Survey Research Methods (2014)
Vol. 8, No. 3, pp. 153-168

ISSN 1864-3361
http://www.surveymethods.org

The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency

Felix Wolter
Department of Sociology
Johannes Gutenberg University Mainz

Bastian Laier
Department of Sociology
Johannes Gutenberg University Mainz

Surveys often contain sensitive questions, i. e., questions about private, illegal, or socially undesirable behavior. When asked directly in standard survey formats, respondents tend to underreport these behaviors, yielding biased results. One method that promises more valid estimates than direct questioning (DQ) is the item count technique (ICT). In this paper, the methodological pros and cons of ICT, as compared to DQ, are weighed up empirically with regard to questions eliciting self-reported delinquency. We present findings from a face-to-face survey of 552 respondents who had all been previously convicted under criminal law prior to the survey. The results show, first, that subjective measures of survey quality such as trust in anonymity or willingness to respond are not affected positively by ICT with the exception that interviewers feel less uncomfortable asking sensitive questions in ICT format than in DQ format. Second, all prevalence estimates of self-reported delinquent behaviors are significantly higher in ICT than in DQ format. Third, a regression model on determinants of response behavior indicates that the effect of ICT on response validity varies by gender. Overall, our results are in support of ICT. This technique is a promising alternative to other specialized questioning techniques such as the much more complicated randomized response technique (RRT).

Keywords: Survey methodology; item count technique; response bias; sensitive questions; social desirability

1 Introduction

The issue of asking sensitive questions in surveys is as old as modern survey methodology: numerous studies dating back to the 1930s and 1940s have investigated the problem that many respondents do not give truthful answers to questions pertaining to private, illegal, or socially undesirable behaviors (Barton, 1958; Hyman, 1944; Katz, 1942). When answering such questions, respondents tend to underreport negatively connoted and to overreport positively connoted behaviors (for reviews, see Krumpal, 2013; Tourangeau and Yan, 2007). For instance, Wolter and Preisendörfer (2013) found that fewer than 60 percent of respondents in a face-to-face survey admitted to having been convicted under criminal law, although everyone in the sample had in fact been convicted. Hadaway, Marler, and Chaves (1993) found that 51 percent of respondents claimed to attend church every week, whereas the actual prevalence was only 28 percent. This well-established response bias leads to two problems. First, prevalence estimates of sensitive characteristics are biased. Second, analyses investigating the effect of variables such as gender, age, or education on sensitive behaviors or attitudes are biased as a result of systematic misreporting depending on the factors investigated (Bernstein, Chadha, & Montjoy, 2001; Wolter & Preisendörfer, 2013).

Among the many methods that have been proposed to reduce response bias — such as the sealed envelope technique (Benson, 1941), the randomized response technique (RRT; Fox and Tracy, 1986; Lensvelt-Mulders, Hox, van der Heijden, and Maas, 2005; Wolter and Preisendörfer, 2013), or wording and filtering techniques—the item count technique (ICT)[1]—is relatively new receiving much attention in the last few years (Blair & Imai, 2012; Glynn, 2013). The idea behind ICT is to anonymize the interview situation so that no one, not even the interviewer, can deduce the "real" answer to the sensitive question from the answer given by the respondent, the aim being that this, in turn, will enhance the respondent's willingness to answer truthfully.[2] The principle

Contact information: Dr. Felix Wolter, Department of Sociology, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 12, D-55128 Mainz, Germany (felix.wolter@uni-mainz.de)

[1] ICT is also called unmatched count technique (Coutts & Jann, 2011) or list experiment (Zigerell, 2011).

[2] More precisely, all the factors that influence respondents to not admit sensitive behaviors in DQ format are expected to be eliminated by ICT. This general hypothesis receives theoretical support

of ICT is to randomly divide the respondents into (at least) two groups. In each group, the respondents receive a list of dichotomous (yes/no) questions and are asked to indicate only the number of questions for which the answer is "yes" and not to answer the questions individually. The lists for the respective two groups differ in that one group receives a list of questions without the sensitive item (short list) while the other group receives the same questions plus the sensitive item (long list). A simple way of arriving at a prevalence estimate of the sensitive item ensues by subtracting the mean sum of questions answered with "yes" in the group with the short list from that of the group with the long list.

Findings from the literature on ICT's effectiveness in eliciting truthful answers to sensitive questions are mixed. Almost all studies take the approach of comparing prevalence estimates of sensitive behaviors obtained by means of ICT with estimates using conventional questioning techniques such as direct questioning (DQ). Following the assumption "more is better", ICT is considered to have performed better than DQ if prevalence estimates of negatively connoted behaviors are significantly higher in ICT format (or, conversely following the assumption "less is better" if prevalence estimates of positively connoted behaviors are lower in ICT format). Applying this rationale, Dalton, Wimbush, and Daily (1994), Wimbush and Dalton (1997), and Streb, Burrell, Frederick, and Genovese (2008), to name but a few, found significantly better estimates of sensitive characteristics using ICT compared with DQ. Coutts and Jann (2011), LaBrie and Earleywine (2000), Rayburn, Earleywine, and Davison (2003b), and Holbrook and Krosnick (2010) also found significantly better ICT estimates for some items, but not for all. Ahart and Sackett (2004), Biemer and Brown (2005), Droitcour et al. (1991), and Biemer, Jordan, Hubbard, and Wright (2005) either found no differences between ICT and DQ formats or, in the case of the former, even found the estimates to be worse.

This paper aims to present new empirical evidence on the effectiveness of ICT in eliciting valid responses to questions about self-reported delinquency. We present the results of an experimental face-to-face survey of a special population of people who all had been convicted of minor offenses such as repeated fare dodging, shoplifting, or driving under the influence under criminal law in the years immediately prior to the survey. The fact that all respondents are convicted "criminals" avoids a problem common to many methodological studies in the field of self-reported delinquency and sensitive questions, namely, that illegal or other sensitive behaviors are often very rare in the general population (or in other populations usually consulted in the literature). Subsequently, the problem normally ensues that very large sample sizes are required for estimations and the analysis of question format differences.[3]

We compare ICT estimates of three illegal behaviors (fare dodging, driving without a driver's license, and driving under the influence) with their corresponding DQ estimates. Furthermore, besides item non-response, we investigate what we termed as subjective measures of survey quality, i. e., variables such as the willingness to respond and the reliability of answers as judged by the interviewers, the perceived trust in anonymity, and discomfort of respondents and interviewers when answering/asking the sensitive questions. All these indicators are expected to show better values in ICT than in DQ format. Another aim of our paper is to provide empirical evidence regarding the issue of systematic misreporting depending on independent variables in both question formats. If the effects of right-hand side variables differ between question formats, then misreporting would occur systematically and analyses investigating determinants of sensitive behavior would be biased. Until just recently, there were no established methods to conduct multivariate analyses with ICT data, however, some new approaches have now been developed (Blair & Imai, 2012; Corstange, 2009; Imai, 2011).

In the following, we will, first, give an overview of the state-of-the-art of the research on ICTs in Section 2. Then we will describe the design of our survey, its ICT procedures, and the variables of interest in Section 3. Section 4 presents empirical results on the questions outlined above followed by a short discussion and desiderata for further research in Section 5.

2 The Item Count Technique (ICT): Methods and Existing Research

To the best of our knowledge, the first mention of ICT was by Smith, Federer, and Raghavarao (1974) under the term "block total response procedure". Subsequently, it was further developed and empirically tested by Miller (1984)[4]. However, it was only in recent years that the technique has

from rational choice theory which states that misreporting occurs if the subjectively expected utility for giving an edited answer is higher than for giving a truthful answer (Esser, 1986, 1991; Stocké, 2004). If, due to techniques such as ICT, the respondent's answer is not deducible, cost/benefit factors that make respondents misreport in DQ formats—such as social desirability concerns or fear of sanctions from third parties—are, supposedly, redendered irrelevant (see also Preisendörfer and Wolter, 2014, for a more general discussion of cost/benefit factors in the context of answering (sensitive) survey questions).

[3]For example, prevalence estimates of cocaine use hover around 1 to 2 percent (Biemer et al., 2005); of intravenous heroin use around 4 percent (DQ) and 0.2 percent (ICT) (Droitcour et al., 1991); of tax evasion around 2 percent (RRT), 4 percent (DQ) and 7 percent (ICT) percent (Krumpal, 2008); and of car theft around 2 percent (DQ) and −1 percent (RRT [sic]) (Durham III & Lichtenstein, 1983).

[4]Cited after Zigerell (2011, p. 552), Droitcour et al. (1991), Hubbard, Casper, and Lessler (1989), and Dalton et al. (1994)

Table 1
ICT Module in the CATI Survey by Holbrook and Krosnick (2010, p. 47)[a]

Interviewer: "Here is a list of four [five] things that some people have done and some people have not. Please listen to them and then tell me *how many* of them you have done." Do not tell me which you have or have not done. Just tell me how many. Here are the four [five] things:

Both samples
- Owned a gun
- Given money to a charitable organization
- Gone to see a movie in a theater,
- Written a letter to the editor of a newspaper

Sample 2 only (long version)
- Voted in the presidential election held on November 7, 2000.

How many of these things have you done?

[a] slightly adapted

attracted extensive attention from researchers in the field of survey methodology.

As already mentioned above, the concept of ICT in surveys is to randomly divide the total sample into (at least) two subsamples. One subsample receives a short list of questions without the sensitive item, the other receives a long list containing the same questions as the short list, plus the sensitive item. The respondent is asked to indicate only the number of applicable items. Hence, it is not disclosed whether the sensitive item has been affirmed or not (unless the respondent negates or affirms all the items on the list, see below). One example of an empirical implementation of ICT is the study by Holbrook and Krosnick (2010) who address participation (voting) in the last presidential election. Here, one would expect overreporting of voting (Belli, Traugott, & Beckmann, 2001; Bernstein et al., 2001). Table 1 gives an overview of the principle of ICT as implemented by Holbrook and Krosnick.

The prevalence $\widehat{\pi}_{\text{ICT}}$ of the sensitive item in the sample can be estimated by subtracting the mean of the short list ($\bar{x}_{\text{SL}}$) from the mean of the long list ($\bar{x}_{\text{LL}}$):

$$\widehat{\pi}_{\text{ICT}} = \bar{x}_{\text{LL}} - \bar{x}_{\text{SL}} \tag{1}$$

Provided that the samples are independent, the sampling variance can be calculated as follows (Tourangeau & Yan, 2007, p. 872):

$$\text{Var}(\widehat{\pi}_{\text{ICT}}) = \text{Var}(\bar{x}_{\text{LL}}) + \text{Var}(\bar{x}_{\text{SL}}) \tag{2}$$

This approach served as a template for many versions of ICT designs which mostly address the rather poor statistical efficiency of classic ICT. Corstange (2009), for instance, suggests adjusting the classic design so that the items on the short list are asked separately in the control group. Recently, Trappmann, Krumpal, Kirchner, and Jann (2014) developed a new version of ICT for quantitative variables, the item sum technique. Trappmann et al. asked respondents in one short-list group to indicate, e. g., the number of hours they had watched TV in the past week; respondents in the corresponding long-list group also received the question of how many hours per week they usually engage in undeclared work.

Some methodological issues of ICT frequently come up in the literature. First, the choice of the non-key or filler items needs careful consideration. Some authors (Biemer et al., 2005; Droitcour et al., 1991) advise keeping the non-key items thematically close to the sensitive item. The short list should contain items with low and high prevalences or highly negatively correlated items in order to avoid a respondent either affirming or denying all items on the list ("ceiling and floor effects"; in both cases, the respondents' protection by the procedure would be nullified, see Blair and Imai, 2012). Moreover, the number of non-key items should not be too low because the more filler items there are, the better the respondent is protected. With respect to the statistical properties of ICT estimates, it is clear that they are less efficient than standard estimates (Droitcour et al., 1991, p. 189). Every non-key item adds additional error variance to the estimate calling for a trade-off between statistical efficiency and respondent protection (Imai, 2011). In the literature, it seems that three to five non-key items are normally used and appear to be the most appropriate. Tourangeau and Yan (2007, p. 872) suggest that the innocuous items should be of low variance (i. e., items with a very low or very high prevalence) in order to improve statistical efficiency. A further proposal is the double-lists design (Droitcour et al., 1991) where an additional list of non-key items is used; what formerly in the classic method is the experimental group (long list) also responds to this supplementary list without the sensitive item and the former control group (short list) responds to the supplementary list including the sensitive item.

Second, one issue that has received attention recently (Holbrook & Krosnick, 2010) is the objection that the long list could—independent of the item content—encourage respondents to report higher numbers of positive responses than those answering the short list. "Yea-sayer" (people tending to uncritically approve everything) could induce such an effect. However, Holbrook and Krosnick (2010, p. 53) report results from an experimental study ($N = 769$) in which respondents received a long list that contained a question about holidays in a non-existent country and in which the (marginal) difference between long and short list was not significant. A generalization of this problem is what Blair and Imai call design effects (Blair & Imai, 2012, 51f.). They

occur if the introduction of the sensitive item affects—for various possible reasons—the response behavior to the filler items. Blair and Imai propose a statistical procedure to test for design effects. Another potential problem is that it is debatable whether the fact that several items are presented to the respondents at once results in a different response than if the same items were presented to the respondents separately. A study by Tsuchiya and Hirai (2010) provides evidence in this regard. They report that ICT (list) responses tend to be fewer than DQ responses. The authors ascribe this effect to a different cognitive process that occurs when answering ICT questions with respondents only checking the items that apply to them, whereas in DQ format the respondents have to check whether an item does *or does not* apply. The additional cognitive category "does not apply" is missing in list designs. The authors compare a wide variety of ICT lists to DQ and report statistically significant differences of mean values between them, mostly for ill-defined (unclear) items. The magnitude of this effect seems to diminish when the items are clear and well-defined.

A further issue is that many researchers are not only interested in estimating the prevalence of sensitive behaviors, but are also in pursuit of statistical analyses of their determinants. However, only quite recently Lensvelt-Mulders (2008, p. 474) proclaimed the impossibility of conducting such analyses with ICT data. Nevertheless, it is indeed possible. One simple option (employed, e. g., by Holbrook and Krosnick, 2010; Janus, 2010) is to use standard OLS regression and interaction effects between the experimental group (short list vs. long list) and independent variables. Blair and Imai (2012, p. 51) note that there are some problems with this approach such as low efficiency and possible predicted probabilities of the sensitive characteristic being below zero or higher than 1. To overcome this, alternative methods have recently been developed (Blair & Imai, 2012; Corstange, 2009; Imai, 2011).

To summarize, methodological and statistical properties of ICT methods have been addressed widely during the last few years and knowledge about them is well developed and subject to ongoing research. Furthermore, particularly in political science, ICTs are now commonly used in several practical applications (Gilens, Sniderman, & Kuklinski, 1998; Gonzalez-Ocantos, de Jonge, Meléndez, Osorio, & Nickerson, 2012; Janus, 2010; Kane, Craig, & Wald, 2004; Kuklinski, Cobb, & Gilens, 1997; Kuklinski, Sniderman, et al., 1997; Martinez & Craig, 2010; Redlawsk, Tolbert, & Franko, 2010). In contrast and somewhat surprisingly, as we and others have already noted with reference to RRT (Umesh & Peterson, 1991; Wolter & Preisendörfer, 2013), there is a substantial gap between the level of methodological refinements, statistical details, use in actual applications, and the simple question whether ICT succeeds at all in avoiding response biases. As already mentioned above, findings are mixed (Tourangeau & Yan, 2007, p. 872). A meta-analysis by Tourangeau and Yan (2007), which, to the best of our knowledge, is the only one investigating the effects of ICTs over several studies, finds a non-significant effect of ICTs compared with DQ and significant heterogeneity between studies. However, this meta-analysis was based on only seven studies, and more empirical evidence on the effectiveness of ICTs has been accumulated during recent years. This calls for an updated meta-analysis, which we, however, would not be able to conduct within the scope of the present paper. Nevertheless, Table 2 gives an overview of all empirical studies that we could find[5] comparing the response validity of ICT estimates with the response validity from DQ estimates.

All studies listed in Table 2 base their evaluation of ICT on the "more (or less) is better" assumption. This is a common procedure in the literature, but one should bear in mind that unless the true value of a sensitive item in a certain population is known, the capacity of ICTs (and other questioning techniques) in producing valid results is, at least partially, unclear, because even if ICT estimates are "better" than DQ ones, they can still be far off the mark from the true value. With respect to other questioning techniques or survey mode effects, validation studies comparing survey estimates with a known true value for the population do exist (see, e. g., Kreuter, Presser, & Tourangeau, 2008; van der Heijden, van Gils, Bouts, & Hox, 2000; Wolter & Preisendörfer, 2013), however, this is not the case with regard to ICT.[6] Consequently, one key desideratum for future ICT research is to conduct validation studies.

Table 2 shows overall evidence that trends towards being in support of ICT vis-à-vis DQ: eight studies find significantly better estimates in ICT format, nine studies find significantly better estimates for some of the multiple items tested, two studies find no difference between DQ and ICT, and two studies find negative effects of ICT. Therefore, 80 percent of all studies report at least partially positive effects of ICTs on response validity, which, in our view, justifies considering ICTs as a promising way to survey sensitive issues in the future. The present study will add further evidence in support of this argument.

If we narrow our scope down to ICT studies on self-reported delinquency, the findings remain mixed. Wimbush and Dalton (1997) report a clearly positive effect of ICT for employee theft, Rayburn, Earleywine, and Davison

[5] We consulted the internet and the SSCI using the search terms "item count technique", "list experiment", "unmatched count technique", and their abbreviations. Furthermore, we checked the bibliographies in the existing literature for empirical studies on ICT.

[6] Comşa and Postelnicu (2013), however, compare self-reported participation in the 2009 European Parliament election—75 percent in DQ and 65 percent in ICT format—with the official participation from the electoral bureau, which is about 28 percent and, thus, quite a way off the survey estimates.

Table 2
DQ versus ICT: Overview of Empirical Studies

Source	Sample	Mode	Sensitive Items	Result
Droitcour et al. (1991)	general population	SAQ	intravenous drug use, receptive anal intercourse	DQ better
Dalton, Wimbush, and Daily (1994)	professional auctioneers	SAQ	proscribed auctioneer behavior	ICT better
Wimbush and Dalton (1997)	employees	SAQ	employee theft	ICT better
Gilens, Sniderman, and Kuklinski (1998)	general population	CATI	attitudes toward blacks	ICT better
LaBrie and Earleywine (2000)	students	SAQ	sexual risk behaviors and alcohol	mixed (ICT better for 3 out of 5 items)
Rayburn, Earleywine, and Davison (2003b)	students	SAQ	hate crime victimization	mixed (ICT better for 13 out of 15 items)
Rayburn, Earleywine, and Davison (2003a)	students	SAQ	anti-gay hate crimes	mixed (ICT better for 2 out of 4 items)
Ahart and Sackett (2004)	students	SAQ	counterproductive behavior[a]	DQ = ICT
Biemer and Brown (2005), Biemer, Jordan, Hubbard, and Wright (2005)	general population	ACASI	cocaine use	DQ better
Anderson, Simmons, Milnes, and Earleywine (2007)	students	SAQ	eating disorders	mixed (ICT better for 5 out of 6 items)
Heerwig and McCabe (2009)	general population[b]	online	support for black president	ICT better
Tourangeau and Yan (2007)	meta-analysis			
Tsuchiya, Hirai, and Ono (2007)	general population[b]	online	blood donation shoplifting	mixed (ICT better for shoplifting)
Krumpal (2008)	general population	CATI	various sensitive behaviors and attitudes	DQ = ICT
Streb, Burrell, Frederick, and Genovese (2008)	general population	CATI	support for female American president	ICT better
Walsh and Braithwaite (2008)	students	SAQ	16 items related to sexual behavior and/or alcohol	mixed (ICT better for 7, worse for 3, and no difference for 6 items)
Holbrook and Krosnick (2010)				
Study 1	general population	CATI	voter turnout	ICT better
Study 2	general population	online	voter turnout	DQ = ICT
Janus (2010)	general population	CATI	attitude to immigration	ICT better
Coutts and Jann (2011)	general population[b]	online	keeping too much change, fare dodging, shoplifting, marijuana use, DUI, infidelity	mixed (ICT better for 1 out of 6 items)
Gonzalez-Ocantos, de Jonge, Meléndez, Osorio, and Nickerson (2012)	general population	FtF	vote buying in Nicaragua	ICT better
Comşa and Postelnicu (2013)	general population	FtF	voter turnout	ICT better but worse than true value
Trappmann, Krumpal, Kirchner, and Jann (2014)	general population	CATI	hours and earning from undeclared work	mixed (ICT better for 1 out of 2 items)

General remark: The term "general population" does not mean representative of the general population but indicates that, rather than a convenience sample, a sample of the general population or some subgroup of the general population was used.
Abbreviations: SAQ = self-administered questionnaire; CATI = computer-assisted telephone interview; ACASI = audio computer-assisted self-interview; FtF = face-to-face; DQ = direct questioning; ICT = item count technique;
[a] "intentional behavior on the part of an organizational member viewed by the organization as contrary to its legitimate interests" (Sackett and DeVore, 2001, cited by Ahart and Sackett, 2004, p. 101). [b] Access panel

(2003a) find a positive ICT effect for two out of four items on anti-gay hate crimes, Krumpal (2008) finds higher, but non-significant ICT estimates for bilking, shoplifting, tax evasion, and spousal violence, and a non-significant lower prevalence for driving under the influence. Coutts and Jann (2011) report a lower prevalence for freeriding in ICT format and higher for shoplifting and drunk driving, however, all differences between question formats are non-significant. Finally, employing the item sum technique, Trappmann et al. (2014) report significantly higher estimates for earnings from undeclared work, but no differences for the number of hours of said work. Overall, we can conclude that the usefulness of ICT in eliciting more valid estimates than DQ with respect to self-reported delinquency has not yet been evaluated and ascertained sufficiently.

Furthermore, multivariate analysis of ICT data has only been considered in some of the more recent studies and is still part of ongoing research. Thus, there is a lack of knowledge regarding determinants of sensitive behavior from the perspective of comparing different questioning techniques. In our analyses, we will present some simple regression models that compare the effects of socio-demographic variables between DQ and ICT formats.

Another issue that has only been addressed with respect to ICT in some explorative or qualitative work (Droitcour et al., 1991; Hubbard et al., 1989) is the above-mentioned subjective measures of survey quality, such as the perception of anonymity by the respondents. Besides viewing them as a separate dimension of "survey quality" in addition to data validity, another rationale behind investigating them is the general hypothesis that misreporting on sensitive questions is caused mainly by the threat of disclosure (Lee, 1993; Lensvelt-Mulders, 2008; Tourangeau & Yan, 2007) which could have negative repercussions for the respondent in the form of substantial sanctions or a loss of social reputation ("social desirability" concerns). Therefore, it is by reducing this threat via the enhancement of subjective feelings of anonymity, protection, and so on that answer validity is expected to improve as a result of using special techniques such as ICT.

3 Methods

3.1 Survey Design and Fieldwork

The results presented in this article stem from a larger project, the main purpose of which was to carry out a double-blind individual validation study in order to compare RRT with DQ (the findings are presented in Wolter, 2012; Wolter and Preisendörfer, 2013). Within the framework of this project, an experimental face-to-face survey in a German metropolitan area was conducted, with a gross sample taken from court records containing the address, age, and criminal offense of persons who had been convicted in the last few years prior to the survey. Only those who had committed "minor" offenses such as shoplifting, driving under the influence, repeated fare dodging on public transportation, drug abuse, or social welfare fraud were included in the sample.[7] The respondents were assigned either with a probability of 0.4 to the DQ format in which all sensitive questions were posed directly, or with a probability of 0.6 to the indirect format in which one set of the sensitive questions were asked using RRT and another with ICT. In this article, we will present the results of the questionnaire module that was devoted to the DQ-ICT comparison.

The survey was given a design and title referring to "quality of life and living conditions" so as to make the appearance of being a conventional population survey. Neither the interviewers nor the respondents were informed about the composition of the sample. The sensitive questions in the RRT and ICT module were "hidden" among other innocuous questions regarding living conditions and quality of life. In order to address ethical and data protection concerns, we added "normal" contact addresses to the gross sample of convicted persons and took several other measures in cooperation with German data protection authorities.[8] These "innocent" addresses are excluded from all analyses reported below.

For fieldwork, 75 interviewers (mostly students) were hired and trained in a half-day interviewer training course. The field phase lasted from February 2009 to October 2010. This long period reflects the difficulties of surveys among special populations such as petty criminals that have already been noted in other studies (Locander, Sudman, & Bradburn, 1976; van der Heijden et al., 2000). The final response rates of the survey are reported in Table 3.

Initially, 3,372 persons were contacted by an advance letter (sent by traditional postal methods) with the researchers' university as the sender. The letter informed the contact persons about the study, assured anonymity and privacy, and announced that an interviewer would personally contact the addressee in the next few days and ask for an interview. 647 of these cases were returned by the postal service due to invalid

[7] In fact, our sample of people with previous convictions represents an arbitrary sample of all convictions in the respective court area. We were not allowed to have any influence over the choice of the actual court records that the court administration provided us with, although it was, however, agreed upon that we were only interested in exclusively receiving cases related to "petty crimes". After receiving the data, we manually excluded cases that raised concerns about interviewer safety. For instance, all cases related to sex crimes (such as sexual harassment) were excluded.

[8] The data protection authorities also agreed to the linkage of court records and survey data (see the remarks on "potential fakes" below); this, however, was done using a third person (depositary), so that the researchers had no data file that contained the identity and the survey data of the respondents (for details, see Wolter, 2012). Being financed by the German Science Foundation (DFG), the study design was also reviewed by their review boards.

Table 3
Response Rates of the Survey

	N	% total	% contacted	% interviewed
Total sample	3,372	100.0		
Not known at this address (incl. death, etc.)	647	19.1		
Not approached by interviewer	479	14.2		
Net sample	2,246	66.6	100.0	
No contact in household	500	14.8	22.3	
No contact with target person	359	10.7	16.0	
Refusal: no time	178	5.3	7.9	
Refusal: other reasons	404	12.0	18.0	
No interview: language difficulties	42	1.3	1.9	
No interview: other reasons, unspecified	183	5.4	8.2	
Interviews conducted	580	17.2	25.8	100.0
Potential fakes	27	0.8	1.2	4.7
Analysis sample	553	16.4	24.6	95.4
DQ interviews	219			
ICT interviews	334			

addresses or marked as invalid by the interviewers because the contact person no longer lived at the indicated location (e. g., due to them being deceased or having moved); 479 cases were not approached by an interviewer because some interviewers decided to leave their job after their initial experience with the survey. These cases were not re-assigned to other interviewers because the advance letter had announced the interviewer by name and contact details; furthermore, since we had assigned all addresses randomly to the interviewers in a regionally limited area, we were confident in assuming that these cases were non-selective dropouts. The remaining 2,246 contact addresses were approached by the interviewers[9] and resulted in 580 interviews. The interviews were conducted using a classic paper-and-pencil format in which the interviewer read out the questions and marked the respondent's answers on the questionnaire. Because it was crucial for the purpose of the survey that precisely the person specified in the court data—and no other member of the household — was interviewed, we compared the respondent's year of birth as ascertained during the interview with the year of birth indicated in the court data. The cases in which the two dates did not match (potential fakes or cases in which other persons had been interviewed) numbered in total (only) 27 and were excluded from the analysis. If we do not count the 27 potential fakes among the valid interviews, the AAPOR response rate (RR2) is 16.4 percent of 3,372 contacts in the gross sample and the cooperation rate (COOP2) is 24.6 percent of 2,246 cases that were contacted.

In the analysis sample, 60.4 percent of the cases are ICT and 39.6 percent DQ interviews. This corresponds almost exactly to the experimental design (40-60 percent DQ-ICT ratio), so dropouts were non-selective regarding question format. Table 4 shows the distribution of the socio-demographic variables gender, age, and education by question format.[10] The results indicate no differences between question formats, showing that randomization worked as intended. Table 4 also contains the "minimum true prevalences" of the three sensitive behaviors of interest as found in the court records (the frame data contains information about the particular offense that caused the conviction). For example, for about 22 percent of respondents in DQ format and 21 percent of re-

[9] The interviewers were instructed to contact the contact person personally at his or her physical address and (if possible) to immediately carry out the interview or make an appointment for it. In the case that no contact was possible on the first attempt, the interviewers were instructed to make at least three contact attempts before utilizing the code "no contact" for the case overall. In the case the contact person was also listed in the official telephone book, interviewers were also allowed to establish contact by telephone. Alerted by low cooperation rates after the beginning of the field phase, we used incentives of 20 euros for participating in the survey, the amount being paid by the interviewer directly after the interview.

[10] Missing values have been deleted listwise for this analysis. The reader might ask why the proportion of female respondents in the sample is only 25 percent. This small proportion reflects the fact that women tend to commit crimes significantly less often than men. In fact, the value of 25 percent corresponds exactly with the official proportion of women among all suspects of criminal offences, as published by German police authorities (Bundeskriminalamt, 2009, p. 72).

Table 4
Distribution of Socio-demographic Variables and "Minimum True Prevalences" of Sensitive Behaviors in the Sample by Question Format (Means)

	DQ	ICT
Gender (1 = female)	0.25	0.25
Age (years)	39.7	39.7
Education (years of general schooling)	10.8	10.7
Fare dodging	0.055	0.051
Driving without driver's license	0.073	0.108
Driving under the influence	0.216	0.213

Note: All differences between question formats are non-significant. The number of cases for the socio-demographic variables is 215 (DQ) and 329 (ICT), and for the sensitive behaviors 218 and 333, respectively.
Abbreviations: DQ = direct questioning; ICT = item count technique.

spondents in ICT format, it is known that the interviewee has driven a car (or other vehicle) under the influence of alcohol and/or drugs in the years prior to the survey (for which he or she has been convicted under criminal law).[11] Here again, differences between question formats are non-significant.

3.2 Questionnaires and ICT Procedure

In both versions of the questionnaire (DQ versus RRT/ICT), an initial set of sensitive questions (four items) was asked using DQ and RRT, respectively (again, see Wolter, 2012; Wolter and Preisendörfer, 2013 for the findings). Then, after several "filler questions", three further sensitive questions were posed. In the DQ questionnaire, the first question on fare dodging was worded (all questions are translated from German):

"Have you ever dodged a fare, that is, deliberately used a bus or train without having a ticket: yes or no?". Then, after three further filler items, the second question on driving under the influence read: "Have you ever driven a car, a motorbike, a scooter, etc. when you were drunk or under the influence of drugs: yes or no?", followed by the third question: "Have you ever driven a car, a motorbike, a scooter, etc. without a valid driver's license: yes or no?"(driving without a license). This was followed immediately by the short lists of the item count design, worded as follows: "In order to shorten the procedure, we are now going to employ a special technique in which several questions are combined together. I am going to hand you lists with four questions, which you should please read first. Then, please tell me only the number of questions that you answer with 'yes', thus, a number between 0 and 4. For a start, let's have a look at an example".[12] At this point, the interviewer handed an example list to the respondent and explained the procedure. Then, the interviewer continued with "Is everything clear? Here we have the actual lists. Please do not tell me the answers to the individual questions, just say the total number of questions you answered with 'yes"'. There were three item lists printed on three separate pages that were handed over by the interviewer one after another (see below for the lists' contents).

Table 5
Wording of the Long Lists in the ICT[a] Design

Long list 1:
- Have you had a traffic accident in the last three months?
- Have you used a taxi in the last seven days?
- *Have you ever driven a car, a motorbike, a scooter, etc. when you were drunk or under the influence of drugs?*
- Have you ever been on a tram?
- Do you travel by bicycle?

Long list 2:
- Have you ever been abroad?
- Have you ever used a taxi?
- Have you been on a plane this week?
- *Have you ever dodged a fare, that is, deliberately used a bus or a train without having a ticket?*
- Did you wash your car in an automatic car wash yesterday?

Long List 3:
- Have you ever been on a night train, i.e., in a sleeping car?
- Have you ever jumped a red traffic light?
- Have you driven yourself this week once or several times?
- *Have you ever driven a car, a motorbike, a scooter, etc. without a valid driver's license?*
- Did you go on a skiing holiday in January this year?

[a] item count technique

In the ICT format, no direct questions on fare dodging, driving under the influence, or driving without a driver's license were asked (all "filler questions", however, were also included in this question mode). At the same point in the

[11] We call these values "minimum true values" because they constitute the lower bound of the real true value for the sample. If a person has been convicted of another crime than those analyzed here, he or she may, before or after the conviction, of course have also committed one of the three delinquent behaviors analyzed here.

[12] We thought very carefully about how to explain or "legitimize" the list principle for DQ respondents. Finally, our approach ("in order to shorten the procedure") worked very well and, indeed, better than anticipated.

questionnaire as in the DQ version, the interviewer read out: "For the next three questions, we are going to use another[13] special technique that again guarantees you complete anonymity. I am going to hand you lists with five questions, which you should please read first. Then, please tell me only the number of questions that you answer with 'yes', thus, a number between 0 and 5. For a start, let's have a look at an example". After this example, the procedure continued as in DQ format, but with the difference that the lists presented also contained the sensitive questions on fare dodging, driving under the influence, or driving without a driver's license.

The content of the filler items was chosen so that two items were expected to have low prevalences and two high. Furthermore, because the three sensitive questions all pertained to the topic of transport or traffic, the filler items also dealt with behavior in the field of transport/traffic. The wording of the long lists is documented in Table 5. The short lists contained the same items without the sensitive item. The interviewers reported no serious problems with the ICT procedure, and there were no respondents who refused to follow it. Although generally preferable due to lower standard errors of the estimates, we decided not to employ the double-lists design of ICT (see section 2) because this would have extended the interview time considerably (three additional lists for every respondent). Furthermore, we were concerned about possible doubts and problems among the DQ respondents who could be confused by a design that requires them to answer some questions directly and others not. In order to not endanger the main purpose of the study, the RRT validation, we did not experiment with complex ICT designs.

In addition to the evaluation of the three sensitive questions, we also wish to focus on subjective measures of survey quality. We will evaluate whether the anonymization of the interview situation by ICT can alleviate problems such as the discomfort of interviewers and respondents, or anonymity and privacy concerns. Table 6 gives an overview of the five indicators used for this evaluation. "Trust in anonymity" and "discomfort of the respondent when answering the sensitive questions" are variables based on the respondents' answers. "Willingness to respond", "credibility of the answers", and "discomfort of the interviewer when asking the sensitive questions" were rated by the interviewers immediately after the interview. The hypothesis is that trust in anonymity, willingness to respond, and credibility of the answers are higher in ICT than in DQ format, and that interviewers and respondents feel less uncomfortable in ICT than in DQ format.

3.3 Statistical Methods

Empirical analysis of the subjective measures of survey quality will be carried out using conventional statistical methods. However, further consideration is needed in relation to the methods for analyzing ICT data. Here, we follow the recommendations given by Blair and Imai (2012, p. 72). We will first explore whether *design effects* of the ICT design could have reduced the validity of the estimates. According to Blair and Imai, design effects occur when the response behavior to the non-key items is affected by introducing the sensitive item. Blair and Imai propose a statistical test for these effects (see Blair and Imai, 2012, 63-65 for details) which we apply to our data using the "list" package of the software R (Blair & Imai, 2013).[14] We then compare prevalence estimates of the three sensitive items by question format. The formulae for calculating the estimates for ICT and their standard errors were already given above; see formulae (1) and (2). To test for significant differences between question formats, we calculate z scores using formula (3).

$$z = \frac{\widehat{\pi}_{\mathrm{ICT}} - \widehat{\pi}_{\mathrm{DQ}}}{\sqrt{\mathrm{Var}(\widehat{\pi}_{\mathrm{ICT}}) + \frac{\widehat{\pi}_{\mathrm{DQ}}(1-\widehat{\pi}_{\mathrm{DQ}})}{n_{\mathrm{DQ}}}}} \tag{3}$$

We then intended to use multivariate regression in order to account for question format-specific differences depending on the socio-demographic variables gender, age, and education. The motivation for analyzing these variables is the fact that socio-demographic factors are usually the first ones to be investigated as determinants of sensitive behavior. For example, several studies have discussed education as an important factor in misreporting in surveys (e. g., Bernstein et al., 2001; Ostapczuk, Musch, & Moshagen, 2009; Preisendörfer & Wolter, 2014). Conventional binary logistic regression

[13] The RRT module in the questionnaire was positioned before the ICT module, so respondents already had experienced RRT. Our design, however, has a slight potential weakness in the fact that the ordering of the questions slightly differs between question formats: DQ respondents answered the three sensitive questions before being introduced to the short lists of ICT while ICT respondents did not. This could potentially induce order effects.

[14] The test proposed by Blair and Imai is based on the comparison of the estimated proportions that a particular number of items in the short-list and the long-list group were included in the respondents' answers. The addition of the sensitive item to the short list should enlarge (or be equal to) the corresponding proportion in the respective answer category in the long-list group (and not lessen it). The calculation is as follows: For both groups, the proportions of respondents that gave at least a corresponding number of yes-answers are calculated (1); see Glynn (2013, 165f.). The subtraction of these proportions in the short-list group from those in the long-list group yields the proportions of respondents in the long-list group who counted the respective number of non-key items and the sensitive item in their answer (2). The subtraction of (2) from (1) for the long-list group yields the estimated proportion of respondents in the long-list group who counted the respective number of non-key-items (and not the sensitive item) in their answer (3). If negative values occur in (2) or (3), design effects are likely to have occurred. The statistical test (a test of two first-order stochastic dominance relationships) verifies whether these possible negative proportions could have arisen by chance; see Blair and Imai (2012, pp. 63-65) for details.

can be used for DQ cases. There are several approaches to multivariate regression using ICT data (see also section 2). We again follow recommendations by Blair and Imai (2012) and use the maximum likelihood (ML) estimator developed in their paper and implemented in R (Blair & Imai, 2013), because empirical evidence and results from simulation analyses by Imai (2011) have shown that this estimation strategy yields more efficient estimates than other approaches. However, we faced several problems in estimating the regression models, probably due to the relatively small number of cases in our study and the low statistical efficiency of ICT data (with four non-key items). Some models did not converge and the results of the ML estimator were unstable and deviated from those of other estimators (proposed by Blair & Imai, 2012). Models that account for floor and ceiling effects did not converge, either. Therefore, we will present a simple model for the variable "driving under the influence" only since results from different estimation techniques were stable for this variable.

Problems and inconsistent results using multivariate regression with ICT data have also been reported by Comşa and Postelnicu (2013). Our conclusion regarding the difficulties that we and other authors have encountered is that regression analysis using ICT data is not (yet) as straightforward as sometimes claimed. Different estimators yield inconsistent and/or unstable results and large sample sizes are necessary.

4 Results

One problem that is often mentioned with reference to sensitive questions is the conjecture that they cause higher rates of item non-response (Lensvelt-Mulders, 2008, 464f.). However, Tourangeau and Yan (2007, p. 862) point out that they "are unaware of any studies that systematically examine this hypothesis". Regarding our data, we find no support at all for the hypothesis that non-response is elevated by sensitive questions: only two respondents, one in DQ and one in ICT format, refused to answer one of the sensitive questions. Also, all respondents followed the list procedure in the DQ version and answered the questions in the short lists. These results are also in favor of ICT because the technique does not yield higher non-response rates than DQ.

Analysis of our subjective measures of survey quality, i. e., trust in anonymity and discomfort when answering the sensitive questions (as indicated by the respondents), willingness to respond, credibility of answers and discomfort when posing the sensitive questions (as indicated by the interviewers) are depicted in Table 7. Here, the results are mixed. Trust in anonymity is slightly higher in DQ than in ICT format (with borderline statistical significance). One explanation for this counter-intuitive result could be that respondents' concerns about anonymity are intensified by the introduction of "special questioning techniques" and by the linguistic framing of the interview situation as "now some sensitive questions".[15] Regarding willingness to respond, credibility of the answers, and respondent's discomfort, we found no differences between question formats. This indicates that the use of ICT does not help improve these subjective factors of survey quality. However, ICT certainly does help the interviewers deal with sensitive questions: whereas in DQ format, 27 percent of interviewers report feeling uncomfortable when asking these questions, in ICT format only about 12 percent did so. All in all, it seems that regarding subjective measures of survey quality, ICT is particularly helpful for interviewers and less so for respondents—an effect that could, nonetheless, improve data quality.

Before turning to the prevalence estimates of self-reported delinquency by question format, Table 8 reports the distribution of answers to the three short and three long lists. What is of particular interest here are possible ceiling effects, i. e., respondents who answer all four non-key items with "yes". In these cases, the anonymization of the ICT procedure is thwarted because five positive answers in the long-list groups indicate the affirmation of the sensitive behavior. The results in Table 8 indicate that the choice of the non-key items worked well: only very low proportions of respondents report four "yes" answers in the short-list groups (1.8 percent, 0.5 percent, and 0.9 percent, respectively) and five "yes" answers in the long-list groups (1.5 percent, 1.5 percent, and 2.4 percent, respectively). In order to test for design effects (see Section 3), we conducted the statistical test proposed by Blair and Imai (2012). The null hypothesis that there are no design effects cannot be refuted for all three items ($p = 0.23$ for fare dodging, $p = 0.20$ for driving without a driver's license, and $p = 0.16$ for driving under the influence). We conducted an additional field experiment in order to test for potential weaknesses of the item count design and asked 95 students in our courses at the University of Mainz to answer the four non-key items from the short list for "driving under the influence" separately, and 90 students to answer the items in the list format (report only the total number of "yes" items) as presented in the ICT design.[16] The result, a mean of 2.17 items answered with "yes" in the separate item group and a mean of 2.18 in the list group ($t = 0.147$, $p < 0.89$), indicates that the response behavior to items presented in the list form as in the ICT design is no different from the one with separate

[15] Also, it must be noted that the two experiments in the study—DQ-RRT-comparison and DQ-ICT-comparison—were conducted on the same set of respondents within the same conditions; the first three items in Table 7 do not exclusively pertain to ICT so the results could potentially be affected by the RRT procedure in the interview. The last two items in Table 7, however, are ICT-specific.

[16] The students received short questionnaires that they filled out in about three minutes at the beginning of the respective lecture. The question format (separate questions—item list) was assigned randomly.

Table 6
Subjective Indicators of Survey Quality in DQ and ICT Formats

Indicator	Coding	Wording
Trust in anonymity (R)	0 = none . . . 4 = very strong	"To what extent do you trust in our measures regarding anonymity and data protection?" 5-point-answer scale.
Willingness to respond (I)	1 = medium/poor 0 = otherwise	"All in all, how was the respondent's willingness to answer the questions?" 3-point answer scale.
Credibility of answers (I)	1 = medium/low 0 = otherwise	"In your view, how credible are the respondent's statements?" 5-point answer scale.
Sensitive Questions: respondent uncomfortable (R)	1 = very/a little 0 = otherwise	"How uncomfortable did you feel answering the questions about traffic (fare dodging, driving without a driver's license, driving under the influence)?" 3-point answer scale.
Sensitive Questions: interviewer uncomfortable (I)	1 = very/a little 0 = otherwise	"How uncomfortable did you feel asking the questions about traffic (fare dodging, driving without a driver's license, driving under the influence)?" 3-point answer scale.

I = Interviewer rating immediately after the interview; R = respondent's answer; DQ = direct questioning; ICT = item count technique.

Table 7
Subjective Indicators of Survey Quality in DQ and ICT Formats

	DQ	ICT	All	t or χ^2
Trust in anonymity [0. . .4]	2.944	2.790	2.851	1.799*
Willingness to respond (1 = medium/poor)	0.065	0.097	0.846	1.736
Credibility of answers (1 = medium/low)	0.158	0.182	0.173	0.534
Sensitive questions: respondent uncomfortable (1 = very/a little)	0.247	0.207	0.222	1.192
Sensitive questions: interviewer uncomfortable (1 = very/a little)	0.274	0.116	0.178	22.413***
N	215	329	544	

Reported are means, a t-test for the metric variable "trust in anonymity", and χ^2 tests for dichotomous indicators.
Abbreviations: DQ = direct questioning; ICT = item count technique.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 8
Distribution of Responses to the Short Lists and Long Lists

No. of "yes" answers	Fare dodging		Driving without license		Driving under influence	
	Short list	Long list	Short list	Long list	Short list	Long list
0	3	0	10	10	3	3
1	18	12	61	53	22	15
2	180	92	98	121	150	98
3	13	190	48	113	41	169
4	4	34	1	31	2	40
5	–	5	–	5	–	8
Total	218	333	218	333	218	333

Table 9
Prevalence Estimates of Self-Reported Delinquency in DQ and ICT Formats (Percentage of "Yes" Answers)

			95% C.I.				95% C.I.		
	DQ	SE	lower	upper	ICT	SE	lower	upper	z
Fare dodging	66.1	3.2	59.7	72.4	79.8	5.4	69.3	90.2	2.197*
Driving without a driver's license	34.4	3.2	28.1	40.8	49.4	7.9	34.0	64.7	1.763
Driving under the influence	45.9	3.4	39.2	52.5	67.9	6.2	55.7	80.1	3.108**
N	218				333				

Reported are prevalence estimates, their standard errors, the 95%-confidence intervals of the estimates, and the z-score of the difference between DQ and ICT estimates.
Abbreviations: DQ = direct questioning; ICT = item count technique; CI = confidence interval.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

items.

All in all, the results thus far are in support of our ICT design: ceiling effects are almost completely avoided with the choice of the filler items, and there is no evidence of design effects, which means that response behavior to the filler items is not affected by introducing the sensitive items in the long lists.

Table 9 documents the main results of our paper, namely, prevalence estimates for the sensitive behaviors in DQ and ICT format. The results show that ICT estimates are higher than DQ estimates for all three items. Whereas 66 percent of the respondents admit to having dodged a fare in DQ format, 80 percent do so in ICT format. 34 percent report having driven a car (or other vehicle) without holding a valid driver's license in DQ format, and 49 percent in ICT format. Finally, 46 percent admit to driving under the influence in the DQ version, and 68 percent in the ICT version.[17] The $z - values$ show that the estimates of the sensitive behaviors in ICT format are significantly higher than in DQ format (on a 5-percent level for fare dodging, on a 10-percent level for driving without a driver's license, and on a 1-percent level for driving under the influence). Therefore, we interpret our findings as a significant result in support of ICT: people more often admit truthfully to delinquent behavior in ICT than in DQ format.

In the analysis of socio-demographic determinants of the three delinquent behaviors, we faced several estimation problems as described in Section 3. Hence, we present one illustrative analysis only (for which we obtained consistent results over different estimation methods) that sheds more detailed light on the results presented above. Table 10 shows the results of two simple regression models—one for DQ and one for ICT cases—for the item "driving under the influence". For both models, the coefficients represent logit coefficients. In both models, we observe a highly significant gender effect: women report less often than men having driven a car (or other vehicle) under the influence of alcohol and/or drugs. However, the magnitudes of the effects suggest that this effect is more pronounced in ICT format.[18] This is confirmed by calculating the predicted probabilities of the model as depicted in Figure 1. Whereas for women there is no substantial difference, for men we find a 40-percentage-point difference between question formats. Conversely, differences in estimated drunk driving prevalence between women and men are less distinctive in DQ compared to ICT format.

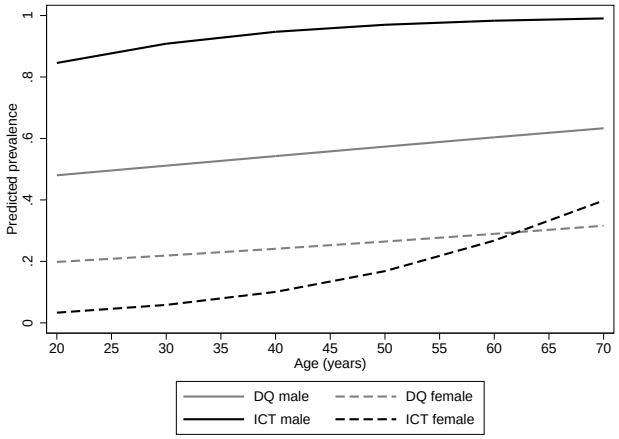

Figure 1. Predicted Probabilities of Driving under the Influence
Note: Prevalence estimates are calculated from the regression models in Table 10 at the mean of years in education.

These results point to a conclusion that we have already drawn with reference to RRT (Wolter, 2012; Wolter & Preisendörfer, 2013): special questioning techniques such as ICT do not exert the same effect on response behavior for all types of respondents. This finding has (at least) two con-

[17] Note that the estimates in both DQ and ICT format are considerably higher than the "minimum true values" gathered from the court records (see Table 4).

[18] A z test comparing the two logit coefficients of the gender effects yields a significant difference on a 5%-level ($z = 2.24$).

Table 10
Determinants of Self-Reported "Driving under the Influence"

	DQ	SE	ICT	SE
Females	−1.32***	0.36	−5.07**	1.64
Age (decades)	0.13	0.09	0.59	0.47
Education (yrs.)	0.09	0.06	0.24	0.29
Intercept	−1.32*	0.64	−2.13	3.50
LL	−139.37		−669.03	
N	215		329	

Unstandardized logit coefficients and their standard errors are reported from a conventional logistic regression for DQ, and from the maximum likelihood estimator (constrained model) provided in the list package for R (Blair & Imai, 2013).
Abbreviations: DQ = direct questioning; ICT = item count technique; LL = log likelihood.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

sequences. First, in developing questioning techniques for asking sensitive questions, research should focus less on the (probably unsolvable) problem of finding a technique that is able to remedy misreporting for all respondents in all interview situations at all times and more on the knowledge of which survey and question technique works best for a certain interview setting with certain respondent characteristics. Second, any analyses on determinants of sensitive behaviors should carefully take into account that response behavior varies according to respondent characteristics, question format, and the interaction between respondent characteristics and question format.

5 Discussion

The aim of this article was to provide empirical evidence on the effectiveness of ICT in avoiding misreporting by survey respondents answering sensitive questions. In an experimental study, we compared the performance of ICT with that of standard DQ with respect to subjective measures of survey quality and prevalence estimates for three sensitive behaviors in the field of self-reported delinquency. Furthermore, we addressed the question of multiple regression analyses on socio-demographic determinants of sensitive behavior using ICT data. Our special sample composition—all respondents had actually been convicted under criminal law—avoids the problem that delinquent or other sensitive behaviors are often rare in the general population which, in turn, requires very high sample sizes for comparative studies of different question formats.

Our analyses yield three main findings. First, subjective measures of survey quality such as trust in anonymity or willingness to respond are not affected positively by ICT except that interviewers feel less uncomfortable than in DQ format when asking sensitive questions. Second, all prevalence estimates of the three delinquent behaviors investigated in this paper are significantly higher in ICT than DQ format. In accordance with the "more is better" assumption and an expected underreporting of these behaviors in DQ format, ICT estimates are more valid. Third, a regression analysis with the item "driving under the influence" shows that the ICT effect varies by gender—it is only for male respondents that we find a clear positive ICT effect. This led to the conclusion that research on determinants of socially loaded behaviors should be more sensitive to the problem of systematic misreporting depending on respondent characteristics and the interview setting.

Overall, we view ICT as a promising alternative to standard questioning techniques and to other special techniques—particularly RRT. The appeal of ICT compared to RRT lies in its simplicity (Blair & Imai, 2012, p. 72) and more encouraging results regarding the effectiveness of ICT in remedying response bias. However, one main drawback, especially for multivariate regression models, is the large sample sizes needed for sufficient statistical power.

For future research, we perceive two key desiderata. First, a comprehensive meta-analysis should investigate the combined effect of the technique on response validity throughout the studies that have been published on the topic (see Table 2), particularly in the last few years. This meta-analysis should also try to identify factors that favor the success of ICT over DQ (such as the sensitivity of questions or design characteristics of ICT procedures). Second, validation studies with known individual true values of sensitive behaviors should be conducted in order to gain further insight into the power of ICT designs. Employing validation studies, research should also focus on the question of whether ICT's effectiveness varies according to different respondent types and interview-situational characteristics.

Acknowledgments

This work was supported by the German Research Foundation (DFG) [Grant PR 237/6 to Peter Preisendörfer] within the DFG priority program "Survey Methodology".

For helpful comments, we would like to thank Peter Preisendörfer, Jürgen Schiener, and two anonymous reviewers of SRM.

References

Ahart, A. M. & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: evaluating the unmatched count technique. *Organizational Research Methods*, *7*(1), 101–114.

Anderson, D. A., Simmons, A. M., Milnes, S. M., & Earleywine, M. (2007). Effect of response format on endorsement of eating disordered attitudes and behaviors. *International Journal of Eating Disorders*, *40*(1), 90–93.

Barton, A. H. (1958). Asking the embarrassing question. *Public Opinion Quarterly*, *22*(1), 67–68.

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.

Benson, L. E. (1941). Studies in secret-ballot technique. *Public Opinion Quarterly*, *5*(1), 79–82.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting. Why it happens and why it matters. *Public Opinion Quarterly*, *65*(1), 22–44.

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

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* (pp. 149–174). DHHS Publication No. SMA 05-4044, Methodology Series M-5. 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 from http : / / cran . r - project . org / web / packages/list/list.pdf

Bundeskriminalamt (Ed.). (2009). *Polizeiliche Kriminalstatistik 2009 Bundesrepublik Deutschland*. Wiesbaden: Bundeskriminalamt.

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.

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.

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.

Durham III, A. M. & Lichtenstein, M. J. (1983). Response bias in self-reported surveys. Evaluating randomized response. In G. P. Waldo (Ed.), *Measurement issues in criminal justice* (pp. 37–57). Beverly Hills: Sage.

Esser, H. (1986). Können Befragte lügen? Zum Konzept des "wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *38*(2), 314–336.

Esser, H. (1991). Die Erklärung systematischer Fehler in Interviews: Befragtenverhalten als "rational choice". In R. Wittenberg (Ed.), *Person—Situation—Institution—Kultur. Günter Büschges zum 65. Geburtstag* (pp. 59–78). Berlin: Duncker and Humblot.

Fox, J. A. & Tracy, P. E. (1986). *Randomized response. A method for sensitive surveys* (Vol. 07-058). Newbury Park: Sage.

Gilens, M., Sniderman, P. M., & Kuklinski, J. H. (1998). Affirmative action and the politics of realignment. *British Journal of Political Science*, *28*(1), 159–183.

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.

Gonzalez-Ocantos, E., de Jonge, C. K., Meléndez, C., Osorio, J., & Nickerson, D. W. (2012). Vote buying and social desirability bias: experimental evidence from nicagarua. *American Journal of Political Science*, *56*(1), 202–217.

Hadaway, C. K., Marler, P. L., & Chaves, M. (1993). What the polls don't show: a closer look at U.S. church attendance. *American Sociological Review*, *58*(6), 741–752.

Heerwig, J. A. & McCabe, B. J. (2009). Education and social desirability bias: the case of a black presidential candidate. *Social Science Quarterly*, *90*(3), 674–686.

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.

Hubbard, L., Casper, R. A., & Lessler, J. T. (1989). Respondent reactions to item count lists and randomized response. *Online Proceedings of the Survey Research Methods Section of the American Statistical Association*, *12*, 544–548. Retrieved from http://www.amstat.org/sections/srms/Proceedings/papers/1989_097.pdf

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.

Kane, J. G., Craig, S. C., & Wald, K. D. (2004). Religion and presidential politics in Florida: a list experiment. *Social Science Quarterly*, *85*(2), 281–293.

Katz, D. (1942). Do interviewers bias poll results? *Public Opinion Quarterly*, *6*(2), 248–268.

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.

Krumpal, I. (2008). Evaluation of the effectiveness of the randomized response technique and the item count method in the telephone survey mode. In S. Balbi, G. Scepi, G. Russolillo, & A. Stawinoga (Eds.), *Proceedings of the 7th international conference on social science methodology, RC33—logic and methodology in sociology (ISA)*. Naples: Jovene Editore.

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

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. R., & Mellers, B. (1997). Racial prejudice and attitudes toward affirmative action. *American Journal of Political Science*, *41*(2), 402–419.

LaBrie, J. W. & Earleywine, M. (2000). Sexual risk behaviors and alcohol: higher base rates revealed using the unmatched-count technique. *Journal of Sex Research*, *37*(4), 321–326.

Lee, R. M. (1993). *Doing research on sensitive topics.* Thousand Oaks: Sage.

Lensvelt-Mulders, G. (2008). Surveying sensitive topics. In E. de Leeuw, J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology.* (pp. 461–478). New York: Lawrence Erlbaum.

Lensvelt-Mulders, G., Hox, J., van der Heijden, P., & Maas, C. (2005). Meta-analysis of randomized response research. Thirty-Five years of validation. *Sociological Methods and Research*, *33*(3), 319–348.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, *71*(354), 269–275.

Martinez, M. D. & Craig, S. C. (2010). Race and 2008 presidential politics in Florida: a list experiment. *The Forum*, *8*(2), Article 4.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior.* Unpublished Dissertation: George Washington University, Washington D.C.

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.

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.

Rayburn, N. R., Earleywine, M., & Davison, G. C. (2003a). An investigation of base rates of anti-gay hate crimes using the unmatched-count technique. *Journal of Aggression, Maltreatment and Trauma*, *6*(2), 137–152.

Rayburn, N. R., Earleywine, M., & Davison, G. C. (2003b). Base rates of hate crime victimization among college students. *Journal of Interpersonal Violence*, *18*(10), 1209–1221.

Redlawsk, D. P., Tolbert, C. J., & Franko, W. (2010). Voters, emotions, and race in 2008: Obama as the first black president. *Political Research Quarterly*, *63*(4), 875–889.

Sackett, P. R. & DeVore, C. J. (2001). Counterproductive behaviors at work. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *International handbook of work psychology* (pp. 145–164). London: Sage.

Smith, L. L., Federer, W. T., & Raghavarao, D. (1974). *A comparison of three techniques for eliciting truthful answers to sensitive questions.*

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

Streb, M. J., Burrell, B., Frederick, B., & Genovese, M. A. (2008). Social desirability effects and support for a female American president. *Public Opinion Quarterly*, *72*(1), 76–89.

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.

Tsuchiya, T. & Hirai, Y. (2010). Elaborate item count questioning: why do people underreport in item count responses? *Survey Research Methods*, *4*(3), 139–149.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A study of the properties of the item count technique. *Public Opinion Quarterly*, *71*(2), 253–272.

Umesh, U. N. & Peterson, R. A. (1991). A critical evaluation of the randomized response method. Applications, validation, and research agenda. *Sociological Methods and Research*, *20*(1), 104–138.

van der Heijden, P. G. M., van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning. Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods and Research*, *28*(4), 505–537.

Walsh, J. A. & Braithwaite, J. (2008). Self-reported alcohol consumption and sexual behavior in males and females: using the unmatched-count technique to examine reporting practices of socially sensitive subjects in a sample of university students. *Journal of Alcohol and Drug Education*, *52*(2), 49–72.

Wimbush, J. C. & Dalton, D. R. (1997). Base rate for employee theft: convergence of multiple methods. *Journal of Applied Psychology*, *82*(5), 756–763.

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 and Research*, *42*(3), 321–353.

Zigerell, L. J. (2011). You wouldn't like me when I'm angry: list experiment misreporting. *Social Science Quarterly*, *92*(2), 552–562.