Polit Behav (2014) 36:659–682
DOI 10.1007/s11109-013-9249-x

ORIGINAL PAPER

# Artificial Inflation or Deflation? Assessing the Item Count Technique in Comparative Surveys

**Chad P. Kiewiet de Jonge · David W. Nickerson**

Published online: 4 August 2013

**Abstract** While the popularity of using the item count technique (ICT) or list experiment to obtain estimates of attitudes and behaviors subject to social desirability bias has increased in recent years among political scientists, many of the empirical properties of the technique remain untested. In this paper, we explore whether estimates are biased due to the different list lengths provided to control and treatment groups rather than due to the substance of the treatment items. By using face-to-face survey data from national probability samples of households in Uruguay and Honduras, we assess how effective the ICT is in the context of face-to-face surveys—where social desirability bias should be strongest—and in developing contexts—where literacy rates raise questions about the capability of respondents to engage in cognitively taxing process required by ICT. We find little evidence that the ICT overestimates the incidence of behaviors and instead find that the ICT provides extremely conservative estimates of high incidence behaviors. Thus, the ICT may be more useful for detecting low prevalence attitudes and behaviors and may overstate social desirability bias when the technique is used for higher frequency socially desirable attitudes and behaviors. However, we do not find strong evidence of variance in deflationary effects across common demographic subgroups, suggesting that multivariate estimates using the ICT may not be biased.

C. P. Kiewiet de Jonge
Political Studies Division, Centro de Investigación y Docencia Económicas (CIDE), Carretera
México-Toluca 3655, Col. Lomas de Santa Fe, Deleg. Álvaro Obregón, C.P. 01210 Mexico,
DF, Mexico
e-mail: chad-kiewiet@cide.edu

D. W. Nickerson (✉)
Department of Political Science, University of Notre Dame, 217 O'Shaughnessy Hall,
Notre Dame, IN 46556, USA
e-mail: dnickers@nd.edu

**Keywords** List experiment · Item count technique · Survey design · Social desirability bias · Uruguay · Honduras

## Introduction

In recent years, a survey experiment known as the item count technique (ICT) or list experiment[1] has become increasingly popular among political scientists as an unobtrusive method to estimate attitudes and behaviors thought to be subject to social desirability pressures. Within American politics, scholars have used the ICT to study attitudes concerning different presidential candidate characteristics (Heerwig and McCabe 2009; Kane et al. 2004; Streb et al. 2008), racism (Kuklinski et al. 1997), and voter turnout (Holbrook and Krosnick 2010). Comparative studies have utilized the technique to estimate levels of vote buying in Latin America (Ocantos et al. 2012), attitudes toward the extension of the franchise to illiterates in Lebanon (Corstange 2009), attitudes about electoral fairness and violence in Tanzania (Weghorst 2010), gang activity in Mexico (Díaz Cayeros et al. 2011), the use of micro-finance loan proceeds in Peru and the Phillipines (Karlan and Zinman 2012), and bribery by multinational firms in Vietnam (Malesky et al. 2011).[2] Given the wide range of sensitive attitudes and behaviors that political scientists seek to study, the ICT is likely to see continued use.

As long as the core assumptions of the technique hold, the ICT can generate valid estimates of attitudes and behaviors normally subject to social desirability pressures by providing respondents with response categories that assure anonymity and leveraging the random assignment of respondents to question formats. The ICT works as follows: respondents are randomly assigned to either the treatment or control group. The control group is shown a list of $X$ nonsensitive items (typically 3 or 4), and the treatment group receives a list of $X + 1$ items that include the same baseline items plus the sensitive attitude or behavior of interest. Respondents are asked to provide the *number* of items on the list that apply to them, not which ones. Since the respondent provides only a number between 0 and $X$ or $X + 1$, the response options offer anonymity for the respondents in the treatment group, thereby decreasing respondent editing of answers due to social desirability concerns.[3] Random assignment assures that the only difference between treatment

[1] The ICT is known by a variety of different names, including block total response, unmatched item count technique, randomized list technique, and the list experiment. The difference in terminology reflects the fact that the technique has developed in different disciplines in relative isolation, with political scientists generally referring to the technique as the list experiment, while those working in sociology, business ethics, and public health using the item count terminology.

[2] The ICT has also been utilized frequently in sociology, psychology, and business ethics. For example, scholars have used the ICT to assess rates of risky behaviors (Anderson et al. 2007; Biemer and Brown 2005; LaBrie and Earleywine 2000; Miller 1984; Zimmerman and Langer 1995), unethical behavior in the workplace (Dalton et al. 1994; Wimbush and Dalton 1997), among many other topics.

[3] Ceiling and floor effects can remove the anonymity of the responses by forcing the respondent to say "All" or "None" of the items. Thus, it is good practice to minimize the variance of responses to the control list and place the mean response at roughly $\frac{X}{2}$.

and control group is the addition of the sensitive item for the treatment group. If the sensitive item is not applicable to any respondent, then the difference between the mean number of items indicated by each group should be zero. If the sensitive item is applicable to some respondents, then these respondents should include it in their counts, and estimates of the sensitive item can be derived by simply differencing means.[4]

A variety of studies have shown that the technique is in many ways superior to the randomized response technique, another leading measurement strategy used to reduce social desirability bias (Droitcour et al. 1991; Hubbard et al. 1989; Coutts et al. 2008). Although some unsuccessful applications of the technique have called its validity into question (Tourangeau and Yan 2007, pp. 872–873), Holbrook and Krosnick's (2010, pp. 44–45) extensive review of the literature suggests that the ICT provides significantly different estimates from direct survey items in the expected direction in two-thirds of times that it has been applied. While the direction of the differences may be "expected" by researchers, the accuracy of estimates provided by ICT is open to question. Assessing the overall validity of the technique is particularly challenging since the baseline rates of most stigmatized attitudes and behaviors are unknown.

A strong requirement of the identification assumptions of the ICT that has not received sufficient attention in the literature is that the differing lengths of the treatment and control lists have no effect on the survey responses independent of the substance of the list. If respondents receiving the longer treatment version of an ICT provide systematically different answers because the list is longer than the baseline list irrespective of the actual treatment item, then the ICT will provide unreliable and potentially biased results. While a few authors have directly (Holbrook and Krosnick 2010) and indirectly (e.g. Tsuchiya et al. 2007) tested the effect of differing list length between treatment and control lists, these tests have used internet surveys based on samples of respondents from the developed world, and such findings may not generalize to other survey modes or populations.

In order to provide a more robust assessment of the list length assumption, we conducted tests of the ICT using face-to-face surveys of nationally representative probability samples of households in Uruguay and Honduras following national elections in each country. The samples were randomly split into three groups—one control group with four baseline items and two treatment groups with an additional item. The first treatment group received an item whose prevalence in the survey samples should be close to zero (being a candidate for office), and the second treatment group received an item whose frequency should be close to universal (knowing an election took place). If the length assumption holds, then the estimates provided by the first treatment group should be indistinguishable from zero, and the estimate provided by the second treatment group should be indistinguishable from one.

[4] Researchers can further stratify the sample so that estimates for different subgroups can be derived in a similar manner. Other researchers have also begun developing multivariate estimators for the ICT (e.g. Glynn 2013; Blair and Imai 2012).

We find little evidence of an upward bias in estimates, but significant under estimation of the occurrence of the common behavior. This result provides further evidence that respondents do not count items in the same way as when they are asked about each item directly (Biemer et al. 2005; Tsuchiya et al. 2007). Nonetheless, the results imply that the ICT does not *overestimate* socially undesirable attitudes and behaviors and may even provide conservative estimates. The ICT may be biased against detecting the existence of low prevalence attitudes and behaviors (see also Droitcour et al. 1991), and may lead to underestimates of high incidence behaviors and attitudes, which may lead scholars to overestimate the extent of social desirability bias. However, we find only weak and non-robust evidence of heterogeneity in deflationary effects across common demographic subgroups. If replicated in future studies, this finding suggests that multivariate analyses using the list experiment as a dependent variable might not produce biased coefficient estimates even though the estimates of overall incidence are likely to be biased downward.

Beyond these substantive findings, this paper makes three additional contributions. First, our validation methodology is unique in asking about two behaviors where we have very strong priors about their incidence (0 and 100 %). Assessing the behavior of the ICT on the extremes is especially useful since the ICT is often used to study rare behaviors like drug use or common behaviors like voter turnout. Second, unlike other tests of the artificial inflation hypothesis that use Internet surveys, our surveys were face-to-face. Self-administered surveys (including Internet surveys) are less useful tests of the ICT since social desirability pressures are lower in these formats than in live interviewer modes (e.g. Droitcour et al. 1991; Holbrook and Krosnick 2010). Additionally, the process by which respondents count list items may be very different for self-administered surveys than in live interviews. Third, unlike nearly all other published assessments of the ICT, we use surveys of countries in the developing world (but see Corstange 2009). To advance survey methodology in the developing world, techniques validated in the developed world must be studied to ensure they provide accurate estimates in settings with lower literacy rates and differing baseline expectations of social desirability bias (cf. Harkness and Van de Vijver 2003).

## Artificial Inflation and Deflation using the ICT

Scholars have explored a number of the underlying attributes of the ICT, including: how respondents react to it (Droitcour et al. 1991; Hubbard et al. 1989; Coutts and Jann 2008); the prevalence and consequences of ceiling effects (Kuklinski et al. 1997); the relative inefficiency of ICT estimates and ways to optimize variance (e.g. Droitcour et al. 1991; Glynn 2013; Tsuchiya et al. 2007); varying list lengths (Tsuchiya et al. 2007); disparities between direct and list counts (Biemer et al. 2005; Tsuchiya and Hirai 2010); and potential response order effects (Tsuchiya and Hirai 2010). Although the ICT provides an intuitive method for decreasing social desirability bias and the existing literature has contributed helpful advice on the proper execution of the technique, its validity remains dependent on a number of assumptions that remain largely untested.[5]

Imai (2011) and Blair and Imai (2012) describe three identification assumptions that must hold for the ICT to produce valid estimates of sensitive items. First, the *Randomization of the Treatment* assumption requires respondents to be randomly assigned to lists. Second, the *No Design Effect* assumption requires that the addition of the treatment item does not influence respondents' answers to the control items. Finally, the *No Liar* assumption specifies that respondents provide a truthful answer to the treatment item (i.e., include it in their count response if the item is applicable or true and not include it in their count response if it is inapplicable or false).[6] While researchers usually do not have difficulty with complying with the *Randomization of the Treatment* assumption, it remains unclear the degree to which the *No Design Effect* and *No Liar* assumptions hold in practice. In the following two sections, we illustrate two ways in which adding the treatment item can violate these assumptions and introduce systematic respondent error in the survey response process.

## Artificial Inflation

If the identifying assumptions hold, the fact that the treatment list is one item longer the control list should in and of itself have no effect on the respondent's answer since the addition of the sensitive item for the treatment group should only elicit a higher count by a respondent if the sensitive item applies to her. However, considerable evidence exists that people vary their substantive answers to survey questions in response to different response categories across a wide variety of topics (Bless et al. 1992; Menon et al. 1995; Schwarz and Bienias 1990; Schwarz et al. 1985; Schwarz and Scheuring 1988). Respondents may use the number of items on the list to derive an initial estimate of how many items "should" apply to them and then adjust this initial starting value, which increases the likelihood of generating errors (Sudman et al. 1996, p. 218). If the process by which respondents form their responses to ICTs is dependent on the number of items on the list, then both the *No Design Effect* and *No Liar* assumptions are violated since such a response process would affect the count of both control items and the treatment item.

[5] Techniques that combat social desirability bias require additional assessment in the developing world setting. Although social desirability bias is a universal source of measurement error (Johnson and Van deVijver 2003), personality-based social desirability scales show variance across different cultural and socioeconomic settings, with greater levels of socially desirable responding occurring in poorer, more collectivist countries as opposed to richer individualistic countries (Johnson and Van deVijver 2003, p. 197–200). As survey research gains increasing importance in the developing world, scholars need to evaluate the usefulness of techniques developed in the industrialized world for countries at lower levels of development (cf. Harkness and Van de Vijver 2003).

[6] Note that neither of these assumptions stipulate that responses to the control items be truthful or accurate; rather, only that the treatment item does not affect control item counts. That is, the assumptions "together eliminate the possibility that the coexistence of the sensitive and control items in a single list influence responses in one way or another" (Imai 2011, p. 409).

For example, consider an ICT that provided a list of political activities and asked how many of these activities the respondent participated in during an electoral campaign. If the respondent believes that he participates in politics at an average level, he might anchor his response on the average number of items on the list and make a few adjustments to this initial estimate based on the actual list items. Thus, instead of actually counting the number of items on the list that applied to him, this respondent would base his answer on an anchoring and adjustment strategy in which the number of response options would partly determine his response. If this respondent were given a five-item treatment list he might answer three (initially anchoring at the mean of 2.5 and rounding up), while if this respondent received a four-point control list he might simply answer two (the mean of this list). To the degree that many respondents engage in this anchoring and adjustment strategy instead of the careful counting assumed by the technique, substantial *artificial inflation* of estimates could result given the differing lengths of the treatment and control lists.

Further, respondents are more likely to use the range of response alternatives as a frame of reference as the cognitive complexity of questions increase (Bless et al. 1992). Given that the ICT tends to require greater cognitive effort than direct questions (particularly in the case of live-interviewer modes) and some respondents engage in satisficing strategies (e.g. Krosnick and Alwin 1987), the number of items presented to respondents may have an effect on answers. Consequently, the additional item presented to the treatment group might artificially inflate responses, independent of the content of the sensitive item.

Some indirect and direct tests cast doubt on this artificial inflation hypothesis. First, the evidence regarding response range effects is largely based on specific behavioral frequencies, whereas the ICT requires respondents to count behaviors or attitudes, which may avoid associated errors. Second, approximately one-third of the applications of the technique have not produced positive and significant results (Holbrook and Krosnick 2010, pp. 44–45), which we would not expect if the artificial inflation hypothesis were true. Third, ICT estimates of behaviors not subject to social desirability bias are not different from estimates from direct questions (Dalton et al. 1994; LaBrie and Earleywine 2000; Tsuchiya et al. 2007). If differing list lengths artificially inflated estimates, there should be inflation for non-sensitive items as well.

The only direct test of the artificial inflation hypothesis provides mixed support. Holbrook and Krosnick (2010, p. 53 fn. 20) conducted a non-representative Internet sample of 1,510 respondents with an ICT asking respondents about everyday behaviors. The treatment item added to the four baseline items was “taken a vacation in the country of Tantatoula,” which is a nonexistent country. There was a positive difference between the mean number of items indicated by control and treatment groups (1.77 and 1.86). Using a two-tailed $t$ test, the difference did not reach statistical significance ($t(1{,}510) = 1.40$), but a one-tailed test nearly reaches traditional thresholds for statistical significance ($p = 0.08$).

## Artificial Deflation

A second way in which respondents could violate the identifying assumptions is if respondents make systematic errors in the counting process such that their counts using the ICT are biased relative to the "true" item counts, *and* that the likelihood of committing such counting errors is positively related to the length of the lists. Such error in the response process would violate both the *No Design Effect* and *No Liar* assumptions since responses to both the control items and the treatment item would be affected by the differences in list lengths. Thus, if respondents systematically underestimate (overestimate) items in the IC lists compared to the "true" count and such underestimation (overestimation) increases as the length of the list increases, the ICT could produce *artificially deflated* (*inflated*) estimates, since treatment lists are longer than control lists.

If we assume that responses to direct questions about non-sensitive items reflect the "true counts" of these items, then there is some evidence in the literature that the ICT could produce negatively biased item counts and deflated estimates. Biemer et al. (2005) compared direct question responses to item count responses by the same respondents and found modest correlations between the two counts, suggesting that a high degree of measurement error may afflict the ICT. Tsuchiya et al. (2007) study of ICT properties using Japanese Internet samples demonstrated that direct questions about non-sensitive items produced higher counts than the standard ICT, and that the difference in estimates between formats increases as list lengths increase. That is, as the cognitive difficulty of counting increases with list length, ICT underestimates frequency more compared to direct questioning.

Although the precise psychological mechanisms by which such downward bias occurs are unclear, additional analysis by Tsuchiya and Hirai (2010) suggests that this deflationary effect is likely due to the fact that respondents are only asked to consider the number of items on the list that apply to them rather than how many apply and do not apply. In contrast, counts based on direct questioning reflect more explicit binary choices about whether the item applies or not. This initial evidence suggests that rather than being subject to artificial inflationary effects, the ICT may be more prone to a deflationary effect that results in underestimates of the treatment item. Further, the extent of the deflationary effect might be related to cognitive abilities, satisficing, and the length of the lists.

Whether this evidence regarding the artificial inflation and deflation hypotheses generalizes to live interviewer modes or developing countries is an open question. A first objection to this position is that nearly all of the indirect and direct evidence in favor of the list length assumption is based on self-administered surveys (Dalton et al. 1994; Holbrook and Krosnick 2010; LaBrie and Earleywine 2000; Tsuchiya et al. 2007). The threat of social desirability bias is much greater with live interview modes since respondents may edit answers to project a positive image to the interviewer (e.g. Holbrook and Krosnick 2010; Tourangeau and Smith 1996). Self-editing responses may cause people to rely on different heuristics in response to different interview modes.

Furthermore, the process of answering ICT questions differs substantially between live interviewer and self-administration modes (Tsuchiya et al. 2007). While respondents can physically count items in telephone and self-administration modes, respondents in face to face interviews are dissuaded from pointing to particular items or counting out loud, since these techniques reveal the counted items to the interviewer, thereby defeating the purpose of the ICT. Thus, the cognitive process is more difficult in the live interview mode, which may lead respondents to rely more on list length when forming their responses.

A second objection against the generalizability of the evidence in favor of the list length assumptions is that the results are all based on samples from the developed world (United States and Japan). The performance of the ICT in countries with lower literacy levels is largely untested (but see Corstange 2009; Ocantos et al. 2012). Illiteracy precludes the use of cognitively easier self-administered survey modes in many developing contexts. Furthermore, the ICT is more cognitively demanding for the illiterate since responses to the ICT must be based on an oral list. As a result, people who struggle to read may rely more on the number of response categories in formulating responses, implying that artificial inflation may be particularly problematic in the developing world setting and among illiterate respondents in particular. Thus, it remains unclear that the ICT travels well beyond the developed world.

## Data and Methods

To test the artificial inflation and deflation hypotheses, we conducted item count experiments using probability samples of households in Uruguay and Honduras in December 2009 and January 2010, respectively, following high profile presidential and Congressional elections. Each study is described below, followed by an analysis of the experimental results, including an examination of potential heterogeneity in effects among different demographic sub-groups.

### Study 1: Uruguay

The survey firm Equipos Mori conducted a face-to-face nationally representative omnibus survey of 900 Uruguayans during December 15–18, 2009 following the second round of the presidential elections in the country. The sample design was a multistage stratified sample of households, and the AAPOR response rate 1 was 31.6 %. Further methodological details of the survey are available in "Appendix 1".

Respondents were randomly assigned to three experimental conditions. Interviewers carried different versions of the questionnaire for each experimental condition and applied them according to a pre-determined randomized list.[7] The ICT question was the following:

> We are interested in knowing how people become involved in politics. I am going to show you a list of political activities and I would like for you to tell me HOW MANY of these activities you completed during the last campaign. Do not tell me which ones, only HOW MANY.

Interviewers showed respondents in the control group a card with the following baseline list, which they also read aloud:

[7] The success of the randomization procedure is discussed further in "Appendix 2".

- I volunteered for the campaign of one of the parties
- I attended a rally
- I tried to persuade a friend to vote for my candidate
- I picked a fight with someone over a candidate

In addition to these baseline items respondents assigned to the first treatment group received a treatment item (placed in the third position) that should be of very low prevalence (1 % claimed to be a candidate when asked directly)[8]:

- I ran for office

The second treatment group received a different additional item (also placed in the third position) whose frequency should approach 100 percent in an election where 89 % of registered voters cast a ballot (in our sample 92 % reported casting a ballot, and 96 % claimed to be aware that the election was taking place):

- I was aware that the elections were taking place[9]

Later in the questionnaire, all respondents were asked about the two treatment items directly. The survey instrument also included a number of other political items related to the presidential election as well as basic demographic questions.

Non-sensitive items were used to establish the baseline rate of behavior to sidestep an unavoidable identification problem. The ICT is used in cases where social desirability bias is expected to bias a person's response to survey questions. This means that direct questions cannot be used to estimate the true incidence of behaviors subject to social desirability pressures. Since it is usually impossible to know objectively the baseline rates of sensitive attitudes and behaviors, this necessitates establishing the reliability of ICT using non-sensitive topics. Fortunately, the available evidence suggests that neither artificial inflation nor deflation should arise more frequently with sensitive items compared to non-sensitive items (Tsuchiya et al. 2007).[10]

## Study 2: Honduras

Our second survey was a nationally representative multistage stratified sample of 1008 Honduran households with the voter registry serving as the sampling frame. This face-to-face omnibus survey was conducted by Borge y Asociados during the

[8] Although the percentage claiming to have been a candidate is higher than expected, the elections in Uruguay included both presidential and parliamentary elections, somewhat increasing the likelihood that the sample would include a candidate.

[9] The wording in Spanish was the following: "Nos interesa saber cómo se involucran las personas en política. Voy a mostrale una lista de actividades políticas y quisiera que me diga cuántas de estas actividades realizó usted durante la última campaña. No me diga cuáles, sólo CUÁNTAS". The baseline response categories were: "Participé como voluntario para la campaña de uno de los partidos," "Participé en una movilización," "Intenté convencer a un amigo de que votara por mi candidato," and "Tuve una pelea con alguien sobre un candidato."

[10] In fact, Tsuchiya et al. (2007) finds deflation to be more of a problem with non-sensitive items than sensitive items.

3 week of January 2010. The AAOPR response rate 1 was 51 % (with a refusal rate of 9 %). Additional methodological details are included in "Appendix 1".

As with the Uruguayan survey, respondents were randomly assigned to either the control group or one of the two treatment groups. Since the survey included two other experiments, there were twelve versions of the survey. Each of the twelve surveys was applied in each of the 84 primary sampling units according to a randomized list developed prior to the fielding of the survey.[11]

The only difference between the experiment conducted in Uruguay and the one conducted in Honduras is that the baseline list differed. The items used in the Honduran experiment were of higher prevalence than those utilized for the Uruguayan survey[12]:

- I voted for a candidate
- I participated in a rally
- I discussed the election with someone
- I saw or read something about the election in the news[13]

The treatment items remained the same, and both were placed in the third position. The Honduran survey did not include direct questions about the treatment items. The questionnaire also asked about a number of political topics associated with the November 2009 presidential elections as well as basic demographic items.[14]

## Results

For both the Uruguayan and Honduran surveys, if artificial inflation were occurring, we would expect that the estimate for being a candidate ("treat low") would be significantly greater than zero. In contrast, if artificial deflation were occurring, we would expect that the estimate for knowing the election was taking place ("treat high") would be significantly less than one.[15] If these inflationary or deflationary effects were related to the cognitive abilities of the respondents or other common demographic variables, we would expect that interactions between such variables and the treatment variables would be significantly different from zero.

[11] The success of the randomization procedure is discussed further in "Appendix 2".

[12] The item counts produced in the Uruguayan survey were quite low, with just over half of respondents indicating zero items. As explained below, this feature of the Uruguayan experiment may partially account for some of the marginal heterogeneity in the results.

[13] In Spanish, the baseline items for Honduras were "Voté por algún candidato," "Participé en una movilización," "Discutí acerca de la eleción con alguien," "Vi o leí algo acerca de la elección en las noticias." The treatment items were "Participé como candidato" and "Estaba al tanto de que las elecciones se iban a llevar a cabo," respectively.

[14] Although the June 2009 coup in Honduras might have increased awareness about the elections, there is no reason to believe that it would have made either of the treatment items sensitive.

[15] As a reviewer helpfully pointed out, artificial deflation might also produce a negative estimate of the treat low condition and artificial inflation could potentially produce an estimate greater than 1 (i.e. greater than 100%) for the treat high condition.

Contrary to the expectations of many critics of the ICT, evidence in favor of the artificial inflation hypothesis is very weak (see Table 1, "Treat Low" rows). In neither Uruguay nor Honduras did the ICT estimate the low incidence behavior to be significantly different from the expected zero. In Uruguay, respondents presented with the baseline list (i.e., control) reported performing 0.69 items on average. In contrast, respondents faced with a list with the low incidence item added (i.e., "I ran for office") reported only 0.77 items on average.[16] This difference of 0.08 is not statistically significant (SE $= 0.09$; $p = 0.18$, one tailed) and could be due entirely to chance.[17] The artificial inflation hypothesis receives even less support in the Honduras survey where respondents presented the control and low-incidence lists both reported 1.90 items on average (s.e. $= 0.12$, $p = 0.50$, one tailed). The precision weighted average of these two differences is 0.05 items (i.e., 5 %, s.e. $= 0.07$), which does not come close to reaching statistical significance ($p = 0.38$, one tailed). Thus, our surveys find little support for the artificial inflation hypothesis. At least for low incidence behaviors, any upward bias from increased list length in and of itself is likely to be small.

The picture is markedly different for the artificial deflation hypothesis (see Table 1, "Treat High" rows). In both surveys, the ICT grossly underestimated the true incidence of knowing the election was taking place. In Uruguay, respondents presented with the longer list containing the high incidence behavior (i.e., knowing the election took place) reported only 1.06 items on average, which was only 0.37 items more than respondents in the control group. The standard error on this difference is only 0.08 and any reasonable confidence interval cannot overlap the assumed value of 1 ($p < 0.01$, one tailed).[18] The Honduran survey finds the same pattern. Respondents with the high incidence treatment list responded with an average of 2.49 items, which differs from the control respondents only by 0.59 items. Again, with a standard error of 0.12 there is no way this result could be confused with the true value, which should approach 1 ($p < 0.01$, one tailed). Given the high profile nature of the election—it occurred after a June 2009 coup d'état against the elected president that had received strong attention both within and outside of the country—it is extremely unlikely our result is due to sampling variability. Thus, it appears that the ICT is subject to a substantial downward bias (i.e., artificial deflation) when estimating high incidence behaviors and attitudes. Although the results do not illuminate the psychological mechanism(s) by which such deflation occurred, the findings are consistent with the evidence in the literature about the underestimation of ICT counts relative to counts based on direct questioning and the positive relationship between such deflationary errors and list length.

[16] Although the low number of items reported by members of both groups in the Uruguayan survey would generally lead to questions about potential floor effects, since the treatment items are non-sensitive, such concerns should be allayed. We thank a reviewer for pointing this out.

[17] Restricting the sample only to those respondents who claimed that they were not a candidate makes the estimated difference between treatment and control lists even smaller (0.056) and the evidence against the artificial inflation hypothesis even stronger (one tailed $p$ value $= 0.26$ for difference of means).

[18] In Uruguay, a later question on the survey directly asked respondents if they knew the election was taking place; 96 % of respondents answered affirmatively.

**Table 1** Treatment results

| | Average items | Treatment estimate (%) | Standard error | $p$ Value |
|---|---|---|---|---|
| *Uruguay* | | | | |
| Control list | 0.69 | | | |
| Test low | 0.77 | 8.1 | (8.90) | 0.181 |
| Test high | 1.06 | 37.1 | (7.90) | <0.001 |
| *Honduras* | | | | |
| Control list | 1.90 | | | |
| Test low | 1.90 | 0.1 | (11.9) | 0.497 |
| Test high | 2.49 | 58.7 | (12.3) | <0.001 |

Standard errors take into account the clustering in the survey design. $p$-values reflect one-tailed tests of whether the estimate is greater than zero or less than one for the test low and test high treatment conditions, respectively

To determine whether artificial inflation and deflation affect some subpopulations more than other, we conducted extensive tests for heterogeneous treatment effects (see Appendix 3). After mining the available data, we could find no robust evidence of treatment heterogeneity in either survey. Furthermore, any hints of populations susceptible to inflation or deflation in one country, could not be replicated in the other country. Even respondents who appear relatively less engaged in the survey process are no more likely to exhibit inflation or deflation in their responses to the item count question. This null finding suggests that susceptibility to artificial deflation in responses is spread uniformly over the population. Although our sample sizes do not allow us to say definitively that heterogeneous responses do not exist, they do allow us to rule out large differences based on gender, age, education, income, and cognitive engagement with the survey. If replicated, these findings would imply that multivariate analysis of list experiments will provide unbiased coefficient estimates and the error will be largely confined to the overall estimate of behavior incidence.

## Discussion

The ICT is a clever way to estimate the frequency of sensitive behaviors or attitudes that respondents might not reveal if asked directly. However, the identifying assumptions of the technique require rigorous testing to understand the conditions under which they are (un)warranted. Our goal in the paper was to test the general hypothesis that biased estimates would arise as a result of the fact that the treatment and control conditions are not perfectly parallel (i.e., they employ lists of different lengths, X and X + 1). In particular, we wanted to see whether the ICT artificially inflated low-frequency behaviors and/or artificially deflated high-frequency behaviors. In both of our surveys, there was scant evidence of artificial inflation and substantial artificial deflation.

Together, our findings do not call into question the broad utility of the ICT, although the results do suggest caution in interpreting ICT estimates of frequent attitudes or behavior. Although the broad generalizability of our findings will require additional testing to determine the causal mechanisms underlying the artificial deflation findings and whether such an effect occurs across different survey modes and subjects, with these caveats in mind, the directional bias we uncovered is good news for researchers studying low incidence behaviors like drug usage (e.g. Biemer and Brown 2005; Droitcour et al. 1991; Miller and Cisin 1984) or criminal behavior (e.g. Tsuchiya et al. 2007) because the ICT appears to offer conservative estimates of rare behaviors. These results are troublesome for researchers studying behaviors like voter turnout (e.g. Holbrook and Krosnick 2010), belief in God (Jackman 2007), or attitudes toward presidential demographic characteristics (Heerwig and McCabe 2009; Kane et al. 2004; Streb et al. 2008) where social desirability bias is hypothesized to cause over-reporting of the behavior. For such topics, lower than expected estimates from ICT might reflect artificial deflation rather than the true incidence of the behavior or attitude.

This finding suggests its own possible strategy for countering this deflationary tendency. Researchers may be able to generate more conservative estimates of social desirability bias by transforming high incidence treatment items into the low incidence inverses of the behavior or attitude of interest. For example, to estimate voter turnout, instead of including “I voted in the last election” as the treatment item, researchers could estimate voter abstention by employing “I abstained in the last election” as the treatment item. Of course, word choice can be critical to proper measurement of concepts in surveys, but this strategy is worthy of further study for instances of researchers using list experiments to estimate socially desirable high incidence behaviors.

Although the results point toward a significant conservative bias, the lack of robust differences across standard demographic subgroups with respect to possible artificial inflation or deflation suggests that multivariate analyses using the list experiment as a dependent variable are unlikely to produce biased estimates. However, testing for possible heterogeneity across subgroups of interest with larger sample sizes, other survey modes and settings, and different survey topics should be a top priority to confirm these findings. Further, since research design necessitated the use of non-sensitive treatment items to test the artificial deflation and deflation hypotheses due to the need for unbiased prior estimates of the treatment items, it will be important for researchers to test these hypotheses using validated data on sensitive items or behaviors. For example, researchers could assess the artificial deflation hypothesis by applying an ICT concerning past drug use among a population of respondents who have participated in drug treatment programs.

Finally, the strong deflationary findings reported here and elsewhere (Biemer et al. 2005; Tsuchiya et al. 2007; Tsuchiya and Hirai 2010) underline the importance of studying techniques for improving the accuracy of respondents’ item counts. Research along these lines has already suggested a few such methods. Tsuchiya and Hirai (2010) argue that instructing respondents to count how many of the items on the list apply *and* do not apply could also prove to be a successful method for increasing count accuracy. Given that deflation of estimates is more severe as list lengths increase (Tsuchiya et al. 2007), developing shorter lists (e.g. 3–4 vs. 4–5) that still protect the anonymity of the respondent should also prove useful in decreasing possible deflation. Glynn (2013) negative correlation design, which calls for baseline items to be negatively correlated with each other across different groups in the sample, is one promising method for effectively reducing list lengths and thereby reducing possible deflation. Another possibility is to train respondents by providing a "practice" ICT before an ICT of interest to the research. Until a large set of best practices for a variety of survey settings are developed, implementation of ICT will often be sub-optimal.

**Acknowledgments** Funding for the surveys was provided by the Kellogg Institute for International Studies and the Institute for Scholarship in the Liberal Arts at the University of Notre Dame. Nickerson is grateful for the Center for the Study of Democratic Politics at Princeton University for the time to work on this project. We thank *Equipos Mori* for fielding the Uruguayan survey and *Borge y Asociados* for conducting the Honduran survey. We would also like to thank Scott Desposato, Macartan Humphries, Jim Kuklinski, and Devra Moeller and anonymous reviewers for helpful comments. We are particularly indebted to the continuing collaboration of Ezequiel Gonzalez Ocantos, Carlos Melendez, and Javier Osorio.

## Appendix 1: Survey Methodology

### Uruguay

*Survey Firm:* Equipos Mori
*Field Dates:* December 15–18, 2009
*Mode:* Omnibus Face-to-Face
*Sampling Universe:* Nationally representative of adults (18+)
*N:* 900
*Sample Design:* The survey utilized a multistage probability sample of households with quotas utilized within households for the final selection of respondents (Sudman 1966). There were 243 final sampling points, with an average of 4 respondents per sampling point. The sample was first stratified into two grand strata—Montevideo and the Interior. Within Montevideo, the sample was further stratified by municipal zones. Within the interior, stratification occurred by population, with cities with populations exceeding 30,000 inhabitants automatically included and lower population cities selected randomly proportional to population size. Within cities (interior) and zones (Montevideo), final sampling points were randomly chosen proportionate to population, households were chosen randomly based on a systematic sampling procedure, and within households respondents were selected using sex and age quotas. For rural areas, departments were selected randomly according to population, and within selected departments and segments, national highways were selected. Highway distances (km markers) were then randomly selected as starting points for the selection of households, which were chosen based on predetermined random procedures. In total 6 rural sampling points were chosen.

*AAPOR Response Rate 1:* 32 %, Refusal Rate: 33 %

*Randomization Design:* The survey battery included one other question that required randomization such that the combination of the different question versions resulted in 6 different questionnaires. Each questionnaire was applied according to a predetermined randomized list.

### Honduras

*Survey Firm:* Borge y Asociados
*Field Dates:* January 16–25, 2010
*Mode:* Omnibus Face-to-Face
*Sampling Universe:* Nationally representative of adults (18+), excluding the sparsely populated department of Gracias a Dios and the Bay Islands.
*N:* 1,008
*Sample Design:* The survey utilized a multistage random sample with 84 final sampling points (segments), including 12 respondents per segment. Sampling proceeded as follows: The sampling frame consisted of the electoral registry, with primary sampling units chosen proportionate to the size of voting centers within department—municipalities. Within municipalities, random selection proceeded by electoral centers, census tracks, and census blocks, with final sampling points (segments or blocks) containing 12 respondents. Households and respondents within households were chosen randomly in such a way that ensured gender balance.

*AAPOR Response Rate 1:* 50 %, Refusal rate: 9 %

*Randomization Design:* The survey battery included two other questions that required randomization such that the combination of the different questions resulted in 12 different questionnaires. Each of the 12 questionnaires was applied according to a predetermined randomized list within each sampling unit, each of which included 12 respondents.

## Appendix 2: Descriptive Statistics and Randomization Balance

See Tables 2 and 3.

**Table 2** Uruguay descriptive statistics and treatment balance

| | | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| Gender (1 = male, 2 = female) | Control | 301 | 1.545 | 0.499 | 1 | 2 |
| | Test low | 300 | 1.540 | 0.499 | 1 | 2 |
| | Test high | 299 | 1.572 | 0.496 | 1 | 2 |
| | | $\chi^2$ = F(1.99, 481.43) | | | $p$ = 0.7276 | |
| Age (1 = 18–29, 2 = 30–49, 3 = 50+) | Control | 301 | 2.166 | 0.770 | 1 | 3 |
| | Test low | 300 | 2.177 | 0.805 | 1 | 3 |
| | Test high | 299 | 2.204 | 0.791 | 1 | 3 |
| | | $\chi^2$ = F(3.92, 948.90) | | | $p$ = 0.5473 | |

**Table 2** continued

| | | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| Education (0 ≤ primary, 1 = primary, 2 = secondary, 3 ≥ secondary) | Control | 300 | 1.787 | 0.781 | 0 | 3 |
| | Test low | 300 | 1.623 | 0.843 | 0 | 3 |
| | Test high | 299 | 1.595 | 0.931 | 0 | 3 |
| | | $\chi^2$ = F(5.82, 1,408.82) | | | $p = 0.0023$ | |
| Income (1 ≤ \$7,100, 2 = \$7,100–16,100, 3 ≥ \$16,100 UYU per month) | Control | 243 | 1.041 | 0.732 | 0 | 2 |
| | Test low | 245 | 0.967 | 0.809 | 0 | 2 |
| | Test high | 248 | 0.960 | 0.768 | 0 | 2 |
| | | $\chi^2$ = F(3.95, 917.21) | | | $p = 0.0739$ | |
| Income missing (0–1) | Control | 301 | 0.193 | 0.395 | 0 | 1 |
| | Test low | 300 | 0.183 | 0.388 | 0 | 1 |
| | Test high | 299 | 0.171 | 0.377 | 0 | 1 |
| | | $\chi^2$ = F(2.00, 483.03) | | | $p = 0.7441$ | |
| Disengagement (# items with missing data) | Control | 301 | 1.950 | 1.577 | 0 | 7 |
| | Test low | 300 | 2.100 | 1.793 | 0 | 12 |
| | Test high | 299 | 2.237 | 1.841 | 0 | 11 |
| | | $\chi^2$ = F(19.92, 4,819.45) | | | $p = 0.7441$ | |

**Table 3** Honduras descriptive Statistics and randomization balance

| | | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| Gender (1 = male, 2 = female) | Control | 336 | 1.509 | 0.501 | 1 | 2 |
| | Test low | 336 | 1.488 | 0.501 | 1 | 2 |
| | Test high | 336 | 1.503 | 0.501 | 1 | 2 |
| | | $\chi^2$ = F(1.80, 149.09) | | | $p = 0.7161$ | |
| Age (1 = 18–29, 2 = 30–49, 3 = 50 +) | Control | 336 | 1.949 | 0.777 | 1 | 3 |
| | Test low | 336 | 1.967 | 0.789 | 1 | 3 |
| | Test high | 336 | 2.009 | 0.797 | 1 | 3 |
| | | $\chi^2$ = F(3.86, 320.21) | | | $p = 0.7947$ | |
| Education (0 ≤ primary, 1 = primary, 2 = secondary, 3 ≥ secondary) | Control | 336 | 1.461 | 0.787 | 0 | 3 |
| | Test low | 336 | 1.360 | 0.736 | 0 | 3 |
| | Test high | 336 | 1.485 | 0.792 | 0 | 3 |
| | | $\chi^2$ = F(5.57, 462.69) | | | $p = 0.1431$ | |
| Income (1 ≤ \$200, 2 = \$200–\$400, 3 ≥ \$400 USD per month) | Control | 331 | 1.958 | 0.686 | 1 | 3 |
| | Test low | 329 | 1.954 | 0.654 | 1 | 3 |
| | Test high | 330 | 1.973 | 0.686 | 1 | 3 |
| | | $\chi^2$ = F(3.80, 315.39) | | | $p = 0.7868$ | |
| Income missing (0–1) | Control | 336 | 0.015 | 0.121 | 1 | 3 |
| | Test low | 336 | 0.021 | 0.143 | 1 | 3 |
| | Test high | 336 | 0.018 | 0.133 | 1 | 3 |
| | | $\chi^2$ = F(1.98, 164.35) | | | $p = 0.8781$ | |

**Table 3** continued

| | | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|---|
| Disengagement (# items with missing data) | Control | 336 | 0.783 | 1.083 | 0 | 5 |
| | Test low | 336 | 0.801 | 1.089 | 0 | 6 |
| | Test high | 336 | 0.878 | 1.164 | 0 | 8 |
| | | $\chi^2$ = F(10.86, 901.19) | | | $p = 0.4256$ | |

## Appendix 3: Analysis of Heterogeneity

To get a sense of what types of respondents are more likely to make errors in responding to the ICT and to the high frequency treatment in particular, we examined whether the treatment effects varied by subgroup using a series of OLS regressions predicting the number of reported items. Variables indicating assignment to the two treatment groups, basic demographic variables (gender, age, education, income),[19] the degree of disengagement with the survey (number of missing values in the instrument), and interactions between the treatment variables and the explanatory variables were included in the analysis.[20] The results for Uruguay and Honduras are reported in Tables 4 and 5 respectively. Coefficients on the terms interacted with each of the treatments (i.e., low propensity behavior treatment list and high propensity behavior treatment list) provide evidence of heterogeneity in response to the treatment. Given the null finding for the low-incidence treatment and the lack of variance on this item, we expect substantially less heterogeneity in the response to the low-incidence treatment.

For the Uruguayan sample, gender and age are unrelated to either artificial inflation ($p = 0.74$ and 0.48, respectively) or artificial deflation biases ($p = 0.49$ and 0.78, respectively). There is a positive effect of education in the low incidence treatment list that reaches marginal levels of statistical significance ($p = 0.09$). If replicated elsewhere, this finding would suggest that greater education increases the likelihood of artificial inflation, contrary to expectation that those with less education would be more likely to inflate their responses. Since education is positively correlated with participating in politics (the substance of the list), it is possible that social norms in relatively educated social circles cause people to provide a higher number in response to a longer list. While the interactions with the three-point income scale are not significant ($p = 0.745$ and 0.12), the interaction with a dummy variable indicating missing data for the income item (due to refusals or don't knows)[21] is highly significant for the test high condition ($p = 0.03$),

[19] The operationalization of all variables is outlined in Appendix 2.

[20] To ease the interpretation of the treatment variables and constants in these models, the explanatory variables were centered at their medians (except gender). Thus, in the models including other explanatory variables, the coefficients for the uninteracted treatment variables (i.e. Test Low and Test High) reflect the average treatment effect (or ICT estimate) for the median respondent. The constant can be interpreted as the average number of items indicated by the median respondent to the control list.

[21] Approximately 18 % of respondents did not answer the income item by either saying don't know or refusing. To account for this large proportion, the income scale runs from 0 to 3, with missing data coded as zero, with a dummy variable and corresponding treatment interactions included test differences between those who responded and did not.

**Table 4** Uruguay demographic interactions

| | Baseline | Gender | Age | Education | Income | Disengage | All |
|---|---|---|---|---|---|---|---|
| Treat low | 0.081 | −0.015 | 0.096 | 0.173* | 0.106 | 0.096 | 0.162 |
| | (0.088) | (0.316) | (0.094) | (0.094) | (0.104) | (0.084) | (0.324) |
| TL × female | | 0.062 | | | | | 0.017 |
| | | (0.186) | | | | | (0.184) |
| TL × age | | | −0.082 | | | | −0.002 |
| | | | (0.117) | | | | (0.122) |
| TL × education | | | | 0.173* | | | 0.173 |
| | | | | (0.101) | | | (0.11) |
| TL × income | | | | | 0.042 | | −0.036 |
| | | | | | (0.131) | | (0.128) |
| TL × income missing | | | | | −0.011 | | −0.142 |
| | | | | | (0.349) | | (0.344) |
| TL × disengagement | | | | | | −0.036 | −0.006 |
| | | | | | | (0.051) | (0.055) |
| Treat high | 0.371*** | 0.6 | 0.380*** | 0.444*** | 0.323*** | 0.406*** | 0.571 |
| | (0.079) | (0.354) | (0.085) | (0.086) | (0.092) | (0.076) | (0.358) |
| TH × female | | −0.146 | | | | | −0.129 |
| | | (0.21) | | | | | (0.202) |
| TH × age | | | −0.036 | | | | 0.06 |
| | | | (0.126) | | | | (0.126) |
| TH × education | | | | 0.09 | | | 0.066 |
| | | | | (0.097) | | | (0.106) |
| TH × income | | | | | 0.211 | | 0.179 |
| | | | | | (0.134) | | (0.134) |
| TH × income missing | | | | | 0.756** | | 0.800** |
| | | | | | (0.341) | | (0.338) |
| TH × disengagement | | | | | | −0.024 | −0.024 |
| | | | | | | (0.05) | (0.053) |
| Female | | −0.069 | | | | | −0.077 |
| | | (0.136) | | | | | (0.132) |
| Age | | | −0.033 | | | | 0.017 |
| | | | (0.088) | | | | (0.089) |
| Education | | | | 0.150** | | | 0.083 |
| | | | | (0.073) | | | (0.082) |
| Income | | | | | 0.09 | | 0.021 |
| | | | | | (0.092) | | (0.088) |
| Missing income | | | | | −0.019 | | −0.199 |
| | | | | | (0.255) | | (0.245) |
| Disengagement | | | | | | −0.140*** | −0.135*** |
| | | | | | | (0.038) | (0.042) |
| Constant | 0.692*** | 0.798*** | 0.697*** | 0.720*** | 0.725*** | 0.675*** | 0.852*** |
| | (0.075) | (0.241) | (0.082) | (0.075) | (0.089) | (0.072) | (0.244) |

**Table 4** continued

| | Baseline | Gender | Age | Education | Income | Disengage | All |
|---|---|---|---|---|---|---|---|
| Observations | 871 | 871 | 871 | 870 | 871 | 871 | 870 |

Estimates are from OLS regressions. Standard errors that take into account the clustered sample design are in parenthesis. Stars reflect whether coefficients are statistically different from 0, except for the non-interacted "Treat High" variables, which reflect whether the coefficients are statistically distinct from 1. Independent variables (excluding dummy variables) are median centered

All significance tests are two tailed

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

**Table 5** Honduras demographic interactions

| | Baseline | Gender | Age | Education | Income | Disengage | All |
|---|---|---|---|---|---|---|---|
| Test low | 0.001 | −0.063 | −0.003 | −0.007 | −0.01 | 0.02 | −0.09 |
| | (0.119) | (0.159) | (0.119) | (0.136) | (0.118) | (0.137) | (0.176) |
| TL × female | | 0.127 | | | | | 0.132 |
| | | (0.207) | | | | | (0.203) |
| TL × age | | | −0.071 | | | | −0.059 |
| | | | (0.116) | | | | (0.12) |
| TL × education | | | | 0.026 | | | 0.071 |
| | | | | (0.124) | | | (0.141) |
| TL × income | | | | | −0.224 | | −0.251 |
| | | | | | (0.158) | | (0.163) |
| TL × income N/A | | | | | −0.38 | | −0.445 |
| | | | | | (0.772) | | (0.763) |
| TL × disengagement | | | | | | −0.022 | −0.020 |
| | | | | | | (0.09) | (0.089) |
| Test high | 0.587*** | 0.720 | 0.586*** | 0.528*** | 0.562*** | 0.636** | 0.674 |
| | (0.123) | (0.16) | (0.122) | (0.134) | (0.118) | (0.138) | (0.178) |
| TH × female | | −0.265 | | | | | −0.243 |
| | | (0.217) | | | | | (0.212) |
| TH × age | | | 0.012 | | | | 0.037 |
| | | | (0.136) | | | | (0.142) |
| TH × education | | | | 0.118 | | | 0.12 |
| | | | | (0.118) | | | (0.133) |
| TH × income | | | | | −0.013 | | −0.059 |
| | | | | | (0.161) | | (0.176) |
| TH × income N/A | | | | | 1.112 | | 0.964 |
| | | | | | (0.736) | | (0.719) |
| TH × disengagement | | | | | | −0.042 | −0.043 |
| | | | | | | (0.086) | (0.083) |
| Female | | −0.129 | | | | | −0.103 |
| | | (0.116) | | | | | (0.116) |
| Age | | | 0.016 | | | | 0.02 |
| | | | (0.087) | | | | (0.093) |

**Table 5** continued

| | Baseline | Gender | Age | Education | Income | Disengage | All |
|---|---|---|---|---|---|---|---|
| Education | | | | 0.018 | | | −0.02 |
| | | | | (0.078) | | | (0.09) |
| Income | | | | | 0.224** | | 0.249** |
| | | | | | (0.111) | | (0.117) |
| Income N/A | | | | | 0.337 | | 0.357 |
| | | | | | (0.669) | | (0.648) |
| Disengagement | | | | | | −0.169*** | −0.175*** |
| | | | | | | (0.059) | (0.06) |
| Constant | 1.898*** | 1.963*** | 1.899*** | 1.890*** | 1.910*** | 2.030*** | 2.111*** |
| | (0.082) | (0.107) | (0.082) | (0.095) | (0.079) | (0.1) | (0.123) |
| Observations | 979 | 979 | 979 | 979 | 979 | 979 | 979 |

Estimates are from OLS regressions. Standard errors that take into account the clustered sample design are in parenthesis. Stars reflect whether coefficients are statistically different from 0, except for the non-interacted "Treat High" variables, which reflect whether the coefficients are statistically distinct from 1. Independent variables (excluding dummy variables) are median centered. All significance tests are two tailed

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

suggesting that those who did not answer the income item were much more likely to count the election awareness item than were those who provided a valid answer on the income scale.[22] Not responding to the income question could reveal the respondent to be the type of person unlikely to reveal sensitive information. The fact that they are substantially more accurate than people who provided income information may suggest that the ICT works as desired for this population, but then one is left with the question about why people who provided income information were not very accurate in providing the number of items. It should be noted, however, that including all the control variables does not appreciably change the estimated magnitude of these detected biases for the educated and income non-responders and only slightly alters the statistical significance (see Table 5, column "All"). Given the standard errors associated with these estimates, one should avoid rushing to too many conclusions about populations for which the ICT is valid based solely on these findings.[23]

To highlight this point, the Honduran sample exhibits no treatment heterogeneity across demographic variables. Although the coefficients on the income variables hint that low-income respondents were more likely to artificially inflate their responses and high-income non-respondents were less likely to deflate theirs, the

[22] Although the relationship does not reach conventional levels of statistical significance, the coefficient of the main test high treatment variable in conjunction with the coefficient of the income test high interaction suggests that those who did not answer the income question were more accurate than were those with the lowest incomes ($p = 0.15$, Wald test).

[23] The robustness of these results is further called into question when removing those respondents who said that they were candidates for office and those who said that they were not aware that the elections were taking place. The only interaction that retains marginal significance with this analysis is the income non-response indicator ($p = 0.08$).

**Table 6** Maximum likelihood estimates

| | Uruguay | | | Honduras | | |
|---|---|---|---|---|---|---|
| | treat low | Treat high | Control list | Treat low | Treat high | Control list |
| Intercept | −1.624 | −0.929 | −1.187*** | −5.745** | 0.435 | −0.048 |
| | (1.139) | (0.781) | (0.178) | (2.391) | (0.379) | (0.058) |
| Female | −0.039 | −0.025 | −0.226** | 1.324 | −0.814* | −0.124 |
| | (0.708) | (0.452) | (0.110) | (1.433) | (0.471) | (0.078) |
| Age | −0.013 | 0.450 | 0.010 | 1.246 | 0.208 | −0.030 |
| | (0.449) | (0.327) | (0.073) | (1.434) | (0.301) | (0.054) |
| Education | 0.950* | 0.354 | 0.207*** | 1.829 | 0.398 | −0.019 |
| | (0.532) | (0.283) | (0.077) | (1.292) | (0.368) | (0.058) |
| Income | −0.145 | 0.810* | 0.041 | −3.894* | −0.087 | 0.198*** |
| | (0.471) | (0.441) | (0.088) | (2.018) | (0.369) | (0.064) |
| Income N/A | −1.235 | 3.224*** | −0.375 | | | |
| | (1.548) | (1.115) | (0.240) | | | |
| Disengagement | −0.317 | −0.027 | −0.299*** | | | |
| | (0.330) | (0.136) | (0.045) | | | |
| Observations | 871 | | | 961 | | |
| Log likelihood | −1,105.153 | | | −1,535.51 | | |

Estimates are from the maximum likelihood procedure described in Blair and Imai (2012). Standard errors are in parentheses. Significance stars are based on two-tailed tests

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

coefficients are not close to significant ($p = 0.16, 0.13$). Gender, age, and education similarly demonstrate no significant effects. Intriguingly, the interaction between gender and the two treatment lists are very similar for both the Uruguay and Honduras samples (positive for the low-incidence behavior and negative for the high-incidence behavior). However, even averaging these estimates the results never come close to the approaching traditional levels of statistical significance (test low combined $p = 0.51$, test high combined $p = 0.18$). So it is possible that the artificial inflation and deflation hypotheses are truer for women than men, but testing that proposition would require much larger datasets. In short, Honduras fails to replicate the interesting education effect found in Uruguay. Thus, both surveys show strong evidence of artificial deflation, but evidence of heterogeneity in exhibition of this bias is weak.

It is also interesting to note that the proxy for survey disengagement is not a moderator for either treatment in either country. If attentiveness to the survey were the cause of artificial deflation, one would expect the interaction between the “treat high” variable and “disengagement” to be highly negative. Instead, the model estimates values that are substantively and statistically indistinguishable from zero. Thus, survey disengagement is not driving the results. Finally, at the suggestion of a reviewer, we also conducted likelihood ratio tests to determine whether the inclusion of the treatment interacted variables improved upon the models only including the non-interacted variables. In neither case does the addition of the interacted variables improve model fit (Uruguay $p = 0.37$; Honduras $p = 0.40$).[24]

As a robustness check, we also estimated the treatment heterogeneity models using the maximum likelihood technique developed by Blair and Imai (2012), which has the potential to increase the efficiency of the estimates (Table 6). Due to convergence difficulties, we had to exclude respondents who did not answer the income item as well as the survey disengagement variable in the Honduras analysis. The only difference for the Uruguay analysis is that income becomes marginally significant for the treat high item. For Honduras, poorer respondents are more likely to inflate responses (treat low) and females are somewhat more likely to report deflated responses (treat high), but both of these effects reach only marginal levels of significance. Given these small differences, the results of this analysis do not change the overall conclusions from the OLS analysis.

Taken together, the analysis demonstrates that any heterogeneity in potential counting errors across demographic subgroups is not robust. If confirmed by other studies, these findings suggest that while ICT estimates are likely to be biased downward, multivariate estimates of treatment item predictors are not likely to be biased since the overall deflationary bias does not vary substantially across subgroups.

## References

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

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

Biemer, P., Kathleen Jordan, B., Hubbard, M., & Wright, D. (2005). A test of the item count methodology for estimating cocaine use prevalence. In J. Kenneth & J. Gfroerer (Eds.), *Evaluating and improving methods used in the national survey on drug use and health*. Rockville: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis, 20*, 47–77.

Bless, H., Bohner, G., Hild, T., & Schwarz, N. (1992). Asking difficult questions: Task complexity increases the impact of response alternatives. *European Journal of Social Psychology, 22*, 309–312.

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

Coutts, E and Jann B. (2008). Sensitive Questions in Online Surveys: Experimental results for the randomized response technique (RRT) and the item count technique (UCT). *ETH Zurich Sociology Working Paper No. 3.* ETH Zurich.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology, 47*(4), 817.

Díaz Cayeros, A., Magaloni, B., Matanock, A., & Romero, V. (2011). Living in fear: Mapping the social embeddedness of drug gangs and violence in Mexico. doi:10.2139/ssrn.1963836.

[24] Technically, likelihood ratio tests are inappropriate with clustered survey data. A Wald test produced similar non- significant results for Uruguay, while a Wald test suggested somewhat better model fit for the Honduras interactions, although none proved significant in the analysis.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. *Measurement errors in surveys*, 185–210.

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

Harkness, J., & Van de Vijver, F. (2003). *Cross-cultural survey methods*. Hoboken: Wiley.

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

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

Hubbard, M.L., Casper, R.A., Lessler, J.T. (1989). Respondent reactions to item count lists and randomized response. Proceedings of the Survey Research Section of the American Statistical Association, pp. 544–548.

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association, 106*(494), 407–416.

Jackman, S. (2007). The social desirability of belief in god. Presentation for the Boston area methods meeting, March 2007.

Johnson, T., & Van de Vijver, F. (2003). Social desirability bias in cross-cultural research. In J. Harkness, F. Van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). Hoboken: Wiley.

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

Karlan, D., & Zinman, J. (2012). List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Development Economics, 98*, 71–75.

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

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997a). Racial attitudes and the new South. *The Journal of Politics, 59*(2), 323–349.

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

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

Malesky, E., Jensen, N., & Gueorguiev, D. (2011). "Rent(s) asunder: Sectoral rent extraction possibilities and bribery by Multinational Corporations. *Working Paper Series*, Peterson Institute for International Economics.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22*(2), 212–228.

Miller, J.D. (1984). A new survey technique for studying deviant behavior. Ph.D. thesis. Washington, DC: George Washington University.

Miller, J. D., & Cisin, I. H. (1984). *The item-count/paired lists technique: An indirect method of surveying deviant behavior*. Washington, DC: George Washington University, Social Research Group.

Ocantos, G., Ezequiel, C. K., de Jonge, C., Meléndez, J. O., & Nickerson, D. W. (2012). Vote buying and social desirability bias: Experimental evidence from Nicaragua. *American Journal of Political Science, 56*(1), 202–217.

Schwarz, N., & Bienias, J. (1990). What mediates the impact of response alternatives on frequency reports of mundane behaviors? *Applied Cognitive Psychology, 4*, 61–72.

Schwarz, N., & Scheuring, B. (1988). Judgments of relationship satisfaction: Inter- and intra-individual comparison strategies as a function of questionnaire structure. *European Journal of Social Psychology, 18*, 485–496.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.

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

Sudman, Seymour. (1966). Probability sampling with quotas. *Journal of the American Statistical Association, 61*(315), 749–771.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., & Smith, T. (1996). Asking sensitive questions: The impact of data collection, question format, and question context. *Public Opinion Quarterly, 60*, 275–304.

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

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

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

Weghorst, K. (2010). Uncovered sensitive political attitudes with list experiments and randomized response technique: A survey experiment assessing data quality in Tanzania. Presented at the 2010 Midwest Political Science Association National Conference.

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

Zimmerman, R. S., & Langer, L. M. (1995). Improving estimates of prevalence rates of sensitive behaviors: The randomized lists technique and consideration of self-reported honesty. *The Journal of Sex Research, 32*(2), 107–117.