $\hat{\pi}$ and its standard error can be calculated using the formulae (1) and (2) below, provided that the short-list and long-list samples are independent. $\overline{x}_{LL}$ and $\overline{x}_{SL}$ represent the mean of the reported numbers in the long-list and short-list group, and $Var\left(\overline{x}\right)$ the sampling variance of the mean estimate.

$$\hat{\pi} = \overline{x}_{LL} - \overline{x}_{SL} \tag{1}$$

$$S.E.\left(\hat{\pi}\right) = \sqrt{Var\left(\overline{x}_{LL}\right) + Var\left(\overline{x}_{SL}\right)} \tag{2}$$

One advantage of the PCT design vis-à-vis ICT is that having one list of persons means that many sensitive items can be asked at once in the same survey. With ICT, a different item list is required for every sensitive item due to anonymity concerns (or, as Grant et al. 2014, p. 6, put it, an additional random split of the sample for every additional sensitive question, when using the same short list for every item). Also, no fabrication of artificial filler items is necessary with PCT, which could, in turn, simplify the answering process for the interviewees because they only have to deal with one question instead of a question list. But this, of course, has to be investigated empirically. One should also note that respondents may not be certain whether the trait being asked about applies to one or more of the uninvolved persons in the list. As cited above, Grant et al. (2014) try to solve this problem by prompting respondents "to guess if you are not sure". If the interviewees follow this instruction, possible errors in judging about the status of the "other persons" represent no problem for the validity of the PCT estimate because, due to the experimental design (random split into short-list and long-list), the errors in both groups will be equal (Grant et al. 2014, p. 19) – provided that there are no design effects (see below). There are, however, some other challenges inherent to the PCT design, namely floor and ceiling effects, statistical power issues, and design effects. These challenges share (at least to some extent) a common cause, namely homophily effects, which I shall discuss first.

Homophily refers to the "similarity between socially connected individuals" (Shakya, Christakis, & Fowler 2017, p. 158). It is a well-established fact that similar people have a higher tendency to be socially connected than dissimilar people. This applies with respect to a variety of socio-demographic, behavioral, and attitudinal characteristics (McPherson, Smith-Lovin & Cook, 2001), including possibly sensitive traits such as marihuana consumption, political orientation, and delinquency (Kandel 1978; South & Felson 1990). One consequence of homophily regarding PCT is that it will affect the composition and characteristics of the "other persons": When asked to think of some people whom they know, respondents probably unconsciously choose people who are similar to themselves, or at least more similar than a random choice would be. As the cited literature shows, this will also hold for the sensitive traits being asked about in the PCT procedure. Another, related argument is that the choice of the "other persons" may be guided by the question content and context (certain stimuli make respondents think of certain types of people). For instance, if the survey question deals with substance abuse, a respondent who smokes marihuana is probably going to imagine a list of "other persons" who are also inclined to smoke marihuana. This conjecture is supported by empirical evidence from social network research on name generators, which shows that question content and context exert an influence on the data generated by name generators in survey settings (e.g., Ferligoj & Hlebec 1999; Shakya et al. 2017). One finding of this research is also that individuals have different networks for different issues: "A person with whom someone discusses politics may not be the person upon whom they rely for assistance with a sick child" (Shakya et al. 2017, p. 158). A third argument for the occurrence of homophily effects (referring to values or attitudes) in PCT designs is derived from research showing that actors often subjectively overestimate the degree to which their acquaintances are similar to them (Huckfeldt & Sprague 1995): "People tend to assume that their friends are like them, when in fact areas of disagreement simply are not discussed" (McPherson et al. 2001, p. 429). Hence, when asked about characteristics of their acquaintances in PCT procedures, respondents may ascribe similar traits to the "other persons" even if this is objectively not the case.

In short, when using PCT we should expect that respondents generate lists of uninvolved persons that, due to homophily, share similar characteristics as themselves. This is probably further reinforced by framing effects of the question content and context, and by a subjectively overestimated degree of similarity by the respondents.

A first consequence of homophily effects with respect to PCT concerns floor and ceiling effects. As already pointed out, floor and ceiling effects occur if respondents either deny or affirm all items (persons) in the list. In this case, the anonymity of the procedure is negated. When using ICT, this can be avoided by a proper choice of the non-key items (negatively correlated high- and low-prevalence items), which is generally under the control of the researcher. When using PCT, floor and ceiling effects are likely to occur more often than with ICT because of homophily. Moreover, they are not as easily controllable as in the basic ICT design, because the choice of the uninvolved persons is not under the control of the researcher – at least in the PCT version proposed by Grant et al. (2014; see the discussion section below for a suggestion of how to possibly advance with this issue). My – preliminary – suggestion regarding the problem of floor and ceiling effects in the PCT design is to instruct respondents in a way that induces them to choose "other persons" that are as different as possible, and to carefully study floor and ceiling issues empirically both in the pretest phase of the survey and with respect to its main results. Also, one should take care not to introduce PCT as a "completely anonymizing technique" to respondents. If floor and ceiling effects occur, respondents may feel cheated by the survey authors.

Another consequence of homophily effects, directly related to the issue of floor and ceiling effects, are issues of statistical power. One main drawback of all ICT designs is that they always produce larger standard errors than conventional estimates. This is obvious, because noise is artificially added to the data. The amount and the statistical properties of this noise affect the statistical efficiency of ICT estimates, which means that design aspects of the ICT/PCT procedure affect statistical efficiency and that there is a trade-off between efficiency and respondent protection (Coutts & Jann 2011; Trappmann et al. 2014). The standard errors of ICT estimates depend on (among other things) the number of non-key items (or the number of uninvolved persons in the PCT procedure), their prevalence, and the covariance between the sensitive item and the filler items (see, for example, Corstange 2009; Trappmann et al. 2014 for a more detailed discussion). For a high level of statistical efficiency, it is desirable that the variance of the short list (non-key items) is small. To achieve this, it is preferable that the number of non-key items or "other persons" is low, that they have a prevalence near 0 or 1 (low variance), that they are negatively correlated with the sensitive item, and also negatively correlated among each other. Homophily among the uninvolved persons and the respondents themselves counteracts these ideals, because it causes high variance in the answers (people will tend to cluster at the minimum and maximum), and thus a large PCT standard error. In the basic ICT design, these features can be controlled by an appropriate and careful choice of the non-key items. For PCT, things are more difficult, because the researcher does not choose the "other persons" whom the respondents are asked to imagine. Hence, it is only the length of the short list (the number of uninvolved persons) that is directly controllable by design. As, for example, Wolter and Laier (2014, p. 155) recommend with respect to the ICT literature, a list length of three to five non-key items seems to be a good choice.

Another problem that could be more pronounced in the case of PCT than with ICT are what Blair and Imai (2012) call design effects. Both ICT and PCT rely on the assumption that respondents' answers to the non-key items or the "other persons" do not change if the sensitive item or the respondent himself is added to the long-list group. If this happens, the mean difference of the short-list and long-list group is not exclusively determined by the sensitive item under concern (the addition of the respondent to the list in the PCT procedure), and the prevalence estimate is biased. When using PCT, the respondent's own status for the sensitive item might, for example, affect his or her assessment of the status of the other persons in the list, causing design effects. This again would be an effect of (misperceived) homophily. With respect to this potential issue, further research including qualitative studies and cognitive pretesting should examine the likelihood of such design effects. Pragmatically, Blair and Imai (2012) propose a statistical test empirically testing whether design effects have occurred.

One further constraint of PCT is that the so-called double list design cannot be implemented in a straightforward manner. Double list designs (Biemer et al. 2005; Droitcour et al. 1991) can improve the efficiency of ICT estimates considerably. The double list procedure administers a second short list of non-key questions to the respondents. Those in the original short-list group receive this second list including the (same) sensitive item; respondents in the former long-list group answer the second list without the sensitive item. The estimates from both lists can then be combined, resulting in lower standard errors than with only one list of innocuous questions. With PCT, this logic does not work because there is only one short list of "other persons". A remedy would be to introduce a second list of (different) people, but this seems to overcomplicate matters.

The issues discussed above reveal that the newly proposed PCT brings some challenges with it requiring further methodological and empirical research on how design aspects of PCT procedures affect the mechanisms at work and the statistical properties of the resulting estimates. This research should clarify whether the gain in simplicity of PCT vis-à-vis ICT outweighs the difficulties inherent to PCT and whether and how these problems can be resolved.

Besides these statistical aspects of ICT and PCT designs, the essential purpose and main goal of using these techniques remains achieving valid survey responses. With respect to ICT, a comprehensive meta-analysis of studies investigating the efficacy of ICT procedures with regard to avoiding or alleviating response bias is, to my knowledge, still lacking. Existing (summary) studies, however, do point, at least partially, to the result that ICT is successful in reducing response bias:[1] A small meta-analysis by Tourangeau and Yan (2007) of seven studies in which ICT was compared to DQ finds an overall positive, but non-significant ICT effect. A literature review by Wolter and Laier (2014) counts 22 comparative studies, of which 17 find results that are at least partially in favor of ICT. Two studies with aggregate external validation data in the field of voting behavior (and self-reporting on it) both find that ICT performs better than DQ with respect to response bias, but ICT estimates are still off the mark with regard to the externally validated true value (Comşa & Postelnicu 2013; Rosenfeld, Imai, & Shapiro 2015).

In terms of PCT, Grant et al. (2014) themselves provide a first empirical assessment of its performance as compared to DQ. In a telephone survey among registered voters in Illinois, respondents were asked about their intentions to vote for certain types of candidates in presidential elections. The PCT design corresponds to the one introduced above in this paper. The authors first find significant

[1] However, it should also be noted that this does not mean ICT should be taken for granted as a universal remedy for all problems induced by sensitive questions. See, for example, Thomas, Johann, Kritzinger, Plescia, and Zeglovits (2017) for a critical study.

evidence for design effects regarding the Republican candidate item (which, for this reason, is not analyzed any further in the rest of the paper), and no evidence for design effects for the other four items. Second, PCT estimates of respondents claiming to be ready to vote for the respective type of presidential candidate are significantly lower than their DQ counterparts regarding the Democrat, female, and Mormon candidate (with a difference of about 20 percentage points). This is in line with the hypothesis that survey respondents, due to social desirability, claim to be open-minded and devoid of prejudice when asked directly, which results in overreporting in this case. For the latter item ("tea party member"), no difference is found between question modes.

Study Design and Methods

Survey Design

The PCT-DQ comparison for attitudes towards refugees was part of a local postal survey in the city of Mainz (Germany). The survey went by the title "Living and Residing in Mainz" and contained questions on a variety of topics: of the two main parts of the questionnaire, one was devoted to environmental problems, the other to attitudes and behaviors regarding foreigners and refugees/asylum seekers. Field work was carried out in autumn 2016. It should be noted with regard to the topic of refugees that within this period of 2015/2016, large numbers of asylum seekers, mainly from Syria and Afghanistan, came to Germany, which, in turn, created considerable concern and tension in the political debate and among parts of the German population.

Because one aim of the survey (not related to the topic of this paper) consisted in obtaining georeferenced data, we employed a special sampling design. Following an idea of Bauer (2014), we conducted a street section sample. Using GIS software for geographical data, we first identified all residential areas within the municipal area of Mainz and then randomly distributed 200 sampling points within these areas. For each of these (preliminary) sampling points, we then established the geographically nearest street sections, street section referring to the section between two street intersections (footways included). We then counted the number of households in each street section, yielding a number of 11,208 households. Another random sample of 68 street sections was then drawn from the original 200 sampling points, containing about 4,000 households.[2] Finally, every second household was

[2] This procedure was necessary because the number of households in each street section was not known in advance. At the same time and for the purpose of other planned (multilevel) analyses, the number of cases in each sampling point had to be sufficiently high. Hence, we applied the two-step procedure of drawing 200 initial sampling points

manually assigned a questionnaire package. The package included a cover letter and a stamped envelope in order to send back the filled-out questionnaires without postage costs. We used the next-birthday method to randomly choose an adult person within each household. This sampling design leads to the selection probability decreasing for persons in larger households. However, I abstain from using design weights for the analyses, since the main goal of this study is the experimental comparison between DQ and PCT.

Out of 2,000 distributed questionnaires, 580 were returned, which corresponds to an AAPOR response rate of 29 percent (RR2). Because this study was a pilot study within the framework of a teaching project with MA students in sociology and, therefore, without funding, we were not able to dispatch follow-up letters or questionnaires to respondents who did not reply after the initial distribution of questionnaires.

The survey featured an experimental split into two subsamples. One half of the respondents were assigned to the PCT version of the questions on refugees, the other half to the DQ version. The DQ version also contained the short list of the PCT design. Normally, one would prefer to form three subgroups (DQ, short list, long list), but due to the financial restrictions of this study, we chose not to in order to ensure a sufficiently high number of cases in each group. However, this means that the samples yielding the DQ and PCT estimates are not independent from one another, which in turn requires special statistical procedures for the empirical analysis (see below).

In the analysis sample, 49 percent of all cases are in the DQ/PCT short-list group and 51 percent are in the PCT long-list group. This corresponds almost exactly to the 50-50 partitioning envisaged by the design. Table 1 reports the distribution of some socio-demographic variables by question format. There are no significant differences between the two experimental groups, meaning the randomization worked as intended. Women are slightly over-represented in the sample, as are people with higher education.

PCT Procedures

The PCT procedure was located roughly in the middle of the questionnaire within a block of various questions on attitudes, contact, and behaviors vis-à-vis refugees and immigrants in general. The PCT questions were devoted to aspects regarding refugees in the city of Mainz. The exact question wording for the long-list group

first, counting the households, and then drawing a subsample in order to meet the predefined distributional criteria by simultaneously not exceeding the projected sample size of 2000 contacts. Counting was carried out manually on location by sociology students.

Table 1 Distribution of Socio-Demographic Variables by Question Mode

	All	DQ	PCT	t	n
Gender (0 = male, 1 = female)	56.0	54.1	57.8	0.88	568
Age	49.6	50.1	49.1	0.56	564
Years of education	14.1	14.2	14.0	0.59	545
Social status (subj., [1…10])	6.3	6.3	6.3	0.36	570
House owner (0 = no, 1 = yes)	39.0	38.9	39.1	0.04	569
Married (0 = no, 1 = yes)	43.3	40.3	46.2	1.41	566

Note: DQ = direct questioning, PCT = person count technique. Reported are percent values (categorical variables) and means (metric variables). Differences between experimental groups were tested using t-tests (assuming equal variances).

and the four sensitive items are depicted in Figure 1 (translated from the German original).

There are three things to note on this design. First, the instruction asked respondents about “preferably diverse persons”. This was done in order to avoid homophily effects and, thus, to reduce the likelihood of floor and ceiling effects. Second, the design asked respondents to write down the initials of the first names of their imagined persons. On the one hand, pretests had shown that this helps respondents in coping with the questioning procedure. On the other hand, it is desirable that respondents do not switch around the people they are thinking of depending on the question content or the respondent’s own opinion (or for other reasons such as lack of knowledge about the persons of whom they initially thought). Of course, this is not a problem as long as the switching behavior is similar in both groups. However, the stimulus of including oneself in the long-list group might result in a different manner of switching and, hence, trigger design effects and biased results. By letting respondents write down the initials of their imagined persons we hoped to avoid this. Third, we did not introduce the PCT procedure as an “anonymizing technique” for “sensitive questions” or the like in order to avoid the respondents framing them in the sense of “the next questions are really sensitive”, which could be detrimental to the aim of achieving valid estimates. Also, this makes the questionnaire instruction more comparable to the short-list version of the PCT procedure. Furthermore, we anticipated that floor and ceiling effects could occur, resulting in a disclosure of the respondent’s individual answer. Introducing PCT as a technique that guarantees anonymity would represent a contradiction if this occurred and could lead to doubts or protests among respondents.

The following questions are about the situation in Mainz.

We are going to use a special questioning technique. For this purpose, please think of three preferably diverse persons among your friends, acquaintances or relatives who you know well and who live in Mainz, too. You can write down the initials of the first name of the three persons in the fields below – this makes things easier, but your notes will remain anonymous.

Initials of my three persons: □ □ □

Now we are going to make a few statements for which you should estimate how many of these three persons plus yourself agree with the respective statement. The answer is thus a number between 0 (applies to no one) and 4 (applies to all three of the persons and yourself). If you are not sure, it is OK to guess, this is not a problem.

[Item 1] "I feel bothered by the refugees in Mainz".

Number of persons who agree: □

[Item 2] "Refugees should not stroll around in the city center of Mainz, but stay in their asylums".

Number of persons who agree: □

[Item 3] "I have a problem with refugees hanging out in my neighborhood".

Number of persons who agree: □

[Item 4] "The opening of a refugee asylum in my neighborhood would bother me".

Number of persons who agree: □

Note: Translated from the German original. Underlining is depicted as in the original.

Figure 1 Wording of the PCT Procedure (Long-list Group)

The wording in the short-list group was identical to the one presented in Figure 1, with the important difference that respondents were asked only about "three people" without themselves and to report a number between zero and three. As the short-list version of the questionnaire also contained the DQ questions of the four sensitive items, immediately after the short-list PCT procedure, the questionnaire read "And now we are interested in your personal opinion on these questions. Please answer with 'yes' or 'no'", followed by the same four items as in the PCT long-list version.

Methods

The survey design with only two (DQ and PCT short list versus PCT long list) instead of three experimental groups means that DQ and PCT estimates are not statistically independent from one another. This must be taken into account when calculating standard errors. Therefore, I calculated the mean estimates for DQ and for the short-list and long-list group, respectively, and used the Stata routine suest (seemingly unrelated estimation) in order to obtain a combined and robust covariance matrix. Tests for mode differences were then performed using this covariance matrix (cf. Weesie 1999).

As explained above, design effects are a potential problem of item count procedures. They occur if the addition of the sensitive item (or the respondent in PCT) to the long list affects the responses to the non-key items ("other persons" in PCT). I will follow the recommendations of Blair and Imai (2012) who propose a statistical test in order to empirically test for design effects. This test basically examines whether implausible negative proportions of respondent types (i.e., respondents with a certain combination of "yes" answers) arise if the sensitive item (the respondent himself or herself) is removed from the respective proportion of respondent type. If such negative proportions occur, the test calculates whether they could have arisen by chance. As the test's logic and computation are complex, I refer to Blair and Imai (2012, pp. 63-65; see also Glynn 2013, pp. 165-167; Wolter & Laier 2014, p. 161) for further details. The test was performed using the "list" package for R by the same authors (Blair & Imai 2013).

Results

A conjecture made by some authors (e.g., Lensvelt-Mulders 2008, p. 464) is that sensitive questions result in higher item-nonresponse rates than non-sensitive questions. If this conjecture holds true and PCT works as intended, nonresponse should be lower when using PCT as compared to DQ. On the other hand, the PCT design requires more cognitive effort on the part of the respondents vis-à-vis answering a conventional survey question, which, in turn, could increase nonresponse rates. Table 2 shows the item-nonresponse rates for each of the four sensitive items regarding attitudes towards refugees in DQ mode and in the two groups of PCT mode.

In DQ mode, nonresponse rates for the four items vary from 2.1 to 2.8 percent, which can be considered low values given that this was a classic self-administered postal survey. This confirms the aforementioned position of Tourangeau and Yan (2007, p. 862) that item-nonresponse generally does not pose a serious problem for sensitive questions. Nonresponse rates for the PCT long-list group are higher and amount to roughly 6 percent. The differences with respect to DQ are all significant

Table 2 Item-Nonresponse Rates by Question Mode

		Item 1	Item 2	Item 3	Item 4
DQ (n = 284)	% NR	2.46	2.82	2.11	2.11
PCT LL (n = 296)	% NR	5.74	6.08	6.08	6.08
PCT SL (n = 284)	% NR	7.04	6.69	6.69	7.04
χ^2 DQ-PCT LL		3.93 *	3.61 +	5.75 *	5.75 *
χ^2 DQ-PCT SL		11.27 ***	8.07 **	11.27 ***	12.25 ***
χ^2 PCT LL-PCT SL		0.41	0.09	0.09	0.22

Note: DQ = direct questioning, PCT = person count technique, LL = long list, SL = short list, NR = nonresponse. Differences were tested using conventional χ^2 tests for differences between experimental modes and McNemar's χ^2 statistic for the DQ-PCT short-list difference. + $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

at least on a 10 percent level. However, the higher nonresponse rates do not seem to be attributable to PCT causing the items to appear more sensitive to the respondents (which in turn could yield higher nonresponse rates), because the nonresponse rates for the PCT short-list version are similar to those from the long-list group and not significantly different from them. Instead, it is the PCT design per se – be it the short or the long list – which boosts nonresponse rates, presumably due to its cognitive demands. Of course, this would be a drawback of this new questioning technique. However, it should again be noted that the survey was self-administered with no interviewer present. Taking this into consideration, nonresponse rates of 6 to 7 percent do not appear to be exceedingly or unreasonably high. Further studies should examine to what degree interviewer-administered survey modes can provide a better approach in order to avoid item-nonresponse in PCT designs.

Before we look at the prevalence estimates for the four sensitive items depending on question mode, Table 3 reports information on the distribution of respondents' answers in the short-list and long-list group, respectively. What is important here are floor and ceiling effects, i.e., respondents denying or affirming all items or persons in the list. Above I made the case for the assumption that, when using PCT instead of ICT, floor and ceiling effects will be more problematic because of homophily effects.

The results in Table 3 clearly confirm this assumption. Floor effects are substantial for all four items, both for the short-list and the long-list groups. Up to 77 percent of respondents report that the sensitive item applies to none of the persons of whom they had been asked to think. For the long-list group, containing the respondent himself or herself, anonymity is no longer ensured. However, given that a "yes" answer to the sensitive item corresponds to expressing a socially undesir-

Table 3 Distribution of Answers in the Short-List and Long-List Group

	Item 1 (%)		Item 2 (%)		Item 3 (%)		Item 4 (%)	
	SL	*LL*	*SL*	*LL*	*SL*	*LL*	*SL*	*LL*
0	56.8	49.5	77.0	73.7	44.5	35.6	27.7	21.6
1	22.7	23.7	16.2	13.0	30.9	30.9	31.1	19.8
2	14.4	11.5	4.2	7.2	16.6	18.4	23.1	25.2
3	6.1	8.6	2.6	2.9	7.9	8.6	18.2	18.4
4	-	6.8	-	3.2	-	6.5	-	15.1
n	264	279	265	278	265	278	264	278

Note: LL = long list, SL = short list.

able attitude, these floor effects are probably less problematic regarding response bias. In this regard, ceiling effects, i.e., respondents reporting “4” for the long list are the main problem, because their sensitive answer is no longer concealed by the PCT design. This holds for approximately 7 (item 1 and 3), 3 (item 2), and 15 (item 4) percent of respondents. While 3 percent (corresponding to 9 out of 278 respondents) appear to be within an acceptable range, 15 percent for item 4 (42 out of 278 respondents) is definitely too high and endangers the main purpose of PCT, namely assuring anonymity. At first glance, this appears to be a major drawback of PCT as compared to the classic ICT design, wherein floor and ceiling effects can be prevented by a careful design of the non-key items. Further studies should investigate possibilities to avoid floor and ceiling effects in PCT designs. For the time being, I suggest following our PCT design reported in Figure 1 above and, at least for now, to not all too loudly hail PCT as a technique that “guarantees complete anonymity”. Future research should also investigate whether the wording of the items affects the tendency for floor and ceiling effects. For example, for item 2 (Table 3), the fraction of “0” answers is by far the highest among the four items. In addition to substantive reasons regarding the level of sensitivity of this item, it can be assumed that this is due to the different cognitive demands processing a single sentence (item 1, 3, and 4) vis-à-vis a normative statement (item 2) requires.

Besides looking at floor and ceiling effects, I performed the aforementioned test for design effects as proposed by Blair and Imai (2012). For none of the four sensitive items could I find evidence for such effects, the p-values for items 1 to 4 are, respectively, p = 0.72, p = 0.69, p = 1.00, and p = 1.00 (the null hypothesis is that there are no design effects; thus, the null cannot be rejected according to the p-values). This can be interpreted as being in favor of PCT, because, at least empirically, based on the Blair-Imai test, there is no evidence that including the respondents themselves in the PCT-long-list changes response behavior to the “other persons” in the list.

Table 4 reports the main results of the study, namely the prevalence estimates of the four sensitive items on attitudes towards refugees in Mainz, according to question formats DQ and PCT. As expressing hostile attitudes towards refugees is considered socially undesirable, higher estimates are taken as more valid than lower ones. Therefore, the DQ-PCT comparison is based on the “more is better” assumption.

The estimates of dismissive attitudes towards refugees are substantially higher in PCT mode than in DQ mode. This holds for all four items. Regarding item 1 and 2 (“I feel bothered by the refugees in Mainz”; “Refugees should not stroll around in the city center of Mainz, but stay in their asylums”), the PCT estimates are three times higher than the DQ ones. However, as the z statistics show, PCT-DQ differences are statistically significant for the first item only, while for item 2 to 4, DQ estimates are not significantly different from their PCT counterparts at conventional levels. An overall test for the DQ-PCT difference, taking into account the four items simultaneously and adjusting for the clustering by respondents, also fails to reach conventional significance levels (diff = 12.21, z = 1.46, p = 0.145). These results are due to the highly inflated standard errors of the PCT estimates. For example, the estimate of 54 percent “yes” answers for item 4 comes with a standard error of more than ten percentage points. As pointed out above, standard errors of ICT estimates will always be higher than those from conventional ones. However, the PCT procedure, as it was implemented in this study, probably aggravates this issue for several reasons. On the one hand and as shown above, there are many respondents who answer “zero” to the person list, and a non-negligible fraction states that the trait applies to all persons in the list. This pattern inflates the variance of the variables, which, in turn, leads to greater standard errors. In classic ICT with non-key items that have either a high or low prevalence, the variance will usually be lower and the standard errors will also follow suit. On the other hand, the correlation between the sensitive item (in PCT: the respondent himself) and the filler items (in PCT: the “other persons”) is probably not negative due to homophily effects, which also boosts standard errors. Furthermore, the prevalence of the sensitive trait itself will also have an impact on standard errors, because the variance of binary variables is a function of their mean and highest for an equal distribution (i.e., a prevalence of 50 percent). These considerations show that careful precautions are required when developing PCT designs. Further studies should go into more depth on these issues and examine the relationship between design features and statistical properties of PCT estimates in a more general perspective.

Despite these challenges and despite the lacking significance for three out of the four items examined in this study, the overall conclusion remains in favor of PCT with respect to its potential and the validity of its estimates: For all items, the

Table 4 Prevalence Estimates of the Sensitive Items by Question Format

		Item 1	Item 2	Item 3	Item 4
DQ	% “yes”	9.75	5.43	23.74	43.88
	s.e.	1.78	1.37	2.55	2.98
	n	277	276	278	278
PCT	% “yes”	29.94	16.47	31.50	53.79
	s.e.	9.43	7.20	9.28	10.44
	n (short list)	264	265	265	264
	n (long list)	279	278	278	278
Difference		20.20	10.72	7.76	9.91
z		2.00 *	1.43	0.75	0.83

Note: DQ = direct questioning, PCT = person count technique. Standard errors and test statistics were calculated taking into account that DQ and PCT estimates are not independent from one another (see the “Methods” section for details). * p < 0.05.

direction of the DQ-PCT comparison points in the anticipated direction. Most of the respondents were able to cope with the PCT instructions without assistance of an interviewer and nonresponse rates were not unreasonably high.

Discussion

The present study evaluated the performance of PCT in a mode-comparing perspective and investigated item-nonresponse and underreporting on four questions regarding hostile attitudes towards refugees in a German city. As far as nonresponse is concerned, the observed rates are higher in PCT mode than in DQ mode. This, however, seems not to be caused by the sensitivity of the questions being asked, but by the PCT procedure itself, which was implemented here in a self-administered postal survey. Despite being higher, nonresponse rates remain at a tolerable level also in PCT mode. With respect to the prevalence of the four sensitive items, all estimates are distinctively higher in PCT mode, though significantly different from DQ for one item only. In this context, very large standard errors of the PCT estimates have been observed, presumably caused by the distribution of answers regarding the “other persons” in the item lists and their correlations among each other and with the respondent himself in the long list. All in all, however, the results show that considerable underreporting of hostile attitudes to refugees occurs when using conventional questioning techniques. Although the findings are not unequivocally in favor of PCT, they suggest considering PCT as a promising alternative in future studies.

Aside from the general difficulties of PCT, this study has some obvious shortcomings, that should be taken into account when judging the results. First, the number of cases (N=580) was low. As the elevated standard errors of the PCT estimates show, a larger sample would have been much more preferable and should be aimed for in future studies. Because of the limited sample size, a two-group design (DQ and PCT short list versus PCT long list) had to be used instead of a three-group design with a random split into DQ-, short-list, and long-list subsamples. This two-group design means, firstly, that DQ and PCT estimates are not statistically independent, which has to be taken into account when performing tests for differences. Secondly, halo effects may affect the results because the experimental stimulus (PCT versus DQ) is confounded with question order. The limited statistical power was also the reason why I restricted the analysis to prevalence estimates and did not conduct a regression analysis on determinants of xenophobic attitudes. Such analysis could have been helpful in judging the external validity of the PCT estimates. Whereas regression analysis is generally possible with ICT (or PCT) data (Blair & Imai, 2012, 2013; Imai 2011), it requires large sample sizes due to the restricted statistical power of PCT data. Further, the elevated item-nonresponse rates of the PCT questions show that self-administered survey modes may not be the best choice when planning to use PCT procedures. Interviewer-administered surveys seem to be preferable in this regard. Another flaw is that validation of the PCT results could only be carried out here on the basis of a "more is better" assumption. As no true values were at hand, higher estimates of hostile attitudes to foreigners were assumed to be more valid. To what degree higher estimates are still off the mark from the true value remains undiscoverable with this approach.

Above, several challenges of PCT have been pointed out, namely floor and ceiling effects, statistical power issues, and design effects. In contrast to the classic ICT design, the researcher has less influence on addressing these issues via a thoughtful design of the non-key items. In what follows, I will propose some modifications or alternatives to the PCT design as it was implemented in the present study, which could (partly) address these issues.

A first modification of the original PCT design aims to give the researcher control over the characteristics of the "other persons". This would help in avoiding floor- and ceiling effects and in making PCT estimates more efficient. I call this design fixed person count technique (FPCT). The simple idea is not to ask respondents to imagine "some people they know", but instead to propose fixed persons by design. A (purely illustrative) example would be to ask respondents to indicate how many of the following persons, including themselves, have already smoked marihuana: Bob Marley, Angela Merkel, and Pope Francis. In this case, the values for

Bob Marley and Pope Francis are more or less fixed and near 1 and 0, respectively. This avoids floor and ceiling effects and improves statistical efficiency. For the sake of anonymity, the Angela Merkel item is more ambiguous. Of course, this is just an illustrative example, as one should not choose such obvious cases as Bob Marley and marihuana consumption. One could easily imagine other possible designs in this regard, for instance, letting respondents imagine a member of a typical group such as a "typical democrat voter" or a "typical primary school teacher". Or, to think of their nearest neighbor, their postman, or their family doctor. A clever choice of these more or less fixed persons might help overcome the problems inherent to the basic PCT design.

Another straightforward modification of PCT is to apply the logic of the above-mentioned item sum technique (IST) for metric sensitive variables to a PCT procedure – the person sum technique (PST) as proposed by Junkermann (2018). PST also asks respondents to imagine one or more other persons they know – as with IST, however, one non-key person will usually suffice. Respondents are then asked to estimate the value of one quantitative sensitive item for the other person in the short-list group. In the long-list group, respondents are asked to add up the value of the other person and their own value. For example, the sensitive item could be the number of cigarettes smoked per day. Respondents in the long-list group are then asked to estimate the number of daily smoked cigarettes for the uninvolved person, and to add this value to the number of cigarettes smoked by themselves.

The research desiderata with respect to PCT are clear-cut. Future studies should, firstly, investigate the (cognitive) mechanisms at work when respondents deal with PCT designs. These studies should focus on, among other matters, homophily effects, isolated persons that have difficulty imagining people they know well, the occurrence of design effects, and what happens if respondents are unsure about the status of the uninvolved person(s) in the list. This entails both qualitative and quantitative work. Second, real validation studies with known true values (from external records, for instance) should be conducted in order to assess the ability of PCT to avoid or at least alleviate response bias. If this is not possible, further studies relying on the "more is better" logic should be conducted – and with larger samples than in the study presented in this paper. Third, empirical studies should also concentrate on experimentally comparing PCT with classic ICT designs. This should be carried out with respect to validity, the anonymity protection subjectively perceived by the respondents, the amount of cognitive burden (is PCT really less demanding than ICT?), and with respect to the trade-off between statistical efficiency, respondent protection, and simplicity of the question procedures. Fourth, further studies on PCT designs should test whether the above introduced FPCT presents a viable alternative to the original PCT design.

References

Allport, G. W. (1954). *Ther Nature of Prejudice*. Reading, MA: Addison-Wesley.

An, B. P. (2015). The Role of Social Desirability Bias and Racial/Ethnic Composition on the Relation Between Education and Attitude Toward Immigration Restrictionism. *The Social Science Journal, 52*(4), 459–467.

Arzheimer, K. (2008). Protest, Neo-Liberalism or Anti-Immigrant Sentiment: What Motivates the Voters of the Extreme Right in Western Europe? *Zeitschrift für vergleichende Politikwissenschaft, 2*(2), 173–197.

Barton, A. H. (1958). Asking the Embarrassing Question. *Public Opinion Quarterly, 22*(1), 67–68.

Bauer, J. (2014). *New Sample Designs. An Improvement and Alternative to Random Route Samples*. Working Paper: LMU München.

Benson, L. E. (1941). Studies in Secret-Ballot Technique. *Public Opinion Quarterly, 5*(1), 79–82.

Biemer, P. P., Jordan, B. K., Hubbard, M., & Wright, D. (2005). A Test of the Item Count Methodology for Estimating Cocaine Use Prevalence. In J. Kennet & J. Gfroerer (Eds.), *Evaluating and Improving Methods Used in the National Survey on Drug Use and Health (DHHS Publication No. SMA 05-4044, Methodology Series M-5)* (pp. 149–174). Rockville: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Blair, G., & Imai, K. (2012). Statistical Analysis of List Experiments. *Political Analysis, 20*(1), 47–77.

Blair, G., & Imai, K. (2013). *Package 'list'. Statistical Methods for the Item Count Technique and List Experiment*: retrieved on http://cran.r-project.org/web/packages/list/list.pdf (2018/01/22).

Bradburn, N. M., & Sudman, S. a. A. (1979). *Improving Interview Method and Questionnaire Design. Response Effects to Threatening Questions in Survey Research*. San Francisco: Jossey-Bass.

Cappelen, C., & Midtbø, T. (2016). Intra-EU Labour Migration and Support for the Norwegian Welfare State. *European Sociological Review, 32*(6), 691–703.

Cea D'Ancona, M. A. (2014). Measuring Xenophobia: Social Desirability and Survey Mode Effects. *Migration Studies, 2*(2), 255–280.

Comşa, M., & Postelnicu, C. (2013). Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique. *International Journal of Public Opinion Research, 25*(2), 153–172.

Corstange, D. (2009). Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis, 17*(1), 45–63.

Coutts, E., & Jann, B. (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods and Research, 40*(1), 169–193.

Creighton, M. J., & Jamal, A. (2015). Does Islam Play a Role in Anti-Immigrant Sentiment? Ab Experimental Approach. *Social Science Research, 53*, 89–103.

Czymara, C. S., & Schmidt-Catran, A. W. (2016). Wer ist in Deutschland willkommen? Eine Vignettenanalyse zur Akzeptanz von Einwanderern. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 68*(2), 193–227.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The Item Count Technique as a Method of Indirect Questioning: A Review of its Development and a Case Study Application. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz., & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 185–210). New York: Wiley.

Ferligoj, A., & Hlebec, V. (1999). Evaluation of Social Network Measurement Instruments. *Social Networks, 21*, 111–130.

Fox, J. A., & Tracy, P. E. (1986). *Randomized Response. A Method for Sensitive Surveys* (Vol. 07-058). Newbury Park: Sage.

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social Desirability Response Effects: Three Alternative Models. *Academy of Management Journal, 26*(2), 321–331.

Glynn, A. N. (2013). What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment. *Public Opinion Quarterly, 77*(Special Issue), 159–172.

Grant, T., Moon, R., & Gleason, S. A. (2014). *Asking Many, Many Sensitive Questions: A Person-Count Method for Social Desirability Bias*: Unpublished Manuscript.

Hoffmann, A., & Musch, J. (2016). Assessing the Validity of two Indirect Questioning Techniques: A Stachastic Lie Detector Versus the Crosswise Model. *Behavior Research Methods, 48*(3), 1032–1046.

Huckfeldt, R., & Sprague, J. (1995). *Citizens, Politics, and Social Communication. Information and Influence in an Election Campaign*. New York: Cambridge University Press.

Hyman, H. (1944). Do They Tell the Truth? *Public Opinion Quarterly, 8*(4), 557–559.

Imai, K. (2011). Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association, 106*(494), 407–416.

Janus, A. L. (2010). The Influence of Social Desirability Pressures on Expressed Immigration Attitudes. *Social Science Quarterly, 91*(4), 928–946.

Junkermann, J. (2018). *Die Person Sum Technique. Ein neues Instrument zur Erhebung quantitativer heikler Items*. University of Mainz.

Kandel, D. B. (1978). Homophily, Selection, and Socialization in Adolescent Friendships. *American Journal of Sociology, 84*(2), 427–436.

Krumpal, I. (2012). Estimating the Prevalence of Xenophobia and Anti-Semitism in Germany: A Comparison of Randomized Response and Direct Questioning. *Social Science Research, 41*(6), 1387–1403.

Krumpal, I. (2013). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity, 47*(4), 2025–2047.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial Attitudes and the "New South". *Journal of Politics, 59*(2), 323–349.

Lensvelt-Mulders, G. J. L. M. (2008). Surveying Sensitive Topics. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 461–478). New York: Lawrence Erlbaum.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*, 415–444.

Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income Measurement Error in Surveys: A Review. *Journal of Official Statistics, 16*(4), 331–361.

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A Randomized-Response Investigation of the Education Effect in Attitudes Towards Foreigners. *European Journal of Social Psychology, 39*(6), 920–931.

Perry, P. (1979). Certain Problems in Election Survey Methodology. *Public Opinion Quarterly, 43*(3), 312–325.

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe. *American Sociological Review, 60*(4), 586–611.

Rosenfeld, B., Imai, K., & Shapiro, J. N. (2015). An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science, 60*(3), 783–802.

Shakya, H. B., Christakis, N. A., & Fowler, J. H. (2017). An Exploratory Comparison of Name Generator Content: Data from Rural India. *Social Networks, 48*, 157–168.

South, S. J., & Felson, R. B. (1990). The Racial Patterning of Rape. *Social Forces, 69*(1), 71–93.

Stocké, V. (2007). The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys. *Journal of Official Statistics, 23*(4), 493–514.

Thomas, K., Johann, D., Kritzinger, S., Plescia, C., & Zeglovits, E. (2017). Estimating Sensitive Bahavior: The ICT and High-Incidence Electoral Behavior. *International Journal of Public Opinion Research, 29*(1), 151–171.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*(5), 859–883.

Trappmann, M., Krumpal, I., Kirchner, A., & Jann, B. (2014). Item Sum – A New Technique for Asking Quantitative Sensitive Questions. *Journal of Survey Statistics and Methodology, 2*(1), 58–77.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association, 60*(309), 63–69.

Weesie, J. (1999). sg121: Seemingly Unrelated Estimation and the Cluster-Adjusted Sandwich Estimator. *Stata Technical Bulletin 52*, 34–47.

Weins, C. (2011). Gruppenbedrohung oder Kontakt? Ausländeranteile, Arbeitslosigkeit und Vorurteile in Deutschland. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 63*, 481–499.

Wolter, F., & Herold, L. (2018). Testing the Item Sum Technique (IST) to Tackle Social Desirability Bias. *SAGE Research Methods Cases*. http://dx.doi.org/10.4135/9781526441928

Wolter, F., & Laier, B. (2014). The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency. *Survey Research Methods, 8*(3), 153–168.

Wolter, F., & Preisendörfer, P. (2013). Asking Sensitive Questions: An Evaluation of the Randomized Response Technique versus Direct Questioning Using Individual Validation Data. *Sociological Methods and Research, 42*(3), 321–353.

Yan, T., Curtin, R., & Jans, M. (2010). Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics, 26*(1), 145–164.

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis. *Metrika, 67*(3), 251–263.

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)