Erschienen in: Public Opinion Quarterly ; 85 (2021), 1. - S. 6-27
https://dx.doi.org/10.1093/poq/nfab002

SENSITIVE QUESTIONS IN SURVEYS
A COMPREHENSIVE META-ANALYSIS OF EXPERIMENTAL SURVEY STUDIES ON THE PERFORMANCE OF THE ITEM COUNT TECHNIQUE

INGMAR EHLER
FELIX WOLTER*
JUSTUS JUNKERMANN

Abstract In research on sensitive questions in surveys, the item count technique (ICT) has gained increased attention in recent years as a means of counteracting the problem of misreporting, that is, the under- and over-reporting of socially undesirable and socially desirable behaviors or attitudes. The performance of ICT compared with conventional direct questioning (DQ) has been investigated in numerous experimental studies, yielding mixed evidence. This calls for a systematic review.

For this purpose, the present article reports results from a comprehensive meta-analysis of experimental studies comparing ICT estimates of sensitive items to those obtained via DQ. In total, 89 research articles with 124 distinct samples and 303 effect estimates are analyzed. All studies rely on the "more (less) is better" assumption, meaning that higher (lower) estimates of negatively (positively) connoted traits or behaviors are considered more valid.

The results show (1) a significantly positive pooled effect of ICT on the validity of survey responses compared with DQ; (2) a pronounced heterogeneity in study results, indicating uncertainty that ICT would

INGMAR EHLER is a research associate in the Department of Sociology at the Technische Universität Kaiserslautern, Kaiserslautern, Germany. FELIX WOLTER is a postdoctoral research associate in the Cluster of Excellence "The Politics of Inequality" and the Sociology Department at the University of Konstanz, Konstanz, Germany. JUSTUS JUNKERMANN is a research associate in the Institute of Medical Sociology at the Martin-Luther-Universität Halle-Wittenberg, Halle, Germany. The authors are very grateful for the anonymous reviewers' helpful comments. This work was supported by the German Research Foundation [DFG; project WO 2242/1-1 to F.W.], as part of its Excellence Strategy. Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder - EXC-2035/1 - 390681379
*Address correspondence to Felix Wolter, Sociology; Cluster of Excellence "The Politics of Inequality," University of Konstanz, Box 32, D-78457 Konstanz, Germany; email: felix.wolter@uni-konstanz.de.

work as intended in future studies; and (3) as meta-regression models indicate, the design and characteristics of studies, items, and ICT procedures affect the success of ICT. There is no evidence for an overestimation of the effect due to publication bias.
Our conclusions are that ICT is generally a viable method for measuring sensitive topics in survey studies, but its reliability has to be improved to ensure a more stable performance.

Introduction

Sensitive questions in surveys, such as those on substance abuse, delinquency, voting, or xenophobia, pose several problems. For one, survey respondents, due to feelings of embarrassment or fear of sanctions, might refuse to answer such questions at all. Moreover, if they do answer, their survey responses are often distorted. Numerous studies dating back to the very beginning of modern survey research (e.g., Hyman 1944) have shown that respondents tend to under-report socially undesirable behaviors, attitudes, or traits (such as stealing or driving under the influence) and to over-report socially desirable ones (such as green behavior or voter turnout). This leads to invalid data and biased estimates in statistical analyses (Tourangeau and Yan 2007; Lensvelt-Mulders 2008; Krumpal 2013).

To cope with these problems, several propositions have been made by survey methodologists on how to best "ask the embarrassing question" (Barton 1958). Conventional questioning and design techniques include "forgiving wording," filtering techniques, anonymity assurances, and the sealed envelope technique (Perry 1979). More sophisticated methods use statistical random noise added to the data to conceal the respondents' answers and thus enhance anonymity. Among them are the randomized response technique (RRT; Warner 1965; Fox and Tracy 1986), the crosswise model (CM; Yu, Tian, and Tang 2008), and the item count technique (ICT), also known as the list experiment or unmatched count technique (Raghavarao and Federer 1979; Droitcour et al. 1991). While RRT and CM use a randomization device administered by the survey respondent or combinations of unrelated survey questions with known aggregate prevalence (such as month of birth), ICT makes use of lists of items. These item lists are answered *en bloc* by the respondents, avoiding disclosure of the answer to the sensitive item in the list.

It has not been definitively established, however, whether any of these techniques actually fulfill their purpose, namely to mitigate or even eliminate misreporting on sensitive questions. In the literature, two types of studies address this issue: external validation studies and comparative studies based on the "more is better" assumption. The former ones use external data with

known true values of the sensitive traits and compare these with survey estimates using RRT, CM, or ICT. Naturally, such studies are very rare because external validation samples and data are difficult to acquire. "More is better" studies carry out survey experiments and compare estimates gathered via RRT, CM, or ICT with those obtained using conventional direct questioning (DQ). If estimates of negatively connoted behaviors or attitudes are higher, they are considered more valid (and, vice versa, "less is better" for positively connoted behaviors where over-reporting is expected). The advantage of the "more is better" approach is that studies can easily be conducted. However, even higher or lower estimates, compared with DQ, can still deviate from the—unknown—true value.

The following article concentrates on comparative ("more is better") experimental studies that evaluate the performance of ICT compared with DQ.[1] Numerous studies have been conducted in this regard and their evidence is mixed. In a 2014 literature review, Wolter and Laier (2014, p. 157) list eight studies with significant findings in favor of ICT, nine studies with mixed evidence, two studies finding no differences between ICT and DQ, and even two studies finding negative effects of ICT on data validity. This heterogeneity within the literature calls for a meta-analysis in which all relevant published studies are summarized and investigated with respect to an overall ICT effect on data validity, heterogeneity of the findings among studies, and publication bias, while also taking into account study or design characteristics as an explanation for this heterogeneity using meta-regression models. As opposed to RRT, for which an older meta-study exists (Lensvelt-Mulders et al. 2005), no such study examining all four aforementioned aspects has been published regarding ICT. Preliminary research in this regard, however, has already been conducted. Tourangeau and Yan (2007) report results from a small-scale meta-analysis encompassing seven studies and find no significant ICT effect. Blair, Coppock, and Moor (2020) report a meta-analysis of 264 ICT–DQ comparisons and find a significant overall effect of ICT. However, the authors only concentrate on studies from the political science literature; they only report descriptive results of the effect sizes without accounting for the clustering in samples and articles; they do not investigate publication bias; and they do not conduct meta-regressions (see also the discussion section and the Supplementary Material for a detailed methodological comparison).

In what follows, we aim to present a comprehensive meta-analysis of all experimental studies that investigate the performance of ICT compared with DQ. The next section will briefly recapitulate and sum up the principles of

1. There are, to our knowledge, two external validation studies regarding ICT (Comşa and Postelnicu 2013; Rosenfeld, Imai, and Shapiro 2015). Both studies find differences between DQ and ICT that favor the latter, but the ICT estimates are still off the mark compared with the known true prevalence.

ICT and its most important variants. Afterward, we present the design and methods of the meta-analysis, namely the data collection and preparation, the effect size measure and other variables, and the statistical methods. The subsequent section of the article presents the results, followed by a discussion in the final section.

The Item Count Technique (ICT)

To our knowledge, the basic idea of ICT was first proposed by Smith, Federer, and Raghavarao (1974) and Raghavarao and Federer (1979) under the term "block total response," though this remained more or less unnoticed by the research community. It was presented again by Miller (1984) and Droitcour et al. (1991). Recently, ICT has gained much attention in research on sensitive questions. The ICT procedure splits the sample into two experimental groups—the short list group (SL) and the long list group (LL). In the SL group, k binary "yes-no" items (non-key or filler items) are presented to the respondent; in the LL group the same k questions are administered in addition to the sensitive question of interest. Respondents are asked not to answer the questions individually but instead to indicate the sum of "yes" answers to the whole list. Hence, the individual answer to the sensitive item remains anonymous (as long as no floor or ceiling effects are observed, as clarified below). For example, Holbrook and Krosnick (2010, p. 47) asked survey respondents: "Here is a list of four 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 [. . .]: 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 [. . .]." In the LL group, the sensitive item "voted in the Presidential election held on November 7, 2000" was added to the list.

Due to the fact that the SL and LL subsamples are randomly split, the answers to the non-key items are theoretically distributed equally in both groups; a simple estimator of the prevalence of the sensitive item $\widehat{\pi}_{ICT}$ can thus be calculated by subtracting the mean of the short list $\overline{x}_{SL}$ from the mean of the long list $\overline{x}_{LL}$ (see, for example, Blair and Imai 2012 for a discussion of other estimators). The standard error is the square root of the sampling variances of the list means (formulae 1 and 2).

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

$$SE(\widehat{\pi}_{ICT}) = \sqrt{Var(\overline{x}_{SL}) + Var(\overline{x}_{LL})} \tag{2}$$

While it is often argued that the ICT procedure is straightforward for survey respondents to understand and easily operable (especially compared with the more complex RRT), it comes with two drawbacks. First, ICT estimates,

due to the statistical noise added by the non-key items, suffer from large standard errors and thus require large sample sizes. Second, the logic of ICT only works if respondents do not deny or confirm all items in the list, in which case the (negative or positive, respectively) answer to the sensitive item is disclosed. Such floor or ceiling effects (Blair and Imai 2012) should be avoided by carefully selecting the non-key items, for example by combining items with high and low prevalence or including negatively correlated items (Glynn 2013).

There are multiple variants of the classic ICT design described above. The double-list design (Droitcour et al. 1991; Coutts and Jann 2011) includes a different SL question block in the LL group and vice versa, which doubles the statistical power. The person count technique (PCT; Grant, Moon, and Gleason 2014; Wolter 2019) uses the acquaintances of the respondent as non-key items instead of actual questions; in the LL group, the respondent himself is added to the list. An example for a PCT LL would be: "Please think of three acquaintances of yours, who you know well and who are not similar to each other. How many of these persons, including yourself, have taken cocaine at least once?" The PCT procedure shortens the time required for each sensitive item compared with the classic ICT design and permits asking several sensitive items consecutively without the need to introduce new item lists for each new sensitive question. Studies using the PCT variant of ICT are included in the meta-analysis. Other design variants, such as the item sum technique (IST; Trappmann et al. 2014) or the person sum technique (PST; Junkermann 2020), enable the researcher to obtain quantitative estimates for sensitive questions, such as hours of undeclared work. Studies using IST or PST, however, were not included in our meta-analysis, since quantitative measures and binary prevalence estimates are not directly comparable.

Design and Methods of the Meta-Analysis

DATA COLLECTION AND DATA STRUCTURE

Our aim was to include all published experimental studies comparing ICT (including PCT) to DQ in a sensitive question setting. We therefore combined a systematic literature search with a snowball approach to also identify gray literature or unpublished studies. Before compiling the studies, the following selection criteria for eligible studies were defined: (1) ICT was used; (2) a DQ–ICT comparison was conducted for identical survey questions; (3) the items were sensitive in some way—that is, the aim of the study had to be testing ICT versus DQ in a sensitive question or social desirability context and not for other purposes; and (4) the standard errors of the difference

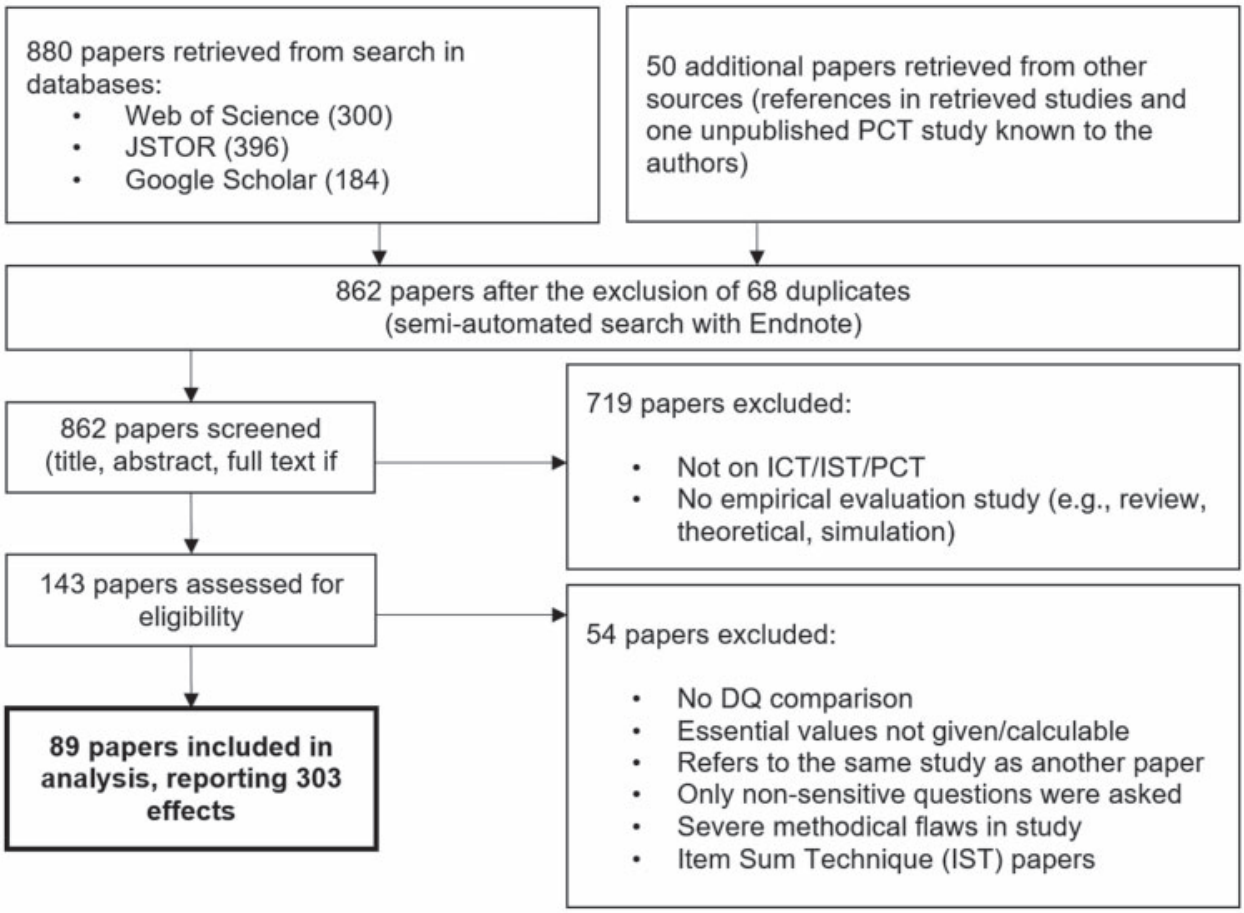

Figure 1. Data collection procedure. "Severe methodical flaws" means a deficient internal validity of the experimental design, e.g., by nonrandom and selective allocation to groups or a confounded experimental stimulus.

between DQ and ICT estimates were reported or calculable from the published results.

The data collection procedure is summarized in figure 1. We searched for literature using the search engines Web of Science, JSTOR, and Google Scholar with predefined search strings (see the Supplementary Material). After this systematic search, we reviewed the bibliographies of the results and found 50 additional studies not captured in our initial search and included one unpublished PCT study known to us. We contacted the authors of the studies for which we could not compute the standard errors and requested the required information. However, we did not get a reply from all contacted authors, and some studies dropped out for this reason. The literature search was finished on November 14, 2018; one study from early 2019 was added afterward.

The final database for the meta-analysis consists of 89 research articles, reporting the results from 124 distinct samples and containing 303 effect

estimates in total. This results in a nested structure of the data; a single article or publication may report results from one or more different samples or studies, which themselves may contain one or several items/DQ–ICT comparisons. This clustering will be taken into account during the statistical analysis.

EFFECT SIZE MEASURE AND STANDARD ERRORS

The effect size measure used in the meta-analysis is the raw mean difference between the ICT prevalence estimate and the DQ prevalence estimate, Δμ. Of course, Δμ has to be adjusted according to the direction of response bias for every item (over-reporting versus under-reporting). For every item, the effect size measure was coded following the hypothesized direction of social desirability ("more is better" or "less is better") by the authors of the respective study (see formulae 3 and 4). For some studies, different directions of social desirability were expected by the study authors for different subgroups of respondents. For example, Vienrich and Creighton (2018) investigated whether respondents supported a complete stop of immigration to the United States. For this item, the authors expected different directions of social desirability for Latino and non-Latino respondents. For these studies, we followed the recommendation by Borenstein et al. (2009, p. 215) and treated the respective subsamples as separate cases (studies) in the database.

$$\Delta\mu = \widehat{\pi}_{ICT} - \widehat{\pi}_{DQ} \text{ if under-reporting is expected ("more is better")} \quad (3)$$

$$\Delta\mu = \widehat{\pi}_{DQ} - \widehat{\pi}_{ICT} \text{ if over-reporting is expected ("less is better")} \quad (4)$$

A peculiarity of using the raw mean difference between ICT and DQ is that it captures only absolute differences measured in percentage points. For instance, if the DQ estimate amounts to 2 percent and the ICT estimate to 4 percent, the raw mean difference is two percentage points, which would be the same as a 52 percent versus 54 percent difference. The raw mean difference treats these differences equally. However, in the first case, the ICT estimate would have doubled (improved by 100 percent), whereas in the second case the improvement would have increased by only 3.8 percent. Preferably, our meta-analysis would use a relative effect size measure, for instance the logged risk ratio $ln\left(\frac{\widehat{\pi}_{ICT}}{\widehat{\pi}_{DQ}}\right)$. This, however, is not suitable in our case. First, the DQ estimates in some studies amount to zero percent and some ICT estimates are even negative. For such cases, no log risk ratio can be calculated and the respective studies (which are certainly non-ignorable for the overall ICT effect) would drop out of the analysis. Second, for extremely low prevalence estimates, in some studies the risk ratio becomes exceedingly high and would distinctively influence the overall pooled effect. This is not necessarily

in favor of ICT because such estimates rather indicate that ICT, in such scenarios, yields imprecise estimates characterized by large standard errors. Therefore, we opt for the risk difference as the effect size measure. We will report, however, a meta-regression model in which we introduce the baseline ICT prevalence estimate as a right-hand-side variable to investigate to what extent the ICT–DQ risk difference varies by the baseline prevalence of the items.

The exact standard errors, z-values, p-values, or confidence intervals of the ICT–DQ difference were reported for 60 out of 303 effects. These were coded into standard errors using conventional formulae. A further 133 ICT–DQ differences provided exact standard errors or variance estimates for the ICT and DQ estimates, respectively. Here, the standard error of the difference was calculated using formula 5.[2] The rest of the studies reported only imprecise test statistics, such as "$p < 0.05$." In these cases, a conservative approach was chosen by coding the upper significance level (e.g., $p = 0.05$). For a few studies that reported merely "non-significant" differences, the authors were contacted and asked to provide the missing information. The authors did so for three effects, but nine effect estimates dropped out of the analysis because no information could be retrieved. Other studies reported X^2 or other statistics (nevertheless providing exact p-values), and these were coded as if they stemmed from z-tests.

$$\widehat{SE}\left(\hat{\pi}_{ICT} - \hat{\pi}_{DQ}\right) = \sqrt{\widehat{SE}\left(\hat{\pi}_{ICT}\right)^2 + \widehat{SE}\left(\hat{\pi}_{DQ}\right)^2} \tag{5}$$

META-REGRESSION VARIABLES

The aim of the meta-regression is to investigate whether characteristics of the studies or the sensitive items exert an influence on the DQ–ICT difference. The motivation for such a meta-regression is threefold. First, theoretical arguments suggest that the performance of ICT should depend on the degree of threat to the respondent or the degree of social desirability concerns induced by the interview situation and the sensitive items asked. Therefore, it can be assumed that the more socially loaded and sensitive the response situation, the stronger the ICT effect will be.[3] We test this

2. If DQ and ICT samples are not independent (for example if the direct question is put to the same respondents answering the SL), DQ and ICT estimates are correlated, which is not accounted for by formula 5 or in most of the studies concerned. For those studies, we had to stick to the unadjusted figures given in the papers, because information to perform the correct calculations was not provided. We are confident, however, that this does not change the main results. We also control for this design feature in the meta-regression.

3. This is the basic argument of the rational choice-theoretic explanation of respondent behavior (Rasinski et al. 1999; Stocké 2007; Preisendörfer and Wolter 2014).

conjecture in our models by investigating several indicators for situational or item sensitivity. Second, the meta-regression will indicate possible biases in the included studies stemming from the way standard errors were reported (or calculated) or from flaws in the experimental design. Third, in ideal circumstances, we can learn practical information from the meta-regression and provide researchers with knowledge about design issues that influence the efficacy of ICT in a positive or negative direction, which can then be incorporated into future studies.

Regarding study characteristics, we first investigate survey mode, which was coded into self-administered (online or paper-and-pencil survey), computer-assisted telephone interviewing (CATI), face-to-face, or "not indicated" for a few cases where the survey mode was not reported. Following the assumption that sensitivity and non-anonymity is highest in face-to-face interviews, moderate in CATI settings, and lowest in self-administered interviews, the ICT effect should be strongest for face-to-face surveys.[4] With respect to the sample, we distinguish national full-population surveys, student samples, special groups (e.g., farmers; professional auctioneers), or other types of samples (e.g., online access panels; snowball samples). We further test whether different implementations of the ICT procedure have an effect on ICT performance. We distinguish between the classic three-group design with separate groups for SL, LL, and DQ, the double-list ICT design as described above, alternating lists where both the SL and the LL were presented alternately to two experimental ICT groups, and studies in which the DQ and SL group were not experimentally separated but instead put together into one group (see footnote 2). According to the theory of response behavior, misreporting should be reduced as anonymity rises. The latter, in turn, is directly affected by the list length of the ICT procedure (the longer the list, the better the respondent protection), so this variable is also added to the meta-regression. Further, we test whether PCT works better or worse than classic ICT and whether the year of publication affects the ICT–DQ difference. With respect to the country or region in which the studies were conducted, one can expect considerable heterogeneity. For instance, the perceptions of question sensitivity, social desirability, general willingness to respond truthfully on surveys, and implementation and understanding of the ICT procedures will likely vary by the geographic region in which the study was carried out. To test for this, we add dummies for seven geographic regions. Finally, a variable indicating the source of the standard errors of the estimated ICT–DQ difference is added to the models. As set out above, not all studies reported exact test statistics for the estimates. If the reporting

4. Supplementary Material, Appendix 2, provides an overview and descriptive statistics of all variables entering the meta-regression.

behavior is related to some sort of "study quality," this could affect the estimated outcome.

Regarding item characteristics, we test for effects of the direction of social desirability (under-reporting versus over-reporting). Recent literature suggests that the response mechanisms causing misreporting operate differently for positively and negatively connoted items (Andersen and Mayerl 2019); hence, different effects of ICT can be expected. We further introduce an indicator for item sensitivity to the models. Here, theory predicts that the ICT effect gets stronger with growing question sensitivity. Following the procedure proposed in the meta-analysis on RRT by Lensvelt-Mulders et al. (2005), an external sensitivity rating by experts was carried out. Eight experts in sensitive question research were asked to rate 194 items on a scale from 1 ("not sensitive at all") to 7 ("very sensitive") (see the Supplementary Material for details). We further test for the effect of the share of related filler items in the item lists to assess the finding of Droitcour et al. (1991, pp. 197ff) that respondents become more distrustful if the non-key questions have no content-related relationship to the sensitive item. We also investigate whether it is detrimental to the success of ICT if the non-key items in the SL are also sensitive. Finally, we conduct an exploratory test for effects of the item content, that is, opinion/attitude versus trait (e.g., a health issue or personal characteristic) versus behavior and the baseline prevalence of the sensitive item. For the latter, the ICT estimate is used, because this estimate is supposedly nearer to the true value than the DQ estimate.

STATISTICAL METHODS

The statistical methods used in this paper follow the recommendations by Borenstein et al. (2009), Tanner-Smith and Tipton (2014), and Veroniki et al. (2016) and estimate random-effects (RE) meta-analysis models. RE models are most appropriate in our case, since we have to assume that different ICT–DQ studies estimate different effects (and not one single effect common to every study, as one would assume in fixed-effects models). In RE models, the variance of the effects is separated into a variance in the true effects across studies and an error variance (the sampling variance within each study picturing the deviation from the study-true value and the estimated effect). The former term is the between-study heterogeneity τ^2, which can be estimated using different computational models (see the above-cited literature for an in-depth discussion). The pooled effect size $\hat{\mu}$ across studies (the pooled DQ–ICT difference in our case) is estimated according to formulae 6 to 9 (Veroniki et al. 2016, p. 59), where y_i is the effect in study i and w_i is the weight assigned to each study. δ_i depicts the deviation of the true

study mean from the overall pooled effect with between-study variance τ^2, and v_i is the sampling variance of study i.

$$\widehat{\mu} = \frac{\sum w_i y_i}{w_i} \tag{6}$$

$$y_i = \mu + \delta_i + \varepsilon_i \tag{7}$$

with

$$\varepsilon_i \sim N(0, v_i)$$

$$\delta_i \sim N\left(0, \tau^2\right)$$

$$Var(y_i) = v_i + \tau^2 \tag{8}$$

$$w_i = \frac{1}{v_i + \tau^2} \tag{9}$$

In the results section, we present different estimates of $\widehat{\mu}$ as a robustness check for the findings. These estimates vary in the way they estimate the between-study heterogeneity τ^2 and how they deal with the clustered data structure. The DerSimonian and Laird (DL) model (DerSimonian and Laird 1986) does not account for the multilevel structure of the data and uses an easily computable analytical model. The three-level restricted maximum likelihood model (3REML) uses a numerical solution to estimate τ^2 and accounts for the three-level data structure (Veroniki et al. 2016). The robust variance estimation (RVE) method (Tipton 2013; Tanner-Smith and Tipton 2014; Fisher and Tipton 2015) accounts for the clustering of items in samples (studies), but not for samples (studies) in articles (for articles that present results from more than one sample or study). RVE provides cluster-robust standard errors for non-independent effect sizes, but with less strong distributional assumptions than the 3REML method. The Stata ado robumeta (Fisher and Tipton 2015) was used for the RVE method and the R package metafor (Viechtbauer 2010) for the 3REML models.

Using these methods, we will first present an estimate for the pooled DQ–ICT difference across studies. This includes estimates of τ^2 and the intraclass correlation I^2, which is the proportion of the between-study variance τ^2 on the overall variance between and within studies. More concisely, I^2 indicates how much of the total variance is induced by variations in the true effects (the DQ–ICT differences across studies). The second analysis step is to compute prediction intervals using conventional calculations, described for example by Borenstein et al. (2009, chap. 17). Prediction intervals use the estimated between-study heterogeneity τ^2 to estimate confidence intervals of the true effect for an imaginary future ICT–DQ difference. Hence, the

Table 1. Meta-analysis results of the DQ–ICT difference

Model	Overall effect (SE)	95% CI of overall effect	τ^2	I^2	95% prediction interval
DerSimonian and Laird (DL)	0.086*** (0.006)	0.075. . .0.098	0.008	0.941	−0.093. . .0.266
Two-level RVE	0.085*** (0.010)	0.065. . .0.105	0.010	0.925	−0.112. . .0.281
Three-level REML	0.085*** (0.011)	0.064. . .0.107	0.008	0.948	−0.106. . .0.276

NOTE.—$N = 303$ effects (DQ–ICT differences).
***$p < 0.001$

interpretation is that the true effect of a future ICT–DQ difference lies with probability $p = 0.95$ within the 95 percent prediction interval. Briefly, prediction intervals tell us something about the reliability of the ICT performance.

The third analysis step will discuss publication bias, for which we use standard methods described in Borenstein et al. (2009). The subsequent results—step 4—report meta-regression models in which the estimated ICT–DQ difference $\widehat{\mu}$ is regressed on the variables outlined in the previous section. Here, to avoid over-specifying the models (given the relatively low number of cases for a three-level dataset), we use the RVE model as a compromise.

Results

The main results of the meta-analysis are shown in table 1. As can be seen, the findings are robust across different estimation methods (with a slightly lower standard error for the DL method, which does not account for the multilevel structure of the data). Over the 303 effects tested, ICT performs on average about 8.5 percentage points better than DQ (based on the "more is better" assumption), which is statistically significant. However, the results also show that the I^2 statistics are greater than 90 percent, meaning that nearly all the overall variance between estimates is due to heterogeneity in the true effects across studies. This is also reflected by the prediction intervals, which all have lower bounds below zero. This means that a researcher cannot be sure that ICT will have the anticipated effect in a future study. Put simply, the main result of the meta-analysis tells us that, on average, ICT produces better results than DQ, but its reliability is unsatisfying: one cannot count on it with certainty the next time it is used.

PUBLICATION BIAS

Publication bias occurs if only a selective subsample of all studies conducted on a specific topic gets published. One normally assumes that studies reporting significant effects and/or large effect sizes are more likely to be published than non-significant studies or those reporting small effects. This can happen for several reasons, for instance if authors do not write down negative results (the "file drawer problem") or journal editors privilege studies reporting significant or "striking" findings. Hence, if the studies entering our meta-analysis are a selective sample of all ICT–DQ studies that have been conducted, the overall ICT could be biased—most likely upward in a direction that favors ICT.

We performed several tests for publication bias. A standard "eyeballing" method is a funnel plot in which the estimated effect sizes of the studies are plotted against their standard errors. The underlying assumption here is that small studies (with large SE) are more prone to get published when reporting significant results. Large studies (with lower SE) are supposedly conducted using high amounts of time and resources, so they are published regardless of statistical significance and likely are not subjected to the same selection process (this assumption is state-of-the-art and we follow Borenstein et al. 2009, p. 281 here, which can be consulted for a more in-depth discussion). A pronounced non-symmetry of the funnel plot may therefore indicate publication bias. Figure 2 shows the funnel plot for our data and yields no evidence of publication bias.[5] This is also confirmed by other tests. The trim-and-fill procedure (Duval and Tweedie 2000) does not impute missing studies on the left-hand side of the funnel plot (which would point to the direction of an overestimation of the ICT effect).[6] Egger's regression test (Egger et al. 1997) also provides no evidence of publication bias ($z = 0.59$, $p = 0.553$). The rank correlation test (Begg and Mazumdar 1994) points into the same direction, as stated in footnote 6: no publication bias overestimating the ICT effect, but one that would underestimate it (Kendall's $\tau = -0{,}11$; $p = 0.006$). In sum, several tests for publication bias show no evidence in this regard. If there is any bias, then it indicates that our overall ICT effect is underestimated by the sample of studies that entered the meta-analysis.

META-REGRESSION

The results of the meta-regressions are displayed in table 2. Because the statistical power of the models is an issue given the relatively low number of

5. For technical reasons, we have to stick to the DL estimator for publication bias analysis. Two outlier studies have also been excluded from the plot.

6. A robustness check with the REML estimator (not documented) even yields a right-hand-side bias, meaning that the publication bias would be the other way round (positive ICT findings missing).

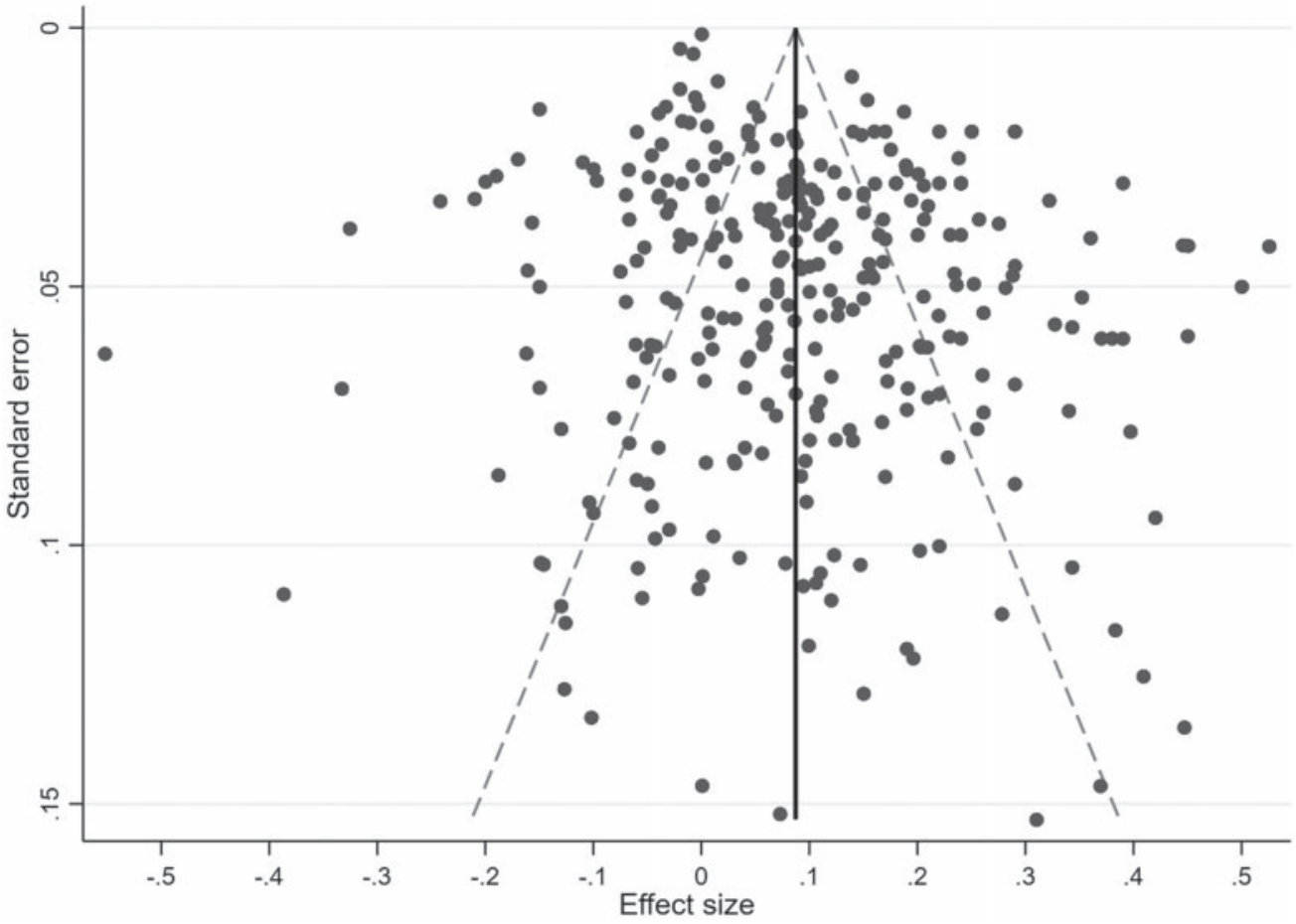

Figure 2. Funnel plot.

cases, we first report the bivariate effects of the variables followed by a full model in which all the covariates are included at once. Since the independent variables reported in each study vary, the number of cases is lower in some models than others. For the full model, on the other hand, we do not include all the independent variables to avoid extensive dropout of studies due to missing information. Hence, three variables (SL containing sensitive items; SL items related to sensitive items; and ICT baseline prevalence) are investigated only with respect to their bivariate effect. The dependent variable is the DQ–ICT risk difference, as described above.

The survey mode has no significant effect on the DQ–ICT difference, with the exception of a marginally significant positive effect of face-to-face (compared with self-administered) in the bivariate model. Regarding the type of sample, there are no differences between national full-population surveys, student or special group samples, and other types of samples. Concerning the ICT list design, the results show that the double-list design performs slightly worse than the classic one (by about five percentage points); however, the

Table 2. Meta-regression results

	Bivariate models			Full model ($n = 290$)	
Variable	b	SE	n	b	SE
Survey mode					
(Self-administered)	(0.077)	(0.013)***	303		
CATI	−0.057	0.052		−0.050	0.052
Face-to-face	0.034	0.020^{+}		0.017	0.033
Not indicated	0.076	*0.105*		0.095	*0.033*
Sample					
(National full pop.)	(0.087)	(0.016)***	303		
Student	−0.021	0.024		−0.023	0.032
Special group (e.g., farmers)	0,020	0.037		0.015	0.035
Other	−0.008	0.024		−0.030	0.024
ICT list design					
(Classic)	(0.104)	(0.016)***	303		
Double-list	−0.057	0.028^{+}		−0.048	0.043
Alternating lists	−0.010	0.031		0.027	0.033
DQ and SL in same group	−0.022	0.021		−0.026	0.024
ICT LL length [3..10]	−0.032	0.011*	294	−0.006	0.013
ICT type PCT (1 = PCT)	0.018	*0.058*	303	0.046	*0.056*
Publication year [0..28]	0.002	0.002	303	−0.004	0.002*
Country					
(USA, AUS, NZ, CA)	(0.066)	(0.013)***	303		
Latin America	0.038	0.028		0.072	0.041^{+}
Europe	0.002	0.027		0.014	0.026
Asia	0.112	0.046*		0.127	0.042**
Africa	0.070	0.026*		0.121	0.048*

(*continued*)

Table 2. (*continued*)

	Bivariate models			Full model ($n = 290$)	
Variable	b	SE	n	b	SE
Middle East	0.065	0.085		0.125	0.087
International	−0.080	*0.027*$^{+}$		−0.030	*0.041*
Standard errors					
(Reported in study)	(0.163)	(0.020)***	303		
From ICT/DQ estimates	−0.123	0.025***		−0.089	0.026**
From test statistics	−0.072	0.024**		−0.030	0.028
Desirability direction (1 = under-report.)	−0.059	0.021**	303	−0.067	0.029*
SL contains sensitive items (1 = yes)	−0.076	0.026*	255	–	–
Item sensitivity (ln)	−0.085	0.028**	298	−0.042	0.035
SL items related to sensit. item (1 = yes)	−0.008	0.022	277	–	–
Item content					
(Opinion/attitude)	0.108	0.017***	303		
Trait	−0.016	*0.033*		−0.034	*0.035*
Behavior	−0.037	0.021^{+}		−0.039	0.023^{+}
Baseline prevalence ICT	0.733	0.090***	280	–	–
Baseline prevalence ICT squared	−0.760	0.102***		–	–
Constant				0.367	0.082***

NOTE.— Depicted are unstandardized regression coefficients from RVE regressions. For polytomous categorical variables, the reference category is shown in brackets and the corresponding b refers to the regression constant. Standard errors in italics suffer from a small number of degrees of freedom and are not trustworthy (Tipton 2015).

$^{+}p < 0.1$

$^{*}p < 0.05$

$^{**}p < 0.01$

$^{***}p < 0.001$

effect is only significant at a 10 percent level in the bivariate model. The coefficient of the LL length (significant only in the bivariate model) is—contrary to our expectations—negative. Hence, the assumption that more non-key items enhance respondent protection and improve ICT performance is not confirmed by the data.[7] Next, we find no evidence that the recently proposed PCT performs worse than the classic design. This assertion, however, is based on a limited number of PCT studies included. The publication year has a significant negative effect in the full model. This means that we cannot infer from the findings that there is a trend toward an improvement of the applied ICT procedures over the years. The geographic regions in which the studies have been conducted show significant and stable effects for Asia and Africa. As stated earlier, we have assumed that geographic differences are plausible for several reasons. One could also suppose a region-specific publication bias to be behind these findings, but this would be pure speculation. Finally, the source of the standard errors as reported/calculated for the meta-analysis also has pronounced effects. Studies reporting detailed standard error statistics find larger DQ–ICT differences. This makes sense if the reporting behavior of the study authors is related to overall study quality, which in turn could improve the ICT performance compared with studies conducted less carefully.

Regarding item characteristics, the direction of social desirability has a stable effect in both models. The direction is negative, meaning that ICT works better in reducing misreporting for positively connoted traits where over-reporting is expected. This is interesting, because the vast majority (73 percent) of studies so far have investigated ICT for under-reporting items. Item sensitivity (as measured by the external rating) has a significant negative effect in the bivariate model. This is contrary to our expectations and means that ICT is less likely to alleviate misreporting to highly sensitive items, but might work better for items of low and medium sensitivity. The variable "item content" shows a marginally significant effect for behavioral items; the direction is negative compared with the reference category, meaning that ICT works better for attitudinal questions.

For the reasons explained above, the three remaining variables are only investigated in the bivariate models. If the SL contains sensitive items in addition to the key sensitive items, the DQ–ICT difference shrinks by $\sim$7 percentage points. This is as expected; obviously, the basic logic of ICT is thwarted if the filler items are also sensitive. Conversely, the conjecture that the SL items should be related in content to the sensitive item is not

7. We checked for non-linearity and found a u-shaped effect, with the ICT effect being lowest for about six to eight LL items and then rising again for longer lists. However, only a handful of studies used a list length of seven or more items, so the data basis is weak and the non-linear effect might have occurred by chance.

confirmed by the data, the effect being practically zero. Finally, the baseline prevalence of the sensitive item, as measured by the ICT prevalence estimate, turns out to have a pronounced inverted u-shaped effect (see an additional plot of the effect in the Supplementary Material). The DQ–ICT risk difference approaches zero for sensitive items with a very low (near-zero) or very high (near-100 percent) prevalence, while the maximum risk difference is around 50 percent. This is in line with theoretical expectations, since ICT estimates, due to their low statistical power, become imprecise for low- and high-prevalence items. We therefore conclude that, in practical terms, ICT is best suited for moderately prevalent sensitive items. From a methodological perspective, experimental studies evaluating ICT by comparing it to DQ suffer from the imprecision of ICT when using items having a prevalence near 0 and 1. Thus, such studies will always have difficulties in finding substantial and significant differences between the estimates.

Discussion

The contribution of this paper consists in synthesizing the mixed results of the empirical literature investigating the performance of ICT in mitigating misreporting to sensitive questions in surveys. Further, meta-regression models enable us to study correlates and causes of the success of ICT and to give practical advice in this regard.

The main overall result is that ICT estimates are 8.5 percentage points higher than DQ estimates, a difference that is statistically significant. Following the "more is better" assumption, this means ICT is a viable method for gathering more valid data than conventional DQ—which is an encouraging finding. We also find no evidence for publication bias, meaning that studies favoring ICT over DQ are no more likely to have been published and to have entered the meta-analysis. The downside of this, however, is that the studies included here exhibit a pronounced heterogeneity in their results, meaning that one cannot be sure that ICT will work as intended for future studies. This is pointed out by a prediction interval overlapping zero. Hence, more knowledge is needed about factors affecting the success of ICT procedures to increase the reliability of this approach. The meta-regression results provide several practical clues in this regard. First, ICT seems to be well suited for all types of survey modes and sample types, since no effects were found here. The same holds for PCT, which seems to work comparably to the classic ICT. Regarding the double-list design, the results indicate that it is slightly less effective than the classic procedure. Furthermore, it is advisable to keep the lists of non-key items short and to ensure their non-sensitive nature.

An important finding is that ICT is more successful in mitigating response bias for socially desirable items where over-reporting is expected (e.g., green behavior or voter turnout). Behavioral items also seem less suited for ICT

than attitudinal ones. Moreover, at least in the bivariate model, the ICT effect gets smaller with growing item sensitivity, a finding contradicting theoretical assumptions. Taken together, the evidence suggests that ICT is not a method best suited to very sensitive and strongly negatively connoted behavioral items (as the classic Barton 1958 item "Did you kill your wife?" illustrates); rather, it is more appropriate for addressing misreporting on moderately sensitive and attitudinal over-reporting questions.

For the reasons stated in the introduction section, our results are not directly comparable to the meta-study by Blair, Coppock, and Moor (2020). It is worth mentioning, however, that the authors arrive at a comparable finding to our study with respect to the overall ICT effect, which is significantly in favor of ICT, stronger for over-reporting items, and associated with a prediction interval overlapping zero (please refer to the Supplementary Material for a more detailed comparison).

With respect to limitations and desiderata of our study, several points should be mentioned. One caveat is that the exchangeability assumption could be questioned—that is, studies entering the analysis are not investigating the same, but completely different, questions (samples, items, etc.). We believe, however, that the assumption is justified insofar as we focus on DQ–ICT differences (with the method being the same in all studies) modeled by multilevel meta-analytic models and meta-regressions that account for study and item characteristics. Hence, we argue that *conditional exchangeability* is given for this approach after adjusting for item and study characteristics (Draper et al. 1993). Second, recent evidence suggests that the "more is better" argumentation might be problematic, since estimates could suffer from false positives, for which Höglinger and Diekmann (2017) and Höglinger and Jann (2018) have provided evidence with respect to the crosswise model. This means that respondents *not* having a certain sensitive trait or having committed a sensitive behavior report affirmatively on it. This would completely thwart the "more is better" argumentation, because differences in estimates by question formats would not depict a reduction in misreporting but an inflated reporting of false positives. Future research should replicate the above-cited findings and also investigate whether RRT and ICT could possibly suffer from false positives.

References

Andersen, Henrik, and Jochen Mayerl. 2019. "Responding to Socially Desirable and Undesirable Topics." *Methods, Data, Analyses (mda)* 13:7–35.

Barton, Allen H. 1958. "Asking the Embarrassing Question." *Public Opinion Quarterly* 22: 67–68.

Begg, Colin B., and Madhuchhanda Mazumdar. 1994. "Operating Characteristics of a Rank Correlation Test for Publication Bias." *Biometrics* 50:1088–1101.

Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20:47–77.

Blair, Graeme, Alexander Coppock, and Margaret Moor. 2020. "When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114:1297–1315.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester: Wiley.

Comşa, Mircea, and Camil Postelnicu. 2013. "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique." *International Journal of Public Opinion Research* 25:153–72.

Coutts, Elisabeth, and Ben Jann. 2011. "Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)." *Sociological Methods and Research* 40:169–93.

DerSimonian, Rebecca, and Nan Laird. 1986. "Meta-Analysis in Clinical Trials." *Controlled Clinical Trials* 7:177–88.

Draper, David, James S. Hodges, Colin L. Mallows, and Daryl Pregibon. 1993. "Exchangeability and Data Analysis." *Journal of the Royal Statistical Society* 156:9–37.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 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*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: Wiley.

Duval, Sue, and Richard Tweedie. 2000. "Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis." *Biometrics* 56: 455–63.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *BMJ* 315:629–34.

Fisher, Zachary, and Elizabeth Tipton. 2015. robumeta: An R-Package for Robust Variance Estimation in Meta-Analysis. https://arxiv.org/abs/1503.02220.

Fox, James Alan, and Paul E. Tracy. 1986. *Randomized Response. A Method for Sensitive Surveys*. Vol. 07-058, *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Newbury Park, CA: Sage.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77 (Special Issue):159–72.

Grant, Tobin, Ruth Moon, and Shane A. Gleason. 2014. "Asking Many, Many Sensitive Questions: A Person-Count Method for Social Desirability Bias." Unpublished manuscript.

Höglinger, Marc, and Andreas Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* (25):131–37.

Höglinger, Marc, and Ben Jann. 2018. "More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model." *PLoS One* 13:e0201770. https://doi.org/10.1371/journal.pone.0201770.

Holbrook, Allyson L., and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports. Tests Using the Item Count Technique." *Public Opinion Quarterly* 74:37–67.

Hyman, Herbert. 1944. "Do They Tell the Truth? "*Public Opinion Quarterly* 8:557–59.

Junkermann, Justus. 2020. "Die Person Sum Technique. Ein neues Instrument zur Erhebung quantitativer heikler Items." In *Devianz und Subkulturen. Theorien, Methoden und empirische Befunde*, edited by I. Krumpal and R. Berger. Wiesbaden: Springer VS.

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

Lensvelt-Mulders, Gerty J. L. M. 2008. "Surveying Sensitive Topics." In *International Handbook of Survey Methodology*, edited by E. D. de Leeuw, J. J. Hox, and D. A. Dillman. New York: Lawrence Erlbaum.

Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. van der Heijden, and Cora J. M. Maas. 2005. "Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation." *Sociological Methods and Research* 33:319–48.

Miller, Judith D. 1984. "A New Survey Technique for Studying Deviant Behavior." *Unpublished dissertation*. Washington, DC: George Washington University.

Perry, Paul. 1979. "Certain Problems in Election Survey Methodology." *Public Opinion Quarterly* 43:312–25.

Preisendörfer, Peter, and Felix Wolter. 2014. "Who Is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys." *Public Opinion Quarterly* 78:126–46.

Raghavarao, Damaraju, and Walter T. Federer. 1979. "Block Total Response as an Alternative to the Randomized Response Method in Surveys." *Journal of the Royal Statistical Society. Series B (Methodological)* 41:40–45.

Rasinski, Kenneth A., Gordon B. Willis, Allison K. Baldwin, Wenchi Yeh, and Lisa Lee. 1999. "Methods of Data Collection, Perceptions of Risks and Losses, and Motivation to Give Truthful Answers to Sensitive Survey Questions." *Applied Cognitive Psychology* 13: 465–84.

Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2015. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60:783–802.

Smith, Linda L., Walter T. Federer, and Damaraju Raghavarao. 1974. "A Comparison of Three Techniques for Eliciting Truthful Answers to Sensitive Questions." *Proceedings of the American Statistical Association (Social Statistics Section)*: 447–52.

Stocké, Volker. 2007. "The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys." *Journal of Official Statistics* 23: 493–514.

Tanner-Smith, Emily E., and Elizabeth Tipton. 2014. "Robust Variance Estimation with Dependent Effect Sizes: Practical Considerations Including a Software Tutorial in Stata and SPSS." *Research Synthesis Methods* 5:13–30.

Tipton, Elizabeth. 2013. "Robust Variance Estimation in Meta-Regression with Binary Dependent Effects." *Research Synthesis Methods* 4:169–87.

———. 2015. "Small Sample Adjustments for Robust Variance Estimation with Meta-Regression." *Psychological Methods* 20:375–93.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–83.

Trappmann, Mark, Ivar Krumpal, Antje Kirchner, and Ben Jann. 2014. "Item Sum—A New Technique for Asking Quantitative Sensitive Questions." *Journal of Survey Statistics and Methodology* 2:58–77.

Veroniki, Areti Angeliki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian P. T. Higgins, Dean Langan, and Georgia Salanti. 2016. "Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis." *Research Synthesis Methods* 7:55–79.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the Metafor Package." *Journal of Statistical Software* 36:1–48.

Vienrich, Alessandra Bazo, and Mathew J. Creighton. 2018. "What's Left Unsaid? In-Group Solidarity and Ethnic and Racial Differences in Opposition to Immigration in the United States." *Journal of Ethnic and Migration Studies* 44:2240–55.

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

Wolter, Felix. 2019. "A New Version of the Item Count Technique for Asking Sensitive Questions: Testing the Performance of the Person Count Technique. "*Methods, data, analyses (MDA)* 13:169–92.

Wolter, Felix, and Bastian Laier. 2014. "The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency." *Survey Research Methods* 8:153–68.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. "Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis. "*Metrika* 67:251–63.