Item Sum: A New Technique for Asking Quantitative Sensitive Questions

Running Header: The Item Sum Technique

word count:

Text, excluding figures, tables, references, and appendices: 5,223

Footnotes: 646

Abstract: 145

Mark Trappmann*

Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany, phone: +49 911 179 3096, fax: +49 911 179 5912, e-mail: mark.trappmann@iab.de

Ivar Krumpal

University of Leipzig, Institute of Sociology, Beethovenstrasse 15, 04107 Leipzig, Germany, phone: +49 341 97 35693, fax: +49 341 97 35669, e-mail: krumpal@sozio.uni-leipzig.de

Antje Kirchner

Institute for Employment Research, Regensburger Str. 104, 90478 Nürnberg, Germany, phone: +49 911 179 6050, fax: +49 +49 911 179 5912, e-mail: antje.kirchner@iab.de

Ben Jann

University of Bern, Institute of Sociology, Lerchenweg 36, 3000 Bern 9, Switzerland, phone: +41 31 631 4831, fax: +41 31 631 4817, e-mail: jann@soz.unibe.ch

MARK TRAPPMANN is a research director at the Institute for Employment Research, Nürnberg, Germany and a professor for sociology at Bamberg University, Germany. IVAR KRUMPAL is a senior researcher at the Institute of Sociology, University of Leipzig, Leipzig, Germany. ANTJE KIRCHNER is a post-doctoral research associate at the University of Nebraska-Lincoln, USA. BEN JANN is an associate professor at the Institute of Sociology, University of Bern, Bern, Switzerland. The authors thank Debra Hevenstone¸ Roger Tourangeau and two anonymous reviewers for helpful comments on an earlier draft of this article. This work was supported by the Institute for Employment Research (IAB) and by the German Research Foundation [VO 684/11 to Thomas Voss and Karl-Dieter Opp, University of Leipzig]. *Address correspondence to Mark Trappmann, Institute for Employment Research, Regensburger Str. 104. 90478 Nürnberg, Germany, phone: +49 911 179 3096, e-mail: mark.trappmann@iab.de.

Abstract

This article contributes to an ongoing debate about how to measure sensitive topics in population surveys. We propose a novel technique that can be applied to the measurement of quantitative sensitive variables: the item sum technique (IST). This method is closely related to the item count technique (ICT), which was developed for the measurement of dichotomous sensitive items. First, we provide a description of our new technique and discuss how data collected by the IST can be analyzed. Second, we present the results of a CATI survey on undeclared work in Germany, in which the IST has been applied. Using an experimental design, we compare the IST to direct questioning. Our empirical results indicate that the IST is a promising data collection technique for sensitive questions. We conclude the article by discussing the limitations of the new technique and outlining possible improvements for future studies.

Keywords: item count technique, item sum technique, sensitive questions, social desirability, undeclared work

1 Introduction

Asking sensitive questions in surveys is a challenge as respondents are required to self-report behaviors or attitudes that potentially violate social norms. Norm violations are often formally or informally sanctioned, so respondents are reluctant to reveal self-stigmatizing information in a survey interview. Therefore, respondents may choose to misreport on sensitive topics and adjust their answers in accordance with the social norm. More specifically, socially undesirable characteristics are likely to be underreported whereas socially desirable characteristics are likely to be overreported. In addition, some respondents may refuse to answer sensitive questions at all. Such systematic misreporting and item nonresponse may introduce considerable bias to the measurement of sensitive topics and lower the overall data quality of a survey study.

There is a wide range of topics referring to taboos, illegal activities or unsocial opinions that can be considered sensitive in a survey interview (see Tourangeau et al. 2000; Tourangeau and Yan 2007; Krumpal 2013). Respondents tend to systematically underreport norm-violations such as illicit drug use, abortion, social fraud, or plagiarism and overreport norm-conforming activities such as election participation, energy conservation, or exercising. To combat misreporting on sensitive topics, survey designers developed various data collection strategies trying to elicit more honest answers from respondents by increasing the anonymity of the question-and-answer process. One prominent example of such "dejeopardizing techniques" (Lee 1993) is the item count technique (Droitcour et al. 1991; Tsuchiya et al. 2007; Holbrook and Krosnick 2010a), which is also known as the "unmatched count technique" (Ahart and Sackett 2004; Dalton et al. 1994), the "block total response" (Smith et al. 1974; Raghavarao and Federer 1979), or the "list experiment" (Kuklinski et al. 1996; Corstange 2009).

The Item Count Technique (ICT)

In the ICT, two subsamples of respondents are generated via randomization. One of the subsamples is confronted with a long list of items (LL), containing a number of innocuous questions plus the sensitive question of interest (Droitcour et al. 1991). The other subsample receives a short list (SL) that only contains the innocuous questions. For example, to study the prevalence of undeclared work, the following list of questions could be used:

1. Do you use public transportation on more than 5 days per week? (LL and SL)
2. Are you covered by liability insurance? (LL and SL)
3. Did you grow up in the countryside? (LL and SL)
4. Did you engage in any undeclared work this year? (LL only)

Respondents are then asked to indicate the number of items that apply to them (i.e. the total number of "yes" answers), without answering each question individually. Hence, unless a respondent indicates that all or none of the items apply, it remains unknown whether the respondent engaged in the sensitive behavior or not. However, the mean difference of answers between the two subsamples provides an estimate for the population prevalence of the sensitive behavior (Droitcour et al. 1991).

Compared to direct questioning, the ICT provides a higher degree of privacy protection and is thus assumed to yield more reliable self-reports. This expectation has been confirmed in several experimental studies comparing the ICT to direct self-reports. In the majority of these studies on topics such as employee theft (Wimbush and Dalton 1997), risky sexual behavior (LaBrie and Earleywine 2000), hate crime victimization (Rayburn et al. 2003), or shoplifting

(Tsuchiya et al. 2007) the ICT yielded higher prevalence estimates for the sensitive behavior than direct questioning (for an overview see Holbrook and Krosnick 2010a).

Compared to other dejeopardizing techniques, such as the Randomized Response Technique (RRT; Warner 1965)[1], the ICT has the advantage that the underlying concept of counting items is relatively simple and the procedure is easy to administer in surveys. Only a moderate cognitive burden is imposed on the respondent, likely increasing the respondent's ability to comply with the interview protocol and to provide more honest self-reports. Recent empirical studies indicate that the ICT outperforms the RRT in reducing social desirability bias in survey measures of sensitive attributes (see Holbrook and Krosnick 2010a, 2010b). One disadvantage of the ICT, however, is the low statistical power. Estimates obtained from the ICT typically have larger standard errors than estimates from the RRT based on the same sample size (see Coutts and Jann 2011).

Our Study

To our knowledge, the ICT has only been applied to dichotomous items. We therefore present a generalization of the ICT that can be used to measure quantitative sensitive characteristics—we call it the "item sum technique" (IST)—and report the results of an empirical application of the new method.

[1] The basic idea of the RRT is to establish a probabilistic relationship between the observed answer and the sensitive question by using a randomizing device (e.g. coins or dice) in the question-and-answer process (see Fox and Tracy 1986 for an overview of different RRT designs and estimation procedures). If respondents understand the procedure and appreciate the induced privacy protection, they should be inclined to provide more honest answers to sensitive questions compared to direct questioning. In empirical practice, however, serious problems with the RRT can occur. A substantial proportion of respondents may not understand or trust the complex probabilistic concept, and thus provide a self-protective answer irrespective of the outcome of the randomizing device (see Holbrook and Krosnick 2010b). This in turn often leads to invalid prevalence estimates of the sensitive characteristics.

The remainder of the article is organized as follows: In section 2 we describe our new technique. Section 3 describes our empirical study in which we applied the new technique. Section 4 presents the results of the study and in section 5 we draw conclusions and discuss limitations. Details on the IST instructions and the employed statistical methods for data analysis can be found in the appendix.

2 The Item Sum Technique

The item sum technique (IST) works as follows. Analogous to the ICT, two random subsamples are generated, whose respondents either receive a long list of questions (LL) or a short list of questions (SL). The long list contains the sensitive question plus at least one innocuous (i.e. non-sensitive) question, the short list only contains the innocuous question(s). The respondents are then asked to report the *sum* of the answers to the questions in their list. While, in theory, there is no restriction on the number of innocuous questions, it is desirable to keep the lists as short as possible. The variance of the sum of the answers usually increases with the number of individual questions, which reduces the statistical efficiency of the procedure. Furthermore, the cognitive demand of summing up the individual answers is increased with each additional item. We therefore suggest using only one innocuous question. Other than in the ICT where the non-sensitive items are binary, a single innocuous question with many possible values should be enough to make privacy protection credible in the IST.

Both the sensitive and the innocuous questions should be *quantitative,* and preferably (but not necessarily) measured on the same scale (e.g. hours or monetary units). Respondents in the first subsample are asked to report the sum of the answers to both questions; respondents in the second subsample directly provide an answer to the innocuous question. For example, to estimate the extent of undeclared work, the following questions could be used:

1. How many hours did you watch TV last week? (LL and SL)
2. On average, how many hours per week do you engage in undeclared work? (LL only)

Because respondents in the LL subgroup only report the sum of hours from both items, the extent of undeclared work remains unknown at the individual level. Assuming that respondents appreciate this privacy protection, the procedure can therefore be expected to elicit more honest answers to the sensitive question than direct questioning.

To estimate the amount of undeclared work from the IST data we can simply compute the mean difference of answers between the two subsamples. Let Y be the observed answer, S be the sensitive variable (e.g. hours of undeclared work), and C be the non-sensitive variable (e.g. hours of watching TV). In the long-list sample we have $Y = S + C$, in the short-list sample we have $Y = C$. Hence, as long as the two samples are unbiased, we can estimate the expected value of S, μ, as the mean difference of Y between the long-list sample and the short-list sample, that is

$$\hat{\mu} = \bar{Y}_{\mathrm{LL}} - \bar{Y}_{\mathrm{SL}}$$

where $\bar{Y}_{\mathrm{LL}}$ is the mean of Y in the long-list sample and $\bar{Y}_{\mathrm{SL}}$ is the mean of Y in the short-list sample. Furthermore, as long as the samples are independent, the variance of $\hat{\mu}$ can be estimated as the sum of the sampling variances of the two group means, that is

$$\hat{V}(\hat{\mu}) = \hat{V}(\bar{Y}_{\mathrm{LL}}) + \hat{V}(\bar{Y}_{\mathrm{SL}})$$

where standard formulas are used for the variances on the right hand side. Methods to estimate regression models for IST data are outlined in appendix 2.

3 Study Design

We implemented the new technique in a nation-wide CATI survey on undeclared work in Germany. The substantive goal of the project was to estimate the amount of undeclared work (particularly in comparison between employees and recipients of unemployment benefits) and analyze its determinants. In October and November 2010 a total of 3,211 interviews were conducted.

Two samples of respondents with different incentives and opportunity structures for undeclared work were used in the survey. The first sample was drawn from the register of employees maintained by the German Federal Employment Agency (FEA). This register includes all employees who are subject to social security contributions, that is, everyone legally employed in Germany except self-employed and civil servants.It consists of people aged 18-70 who were employed in December 2009. The second sample was drawn from the FEA register of basic income support recipients. It consists of people aged 18-64 who received basic income support ("Unemployment Benefit II", short UB II) in June 2010. For both samples, the latest available registers were used. Nonetheless, employment status or UB II eligibility may have changed until the date of the interview. That is, part of the respondents in the employees sample or the benefit recipients sample were no longer employed or receiving benefits when being interviewed.

The FEA registers only contain telephone numbers for about forty percent of the employees and ninety percent of benefit recipients. Furthermore, it is known from past surveys that some of these numbers are out of date. We therefore tried to complete the numbers using public telephone directories. Nonetheless, 31.8 percent of the employees sample and 8.3 percent of the benefit recipients sample remained without telephone number. All other respondents were sent a letter announcing the survey. During fieldwork, 17.5 percent of the phone numbers in the employees sample and 17.2 percent of the phone numbers in the benefit recipients sample

turned out to be invalid and could not be replaced by a working phone number from the public directories. Among those with a working phone number, about 26 percent agreed to participate in the survey. However, due to the large proportion of missing or invalid phone numbers, the overall response rates were 16.3 and 18.8 percent in the two samples, respectively (RR1 according to AAPOR 2011). Table 1 provides an overview of the sample sizes and response rates.

Table 1: Sample sizes and response rates

	Employees	Benefit recipients	Total
Gross sample	9,996	8,999	18,995
Net sample (with phone number)	6,820	8,250	15,070
Invalid phone number	1,196	1,422	2,618
Non-contact	813	1,564	2,377
Refusal	3,094	3,173	6,267
Ineligible (deceased, moved abroad, outside age range, speaks no German)	104	493	597
Interview completed	1,613	1,598	3,211
Response Rate (AAPOR RR1)	16.3%	18.8%	17.5%

In both samples about one third of the respondents were randomly assigned to direct questioning (DQ), one third to the short-list IST group, and one third to the long-list IST group. Some of the respondents assigned to the IST groups, however, opted out of being questioned using a special technique and were then assigned to the survey with direct questioning.[2] Table 2 gives an overview of the number of respondents in each experimental group and the number of respondents from the IST groups who opted for direct questioning.

[2] To be precise, respondents were given the option to switch to direct questioning after refusing to answer an RRT question measuring the prevalence of illicit work (as a binary variable) earlier in the survey. They were given this option in order to prevent item nonresponse or even interview break-offs. In the subsequent IST experiment, these non-compliers were kept in the direct questioning condition so as not to confront them with a second privacy preserving technique.

Chi-squared tests and two-sided t-tests (assuming unequal variances) indicate that the "defiers" (those who opted for direct questioning) do not significantly differ from the "compliers" (those who stayed with the IST) with respect to gender ($\chi^2 = 0.60$, $p = 0.44$), their attitudes towards undeclared work ($t = -1.32$, $p = 0.19$), the presumed prevalence of undeclared work among friends ($t = 0.44$, $p = 0.65$), and the perceived risk of being caught and sanctioned conducting undeclared work ($t = 1.00$, $p = 0.32$). Defiers, however, are more likely to receive benefits than compliers ($\chi^2 = 33.20$, $p < 0.01$). Because the defiers might also differ from compliers with respect to other, possibly unobserved characteristics, we have to be careful about how to treat them in the data analysis (see below).

Table 2: Number of respondents per experimental condition

	Employees	Benefit recipients	Total
Direct questioning (DQ)	565	580	1,145
Short list IST group			
– Remained with IST	496	460	956
– Opted for DQ[a]	38	90	128
Long list IST group			
– Remained with IST	459	377	836
– Opted for DQ[a]	55	91	146
Total	1,613	1,598	3,211

[a] Note: Cases labeled "Opted for DQ" opted for direct questioning in a prior RRT experiment and were kept in the DQ condition in the IST experiment.

In the DQ mode respondents received a filter question asking whether they engaged in undeclared work in the past year (preceded by a confirmation that responses will be handled confidentially). Depending on the answer to the filter question, respondents were then led to questions about the weekly hours of undeclared work and the monthly income from undeclared work. For respondents who answered "no" to the filter question in the DQ mode, the two variables were set to zero.

Within the IST condition, the first group received two long lists (LL). Each of the long lists contained a sensitive question and an innocuous question. In the first list, the question about weekly hours of undeclared work was paired with a question about the number of hours the respondent watched TV last week; in the second list, the question about monthly earnings from undeclared work was paired with a question about the monthly costs for housing (both in euro). Respondents were asked to get paper and pencil and write down their individual responses to each question, but not to tell them to the interviewers. Respondents were then asked to report the sum of the two answers for each list. Prior to the first questions in the IST format respondents were given an example requiring them to report the sum of two answers which could be validated because the same items were answered in previous parts of the questionnaire. Thus, interviewers could ensure that respondents understood the task. In case that the reported sum did not correspond to the sum of the two previous answers, the respondent received additional explanations (see appendix 1 for the wording in the CAPI instrument). The second IST group received two short lists (SL), each containing just the innocuous question regarding hours of TV or housing costs, respectively. See Figure 1 for the wording of all questions (translated from German).

Figure 1: Wording of the items regarding undeclared work (translated from German)

Direct Questioning (DQ) *S1: On average, how many hours per week do you engage in undeclared work?* *S2: On average, how much do you earn per month from undeclared work?*
Long lists (LL) *C1: How many hours did you watch TV last week?* *S1: On average, how many hours per week do you engage in undeclared work?* *C2: How high are your monthly costs for your apartment or your house? Monthly costs can include rent, utilities, coop or condo fees, and mortgage.* *S2: On average, how much do you earn per month from undeclared work?*
Short lists (SL)

C1: How many hours did you watch TV last week?
C2: How high are your monthly costs for your apartment or your house? Monthly costs can include rent, utilities, coop or condo fees, and mortgage.

Note: Monetary units are in euros.

Apart from the experimental manipulation of the sensitive questions, all respondents received the same questionnaire covering items on demographics, employment, social networks, opportunity structures, attitudes and norms. Before asking the questions on undeclared work, a definition of undeclared work based on the German legal context was provided to the respondents.[3]

4 Results

This section provides an analysis of the item sum data from the CATI survey. The goal is to evaluate how the reported amount of undeclared work depends on the data collection method (direct questioning vs. IST) and the labor market status (employees sample vs. benefit recipients sample).

Table 3 reports the estimated mean hours of and earnings from undeclared work for both samples across experimental conditions. The results are based on the *realized* assignment to either direct questioning (DQ) or the IST and, hence, ignore the fact that some respondents were initially assigned to a different mode. We therefore further subdivide the DQ group into those who were originally assigned to DQ and those who opted for DQ.

[3] Undeclared work (including moonlighting) is defined as paid work that is hidden from (tax) authorities due to various reasons. Illegal (e.g. drug trafficking) or unpaid activities (e.g. neighborly help) are usually not included in the definition of "undeclared work" or "informal employment" (cf. Schneider and Enste 2007; Pedersen 2003; Williams 2009).

Table 3: Means of hours of undeclared work per week and monthly earnings from undeclared work depending on questioning mode and sample (standard errors in parentheses)

	Hours of undeclared work		Earnings from undeclared work	
	Employees	Benefit recipients	Employees	Benefit recipients
Direct questioning (DQ)	0.07	0.14	1.8	3.4
	(0.03)	(0.06)	(0.7)	(1.2)
– Assigned to DQ	0.07	0.19	1.9	4.4
	(0.04)	(0.08)	(0.8)	(1.5)
– Opted for DQ	0.05	0.00	1.1	0.0
	(0.05)	(-)	(1.1)	(-)
Item sum technique (IST)	0.85	-0.17	113.8	83.4
	(0.70)	(1.06)	(40.1)	(27.4)

Direct questioning leads to an estimate of 0.07 hours of undeclared work per week for employees and 0.14 hours for benefit recipients. Using IST, the estimate for employees rises to 0.85 hours while for benefit recipients a negative estimate of -0.17 hours results. A negative value for the number of hours of undeclared work does not make sense, of course. However, note that the estimate is not significantly different from zero. For mean earnings from undeclared work, we get a DQ estimate of 1.8 euros per month for employees and 3.4 euros per month for benefit recipients. If using the IST, the estimates rise substantially to 113.8 and 83.4 euros per month, respectively.

In Table 4 the differences between the estimates from direct questioning and the IST are shown. The first row ("naive estimate") contains the differences between the raw estimates as reported in Table 3. For hours of undeclared work the effect of the questioning method does not appear to be significant (with a p-value of 0.26 in the employees sample and 0.77 in the benefits recipients sample, respectively). For earnings from undeclared work, however, the IST yielded significantly higher estimates than direct questioning in both samples ($p < 0.01$).

Table 4: Differences between direct questioning and the IST (standard errors in parentheses)

	Hours of undeclared work		Earnings from undeclared work	
	Employees	Benefit recipients	Employees	Benefit recipients
Naive estimate	0.78	-0.31	112.0*	80.0*
	(0.70)	(1.06)	(40.1)	(27.4)
ITT estimate	0.70	-0.32	99.7*	62.5*
	(0.62)	(0.87)	(36.4)	(23.0)
IV estimate	0.78	-0.40	111.9*	77.9*
	(0.70)	(1.08)	(41.0)	(27.1)

Standard errors for the ITT and IV estimates were obtained by the bootstrap method (1000 replications, stratified by assigned experimental condition); for estimation methods see appendix 2.

* $p < 0.01$ (two-sided t-tests)

A comparison of the naive estimates might provide biased estimates of the effects of the questioning method because the group of respondents who opted out of the IST may be selective. Indeed, if we compare the results for those who were originally assigned to DQ with those who opted for DQ, we see that the latter have lower estimates for the sensitive questions (Table 3). One approach to deal with such treatment assignment non-compliance bias in randomized experiments is to compute the so-called intention-to-treat effect (ITT; see e.g. Hollis and Campbell 1999, Newell 1992): Instead of measuring the effect of actually receiving the treatment, the effect of being *assigned* to the treatment is estimated. The ITT is a conservative estimate for the causal treatment effect. That is, because only a fraction of the assigned treatment group is actually treated, the ITT is a weighted average of the true treatment effect and a zero effect. Hence, other than the naive approach, the ITT protects from possible overestimation of the causal effect of the treatment.

ITT estimates of the effect of the questioning method can be found in the second row of Table 4 (although the ITT principle is conceptually simple, its application is somewhat involved in our situation; see appendix 2 for details). For hours of undeclared work the ITT estimates of

the effect of the IST are almost identical to the naive estimates, supporting our earlier finding that in both samples the IST did not yield significantly higher estimates than direct questioning. For the reported earnings from undeclared work, however, the ITT estimates are smaller than the naive estimates. Yet, the effects are still significant at the one percent level. Hence, also if we employ the conservative ITT approach we find that the IST had a substantial effect on the reported earnings in both samples.

As indicated above, treatment effects may be underestimated by the ITT procedure. We can improve on the ITT by using an instrumental variables (IV) approach, in which the realized treatment is instrumented by the assigned treatment. If a treatment effect is homogeneous (i.e. the same for everyone), then such an approach yields a consistent estimate of the causal effect. Alternatively, in the case of a heterogeneous treatment effect, the so-called local average treatment effect (LATE) is estimated (Angrist et al. 1996). This is the average treatment effect for the subpopulation of those who actually received treatment.

The last row of Table 4 displays the IV estimates of the effect of the IST on response behavior for our data (for methods again see the appendix). As can be seen, the results from the IV procedure are almost identical to those from the naive comparison of the direct questioning estimates and the IST estimates. Hence, we conclude that noncompliance with the treatment assignment did not substantially bias the data, so that the findings based on the naive estimates appear valid. To summarize, irrespective of the employed estimation method, we find that in three out of four independent comparisons the IST yielded larger estimates than direct questioning. In two out of these three cases the difference is statistically significant.

The differences reported in Table 4 are average effects over the respondents in our experimental groups. The performance of the IST, however, might depend on cognitive skills

of the respondents. Therefore, using the regression technology outlined in the appendix, we evaluated whether the treatment effects vary by indicators for old age and low education (in lack of better data we use old age and low education as proxies for cognitive skills). No significant interaction effects were found in these models (not shown).

5 Discussion and Conclusions

In this paper, we presented a new method, the Item Sum Technique (IST), for the measurement of quantitative sensitive characteristics. Compared to alternative methods, such as quantitative RRT schemes[4], the IST has several advantages: (1) a randomizing device is not required; (2) the cognitive effort demanded from respondents is relatively low; (3) implementation is easily possible in both interviewer- and self-administered interviews. The experimental evidence of our empirical study suggests that the IST is a promising data collection technique. It yielded significantly higher estimates of earnings from undeclared work than direct questioning in both the employees sample and the benefit recipients sample. Also for hours of undeclared work, estimates from the IST were higher than from direct questioning in one of the two samples, although not significantly so. Survey researchers aiming at measuring sensitive behaviors at a quantitative scale could therefore benefit from using the IST. Nonetheless, our study can only be regarded as a first step in the development and evaluation of the new technique.

[4] Almost all empirical implementations of the RRT focus on dichotomous sensitive variables. Although RRT schemes tailored to quantitative sensitive characteristics have been proposed in the literature (e.g. Himmelfarb and Edgell 1980; Eichhorn and Hayre 1983; Gjestvang and Singh 2007), there is little evidence on how these techniques perform in practice. Due to their complexity, it can be expected, however, that RRT schemes for quantitative variables are even more difficult to implement than dichotomous RRT schemes.

One issue of our study is that a considerable share of respondents did not remain in the treatment condition they were initially assigned to, thus compromising the randomization of experimental groups. Using intention-to-treat and instrumental variables strategies, however, we believe that we convincingly demonstrated that our findings are robust despite this problem.

A more serious concern is that in the direct questioning condition the quantitative sensitive questions were preceded by a filter question on whether any illicit work was carried out at all, while no such filter question was present in the item sum condition.[5] Evidence suggests that filtering a quantitative question may lead to underestimation of the quantity of interest. For example, in a study on crime victimization, Knäuper (1998) found that direct questioning estimates were twice as high compared to estimates obtained from a filtered question. Such biases are most likely due to differential interpretations of a construct depending on question format in cases where the construct is not clearly defined. In our IST study, however, an explicit definition of undeclared work was given directly before the relevant questions were asked (see section 3). Furthermore, for earnings from undeclared work our IST estimates are magnitudes higher than the direct questioning estimates (by a factor of 63 in the employees sample; by a factor of 25 in the benefit recipients sample; see Table 3). Thus, we do not believe that the filtering could explain the differences observed in our study.

One can also speculate that carry over effects from the preceding RRT experiment on the subsequent IST answers might be an issue in our study. Variations in the question wording or

[5] The IST experiment described in our article was preceded by an experiment using direct questioning versus RRT to estimate the prevalence of undeclared work. Respondents from the direct questioning group skipped the subsequent quantitative questions if they answered that they did not carry out any undeclared work. In contrast, all respondents from the RRT group were directed to the quantitative questions (and randomized into either the short list or long list of the IST), since filtering based on RRT answers is not possible.

question format of preceding items could in principle influence respondents' cognitions as well as their answers to subsequent items (for an overview see Schwarz and Sudman 1992). Yet, in our study we have no strong theoretical reasons to believe that our IST estimates would be biased in a specific direction. Nonetheless, to further optimize the questionnaire design of IST experiments in future studies, one should make sure that all preceding items have identical format and wording for all experimental groups.

What concerns us most is that the IST was only successful for one of the two questions. For earnings from undeclared work, the IST impressively outperformed direct questioning, but no significant differences were found for the question on the number of hours of undeclared work. We do not believe that this null-result is due to a lack of statistical power. First, relative differences as observed for earnings (which one would expect if working hours and earnings are proportional) would have been easily detected given the power of the study. Results from a simulation based on the characteristics of our data (not shown) indicate that the power to detect such a difference at the 5% level would have been about 97%. Second, in both the employees sample and the benefit recipients sample, the effect is absent, which makes us believe that the pattern is systematic. A candidate explanation may lie in the choice of innocuous items. While the item on housing costs appears unproblematic, the question on the number of hours a respondent watched TV, which was paired with the question on hours of undeclared work, might not have been an optimal choice. There is evidence that answers to this question strongly depend on question format (Schwarz et al. 1985). In addition, the question on watching TV might be considered sensitive by some respondents. In this case, respondents would tend to underreport their TV consumption if asked directly, but possibly they would underreport even more in the IST condition since the sum of *two sensitive* variables has to be reported. Perhaps an explanation might also be that there were learning effects across the two IST questions.

As illustrated by the above qualifications, further experimental research is needed to fully understand the mechanisms at work when respondents are confronted with sensitive questions in the item sum format. Obtaining unbiased estimates for sensitive variables by the IST rests on a number of assumptions. In the following we outline how an implementation of the IST that maximizes the credibility of these assumptions could look like and develop ideas about how the validity of the assumptions could be evaluated (see Blair and Imai 2012 or Glynn 2013 for similar discussions in the context of the item count technique).

First, it is necessary to assume that the respondents comply with the design. Therefore, careful cognitive pretests are necessary to make sure that respondents fully understand the procedure. In addition, measures could be implemented to minimize nonresponse or noncompliance with the IST (cf. de Leeuw et al. 2003: 154 f. for an overview). These could, for example, include interviewer probing after an initial "do not know"-answer.

Second, it is crucial that the answers to the non-sensitive item are independent of the question format, that is, that the answers do not depend on whether they are given directly or serve as a summand in the item sum format ("no design effect", see Blair and Imai 2012). It is important, therefore, that the innocuous item is truly non-sensitive and not affected by social desirability bias itself. Furthermore, the summation task should be made as easy and convenient as possible. Although adding two numbers might appear simple, the ability to perform such a task error free is likely to depend on cognitive capabilities.[6] Even if most respondents exhibit sufficient cognitive skills to correctly add the items, they might engage in satisficing, most likely leading to rounding errors (see Tourangeau et al. 2000 for an overview

[6] We assume that the task is easier for respondents if the items are on the same scale. Also, the items should be disjoint to prevent confusion over whether the overlap must be subtracted from the sum or not.

on rounding). Such "heaping" might bias the results if the net effect of rounding is different between the short list and the long list, for example due to differing distributions of the true values around focal rounding points. In order to reduce summation as well as rounding errors, in our telephone survey we asked respondents to write down the answers to the individual items before summing them up. This is obviously not to be recommended in face-to-face settings.

The "no design effect" assumption can be evaluated, for instance, by administering a survey in which one experimental group is asked to answer two separate non-sensitive questions and the other group is asked to report the sum of the two questions. If the "no design effect" assumption holds, the results in the two groups should be the same (for the ICT, see Tsuchiya et al. 2007; Tsuchiya and Hirai 2010).

Third, careful power analyses are necessary to determine sufficient sample sizes. For this purpose it is helpful if the distribution of the innocuous item is known in the survey population and at least crude ideas about the distribution of the sensitive item and its covariance with the innocuous item exist. The variance of the innocuous item plays a crucial role in the trade-off between privacy protection and statistical efficiency. If the variance is too small, then it does not sufficiently protect the privacy of respondents; if it is too large, then the estimation becomes inefficient[7]. In addition, the covariance between the items also matters: a negative covariance between the non-sensitive item and the sensitive item reduces the total

[7] Another prerequisite for sufficient privacy protection is that the sum of the two items does not exceed the possible maximum value of the innocuous item. The innocuous items used in our study seem unproblematic in this regard. Monthly housing costs have no upper limit and follow a skewed distribution so that extreme outliers are rare but credible. Watching TV has a theoretical upper limit of 168 hours per week, but the sum of undeclared work and watching TV will not reach this limit (at least if they are disjoint) because both have to fit into the overall time budget of a person.

variance and therefore increases efficiency (cf. advice by Glynn 2013 for ICT designs). We conducted some preliminary simulations to evaluate the bias-variance trade-off for the IST compared to direct questioning (not shown). Assuming that the standard deviation of the innocuous variable is about five times the standard deviation of the sensitive variable (which roughly corresponds to our data for the second item) and given a fixed sample size of 500 observations for both questioning techniques, the IST has a lower mean squared error (MSE) than direct questioning if the direct questioning results are biased by half a standard deviation of the sensitive variable or more. If the sample size for the IST is doubled (as in our case), then the MSE of the IST is smaller than the MSE of direct questioning if the bias is about 30 percent of a standard deviation or more. Of course, the variance of the innocuous variable has a strong effect on these results. For example, if both variables have the same variance, then the bias can be as low as 10 percent of a standard deviation before the MSE of the IST exceeds the MSE of direct questioning (given a fixed sample size of 500 for both questioning techniques). The correlation between the sensitive variable and the innocuous variable has substantial effects only if the variances of the two variables are similar. We conclude that, as long as the relative variance of the innocuous variable is not too large, the IST provides estimates that are superior to direct questioning even if social desirability bias is only moderate. Furthermore, for ease of statistical modeling it is convenient if the innocuous item has a distribution that is approximately normal. More research is needed that experiments with these issues to find an optimal trade-off between perceived privacy protection and statistical efficiency. For example, it might be a promising approach to use non-sensitive items whose variance (or covariance with the sensitive item) is subjectively overestimated by survey respondents (cf. Diekmann 2012 for RRT). A different strategy for a more efficient estimation might be to select non-sensitive items that can be predicted with high accuracy from other variables collected in the same survey. In order to increase statistical power, future studies could also consider the use of a double list design (similar to the double list variant of

the ICT where both groups receive a long list and a short list with varying non-sensitive items; see Droitcour et al. 1991; Biemer et al. 2005) requiring a smaller sample size to achieve a given level of statistical power.

Finally, an ideal study to evaluate the IST would not rely on the "more is better" assumption but instead use validation data with known true scores for the sensitive variable. Opportunities for validation studies are notoriously difficult to find and data protection issues arise if individual-level data from different sources is linked without the informed consent of the respondents. In the past, validation studies for dejeopardizing techniques have successfully made use of register samples that were constant with respect to the dependent variable, thus avoiding individual linkage of survey and register data (compare the validation studies for RRT reviewed in Lensvelt-Mulders et al. 2005). This strategy might be permissible from a data protection perspective in the case of quantitative sensitive items as well. However, samples created in this way would be very specific as everyone in the study would have the exact same true value for the quantitative item of interest. This might raise questions about the generalizability of such results. Another approach might be to conduct validation studies in which the distribution of the sensitive item among respondents is known, but the data are not linked at the individual level (for similar approaches in a different context see Kreuter et al. 2010; Sakshaug and Kreuter 2012). In such studies at least an overall evaluation of the validity of the IST would be possible.

Appendix 1: IST long list instructions (translated from German)

Introduction
Thank you very much. We are now done with the coin flip questions. Please answer each of the following questions truthfully. However, please keep the pencil and the piece of paper ready. I will now read two blocks of two questions each to you. Each question within a block has to be answered with a number. However, it is also possible that you will answer one or both questions with zero. Please note the answer to each question on your piece of paper. Afterward, please add the numbers from both answers together and tell me the total result. However, please do not tell me how you answered the individual questions so that I, as interviewer, do not know how you came to your result.

Test question
Let me give you a short example. I will now read two questions from one block to you: 1. Question: How many persons permanently live in your household, including yourself? Please do not tell me the answer, but write it down. *Interviewer: Leave enough time to the respondent to write down the answer.* 2. Question: How many persons living in your household are aged 18 or older, including yourself? Please also write down this answer. *Interviewer: Leave enough time to the respondent to write down the answer.* Thank you very much. What is your result? *Interviewer: Write down result and compare to S1 and S2!* Did you understand the procedure?

Interviewer question	
Did the respondent understand the system?	
1	*Yes*
2	*No (or: the result cannot be determined)*

9	*NA*

Additional instructions if respondent did not understand
I would like to briefly explain the method to you one more time. Assume there are three people living in your household. In this case you would have to write down a three for the question "How many persons permanently live in your household, including yourself?" Further, assume that two of these three people are 18 years old or older. In this case you would have to write down the number two for the question "How many persons living in your household are aged 18 or older, including yourself?" Now, add the two noted numbers together and tell me the result (*Interviewer: short pause*): In this case the number five. For the following questions, please proceed in the same manner. However, do not tell me the individual answers, but only the result at the end of the question block. Do you still have questions?

Appendix 2: Regression estimates for the IST

Let L be an indicator for the long-list sample (i.e. L_i equals 1 if respondent i belongs to the long-list sample and else is 0). In the long-list sample the respondents are asked about the sum of a sensitive item S (e.g. hours of undeclared work) and a non-sensitive item C (e.g. hours of watching TV). In the short-list sample the respondents are only asked about the non-sensitive item C. That is, the answers of the respondents can be written as

$$Y_i = \begin{cases} S_i + C_i & \text{if } L_i = 1 \text{ (long-list sample)} \\ C_i & \text{else (short-list sample)} \end{cases}$$

Furthermore, let $X_i = (X_{1i}, \dots, X_{ki}, 1)'$ and $Z_i = (Z_{1i}, \dots, Z_{mi}, 1)'$ be two vectors of covariates (each including a constant; typically, $Z = X$) and assume that S and C can be modeled as

$$S_i = X_i'\beta + \upsilon_i, \;\; E(\upsilon_i) = 0 \quad \text{and} \quad C_i = Z_i'\gamma + \nu_i, \;\; E(\nu_i) = 0$$

where β and γ are coefficient vectors and υ and ν are random errors. It follows that

$$Y_i = L_i S_i + C_i = L_i(X_i'\beta + \upsilon_i) + (Z_i'\gamma + \nu_i) = L_i X_i'\beta + Z_i'\gamma + \varepsilon_i$$

with $\varepsilon_i = L_i\upsilon_i + \nu_i$ and, hence, $E(\varepsilon_i) = 0$. For example, if $X = Z = 1$, that is, if there are no covariates, then an estimate for the mean of S can be gained by regressing Y on L using the least squares method (assuming heteroscedastic errors to account for the fact that the error variance depends on L). The slope coefficient of the fitted model, $\hat{\beta}$, is an estimate of $E(S)$,

the intercept, $\hat{\gamma}$, is an estimate of $E(C)$. More generally, least-squares regression of Y on $L \cdot X$ and Z (again assuming heteroscedastic errors) provides estimates of effects of covariates X on S; the coefficients we are interested in are the ones attached to the interaction term $L \cdot X$. Alternatively, assuming bivariate normality of υ and ν (with variances σ_υ^2 and σ_ν^2 and correlation ρ), coefficients can also be estimated by the maximum-likelihood method, based on the log-likelihood

$$\ln \mathcal{L} = \sum_{i=1}^{n} \left\{ L_i \cdot \ln \left[\frac{1}{\sigma_\varepsilon} \phi \left(\frac{Y_i - X_i'\beta - Z_i'\gamma}{\sigma_\varepsilon} \right) \right] + (1 - L_i) \cdot \ln \left[\frac{1}{\sigma_\nu} \phi \left(\frac{Y_i - Z_i'\gamma}{\sigma_\nu} \right) \right] \right\}$$

where $\phi(.)$ is the density function of the standard normal distribution and n is the sample size. Formally, $\sigma_\varepsilon^2 = \sigma_\upsilon^2 + \sigma_\nu^2 + 2\rho\sigma_\upsilon\sigma_\nu$, but σ_ν and ρ cannot be identified separately in this model so that σ_ε is estimated directly. The maximum-likelihood approach will yield identical point estimates for β as the least-squares procedure, although standard error estimates may differ. The results in this paper were computed using the least-squares procedure.

The presented method can easily be extended to include a third sample of respondents for which the sensitive item was measured via direct questioning. Let D be an indicator for the direct-questioning sample (i.e. D_i equals 1 if respondent i belongs to the direct-questioning sample and else is 0). Because S is sensitive, S^* is measured in the direct-questioning sample, which is equal to S plus social-desirability bias. In the direct-questioning sample, let $Y_i = S_i^* = X_i'\beta^* + \upsilon_i$, where β^* is a coefficient vector that includes social-desirability bias, then we can write the regression model across the three subsamples as

$$Y_i = L_i X_i'\beta + (1 - D_i) Z_i'\gamma + D_i X_i'\beta^* + \varepsilon_i$$

or, equivalently, as

$$Y_i = L_i X_i'(\beta - \beta^*) + (1 - D_i) Z_i' \gamma + (D_i + L_i) X_i' \beta^* + \varepsilon_i$$

In the latter form, the coefficients attached to the interaction term $L \cdot X$ provide an estimate of the (negative of the) bias in β^*. For example, if regressing Y on L, $(1 - D)$, and $(D + L)$ (i.e. if no covariates are taken into account), then the coefficient attached to L provides an estimate of the effect of the item sum technique (i.e. the degree to which the IST leads to a higher average value of the sensitive variable than direct questioning), the coefficient attached to $(1 - D)$ reflects the mean of the non-sensitive item C, and the coefficient attached to $(D + L)$ is an estimate of the mean of the sensitive variable based on direct questioning.

In our study, some of the respondents initially assigned to the IST mode opted for direct questioning. That is, there is noncompliance with the treatment assignment. To gain an intention-to-treat estimate (ITT) of the effect of the IST, we employ a two-step procedure, where in the first step we fit model $C_i = Z_i' \gamma + \nu_i$ using the observations from the (realized) short-list group. In the second step, we residualize the (realized) long-list observations using the least-squares estimate $\hat{\gamma}$ from the first step and then fit model

$$\tilde{Y}_i = \tilde{L}_i X_i'(\beta - \beta^*) + X_i' \beta^* + \varepsilon_i \quad \text{with} \quad \tilde{Y}_i = \begin{cases} Y_i - Z_i \hat{\gamma} & \text{if } L_i = 1 \\ Y_i & \text{else} \end{cases}$$

based on the respondents who were initially assigned to the long-list sample or to direct questioning, where $\tilde{L}$ is an indicator for initial assignment to the long-list sample. The least-squares estimate of $(\beta - \beta^*)$ is a consistent estimate of the intention-to-treat effect of the

IST[8], but note that standard errors are biased because the additional uncertainty introduced by the variance of $\hat{\gamma}$ is not taken into account. We therefore apply the non-parametric bootstrap procedure across the two steps to compute the standard errors (Davison and Hinkley 1997).

The ITT is a conservative estimate of the causal treatment effect. We can improve on the ITT by fitting

$$\tilde{Y}_i = L_i X_i'(\beta - \beta^*) + X_i'\beta^* + \varepsilon_i$$

in the second step above (i.e. using L_i instead of $\tilde{L}_i$ in the equation), while instrumenting L_i with $\tilde{L}_i$ based on a two-stage least squares procedure. Since $\tilde{L}_i$ is randomized, it is a valid instrument. Such an instrumental variables (IV) estimate of $(\beta - \beta^*)$ provides a consistent estimate of the local average treatment effect (LATE) of the IST.

[8] A necessary assumption for the ITT estimate to be consistent is that the respondents opting out of the short-list sample (who are completely discarded in our ITT estimation procedure) are not systematically different from the respondents opting out of the long-list group. This assumption appears plausible in our case because respondents did not know to which of the two IST groups they were assigned to when deciding to opt for direct questioning (we also tested the assumption by comparing the outcomes for the two subgroups; no significant differences were found).

References

Ahart, A. M., and 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, 101-114.

Angrist, J.D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, 91, 444-455.

Biemer, P. P., Jordan, B.K., Hubbard, M. L., and Wright, D. (2005), "A Test of the Item Count Methodology for Estimating Cocaine Use Prevalence." in *Evaluating and Improving Methods Used in the National Survey on Drug Use and Healt,* eds. J. Kennet and J. Gfroerer, Rockville: Substance Abuse and Mental Health Service Administration, Office of Applied Studies.

Blair, G., and Imai, K. (2012), "Statistical Analysis of List Experiments", *Political Analysis*, 20, 47-77.

Corstange, D. (2009), "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT", *Political Analysis*, 17, 45-63.

Coutts, E., and 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, 169-193.

Dalton, D. R., Wimbush, J. C., and Daily, C. M. (1994), "Using the Unmatched Count Technique (UCT) to Estimate Base Rates for Sensitive Behavior," *Personnel Psychology*, 47, 817-828.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and their Application*, Cambridge: Cambridge University Press.

De Leeuw, E. D., Hox, J. Huisman, M. (2003), "Prevention and Treatment of Item Nonresponse", *Journal of Official Statistics,* 19, 153-176.

Diekmann, A. (2012), "Making Use of "Benford's Law" for the Randomized Response Technique," *Sociological Methods and Research*, 41, 325-334

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsely, T. L., Visscher, W., and 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 *Measurement Errors in Surveys,* eds. P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz and S. Sudman, New York: Wiley, pp. 185-210

Eichhorn, B. H., and Hayre, L. S. (1983), "Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data," *Journal of Statistical Planning and Inference*, 7, 307-316.

Fox, J. A., and Tracy, P. E. (1986), *Randomized Response: A Method for Sensitive Surveys,* Berverly Hills: Sage.

Gjestvang, C. R., and Singh, S. (2007), "Forced Quantitative Randomized Response Model: A New Device," *Metrika*, 66, 243-257.

Glynn, A.N. (2013), "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment", *Public Opinion Quarterly*, 77, 159-172.

Himmelfarb, S., and Edgell, S. E. (1980), "Additive Constants Model: A Randomized Response Technique for Eliminating Evasiveness to Quantitative Response Questions," *Psychological Bulletin*, 87, 525-530.

Holbrook, A. L. and Krosnick, J.A. (2010a), "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique," *Public Opinion Quarterly*, 74, 37-67.

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

Hollis, S. and Campbell, F. (1999), "What is meant by intention to treat analysis? Survey of published randomised controlled trials," *British Medical Journal*, 319, 670-674.

Knäuper, B. (1998), "Filter Questions and Question Interpretation: Presuppositions at Work", *Public Opinion Quarterly*, 62, 70-78.

Kreuter, F., Müller, G., and Trappmann , M. (2011), "Nonresponse and Measurement Error in Employment Research. Making Use of Administrative Data", *Public Opinion Quarterly*, 74, 880–906.

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

Kuklinski, J. H., Sniderman, P. M., Knight, K., Piazza, T., Tetlock, P. E., Lawrence, G. E., and Mellers, M. (1996), "Racial Prejudice and Attitudes toward Affirmative Action," *American Journal of Political Science*, 41, 402–19.

LaBrie, J. W., and Earleywine, M. (2000), "Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-count Technique," *Journal of Sex Research*, 37, 321-326.

Lee, R. M. (1993), *Doing Research on Sensitive Topics*, London: Sage.

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., and Maas, C. J. (2005), "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation", *Sociological Methods & Research*, 33, 319-348.

Newell, D. J. (1992), "Intention-to-Treat Analysis: Implications for Quantitative and Qualitative Research," *International Journal of Epidemiology*, 21, 837-841

Pedersen, S. (2003), *The Shadow Economy in Germany, Great Britain and Scandinavia: A Measurement Based on Questionnaire Service*, Copenhagen: The Rockwool Foundation Research Unit.

Raghavarao, D., and Federer, W. T. (1979), "Block Total Response as an Alternative to the Randomized Response Method in Surveys," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 41, 40-45.

Rayburn, N. R., Earleywine, M., and Davison, G. C. (2003), "Base Rates of Hate Crime Victimization among College Students," *Journal of Interpersonal Violence*, 18, 1209-1221.

Sakshaug, J.W., and Kreuter, F. (2012), "Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data", *Survey Research Methods*, 6, 113-122.

Schneider, F., and Enste, D. H. (2007), *The Shadow Economy. An International Survey*, New York: Cambridge University Press.

Smith, L. L., Federer, W. T., and Raghavarao, D. (1974), "A Comparison of Three Techniques for Eliciting Truthful Answers to Sensitive Questions," in *American Statistical Association Proceedings of the Social Statistics Section,* pp. 447-452.

Schwarz, N., Hippler, H. J., Deutsch, B. and Strack, F. (1985), „Response Categories: Effects on Behavioral Reports and Comparative Judgments", *Public Opinion Quarterly*, 49, 388-395.

Schwarz, N., Sudman, S. (1992). *Context effects in social and psychological research.* New York: Springer.

The American Association for Public Opinion Research (2011), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (7th ed.), AAPOR, Lanexa.

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

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

Tsuchiya, T., and Hirai, Y. (2010), "Elaborate Item Count Questioning: Why Do People Underreport in Item Count Responses?" *Survey Research Methods*, 4, 139-149.

Tsuchiya, T., Hirai, Y., and Ono, S. (2007), "A Study of the Properties of the Item Count Technique," *Public Opinion Quarterly*, 71, 253-272.

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

Williams, C. C. (2009), "Formal and Informal Employment in Europe: Beyond Dualistic Representations," *European Urban and Regional Studies*, 16, 147-159.

Wimbush, J. C., and Dalton, D.R. (1997), "Base Rate for Employee Theft: Convergence of Multiple Methods," *Journal of Applied Psychology*, 82, 756-763.