

Solutions to Turnout Over-Reporting:  
What Is Out There, What Works, and Can We Do Better?

Chi-lin Tsai

A thesis submitted for the degree of PhD

Department of Government

University of Essex

Date of submission: 3 October 2016



## **Summary**

Valid measurement of voter turnout is crucial to electoral studies. One major problem in obtaining valid turnout measurements is over-reporting, i.e. survey respondents who did not vote report having voted. Aiming to identify effective solutions to turnout over-reporting, this doctoral thesis consists of four separate but interrelated papers, plus introductory and concluding chapters. The introductory chapter reviews the causes and consequences of turnout over-reporting, providing the basis for an in-depth research into solutions. Each of the papers then addresses a question about solutions. Paper 1 critically re-examines an influential study of turnout over-reporting. The examination results highlight the need for better solutions to over-reporting. Addressing the question of “What is out there?”, Paper 2 conducts a meta-analysis of studies that have experimented on innovative solutions to turnout over-reporting. Addressing the question of “What works?”, Paper 3 experimentally compares two promising solutions – item-count and pipeline techniques – and finds that the former is, overall, better than the latter for preventing turnout over-reporting. Addressing the question of “Can we do better?”, Paper 4 improves the design and analysis of the item-count technique, making it an even better solution to turnout over-reporting. From the results of these research papers, the concluding chapter considers the implications for developing effective solutions to turnout over-reporting, and laying the foundations for future advances in the

## *Summary*

measurement of turnout. Furthermore, the concluding chapter also discusses how the results of this doctoral research can contribute beyond election studies, towards scientific studies on a wide range of topics on which people often misreport.

## Contents

<b>Introduction .....</b>	<b>1</b>
Problems in measuring turnout .....	2
Turnout over-reporting .....	7
1. Sensitive questions and misreporting .....	7
2. Causes of turnout over-reporting .....	10
3. Mechanisms for turnout over-reporting .....	13
4. Variations in turnout over-reporting .....	15
5. Consequences of turnout over-reporting .....	16
Roadmap of the thesis .....	18
Supplementary materials .....	21
References .....	21
<b>Do Turnout Over-Reporters Almost Never Vote? .....</b>	<b>29</b>
Introduction .....	30
Typical case or anomalous instance? .....	31
1. Re-examination of the 1976 ANES .....	32
2. Comparison of ANESs .....	34
Ineligible electors or misclassified voters? .....	36
1. Corroborative evidence based on political knowledge .....	37
2. Projections based on other ANESs .....	38
Discussion and conclusion .....	40
Supplementary materials .....	43
Appendix .....	43
1. Data .....	43
2. Variables .....	43
3. Coding .....	45
4. Simulation .....	56
References .....	58
<b>A Meta-Analysis of Solutions to Turnout Over-Reporting.....</b>	<b>61</b>
Introduction .....	62
Data and methods .....	63
Solutions to turnout over-reporting .....	67

1. Question wording.....	68
2. Response options .....	76
3. Questioning format .....	79
4. Question order.....	81
5. Survey mode .....	88
Overall effects of the solutions .....	92
Discussion and conclusion .....	95
Supplementary materials.....	99
References.....	99
<b>A Comparison between Two Solutions to Turnout Over-Reporting.....</b>	<b>105</b>
Introduction.....	106
Development of the techniques.....	109
1. Item-count techniques .....	109
2. Pipeline techniques .....	111
3. ICT vs PLT .....	114
Experiment design.....	118
Randomisation checks.....	122
Assessment criterion .....	122
Experiment results.....	123
1. Comparing aggregate turnout estimates.....	123
2. Investigating the design of PLT .....	124
3. Assessing the cost of ICT .....	125
4. Evaluating the impact of non-key items .....	127
5. Examining mechanisms underlying turnout over-reporting .....	129
Conclusion .....	131
Supplementary materials.....	134
Appendix.....	134
1. Experiment administration.....	134
2. Meta-questions .....	135
3. Random assignment .....	137
4. ICT diagnostics .....	141
5. Impact of the drop-outs .....	143
References.....	145

<b>Not Just Valid but Precise.....</b>	<b>149</b>
Introduction.....	150
Estimators for the standard item-count technique.....	152
Estimation with auxiliary information .....	158
1. Derivation of the likelihood function.....	158
2. Selection of the auxiliary variable .....	160
3. Properties of TML.....	162
4. Relationship with other methods.....	164
Simulation studies .....	165
1. General settings.....	166
2. Simulations without covariates .....	167
3. Simulations with covariates .....	170
Comparison based on real data .....	171
Limitation.....	175
Conclusion .....	176
Supplementary materials.....	177
Appendix.....	177
1. Notation.....	177
2. Log-likelihood function .....	178
3. Expectation-maximisation algorithm.....	178
4. Properties .....	180
5. Connections.....	183
6. Simulation data .....	184
References.....	188
<b>Conclusion .....</b>	<b>191</b>
Summary and discussion of research findings.....	191
Implications and recommendations for future research.....	194
Appendix.....	200
1. A proposal to experiment with a solution to turnout over-reporting .....	200
2. Improvements to the measurement of turnout intention .....	204
References.....	205

## Figures

<b>Introduction .....</b>	<b>1</b>
Figure 1. Mapping problems onto the total survey error framework.....	4
Figure 2. Prejections for turnout overestimation due to over-reporting .....	5
Figure 3. Public perceptions of voting as a civic duty .....	10
Figure 4. Impression management as a mechanism for turnout over-reporting .....	15
<b>Do Turnout Over-Reporters Almost Never Vote? .....</b>	<b>29</b>
Figure 1. Demonstration of the anomalousness in the 1976 ANES.....	35
Figure 2. Assessment of the possibility of validation misclassification .....	37
Figure 3. Simulation of who over-reported in the 1976 ANES .....	39
<b>A Meta-Analysis of Solutions to Turnout Over-Reporting.....</b>	<b>61</b>
Figure 1. Catalogue of solutions to turnout over-reporting .....	68
<b>A Comparison between Two Solutions to Turnout Over-Reporting.....</b>	<b>105</b>
Figure 1. Experiment design .....	118
Figure 2. Turnout estimates by techniques and waves.....	124
Figure 3. Median response time .....	127
Figure 4. Relative importance of over-reporting mechanisms.....	131
Figure A1. Weighted distribution of personal income by groups in Wave 1 .....	140
Figure A2. Turnout estimates based on the complete and drop-out data.....	144
<b>Not Just Valid but Precise.....</b>	<b>149</b>
Figure 1. Simulation results without covariates.....	168
Figure 2. Simulation results with covariates .....	171

## Tables

<b>Introduction .....</b>	<b>1</b>
Table 1. Bias of turnout over-reporting .....	18
<b>Do Turnout Over-Reporters Almost Never Vote? .....</b>	<b>29</b>
Table 1. Summary of the results of the 1977 validation exercise .....	32
Table 2. Re-Examination of who over-reported in the 1976 ANES .....	34
Table A1. Information on data .....	43
<b>A Meta-Analysis of Solutions to Turnout Over-Reporting.....</b>	<b>61</b>
Table 1. Experimental studies of solutions to turnout over-reporting .....	64
Table 2. Effects of solutions to turnout over-reporting.....	93
<b>A Comparison between Two Solutions to Turnout Over-Reporting.....</b>	<b>105</b>
Table 1. Comparisons between ICT and PLT .....	117
Table 2. Differences in turnout estimates across techniques and waves .....	124
Table 3. Decomposition of turnout estimates .....	130
Table A1. Sample sizes .....	135
Table A2. Randomisation checks.....	138
Table A2. Estimated probabilities of ICT responses .....	142
<b>Not Just Valid but Precise.....</b>	<b>149</b>
Table 1. Classification of respondents by the observed variables $T_i$ and $Y_i$ .....	155
Table 2. Likelihood functions of IML for different types of respondents .....	156
Table 3. Likelihood functions of TML for different types of respondents .....	159
Table 4. Likelihood function of TML with a perfect auxiliary variable.....	163
Table 5. Comparison between IML and TML .....	174
Table A1. Variables for simulations .....	185



## INTRODUCTION

---

CHI-LIN TSAI

A sample survey is an important means of data collection for social science. It opens up avenues for attitudinal and behavioural studies, where respondents' self-reports are often the only data obtainable. Even when alternative means are available (e.g. census), a sample survey is still often preferable, because it is fast, flexible and cost-effective. The widespread use of survey data in social science renders the validity of survey measurement a matter of the utmost importance.

In political science, citizens' participation in elections has always received close attention. From 2001 to 2010, nearly one hundred papers were published in top journals addressing questions about turnout, and more than 80% of these papers were based on survey data (Smets and Ham 2013: 348). Studies that do not directly concern turnout also need survey respondents' self-reports to identify who voted and who did not, so as to address questions of greater interest, such as voters' party/candidate choices, or non-voters' possible impacts on election results. Our understanding of elections is largely founded on surveys and rests heavily on the validity of self-reported turnout.

Nonetheless, survey data are not error-free. It has long been known that surveys over-estimate turnout rates. Even with rigorously designed surveys, e.g. the American National Election Study (ANES), the British Election Study (BES) and Taiwan's Election and Democratisation Study (TEDS), it is not uncommon for estimates to be higher than official turnout rates by double digits (McDonald 2003: 181; Swaddle and Heath 1987: 539; Wu 2006: 233). This issue raises concerns about the reliability of

survey data on turnout, thus posing a threat to electoral studies.

There are several reasons why surveys over-estimate turnout rates; one important reason is '*turnout over-reporting*', i.e. *respondents did not vote but report having done so*. This doctoral thesis comprises four papers that study solutions to turnout over-reporting from different aspects, aiming to improve survey measurement of turnout, and thus to solidify the foundations for electoral studies. Furthermore, though focusing on turnout, the thesis also aims to offer a frame of reference for the measurement of other sensitive issues on which respondents often misreport (e.g. drug use, alcohol abuse), and hence to contribute towards research beyond political science.

This Introduction provides an overview of turnout over-reporting and of the four papers that constitute the thesis. It begins by explaining the difference between over-reporting and other problems with survey measurement of turnout. It then proceeds to discuss the causes and consequences of turnout over-reporting, providing the basis for an in-depth study of solutions. The Introduction concludes with an outline of the structure of the thesis, summarising the contributions of each of the four papers.

## **Problems in measuring turnout**

The quality of survey data on turnout is a function of several factors. Though this thesis focuses on over-reporting, it is worth reviewing other significant problems in measuring turnout by survey, in order to have an overall picture in mind before further discussion about over-reporting. Those problems are the Hawthorne effect, denominator difference and voter over-representation. Like over-reporting, they cause surveys to over-estimate turnout rates, raising concerns about data quality (Burden 2000; Clausen 1968: 590–594; Health and Taylor 1999; McDonald and Popkin 2001; Traugott and Katosh 1979).

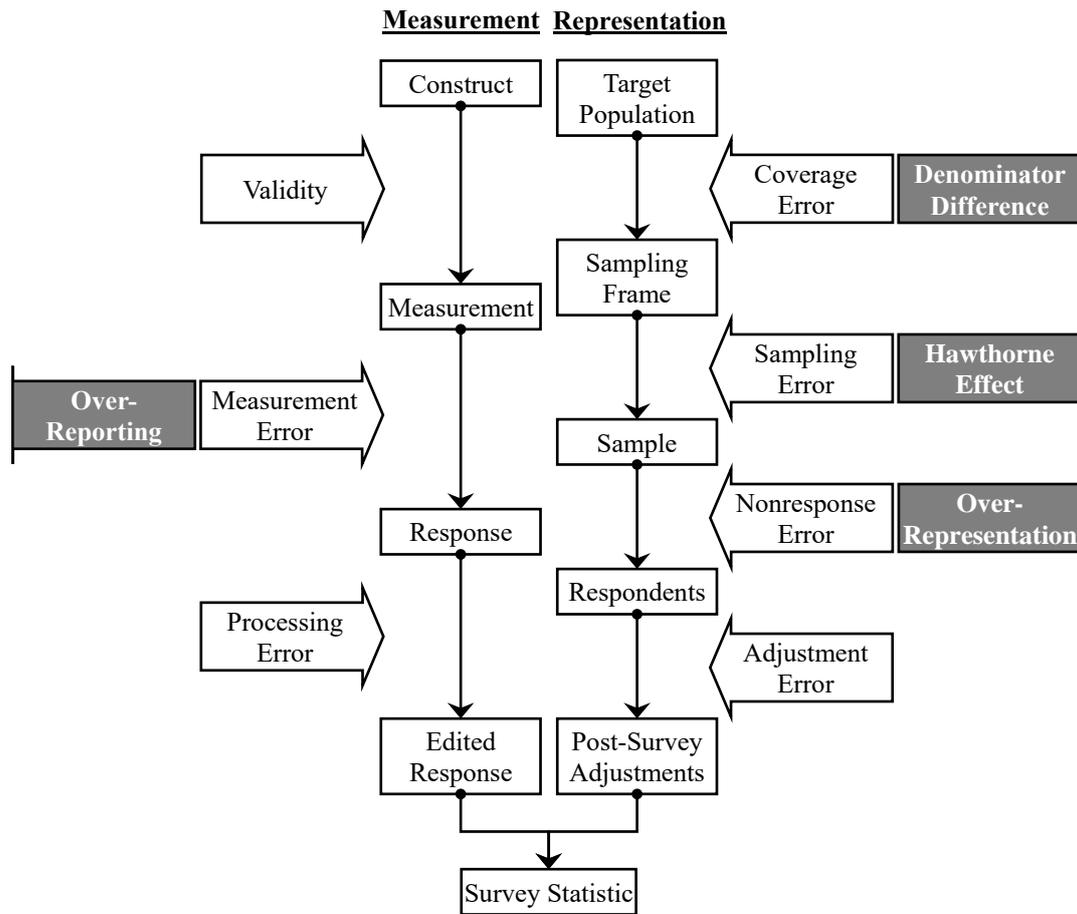
The Hawthorne effect refers to the tendency for individuals to change their behaviour in reaction to their awareness of being observed (James and Vo 2010: 561). This tendency causes respondents who were interviewed in pre-election surveys to vote at higher rates than they would otherwise have done (Kraut and McConahay 1973; Yalch 1976). As a result, post-election surveys that involve re-interviewing those respondents over-estimate turnout rates.

Secondly, the denominator in turnout calculation matters as well. For example, the U.S. Bureau of the Census uses the population aged 18 or older as its denominator to calculate turnout rates, whereas the sampling frame of the ANES disregards not only Americans younger than 18, but also adults who are ineligible to vote (McDonald 2003: 180). Since the ANES calculates turnout rates using a smaller denominator, a difference between its estimates and official turnout rates is only to be expected.<sup>1</sup>

Thirdly, there is a positive correlation between a willingness to vote and a willingness to participate in electoral surveys (Voogt and Saris 2003). Voters therefore tend to be over-represented in samples (and equivalently, non-voters are under-represented), which consequently causes surveys to over-estimate turnout rates (Tsai 2010).

<sup>1</sup> The ANES puts extra effort into filtering out ineligible electors, whereas the U.S. Bureau of Census does not, and so, in this example, *ceteris paribus*, survey estimates should be more accurate than official statistics. Nonetheless, this is not always the case. In countries where the government knows the exact size of the electorate (e.g. Taiwan), the denominators used in the calculation of official turnout rates are impeccable. Surveys that are unable to organise their sampling frames in accordance with government-defined denominators produce inaccurate estimates of turnout rates. For this reason, the denominator difference has been seen as a problem for surveys in measuring turnout.

**Figure 1. Mapping problems onto the total survey error framework**



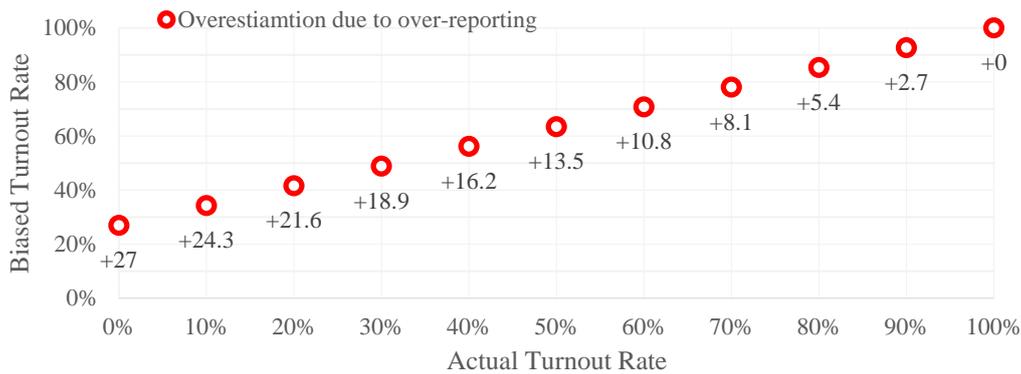
NOTE: This figure is adapted from Groves et al. (2004: 48). See pages 39–65 of their book for definitions of each term in the figure.

Together with over-reporting, these are four major problems in survey measurement of turnout. Each of them relates to a source of survey errors, as shown in Figure 1.<sup>2</sup> The denominator difference refers to the discrepancy between the sampling frame and the target population, hence relating to coverage error. The Hawthorne effect

<sup>2</sup> The term ‘error’ refers to “deviations of what is desired in the survey process from what is attained”. (Groves et al. 2004: 40) In practice, errors are inevitable. For example, due to the nature of random sampling, the composition of a survey sample is never identical to that of the target population; some surveys over-sample a certain type of individuals, whereas others under-sample them. However, as long as errors are not systematic (i.e. not always over-sampling or under-sampling that type of individuals), errors tend to cancel each other out in the long run, and so, theoretically, survey statistics remain ‘unbiased’. In contrast, the term ‘bias’ refers to a systematic error that renders statistics misleading from a long-term perspective. The four problems in measuring turnout are systematic errors, resulting in an upward bias in survey statistics.

may be categorised generally as a sampling error, since it spoils a sampling design that would otherwise produce a representative sample.<sup>3</sup> Voter over-representation results from non-voters' relative unwillingness to participate in surveys, so it is a kind of nonresponse error (specifically, unit nonresponse, i.e. sampled units do not respond to any part of the questionnaire for reasons such as a refusal to participate in a survey).

**Figure 2. Projections for turnout overestimation due to over-reporting**



**NOTE:** This figure assumes the proportion of over-reporters among actual non-voters is 27%.

In contrast to the three problems above that relate to the representational dimension of a survey, over-reporting relates to the measurement dimension, falling precisely within the definition of measurement error: “A departure from the true value of the measurement as applied to a sample unit and the value provided” (Groves et al. 2004: 52).<sup>4</sup> Analysing 49 surveys from six countries, Selb and Munzert (2013: 190) found that, on average, 27% of non-voters over-reported their turnout, and this proportion ranged from 10% to 68%. Based on their estimate, Figure 2 shows rough

<sup>3</sup> Ideally, a proper sampling design should allow a post-election survey to re-interview participants in pre-election surveys, and then combine them with new participants to form a representative sample. However, due to the Hawthorne effect, those re-interviewed vote at higher rates than they would have done otherwise, which spoils the sampling design, and hence sample representativeness. In this regard, I consider the Hawthorne effect to be a sampling error.

<sup>4</sup> Though relating to different survey errors, the four problems in measuring turnout are not unrelated. For example, Tourangeau, Groves, and Redline (2010: 431) found a strong link between nonresponse errors and measurement errors, so they conclude that “non-voters were both less likely to take part in the survey and more likely to misreport if they did take part”.

projections for turnout overestimation due to over-reporting. For an election where 50% of electors voted, over-reporting can cause a survey to over-estimate the turnout rate by 13.5 percentage points. As will become clear, over-reporting does not occur randomly; some respondents are more likely to over-report than others. Non-random occurrences, together with the appreciable number of over-reporters, make over-reporting a non-negligible problem that needs to be addressed.

All four problems bias turnout-rate estimates upwards, limiting the usefulness of surveys for studies on the aggregate-level turnout. In a comparison, over-reporting is the most worrisome problem among the four, as it also threatens the validity of individual-level studies that rely on surveys to separate voters and non-voters from each other. Despite spoiling sample representativeness, the Hawthorne effect, denominator difference and voter over-representation are of little consequence to measurement quality. Respondents' self-reports of turnout should remain valid for separating voters from non-voters, even when these three problems occur. In contrast, over-reporting leads researchers to misclassify a proportion of non-voters as voters, and consequently to draw invalid conclusions. Unlike aggregate-level studies, which can substitute official turnout rates for biased survey statistics ([Geys 2006](#)), individual-level studies have no alternative but to use respondents' self-reported turnout most of the time. In this regard, though all four problems in measuring turnout are harmful to scientific research, over-reporting is more harmful than the others, and hence in the most urgent need of solutions.

## Turnout over-reporting

The question of turnout, though seemingly innocuous, is in fact sensitive, and hence subject to over-reporting. In this section, I define the concept of sensitivity, explain what makes turnout a sensitive question, discuss why and how people over-report turnout and consider the consequences of turnout over-reporting.

### 1. Sensitive questions and misreporting

Over-reporting, or more generally misreporting, is a major challenge confronting studies that involve asking people sensitive questions. According to [Tourangeau, Rips, and Rasinski \(2000: 257–259\)](#), a question is sensitive if it is intrusive or involves a risk of disclosure. Income, for instance, is a sensitive question in the sense of intrusiveness. People tend to evade or even misreport on questions about income, not necessarily because they are poor or rich, but because income itself is private. A question about such a private topic is sensitive on its own, since it poses a so-called ‘intrusive threat’ to everyone being asked, regardless of their true answers ([Lee and Renzetti 1990](#)). Likewise, questions such as sexual experiences ([Berg and Lien 2006](#)) and physical conditions ([Baldwin, Ginsberg, and Harkaway 2003](#)) are also intrusive, and hence sensitive and susceptible to misreporting.

The second dimension of sensitivity is the risk of disclosure. Sensitivity in this sense is determined not only by how people feel about a question but, more crucially, by what their true answers are. In this regard, a question that is sensitive for some people is unnecessarily sensitive for others. For example, people who have submitted distorted income reports to tax authorities may be more likely to consider income a sensitive question, and feel particularly compelled to misreport on it in surveys, for fear of self-incrimination as a consequence of truth-telling ([Hurst, Li, and Pugsley 2014:](#)

19). This explains why people tend to under-report law-breaking behaviour, e.g. drug use (Morrall, McCaffrey, and Iguchi 2000) or election fraud (Gonzalez-Ocantos et al. 2012), and over-report law-abiding behaviour, e.g. safety-belt use (Li, Kim, and Nitz 1997).

Even though there is no risk of self-incrimination, people may still consider a question sensitive and misreport on it anyway if they worry that their true answers are socially undesirable. This social desirability concern is a special case of the risk of disclosure. It raises fears about a specific consequence of truth-telling – social disapproval (Tourangeau and Yan 2007: 860). Because of that concern, people over-report socially desirable behaviour, e.g. church attendance (Hadaway, Marler, and Chaves 1998), and under-report socially undesirable behaviour, e.g. alcohol consumption (Stockwell et al. 2004), cigarette consumption (Warner 1978), abortion (Jones and Forrest 1992) and arrest experiences (Krohn et al. 2013). People even misreport on seemingly innocuous questions, such as body weight, as long as they believe that there is a social norm regarding an ideal body image (Gil and Mora 2011).

Taken together, a question is sensitive if it evokes feelings of unease about the consequences of truth-telling (intrusive threat, self-incrimination, social disapproval etc.). From the perspective of subjective expected utility theory, truth-telling is therefore a risk-taking behaviour, and so, accordingly, misreporting is a risk-avoiding strategy. This utilitarian perspective gains credibility from experimental studies that have demonstrated a positive correlation between the temptation to misreport and the risk of truth-telling (operationalised by, for example, the degree of response privacy) (Nathan et al. 1990; Rasinski et al. 1994; see Tourangeau and Yan 2007: 877 for a summary). Nonetheless, those experiments did not provide many insights into how respondents to sensitive questions assess the risk of truth-telling and decide whether to

misreport or not.

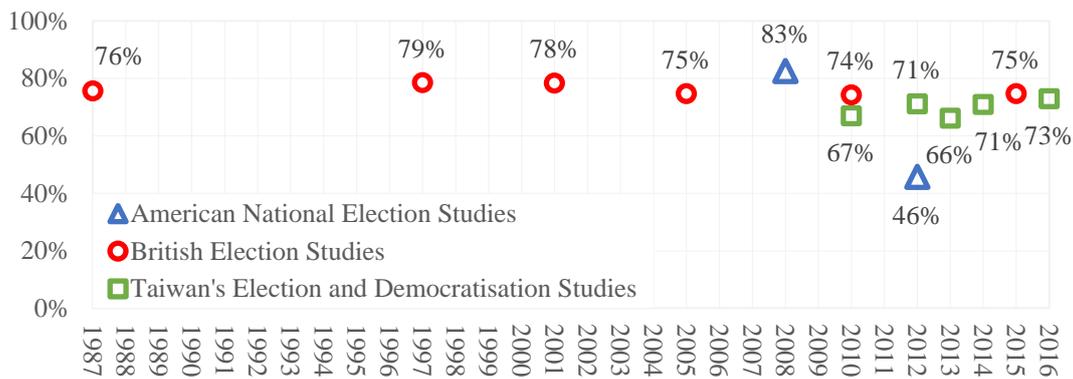
There are four major steps in the question-response process: (a) comprehension of the question, (b) retrieval of relevant information, (c) judgements based on information, and (d) finalising and reporting an answer (Tourangeau, Rips, and Rasinski 2000: 7–8). Schaeffer (2000: 118–120) argues that feelings of unease evoked by sensitive questions (e.g. social desirability concerns) can trigger misreporting in any step of the process under or beyond the respondent's control. In the comprehension step, people may misinterpret sensitive questions in a way that distance themselves from undesirable truths. For example, those who occasionally abuse alcohol may interpret the question “Have you ever consumed an excessive amount of alcohol?” as if it was asking “Do you think of yourself as a chronic alcoholic or not?” By denying chronic alcohol abuse, they distance themselves from the fact of their occasional alcohol abuse. Furthermore, people may be tempted to search for desirable information and ignore undesirable information in the retrieval step, or they may be tempted to attach more weight to desirable information than to undesirable information in the judgement step. Worse still, people may partially or completely skip the retrieval and judgement steps, if they (mis-)believe that sensitive questions do not apply to them at all (e.g. “I could not possibly have anything to do with alcohol abuse!”). Even if people know exactly what their true answers are, they may still decide to misreport in the final step, in order to avoid risks of truth-telling.<sup>5</sup>

<sup>5</sup> Two things are worth clarifying. First, although all of the examples mentioned in this section are about behavioural or factual topics, people do in fact misreport attitudes as well. For example, Kuklinski, Cobb, and Gilens (1997) found that noticeable numbers of white Americans who were hostile towards black people misreported their racial attitudes in surveys, for fear of being accused of racism. Second, the concepts and ideas discussed in this section are applicable to contexts beyond survey interviews. For example, Allen (2007: 624–628) applied the aforementioned utilitarian perspective to the under-reporting of rape, explaining why victims are reluctant to report the crime to the police.

## 2. Causes of turnout over-reporting

“Did you vote or not?” This question seems innocuous, but it actually falls within the definition of a sensitive question, particularly in the sense of disclosure risks (i.e. the second dimension of sensitivity). In countries where electors are obliged to vote, questions about turnout are highly sensitive for those who did not vote, since no one wishes to disclose their unlawful behaviour when it is not absolutely necessary to run the risk of disclosure (e.g. self-incrimination). Turnout over-reporting is therefore a strategy for non-voters to avoid risk.

**Figure 3. Public perceptions of voting as a civic duty**



**NOTE:** Numbers in the figure indicate the percentages of respondents in a survey who considered voting to be a citizen’s duty. TEDS phrases the question as: “Different people have different opinions about voting. Some people think that voting is a responsibility, and you should vote even if you don’t like any candidates or parties. Other people think that it is all right either to vote or not to vote, and the decision depends on how you feel about the candidates or parties. Do you think that voting is a responsibility, or do you think that it is all right to vote or not to vote?” BES: “How much do you agree or disagree with the following statements? ‘It is every citizen’s duty to vote in an election.’” The 2008 ANES: “Generally speaking, do you believe that you have a duty to vote in every national election, or do you believe that you do not have a duty to vote in every national election?” The 2012 ANES: “Different people feel differently about voting. For some, voting is a duty – they feel they should vote in every election no matter how they feel about the candidates and parties. For others voting is a choice – they feel free to vote or not to vote, depending on how they feel about the candidates and parties. For you personally, is voting mainly a duty, mainly a choice, or neither a duty nor a choice?”

In countries where voting is not compulsory, even though self-incrimination is not a matter of concern, turnout is still a sensitive topic to talk about for non-voters, because of their nonconformity to norms of voting. As shown in Figure 3, people from different

counties generally agree that it is every citizen's duty to vote in an election. The act of voting is widely regarded as a characteristic of a good citizen (Clarke et al. 2004: 274). A failure to vote is to neglect a citizen's duty, which is socially undesirable, and hence too sensitive to admit in surveys. For this reason, the literature has used social desirability to describe respondents' motivation for turnout over-reporting (e.g. Belli, Traugott, and Beckmann 2001: 493; Parry and Crossley 1950: 72; Tittle and Hill 1967: 104–105; Weiss 1968: 623).

Traugott (1989: 25–26) criticises social desirability as an oversimplified explanation for turnout over-reporting, because he found that, in an experiment carried out in the 1984 ANES, telephone interviewing did not elicit more accurate responses to the turnout question than did face-to-face interviewing. This criticism is not convincing, because it rests on a questionable assumption that the social desirability concern is lower in telephone interviews than in face-to-face interviews. On the one hand, respondents interviewed by phone do not have to answer sensitive questions in front of an interviewer. This does help to relieve some of the social desirability concern (Groves et al. 2004: 157). However, on the other hand, the absence of an interviewer on the spot also causes difficulties in establishing a rapport with respondents. As suggested by Holbrook, Green, and Krosnick's (2003: 110) review of experiments on survey modes, "Any benefit of increased privacy over the telephone for the accuracy of reports of sensitive topics is less than the advantage of greater rapport developed in face-to-face interviews." In this regard, telephone interviewing is no better than face-to-face interviewing with respect to reduction in the social desirability concern. This flatly contradicts Traugott's assumption, and thus calls into question his criticism.

Nonetheless, Traugott's emphasis on the need for more sophisticated explanations of turnout over-reporting is not misplaced. Refining the social desirability explanation,

Deufel and Kedar (2010: 290–291) hold that the effect of social desirability on turnout over-reporting is not linear but has an inverted U shape. While respondents who have the least social desirability concern seldom over-report their turnout, those who have the greatest social desirability concern have low levels of over-reporting too, because most of them actually turn out to vote. It is the respondents with moderate social desirability concern who over-report at the highest rates, since that moderate concern may be insufficient motivation to vote but sufficient to be a temptation to over-report.

Silver, Anderson, and Abramson (1986) found that highly educated respondents are most inclined to overreport their turnout. Since the highly educated are precisely those who are more likely to be aware of the ‘correct’ or socially approved responses, the researchers consider social desirability a possible explanation for their finding. However, they further argue that highly educated respondents, as a class of high-status people, tend over-report turnout because norms of voting “are [currently] consistent with their class interests”, and they “express their satisfaction with the [this] status quo” by reporting having voted, even when they actually have not (Bernstein, Chadha, and Montjoy 2001: 25–26; Silver, Anderson, and Abramson 1986: 623). In this regard, expression of satisfaction with the status quo is also a plausible cause of turnout over-reporting.

Bernstein, Chadha, and Montjoy (2001) consider feelings of guilt to be a more powerful explanation of turnout over-reporting. They contend that neither social desirability concern nor expression of satisfaction can explain why African-Americans and Latinos are inclined to over-report their turnout – it is unlikely that those ethnic minorities are either more aware of socially desirable responses or more satisfied with the status quo. Instead, African-Americans and Latinos must be more aware of hard-fought struggles for minorities’ suffrage in history, and must more often be told of their

consequent duty to exercise their right to suffrage. For those ethnic minorities, nonvoting is morally unacceptable and can evoke feelings of guilt. In order to avoid confessing that guilt publicly, minorities therefore tend to over-report turnout.

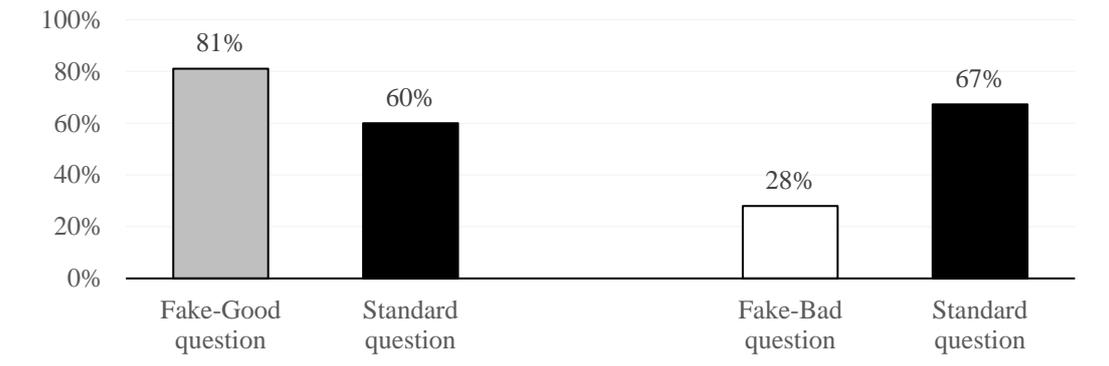
Generalising the feelings-of-guilt explanation further based on [Stryker's \(1980\)](#) identity theory, [Brenner \(2012\)](#) describes turnout over-reporting as the enactment of a political identity. As he elaborates, people take opportunities in their daily lives to validate, affirm and enact their political identities. Voting, for instance, is an opportunity, so is reporting having voted. Non-voters have missed an opportunity in an election; a survey interview offers them a second chance to show their commitments to political identities. Many of them thus seize the chance and over-report their turnout. Brenner argues that feelings of guilt are a form of psychological pressure resulting from a failure to demonstrate an identity commitment. He, therefore, considers enactment of identities to be a more general explanation for turnout over-reporting than other explanations.

### **3. Mechanisms for turnout over-reporting**

As reviewed above, four factors have been identified as causes of turnout over-reporting: social desirability concern, expression of satisfaction with the status quo, feelings of guilt, enactment of political identities. There are two mechanisms whereby these factors lead respondents to over-report turnout: self-deception and impression management ([Paulhus 2002](#); [Stocké 2007: 238–239](#)). Self-deception refers to the distortion of one's self-image towards a desirable criterion. This is mostly an automatic mechanism, activated beyond respondents' control, regardless of whether or not others can observe respondents' answers ([Holtgraves 2004](#)). Self-deception in this context is likely to be shaped by past behaviour. In the case of turnout, respondents who usually

vote are more likely to have voting as part of their self-image, and so are more likely to deceive themselves into thinking that they voted last time around. Cahalan (1968: 621) considers self-deception a possible mechanism for turnout over-reporting, since he found that the accuracy of responses to the turnout question is minimal in relation to interviewer effects.

In contrast, impression management is largely a controlled mechanism – non-voters over-report on the turnout question in surveys, in order to create a positive impression in others. The 2000 ANES pilot study provides some evidence for this mechanism. The study re-interviewed some of the respondents to the 1998 ANES. Half of them were asked a ‘fake-good’ question: “If you were trying to make the *best* impression on me that you could, would you say that you voted in the 1998 Congressional elections?” The rest were asked a ‘fake-bad’ question: “If you were trying to make the *worst* impression on me that you could, would you say that you voted in the 1998 Congressional elections?” As shown in Figure 4, in the 1998 ANES, around two-thirds of those respondents, regardless of group, reported having voted in the election. However, in the pilot study, more than 80% in the fake-good group reported having voted, whereas less than 30% in the fake-bad group reported having done so. These results suggest that respondents do consider misreporting to be a strategy to manage impressions, and impression management is a plausible mechanism for turnout over-reporting.

**Figure 4. Impression management as a mechanism for turnout over-reporting**

**NOTE:** Numbers in the figure are the percentages of respondents who reported having voted in the 1998 American Congressional Election. The standard turnout question was fielded in the 1998 ANES post-election survey, and the wording was: “In talking to people about elections, we often find that a lot of people were not able to vote because they weren’t registered, they were sick, or they just didn’t have time. How about you – did you vote in the elections this November?” The fake-format questions were fielded in the 2000 ANES pilot survey. The fake-good question was: “In talking to people about elections, we often find that a lot of people were not able to vote because they weren’t registered, they were sick, or they just didn’t have time. If you were trying to make the best impression on me that you could, would you say that you voted in the 1998 Congressional elections?” The fake-bad wording replaced “If you were trying to make the best impression on me” with “If you were trying to make the worst impression on me.”

#### 4. Variations in turnout over-reporting

The tendency towards turnout over-reporting varies with contextual conditions and personal attributes. In high-salience elections (e.g. elections where high participation is the norm), the tendency is greater than in low-salience elections, since electoral salience intensifies the effects of factors that cause turnout over-reporting (e.g. increasing the perception that voting is socially desirable) (Karp and Brockington 2005; Belli, Traugott, and Beckmann 2001: 494). Furthermore, the impact of electoral salience on turnout over-reporting is moderated by political awareness. Respondents with higher political awareness have greater ability to distinguish between different levels of electoral salience, so their tendencies towards turnout over-reporting are more sensitive to the impact of electoral salience, varying more greatly between low- and high-salience elections (Górecki 2011a; 2011b).

Memory and time are worthy of discussion as well. Failure of memory alone is not usually considered a primary cause of turnout over-reporting because, if it were, turnout under-reporting should occur as frequently as over-reporting, yet turnout over-reporting in fact occurs far more often than under-reporting (Selb and Munzert 2013: 191). The impact of memory on turnout over-reporting is indirect. When the facts are unambiguous, respondents are basically honest in their responses, because of the drive for honesty (e.g. to comply with norms of honesty, or to avoid psychic costs from lying); when memories fade with time, the drive for honesty become weak and cannot counterbalance the drive for over-reporting (social desirability concern etc.) (Stocké and Stark 2007: 240–241). Hence, the tendency to turnout over-reporting varies according to the time of survey fieldwork: the longer the time the interview is conducted after Election Day, the more frequently turnout over-reporting occurs (Abelson et al. 1992: 151–152; Belli, Traugott, and Beckmann 2001: 497; Belli et al. 1999: 101–105; Cahalan 1968: 609–610).<sup>6</sup>

## 5. Consequences of turnout over-reporting

Turnout over-reporting means, by definition, non-voters reporting having voted, which in consequence leads surveys to over-estimate turnout rates and limits the usefulness of surveys for studies on aggregate-level turnout. For individual-level studies, over-reporting results in bias in statistical modelling of turnout (Mircea and Postelnicu 2013: 169). Bernstein, Chadha, and Montjoy (2001: 22) provide a general rule for assessment of over-reporting bias (for similar intuition, also see Weiss 1968: 626–629):

<sup>6</sup> The impacts of memory and time, though indirect, are by no means negligible. These impacts have been acknowledged not only by studies of turnout over-reporting, but also by studies of misreporting on other topics. Atkeson (1999: 201–207), for example, demonstrates that memory and time (interacting with social-psychological factors) are crucial to survey respondents' over-reporting of the primary vote for party nominees.

*Using reported votes in place of validated votes substantially distorts standard multivariate explanations of voting, increasing the apparent importance of independent variables that are related in the same direction to both over-reporting and voting, and sharply decreasing the apparent importance of independent variables related in opposing directions to those two variables.*

This rule distinguishes between two types of independent variables: ‘relate-in-the-same-direction’ and ‘relate-in-the-opposing-direction.’ As demonstrated in Table 1, when a variable,  $X$ , positively correlates with both turnout and over-reporting ( $\tau = 0.60$ ), surveys exaggerate the relationship between  $X$  and turnout ( $\tau = 0.63$ ). Variables such as education and partisanship are subject to this kind of bias (e.g. [Bernstein, Chadha, and Montjoy 2001: 39](#); [Presser and Traugott 1992: 77](#)). In contrast, if a variable,  $Z$ , negatively correlates with turnout ( $\tau = -0.60$ ) but positively correlates with over-reporting ( $\tau = 0.60$ ), the relationship between  $Z$  and turnout ( $\tau = 0.03$ ) is underestimated when analysis is based on surveys. Race (e.g. blacks vs others) is a variable subjected to bias of this kind (e.g. [Bernstein, Chadha, and Montjoy 2001: 39](#); [Deufel and Kedar 2010](#); [Hill and Hurley 1984: 204](#); [Sigelman 1982: 53](#)).

Both kinds of bias distort our understanding of who votes and why, and hence they have profound political and social consequences for the real world. For example, African-Americans’ turnout is often exaggerated because of over-reporting bias. Racial differences in turnout thus tend to be downplayed, and so is the need for policies and resources to eliminate differences. In contrast, educational differences in turnout are often over-stated because of over-reporting bias. That can mislead policymakers to over-allocate resources for mobilisation of the less education, and consequently, those who are really in need are crowded out.

**Table 1. Bias of turnout over-reporting**

		Relate-in-the-Same-Direction		Relate-in-the -Opposing-Direction	
		Low X	High X	Low Z	High Z
True Non-Voter		40%	10%	10%	40%
True Voter		10%	40%	40%	10%
		$Tau-b = +0.60$		$Tau-b = -0.60$	
&					
		Low X	High X	Low Z	High Z
Truth-Reporter		40%	10%	40%	10%
Over-Reporter		10%	40%	10%	40%
		$Tau-b = +0.60$		$Tau-b = +0.60$	
↓					
		Low X	High X	Low Z	High Z
True Non-Voter	Truth-Reporter	32%	2%	8%	8%
	Over-Reporter	8%	8%	2%	32%
True Voter		10%	40%	40%	10%
↓					
		Low X	High X	Low Z	High Z
	Self-Reported Non-Voter	32%	2%	8%	8%
	Self-Reported Voter	18%	48%	42%	42%
		$Tau-b = +0.63$		$Tau-b = +0.03$	

**NOTE:** Data are artificially generated for illustrative purposes. Education is an example of the ‘relate-in-the-same-direction’ variable. We can consider ‘Low X’ and ‘High X’ as less educated and highly educated, respectively. Though labelled as if they are ordinal variables, X and Z can be also dummy variables. For instance, in the column for ‘Relate-in-the-Opposing-Direction’, we can consider ‘Low Z’ and ‘High Z’ as whites and blacks, respectively.

## Roadmap of the thesis

Obviously, turnout over-reporting matters. Even worse, in spite of Bernstein, Chadha, and Montjoy’s rule, the consequences of turnout over-reporting are still largely variable-specific and model-specific – the use of a given variable in a given model to identify a class of over-reporters has little utility in estimating the general bias of over-reporting (Tittle and Hill 1967: 106; Tsai 2011). An ounce of prevention is worth a pound of cure. This doctoral thesis consists of four papers; each addresses an important question about the prevention of turnout over-reporting. Initially, the first paper highlights the need for methods to prevent turnout over-reporting. The remaining three

papers then address the questions of “*What methods are out there?*”, “*What works?*” and “*Can we do better?*”

### **Paper 1. Do Turnout Over-Reporters Almost Never Vote?**

Knowing who over-reports is crucial to preventing over-reporting. Though most studies have inferred that turnout over-reporters should resemble actual voters more closely than non-voters, [Presser and Traugott’s \(1992\)](#) observation points to an opposite conclusion: over-reporters almost never vote. Re-examining their study, the first paper of this thesis reveals that Presser and Traugott’s finding is lacking in generality, hence it is insufficient to overturn the established view about the resemblance between over-reporters and actual voters. This implies that it is unrealistic to expect a simple method for distinguishing over-reporters from actual voters, let alone for correcting over-reporting bias after it has occurred. Therefore, instead of post-correction, it is prevention that is a fundamental solution to turnout over-reporting.

### **Paper 2. A Meta-Analysis of Solutions to Turnout Over-Reporting**

Over recent decades, survey methodologists have developed various methods for preventing turnout over-reporting. Numerous experiments have been carried out to examine those methods, but a systematic review is still lacking. To fill this gap, in the second paper of this thesis, I conduct a meta-analysis of those experiments. The purpose of the second paper is twofold. First, by listing a catalogue of methods for preventing turnout over-reporting, this paper aims to offer a useful guide for those who intend to measure turnout by opinion polls. Second, by discussing the pros and cons of available solutions, this paper also aims to lay the foundations for future advances in survey measurement of turnout.

### **Paper 3. A Comparison between Two Solutions to Turnout Over-Reporting**

Whilst the second paper of this thesis broadly reviews various methods for preventing turnout over-reporting, the third paper focuses on two promising methods – pipeline technique and item-count technique. These methods are inspired by two opposite rationales. The item-count technique embodies the idea that confidentiality promotes candour, whereas the pipeline technique leads respondents to believe that nothing is confidential and thus honesty is the best policy. These two techniques have separately proven to be effective against turnout over-reporting in different studies, but it remains unclear which technique is better. In order to answer this question, I conduct Web-survey experiments and analyse the results in the third paper.

### **Paper 4. Not Just Valid but Precise**

According to the third paper of this thesis, the item-count technique is, overall, better than the pipeline technique as a method for preventing turnout over-reporting. The fourth paper thus aims to further improve the item-count technique. In order to overcome the main drawback of the item-count technique, i.e. statistical inefficiency, I develop a new statistical estimator that supplements the technique with auxiliary information. The proposed estimator takes advantage of the item-count technique to produce valid estimates, and also takes advantage of auxiliary information to improve the precision of estimation. Monte Carlo simulations provide evidence for the usefulness of the proposed estimator.

Together, the four papers provide clear guidance both to designers and users of election survey data about the solutions to and likely impact of turnout over-reporting.

## Supplementary materials

Supplementary materials are freely available online at: <https://goo.gl/hTrESH>

## References

- Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. "Attempts to Improve the Accuracy of Self-Reports of Voting." In *Questions about Questions*, ed. Judith M. Tanur. New York: Russell Sage Foundation.
- Allen, W. David. 2007. "The Reporting and Under-Reporting of Rape." *Southern Economic Journal* 73(3): 623-641.
- Atkeson, Lonna Rae. 1999. "'Sure, I Voted for the Winner!' Overreport of the Primary Vote for the Party Nominee in the National Election Studies." *Political Behaviour* 21(3): 197-215.
- Baldwin, Kelly C., Phillip C. Ginsberg, and Richard C. Harkaway. 2003. "Under-Reporting of Erectile Dysfunction among Men with Unrelated Urologic Conditions." *International Journal of Impotence Research* 15(2): 87-89.
- Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17(4): 479-498.
- Belli, Robert, F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63(1): 90-108.
- Berg, Nathan, and Donald Lien. 2006. "Same-Sex Sexual Behaviour: US Frequency Estimates from Survey Data with Simultaneous Misreporting and Nonresponse." *Applied Economics* 38(7): 757-769.
- Bernstein, Robert, Antia Chadha, and Robert Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quarterly* 65(1): 22-44.
- Brenner, Philip S. 2012. "Overreporting of Voting Participation as a Function of Identity." *Social Science Journal* 49(4): 421-429.
- Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8(4): 389-398.
- Cahalan, Don. 1968. "Correlates of Respondent Accuracy in the Denver Validity Survey." *Public Opinion Quarterly* 32(4): 607-621.
- Clarke, Harold D., David Sanders, Marianne C. Stewart, and Paul F. Whiteley 2004. *Political Choice in Britain*. Oxford: Oxford University Press.

- Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 32(4): 588-606.
- Deufel, Benjamin J., and Orit Kedar. 2010. "Race and Turnout in U.S. Elections: Exposing Hidden Effects." *Public Opinion Quarterly* 74(2): 286-318.
- Geys, Benny. 2006. "Explaining Voter Turnout: A Review of Aggregate Level Research." *Electoral Studies* 25(4): 637-663.
- Gil, Joan, and Toni Mora. 2011. "The Determinants of Misreporting Weight and Height: The Role of Social Norms." *Economics and Human Biology* 9(1): 78-91.
- Gonzalez-Ocantos, Ezequiel, Chad Kiewiet de Jonge, Carlos Meléndez, Javier Osorio, and David W. Nickerson. 2012. "Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua." *American Journal of Political Science* 56(1): 202-217.
- Górecki, Maciej A. 2011a. "Electoral Salience and Vote Overreporting: Another Look at the Problem of Validity in Voter Turnout Studies." *International Journal of Public Opinion Research* 23(4): 544-557.
- Górecki, Maciej A. 2011b. "Why Bother Lying When You Know So Few Care? Party Contact, Education and Over-reporting Voter Turnout in Different Types of Elections." *Scandinavian Political Studies* 34(3): 250-267.
- Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: John Wiley and Sons.
- Hadaway, C. Kirk, Penny Long Marler, and Mark Chaves. 1998. "Over-Reporting Church Attendance in America: Evidence That Demands the Same Verdict." *American Sociological Review* 63(1): 122-130.
- Health, Anthony, and Bridget Taylor. 1999. "New Sources of Abstention." In *Critical Elections*, eds. Geoffrey Evans and Pippa Norris. London: Sage.
- Hill, Kim Quaile, and Patricia A. Hurley. 1984. "Nonvoters in Voters' Clothing: The Impact of Voting-Behavior Misreporting on Voting-Behavior Research." *Social Science Quarterly* 65(1): 199-206.
- Holbrook, Allyson L., Melanie C. Green, and Jon A. Krosnick. 2003. "Telephone versus Face-to-face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response bias." *Public Opinion Quarterly* 67(1): 79-125.
- Holtgraves, Thomas. 2004. "Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding." *Personality and Social Psychology Bulletin* 30(2): 161-172.
- Hurst, Erik, Geng Li, and Benjamin Pugsley. 2014. "Are Household Surveys Like Tax Forms? Evidence from Income Under-Reporting of the Self-Employed." *Review of Economics and Statistics* 96(1): 19-33.

- James, Lisa M. and Hoa T. Vo. 2010. "Hawthorne Effect." In *Encyclopaedia of Research Design*, ed. Neil J. Salkind. Thousand Oaks, CA: Sage Publications Ltd.
- Jones, Elise F., and Jacqueline Darroch Forrest. 1992. "Under-Reporting of Abortion in Surveys of United-States Women: 1976 to 1988." *Demography* 29(1): 113-126.
- Karp, Jeffrey A., and David Brockington. 2005. "Social Desirability and Response Validity: A Comparative Analysis of Overreporting Voter Turnout in Five Countries." *Journal of Politics* 67(3): 825-840.
- Kraut, Robert E. and John B. McConahay. 1973. "How Being Interviewed Affects Voting: An Experiment." *Public Opinion Quarterly* 37(3): 398-406.
- Krohn, Marvin D., Alan J. Lizotte, Matthew D. Phillips, Terence P. Thornberry, and Kristin A. Bell. 2013. "Explaining Systematic Bias in Self-Reported Measures: Factors that Affect the Under- and Over-Reporting of Self-Reported Arrests." *Justice Quarterly* 30(3): 501-528.
- Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the 'New South'." *Journal of Politics* 59(2): 323-349.
- Lee, Raymond M., and Claire M. Renzetti. 1990. "The Problems of Researching Sensitive Topics." *American Behavioural Scientist* 33(5): 510-528.
- Li, Lei, Karl Kim, and Lawrence Nitz. 1997. "Predictors of Safety Belt Use among Crash-Involved Drivers and Front Seat Passengers: Adjusting for Over-Reporting." *Accident Analysis and Prevention* 31(6): 631-638.
- McDonald, Michael P. 2003. "On the Overreport Bias of the National Election Study Turnout Rate." *Political Analysis* 11(2): 180-186.
- McDonald, Michael P., and Samuel L. Popkin. 2001. "The Myth of the Vanishing Voter." *American Political Science Review* 95(4): 963-974.
- Mircea, Comşa, and Camil Postelnicu. 2013. "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique." *International Journal of Public Opinion Research* 25(2): 153-172.
- Morrall, Andrew R., Daniel McCaffrey, and Martin Y. Iguchi. 2000. "Hardcore Drug Users Claim to Be Occasional Users: Drug Use Frequency Under-Reporting." *Drug and Alcohol Dependence* 57(3): 193-202.
- Nathan, Gad, Monroe G. Sirken, Gordon B. Willis, and James L. Esposito. 1990. "Laboratory Experiments on the Cognitive Aspects of Sensitive Questions." Paper Presented at the International Conference on Measurement Errors in Surveys, Tucson, AZ, November 11-12, 1990.
- Parry, Hugh J., and Helen M. Crossley. 1950. "Validity of Responses to Survey Questions." *Public Opinion Quarterly* 14(1): 61-80.

- Paulhus, Delroy L. 2002. "Socially Desirable Responding: The Evolution of a Construct." In *The Role of Constructs in Psychological and Educational Measurement*, eds. Henry I. Braun, Douglas N. Jackson, and David E. Wiley. Mahwah, NJ: Erlbaum.
- Presser, Stanley, and Michael Traugott. 1992. "Little White Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting." *Public Opinion Quarterly* 56(1): 77-86.
- Rasinski, Kenneth A., Alison K. Baldwin, Gordon B. Willis, and Jared B. Jobe. 1994. "Risk and Loss Perceptions Associated with Survey Reporting of Sensitive Behaviours." In Proceedings of the Survey Research Methods Section, American Statistical Association: 497-502. Retrieved from [http://www.amstat.org/sections/srms/Proceedings/papers/1994\\_082.pdf](http://www.amstat.org/sections/srms/Proceedings/papers/1994_082.pdf)
- Schaeffer, Nora Cate. 2000. "Asking Questions about Threatening Topics: A Selective Overview." In *The Science of Self-report: Implications for Research and Practice*, eds. Arthur A. Stone, Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. New York, NY: Lawrence Erlbaum Associates, Inc.
- Selb, Peter, and Simon Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 32(1): 186-196.
- Sigelman, Lee. 1982. "The Nonvoting Voter in Voting Research." *American Journal of Political Science* 26(1): 47-56.
- Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting." *American Political Science Review* 80(2): 613-624.
- Silver, Brian D., Paul R. Abramson, and Barbara A. Anderson. 1986. "The Presence of Others and Overreporting of Voting in American National Elections." *Public Opinion Quarterly* 50(2): 228-239.
- Smets, Kaat, and Carolien van Ham. 2013. "The Embarrassment of Riches? A Meta-analysis of Individual-level Research on Voter Turnout." *Electoral Studies* 32(2): 344-359.
- Stocké, Volker, and Tobias Stark. 2007. "Political Involvement and Memory Failure as Interdependent Determinants of Vote Overreporting." *Applied Cognitive Psychology* 21(2): 239-257.
- Stocké, Volker. 2007. "Response Privacy and Elapsed Time Since Election Day as Determinants for Vote Overreporting." *International Journal of Public Opinion Research* 19(2): 237-246.
- Stockwell, Tim, Susan Donath, Mark Cooper-Stanbury, Tanya Chikritzhs, Paul Catalano, and Cid Mateo. 2004. "Under-Reporting of Alcohol Consumption in Household Surveys: A Comparison of Quantity-Frequency, Graduated-Frequency and Recent Recall." *Addiction* 99(8): 1024-1033.

- Stryker, Sheldon. 1980. *Symbolic interactionism: A Social Structural Version*. Caldwell, NJ: Blackburn Press.
- Swaddle, Kevin, and Anthony Heath. 1989. "Official and Reported Turnout in the British General Election of 1987." *British Journal of Political Science* 19(4): 537-551.
- Taiwan's Election and Democratisation Study (teds.nccu.edu.tw). 2010. TEDS 2010C: Taipei, Taichung, and Kaohsiung Cities Mayoral Elections (dataset). Taipei, Taiwan: Election Study Centre, National Cheng-chi University (distributor).
- Taiwan's Election and Democratisation Study (teds.nccu.edu.tw). 2012. TEDS 2012: The Survey of the Presidential and Legislative Elections (dataset). Taipei, Taiwan: Election Study Centre, National Cheng-chi University (distributor).
- Taiwan's Election and Democratisation Study (teds.nccu.edu.tw). 2013. TEDS 2013: Benchmark Survey (dataset). Taipei, Taiwan: Election Study Centre, National Cheng-chi University (distributor).
- Taiwan's Election and Democratisation Study (teds.nccu.edu.tw). 2014. TEDS 2014: The Survey of the Nine-in-One Local Elections (dataset). Taipei, Taiwan: Election Study Centre, National Cheng-chi University (distributor).
- Taiwan's Election and Democratisation Study (teds.nccu.edu.tw). 2016. TEDS 2016: The Survey of Presidential and Legislative Elections (dataset). Taipei, Taiwan: Election Study Centre, National Cheng-chi University (distributor).
- The American National Election Studies (www.electionstudies.org). 2008. The ANES 2008 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies (www.electionstudies.org). 2012. The ANES 2012 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies (www.electionstudies.org). 1998. The ANES 1998 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies (www.electionstudies.org). 2000. The ANES 2000 Pilot Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The British Election Study (www.britishelectionstudy.com). 1987. The 1987 BES Cross-Section Survey (dataset). 2nd Edition. Colchester, Essex: UK Data Archive (distributor).
- The British Election Study (www.britishelectionstudy.com). 1997. The 1997 BES Cross-Section Survey (dataset). 2nd Edition. Colchester, Essex: UK Data Archive (distributor).

- The British Election Study ([www.britishelectionstudy.com](http://www.britishelectionstudy.com)). 2001. The 2001 BES Cross-Section Survey (dataset). Colchester, Essex: UK Data Archive (distributor).
- The British Election Study ([www.britishelectionstudy.com](http://www.britishelectionstudy.com)). 2005. The 2005 BES Pre-Election Cross-Section Survey (dataset). Colchester, Essex: UK Data Archive (distributor).
- The British Election Study ([www.britishelectionstudy.com](http://www.britishelectionstudy.com)). 2014. The 2010 BES Cross-Section Survey (dataset). Colchester, Essex: UK Data Archive (distributor).
- The British Election Study ([www.britishelectionstudy.com](http://www.britishelectionstudy.com)). 2015. The 2015 BES Cross-Section Survey (dataset). 2.2 Version. Colchester, Essex: UK Data Archive (distributor).
- Tittle, Charles R. and Richard J. Hill. 1967. "The Accuracy of Self-Reported Data and Prediction of Political Activity." *Public Opinion Quarterly* 31(1): 103-106.
- Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5): 859-883.
- Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, U.K.: Cambridge University Press.
- Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(3): 413-432.
- Traugott, Michael W., and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behaviour." *Public Opinion Quarterly* 43(3): 359-377.
- Traugott, Santa. 1989. "Validating Self-Reported Vote: 1964-1988." Paper presented at the Annual Meeting of the American Statistical Association, Washington D.C., August 7-10th, 1989
- Tsai, Chi-lin. 2010. "Don't Ask Me! I'm Not Interested in Politics: An Analysis of Topic Effect and the Turnout Overestimate in TEDS." *Journal of Electoral Studies* 17(2): 135-175. (In Chinese).
- Tsai, Chi-lin. 2011. *Who Over-Reports and the Consequences of Turnout Over-Reporting: An Analysis of the British Election Study 2005*. MSc Dissertation, Department of Government, University of Essex, UK.
- Voogt, Robert J.J., and Willem E. Saris. 2003. "To Participate or Not to Participate: The Link between Survey Participation, Electoral Participation, and Political Interest." *Political Analysis* 11(2): 164-179.
- Warner, Kenneth E. 1978. "Possible Increases in Under-Reporting of Cigarette Consumption." *Journal of the American Statistical Association* 79(362): 314-318.

Weiss, Carol H. 1968. "Validity of Welfare Mothers' Interview Responses." *Public Opinion Quarterly* 32(4): 622-633.

Wu, Chung-li. 2006. "Vote Misreporting and Survey Context: The Taiwan Case." *Issues and Studies* 42(4): 223-239.

Yalch, Richard F. 1976. "Pre-election Interview Effects on Voter Turnout." *Public Opinion Quarterly* 40:331-336.



## DO TURNOUT OVER-REPORTERS ALMOST NEVER VOTE?

### A FURTHER EXAMINATION USING TURNOUT VALIDATION DATA FROM AMERICAN NATIONAL ELECTION STUDIES

---

CHI-LIN TSAI

**Abstract** Many respondents in post-election surveys report having voted when they actually have not. While most studies have discussed who over-reports based on indirect evidence, [Presser and Traugott \(1992\)](#) attempted to resolve the question based on direct observation. Their finding – that *over-reporters almost never vote* – seriously challenged the established view and has strongly influenced subsequent studies. To examine that influential finding further, this research note re-analyses [Presser and Traugott’s](#) research data – the 1976 American National Election Study (ANES) – and compares them with other ANESs from 1972 to 1990. The re-examination reveals that the 1976 ANES was, in parts, highly unusual. That raises concerns about the generality and, even worse, the validity of [Presser and Traugott’s](#) finding on who over-reports. This research note urges caution regarding their finding, and stresses the need for more data collection and analysis in order to reach a robust conclusion about who over-reports.

## **Introduction**

Turnout over-reporting is a survey measurement error that occurs when non-voters claim to have voted at interviews. Those who voted at least intermittently, if not habitually, have been considered prime suspects for over-reporting (Belli, Traugott, and Beckmann 2001: 495; Tittle and Hill 1967: 105). This is because, when intermittent and habitual voters are recalling an election from which they happened to abstain, their experiences of voting can cause confusion (Abelson, Loftus, and Greenwald 1992: 151), and their motivations for voting in other elections (e.g. to conform to the norms of voting, or to avoid guilt feelings about abstention) could pressure them to report having voted when they failed to do so (Bernstein, Chadha, and Montjoy 2001; Brenner 2012). This line of reasoning has gained support from a large number of empirical studies that found over-reporters resembling voters more closely than non-voters in demographic and attitudinal characteristics (Ansolabehere and Hersh 2011: 273, 2012: 453; Belli, Traugott, and Beckmann 2001: 497; Hill and Hurley 1984: 202; Kitt and Gleicher 1950: 407–8; Traugott and Presser 1992: 11; Sigelman 1982: 55; Silver, Anderson, and Abramson 1986: 613; Tittle and Hill 1967: 104–5; Tourangeau, Groves, and Redline 2010: 426; Weiss 1968: 626, yet cf. Cahalan 1968–1969: 615; Swaddle and Heath 1989: 550; Traugott and Katosh 1979: 376). However, in the absence of any direct observation of over-reporters' turnout histories, none of those studies has provided clear evidence about who over-reports.

Presser and Traugott (1992 [1983]) made an important attempt to fill this gap. They used data from the 1976 American National Election Study (ANES) to trace respondents' official turnout records in three subsequent elections, and found that "Misreporters [over-reporters] almost never voted." This finding is noteworthy not

because it challenges the established view of over-reporters, but because it challenges the view by direct observational evidence. Indeed, one single study is insufficient to overturn an established view. However, because of the very nature of the evidence, Presser and Traugott's finding is seemingly so irrefutable as to weigh disproportionately heavily with subsequent studies. Over the past thirty years, what subsequent studies have done is to hold the established view of over-reporting, while taking Presser and Traugott's finding as a valid counterpoint (e.g. [Belli, Traugott, and Beckmann 2001: 496](#); [Brenner 2012: 421](#); [Darmofa 2010: 155](#); [Harbaugh 1996: 65](#); [Holbrook and Krosnick 2013: 107](#); [Traugott 1989: 26](#)). To date, Presser and Traugott's finding is still at a point that most studies do not accept it, yet none can ignore it either.

This research note set outs to resolve this contradiction. While recognising Presser and Traugott's contribution, my research results caution against overemphasising their finding and for the use of turnout validation data from the 1976 ANES panel study because, on close examination, what they found appears to be more like an anomaly from an unusual dataset, rather than a typical case of over-reporting. The following analysis demonstrates how anomalous Presser and Traugott's finding was, explores possible explanations for that anomaly, and concludes with a discussion of the implications for future studies.

### **Typical case or anomalous instance?**

The 1976 ANES was designed both to collect a representative sample of U.S. citizens for research into the 1976 U.S. national election, and to complete the 1972–1974–1976 panel study. An exercise to check whether respondents in this ANES series voted in the 1972, 1974 and 1976 national elections was not carried out right after each election, but rather in August–October 1977. It ended up with 1,093 respondents whose voter

registrations were not found in government records (Table 1).<sup>1</sup> These ‘*not-found*s’ were included in the data as ineligible electors (hence non-voters) in *all* three elections. Analysing this data, [Presser and Traugott](#) surprisingly found that, of the respondents whose claims to have voted in the 1976 election could not be confirmed, 87.9% voted in neither 1972 nor 1974 (Table 2A), which led to the conclusion that most over-reporters appeared to be habitual non-voters. Since this conclusion was based on a straightforward observation, it seemed so irrefutable as to cast serious doubt on the established view that over-reporters are habitual or intermittent voters.

**Table 1. Summary of the results of the 1977 validation exercise**

Validation results	1976 election	1974 election	1972 election
R's voter registration record wasn't found	1,093	1,093	1,093
R's voting record couldn't be checked	39	268	334
R was confirmed as a non-voter	585	987	842
R was confirmed as a voter	1,786	1,155	1,234
Total	3,503	3,503	3,503

**NOTE:** This table is based on the entire validation sample of the 1972–1974–1976 ANESs. Numbers are unweighted sample sizes.

## 1. Re-examination of the 1976 ANES

Although [Presser and Traugott](#) counted the not-founds among non-voters in their analysis of the 1976 ANES, another paper which they co-authored with [Santa Traugott](#) (1990: 8) in fact argued, “*Failure to find a registration record does not have the same analytical status as finding a record that is marked in an inconsistent manner relative to the respondent’s report.*” Sticking to this principle, the following analysis divides

<sup>1</sup> Table 1 also shows other results of the validation exercise. It is noteworthy that the number of respondents whose registrations were found but whose *voting* records could not be checked increased as time elapsed after Election Day: only 39 respondents’ voting records for the 1976 election could not be checked; the number soared to 268 for the 1974 election and 334 for the 1972 election. This was perhaps a consequence of not conducting a validation exercise promptly after each election. See Appendix 3 for details of variable coding.

over-reporters into two types:

1. **Not-Found Type (NF)**: those who claim to have voted in an election but whose voter registration records are not found.
2. **Non-Voter Type (NV)**: those who claim to have voted in an election but who are clearly marked as non-voters in official records.

Table 2 shows that the majority of over-reporters in the 1976 ANES were of the NF type.<sup>2</sup> As mentioned above, the not-found were labelled as ineligible electors in all three elections, so all NF-over-reporters in the 1976 ANES were by definition non-voters in the 1972 and 1974 elections (Table 2B). Since NF-over-reporters ( $n=394$ ) numbered nearly three times as many as NV-over-reporters ( $n=147$ ) in the 1976 ANES, combining two types of over-reporters for analysis inevitably leads to the conclusion that over-reporters almost never voted. In contrast, separating the non-found from non-voters yields a very different result: 44.8% of NV-over-reporters voted in at least one of the two elections (Table 2C), i.e. many NV-over-reporters in the 1976 ANES appeared to vote intermittently though not habitually, and this result is closer to the established view of over-reporters.

<sup>2</sup> Throughout this note, when the term ‘over-reporters’ is mentioned without the prefixes ‘NF’ or ‘NV’, it refers to *overall* over-reporters.

**Table 2. Re-Examination of who over-reported in the 1976 ANES**

	1976 ANES	1972 and 1974 elections	N	%
A.	Overall over-reporters	Didn't vote or <i>the not-found</i> s in both	474	87.94%
		Voted in one; didn't vote in the other	47	8.72%
		Voted in both	18	3.34%
B.	Not-found-type over-reporters	Not-found	394	100.00%
		Didn't vote in both	0	-
		Voted in one; didn't vote in the other	0	-
		Voted in both	0	-
C.	Non-voter-type over-reporters	Not-found	0	-
		Didn't vote in both	80	55.17%
		Voted in one; didn't vote in the other	47	32.41%
		Voted in both	18	12.41%

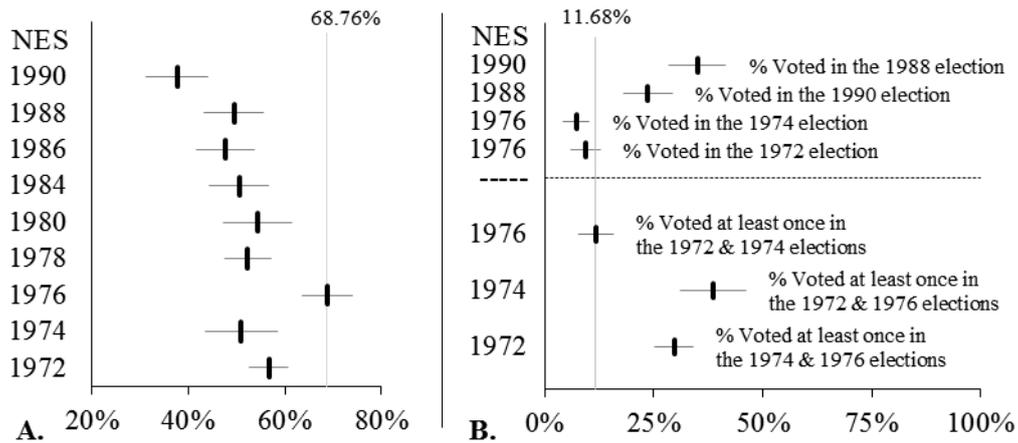
**NOTE:** This table is based on the 1976 ANES and cross-sectional sample weight (V764008). Part A replicates [Presser and Traugott's Table 1 \(1992: 80\)](#) in which the not-found were counted among non-voters; those whose registration records were found but whose voting records could not be checked are excluded from the analysis. Respondents in Part A are divided into parts B and C.

## 2. Comparison of ANESs

It is now clear that [Presser and Traugott's](#) unusual finding on who over-reported was associated with two factors: (1) the not-found were counted as non-voters in all three elections from 1972–1976, and (2) NF-over-reporters heavily outnumbered NV-over-reporters in the 1976 ANES. Leaving aside the first factor, further investigation into the second factor (Figure 1A) reveals that the proportion of NF-over-reporters among overall over-reporters in the 1976 ANES (68.8%) was highly unusual – almost 20 points higher than the average of other ANESs (49.9%).<sup>3</sup> So unusual was the 1976 ANES that any finding based on it might lack generality.

<sup>3</sup> The ANES conducted turnout validation exercises for its 1964, 1972, 1974, 1976, 1978, 1980, 1984, 1986, 1988 and 1990 studies. This research note does not use the 1964 ANES for three reasons. First, the validation exercise for the 1964 ANES was not well documented. Even the ANES team nowadays has little information about that exercise ([Traugott 1989: 11](#)). Second, perhaps because of its inexperience, the 1964 ANES did not seem to be well-prepared for turnout validation. A significant number of respondents were excluded from the validation exercise due to the lack of their names. Last but not least, the not-found in the data of the 1964 ANES were grouped with other types of ineligible electors and cannot be separated out for analysis. As for the other validation data used in this research note, it is noteworthy that the ANES from 1984 stopped checking the turnout of respondents who admitted not having registered *and* not having voted. In order to ensure comparability, I exclude this kind of respondents from the samples of the 1972–1980 ANESs. (In the absence of the variable of self-reported registration, the 1974 ANES can only be partly adjusted.)

**Figure 1. Demonstration of the anomalousness in the 1976 ANES**



**NOTE:** Figure 1A shows the percentages of NF-over-reporters among overall over-reporters. Figure 1B shows the percentages of overall over-reporters in an election who voted in other elections. The horizontal lines across point estimates are 90% confidence intervals. These two figures use the cross-sectional samples of a series of ANESs. While Figure 1A involves nine ANESs, Figure 1B involves only five, because only these five ANESs checked respondents’ turnout in more than one election. As the ANES suggests, the estimates for 1976 and 1974 are computed with sample weights (V764008 and V764006). In order to ensure comparability, the data for 1972–1980 are adjusted to match the validation designs of the 1980–1990 ANESs (see Footnote 3) and consequently the estimates here differ slightly from those in Table 2.

Figure 1B reinforces this concern, showing that over-reporters in the 1976 ANESs were less electorally engaged than their counterparts in other ANESs.<sup>4</sup> Merely 11.7% of over-reporters in the 1976 ANES voted at least once in the 1972 and 1974 elections, whereas 29.7% of over-reporters in the 1972 ANES voted at least once in the 1974 and 1976 elections, and 38.6% of over-reporters in the 1974 ANES voted at least once in the 1972 and 1976 elections. Moreover, only 9.4% and 7.2% of over-reporters in the 1976 ANES voted in the 1972 and 1974 elections, respectively, while 23.7% of over-reporters in the 1988 ANES voted in the 1990 election, and 35% in the 1990 ANES voted in the 1988 election. These comparisons once again demonstrate the anomalous nature of the 1976 ANES, substantiating the concern that the finding based on such

<sup>4</sup> This section aims to demonstrate how anomalous the 1976 ANES was, rather than arguing whether it is legitimate to count the not-found as non-voters. Hence, Figure 1B still groups these two types of respondents together, but the next section will discuss this categorisation issue in more detail.

anomalous data (i.e. over-reporters almost never voted) was not a typical case, thus providing insufficient evidence to overturn the established view.

### **Ineligible electors or misclassified voters?**

Why was [Presser and Traugott](#)'s finding on who over-reported in the 1976 ANES so anomalous? Sampling error is certainly a possible explanation, but did non-sampling errors come into play as well? Is it possible that some of the not-found did vote in some elections in 1972–1976, but the 1977 validation exercise failed to locate their voter registrations, misclassified them as ineligible electors, and consequently led to that anomalous finding?

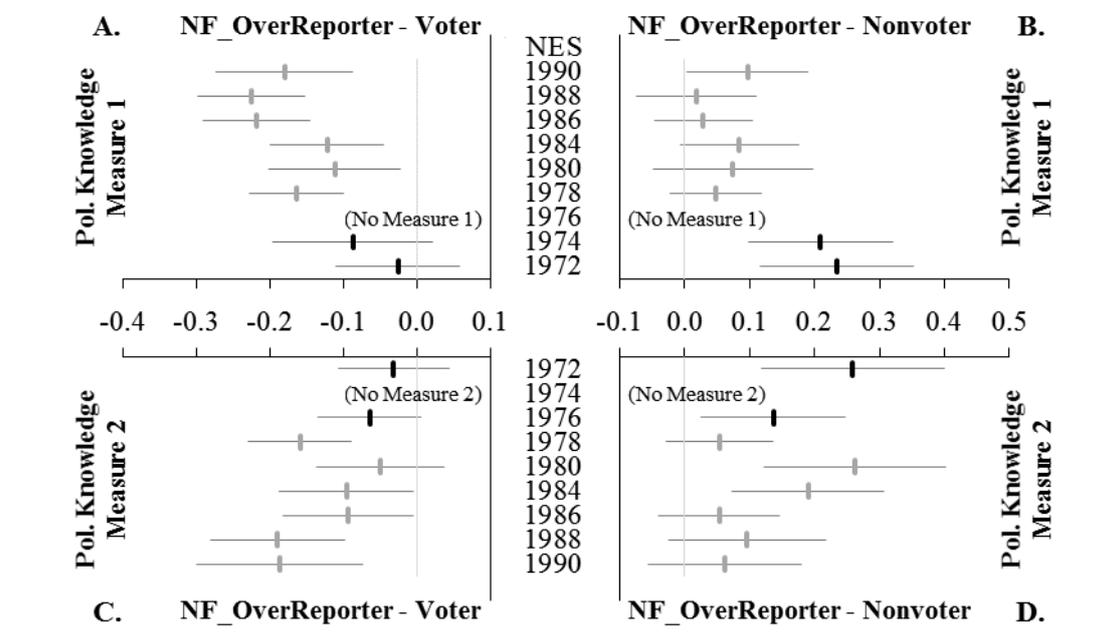
In fact, some researchers do think it presumptuous to assume the not-found to be ineligible electors ([Clausen 1968](#); [Traugott 1989](#); [Berent, Krosnick, and Lupia 2011](#)). Strictly speaking, we do not know whether the not-found registered to vote. It is possible that “a record cannot be located because of problems of name or address spelling, incorrect addresses, or a misunderstanding about the locality in which the respondent is actually registered” ([Presser, Traugott, and Traugott 1990: 8](#)). In this regard, the unusually high proportion of NF-over-reporters in the 1976 ANES might be a sign that some of the not-found were actual voters in the 1976 election but were misclassified as ineligible electors due to missing records. Furthermore, regardless of whether the not-found voted in 1976 or not, counting them as ineligible electors in the 1972 and 1974 elections was in itself risky, because the long gaps between the validation exercise in 1977 and the polling days of those two elections could cause even more difficulties when looking into past records ([Traugott 1989: 7](#)). Both considerations point towards validation misclassification as a plausible explanation for why the finding on who over-reported in the 1976 ANES was so anomalous. The

following analysis examines this explanation in more detail.

### 1. Corroborative evidence based on political knowledge

Political knowledge has been identified as a key determinant of turnout (Larcinese 2007). Given that respondents can hardly exaggerate their political knowledge, analysis of the resemblance between the not-found's and other respondents' levels of political knowledge may provide information to assess whether the not-found's in the 1977 validation exercise were misclassified voters. The analysis focuses on the not-found's who reported having voted (i.e. NF-over-reporters), since there is little doubt that the not-found's who confessed to abstention were non-voters (Selb and Munzert 2013: 191; Traugott 1989: 15).

**Figure 2. Assessment of the possibility of validation misclassification**



**NOTE:** These figures are based on a series of ANES cross-sectional samples. The estimates are percentages of NF-over-reporters answering correctly to a measure minus the percentage of members in a reference group answering correctly to that measure. The horizontal lines are 90% confidence intervals. As the ANES suggests, the estimates for 1976 and 1974 are computed with sample weights (V764008 and V764006). In order to ensure comparability, the data for 1972–1980 are adjusted to match the validation designs of the 1980–1990 ANESs (see Footnote 3).

The analysis uses two different measures of political knowledge.<sup>5</sup> Analysis of Measure 1 (Figure 2A & B) shows that, in the 1978–1990 ANESs, NF-over-reporters' levels of political knowledge were similar to those of non-voters, but significantly lower than those of actual voters at the 0.1 level; the results of the 1972 and 1974 ANESs were the exact opposite. Analysis of Measure 2 (Figure 2C & D) also shows that NF-over-reporters in the 1972 and 1976 ANESs were more politically knowledgeable than non-voters and akin to voters.<sup>6</sup> Overall, Figure 2 demonstrates that, in contrast to their counterparts in the 1978–1990 ANESs, NF-over-reporters in the 1972–1976 ANESs bore a greater resemblance to voters than to non-voters in terms of political knowledge. Given that political knowledge correlates with turnout and cannot be easily exaggerated, this result corroborates the suspicion that some actual voters in the 1972–1976 elections were misclassified as ineligible electors due to the failure of the 1977 validation exercise to locate their voter registrations.

## **2. Projections based on other ANESs**

Had the problem of validation misclassification been minimised, how might the finding on who over-reported in the 1976 ANES were different? I try to answer this counterfactual question by imagining two scenarios. First, imagine that fewer actual voters in the 1976 election were misclassified due to missing records. In this scenario, the proportion of NF-over-reporters in the 1976 ANES should have been smaller than it was (68.8%; see Figure 1A). Reducing the proportion to a relatively normal level

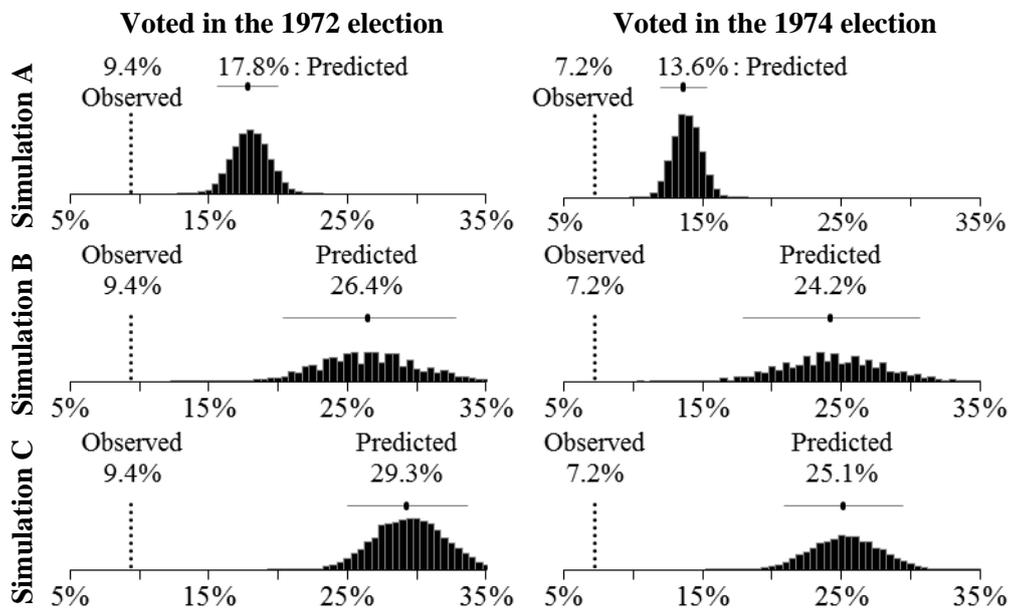
<sup>5</sup> Measure 1 asked respondents whether they remembered at least one congressional candidate's name and political party in their constituencies. Measure 2 asked about whether respondents knew the majority party in Congress before the election. These two measures were used since they were the most common measures of political knowledge in the ANESs from 1972 to 1990.

<sup>6</sup> The patterns in Figure 2C and D are less clear than those in Figure 2A and B, though still recognizable. One possible explanation is that Measure 2 had less discriminatory power – the rate of correct answers to Measures (33.2% – 76.6%) was significantly higher than that to Measure 1 (20.91% – 39.00%) in every ANES in the analysis.

(≈49%), Simulation A in Figure 3 shows that 17.8% of 1976 ANES over-reporters would have been found to have voted in the 1972 election, and 13.6% in the 1974 election. Both are almost twice as high as the original observations (see Appendix 4 for details about simulations).

Second, imagine that the validation exercise had not been delayed to 1977 and thus fewer actual voters in the 1972 and 1974 elections were misclassified. The 1988 ANES offered information about how validation misclassification was associated with a delay in validation. Given that information, Simulation B shows that, had validation exercises been promptly carried out after each election, the 1976 ANES over-reporters' turnout in the 1972 and 1974 elections would have been found to be 26.4% and 24.2%, respectively. These numbers are even higher, 29.3% and 25.1%, when the first scenario is considered simultaneously (Simulation C).

**Figure 3. Simulation of who over-reported in the 1976 ANES**



NOTE: These figures are based on the 1976 ANES cross-sectional samples. The observed percentages of over-reporters in the 1976 ANES who voted in the 1972 and/or 1974 elections are identical to the estimates shown in Figure 1A. Simulation results are shown as histograms. The mean and range between the 5th and 95th percentiles of each set of simulation results are also presented (see Appendix 4).

Overall, these simulations suggest that, if [Presser and Traugott](#) had taken the possibility of validation misclassification into account, they would have found over-reporters in the 1976 ANES to be more electorally engaged than the raw data indicated, and thus might not have arrived at the conclusion that over-reporters almost never voted.

## **Discussion and conclusion**

Whilst most studies have considered habitual or intermittent voters as prime suspects for turnout over-reporting, [Presser and Traugott](#)'s observation from the 1976 ANES pointed to an opposite conclusion: over-reporters almost never vote. Their study has been quoted as evidence against the established view, but this research note advises caution. Closer examination reveals that [Presser and Traugott](#)'s finding was a direct result of an abnormal feature of the 1976 ANES – an unusually high proportion of respondents who claimed to have voted but whose voter registrations were not found (i.e. not-found-type over-reporters). That raises a concern that a finding based on such unusual data might lack generality, and hence could not form a sufficient basis for overturning the established view. Further investigation into that abnormal feature suggests that some actual voters might have been misclassified as ineligible electors due to difficulties in finding their voter registrations. When this possibility is taken into account, over-reporters in the 1976 ANES appear to have been more electorally engaged than [Presser and Traugott](#)'s observation suggests. As a result, there is little ground for believing that over-reporters almost never vote.

The analysis shows no indication that over-reporters are habitual voters either – if anything, over-reporters seem more like intermittent voters. However, there are insufficient data to reach a firm conclusion. Most opinion polls either did not check respondents' turnout records or only checked one election. Only a very few polls

checked two or three elections (e.g. the 1976 ANES), and based on such a small number of elections it is hard to judge whether over-reporters are habitual or intermittent (non-)voters.<sup>7</sup> In order to resolve the question of who over-reports, it is essential to check respondents' turnout records for a longer series of elections. To simplify this task, pollsters may draw inspiration from [Ansolabehere and Hersh's \(2012\)](#) research that sought to cooperate with commercial data vendors who had already collected turnout records from governmental sources.<sup>8</sup>

Collecting new data is vital, so is prudence in using existent data. The present research note finds that the turnout validation data for the 1976 ANES contain an unusually high and possibly erroneous proportion of the not-found-type over-reporters, so researchers should exercise caution when using that dataset to study over-reporting. Discarding the 1976 ANES from the analysis is an option, but a more sensible approach would be to analyse the not-found and other respondents – particularly non-voters – separately, so that the dataset can still be used, and the problem can be explicitly monitored.<sup>9</sup> In fact, this advice is not specific to addressing the problem of the 1976 ANES; it applies to other validation data as well, since it has been argued that the non-found and non-voters do not have the same analytical status. Whenever it is appropriate, treating the not-found as a separate category for analysis is generally recommended.

<sup>7</sup> Note that, if an opinion poll checks respondents' turnout records for  $n$  elections, it allows us to assess how often over-reporters voted in  $n-1$  elections, because one election should be used to define whether respondents over-reported or not. Therefore, when a poll checked only one election, its data can tell us who over-reported, but nothing about over-reporters' turnout histories.

<sup>8</sup> [Ansolabehere and Hersh \(2012\)](#) demonstrate how academic opinion polls and commercial data vendors can cooperate to facilitate less costly and more reliable turnout validation. However, it seems that their study only checked respondents' turnout records for one election, and so did not provide additional information about how often over-reporters vote.

<sup>9</sup> Considering the distinctive design of the 1976 ANES, some studies chose to completely or partly exclude it from their analysis, in order to prevent any comparability issue (e.g. [Belli, Traugott, and Beckmann 2001: 482](#); [Traugott 1989: 7](#)).

The analysis results of this research note have implications for not only who over-reports turnout, but also why. [Presser and Traugott's](#) study greatly reduced the appeal of misremembering as a possible explanation of over-reporting because, if over-reporters almost never vote, it is difficult to imagine how their memories could mislead them into over-reporting ([Belli, Traugott, and Beckmann 2001: 496](#)). Nonetheless, considering that [Presser and Traugott's](#) finding might not be a typical case, misremembering should still not be dismissed out of hand as irrelevant to over-reporting.

Furthermore, with [Presser and Traugott's](#) study being found to lack generality, the established view about the resemblance between over-reporters and actual voters goes largely unchallenged at the present time. That implies a gloomy prospect for developing a simple correction to turnout over-reporting. It seems unrealistic to expect to distinguish over-reporters from actual voters based on just one or two survey variables ([Hill and Hurley 1984: 206](#)). Additionally, "Turnout validation is anything but an easy cut into the over-reporting" ([Traugott 1989: 3](#)). Therefore, there is a great need to develop more sophisticated solutions to over-reporting.

The ANES' as well as [Presser and Traugott's](#) endeavours to resolve the question about who over-reports turnout are laudable. Their research, though imperfect, brought that important question to the attention of academic circles. The present research note points out the limitations of [Presser and Traugott's](#) research data and results. Future studies should work to overcome those limitations in order to fully resolve the question about who over-reports.

## Supplementary materials

Supplementary materials are freely available online at: <https://goo.gl/hTrESH>

## Appendix

### 1. Data

Data for this paper are sourced from the ANES website ([www.electionstudies.org](http://www.electionstudies.org)).

**Table A1. Information on data**

ANES	Data Version	Response Rate	Refusal Rate	Sample Size
1990	03rd May 1999	70.6%	20.3%	1,980
1988	21st May 1999	70.5%	22.2%	2,040
1986	06th May 2002	67.7%	25.6%	2,176
1984	03rd May 1999	72.1%	20.7%	2,257
1980	03rd May 1999	71.8%	20.8%	1,614
1978	03rd May 1999	68.9%	22.7%	2,304
1976	03rd May 1999	70.4%	-	2,248
1974	03rd May 1999	70.0%	16.5%	1,575
1972	03rd May 1999	75.0%	14.5%	2,705

**NOTE:** This table is taken from the ANES webpage ([www.electionstudies.org/overview/dataqual.htm](http://www.electionstudies.org/overview/dataqual.htm)). ANES calculated the response and refusal rates in the 1976 survey based on ‘the 1976 Single-Year File’, for which the sample size was 2,248. In a dataset released later, called ‘the 1972–1974–1976 Merged File’, 11 additional respondents were added to the 1976 cross-sectional sample ([www.electionstudies.org/studypages/anes\\_mergedfile\\_1972to1976/anes\\_mergedfile\\_1972to1976.htm](http://www.electionstudies.org/studypages/anes_mergedfile_1972to1976/anes_mergedfile_1972to1976.htm)). The present research note uses the 1972–1974–1976 ANES Merged File, since it is better organised. Moreover, in this table, the sample sizes for the 1974 and 1976 ANESs are only for fresh cross-sectional samples. When panel samples are included, the sample sizes for the 1974 and 1976 ANESs become 2,099 and 2,916, respectively.

### 2. Variables

*Voter turnout:*

The supplementary file – Turnout Validation of the ANESs from 1972 to 1990 – describes the variables of voter turnout in great detail.

*Political knowledge:*

Measure 1. The wording varied slightly from one ANES to another but, in general, it

was, “Do you happen to remember the names of the candidates for congress – that is, for the House of Representatives in Washington – who ran in this district this November?” The main variables in the ANESs are V720945, V742214, V780115, V800823, V840737, V860103, V880565 and V900107. (The 1976 ANES did not contain this variable.) Some other variables are also used to check whether respondents really remembered candidates’ names (see supplementary materials.) The values of these variables are dichotomised into ‘remember versus forget’.

Measure 2. The wording was, “Do you happen to know which party had the most members in the House of Representatives in Washington before the election (this/last) month?” The variables in the ANESs are V720950, V763683, V780500, V801028, V841006, V860349, V880878 and V900402. (The 1974 ANES did not contain this variable.) The values of these variables are dichotomised into ‘correct versus incorrect’.

- ‘Don’t know’ answers are counted as a negative response in both measures. Other vague and nonresponse answers are coded as missing values.

### **3. Coding**

This document depicts the turnout-validation exercises of the American National Election Studies from 1972 to 1990. The depiction of each validation exercise consists of three parts: (1) the wording of key items, (2) a flowchart showing the connections between items, and (3) the rules for coding the turnout variables for this research note.

Two points are worth mentioning prior to the depiction. First, the research note sets out to examine Presser and Traugott's (1992) research finding on who over-reports turnout, so it is crucial to know how they coded turnout variables for analysis. Unfortunately, their paper did not detail their coding rules, so I deduced that information through trial and error. After applying different rules for coding turnout variables, I eventually found a set of rules that allows me to replicate Presser and Traugott's tables. Second, Presser and Traugott counted the not-founds among non-voters, but I treat them separately. As for coding other categories, such as voters and missing values, I follow Presser and Traugott's coding rules precisely.

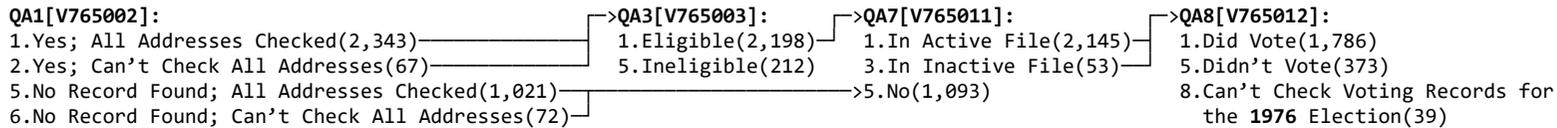
In what follows, the numbers in parentheses are unweighted respondent counts, words in brackets are variable names given by ANES datasets, and words in braces are categorical definitions. The labels of some categories are abbreviated due to space limitations. (For full labels, refer to relevant documents on the ANES website: [www.electionstudies.org](http://www.electionstudies.org)).

■ 1972-1974-1976 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1972-1974-1976 ANES:

- QA1. Was there a voter registration record available for the person whose name is on this label?
  - QA3, 10, 14. Was this person registered, or otherwise eligible, to vote in the [election date], general election?
  - QA7, 11, 15. Did this person's name appear on a list, file card, or some other record indicating eligibility to vote on [election date]?
  - QA8, 12, 16. Does the record indicate that this person did vote in the [election date], general election?
- [[Http://www.electionstudies.org/studypages/1976prepost/1976prepost\\_qnaire\\_vv.pdf](http://www.electionstudies.org/studypages/1976prepost/1976prepost_qnaire_vv.pdf)]  
 [[Http://www.electionstudies.org/studypages/anes\\_mergedfile\\_1972to1976/anes\\_mergedfile\\_1972to1976\\_vardoc\\_codebook.pdf](http://www.electionstudies.org/studypages/anes_mergedfile_1972to1976/anes_mergedfile_1972to1976_vardoc_codebook.pdf)]

► VALIDATION OF WHETHER RESPONDENTS IN THE 1976 ANES VOTED IN THE 1976 ELECTION:



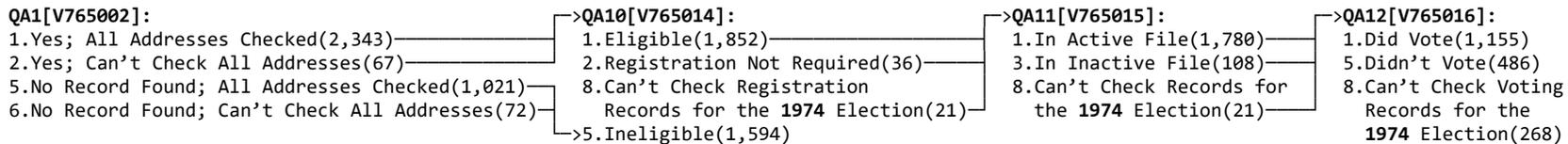
► PRESSER & TRAUOGOTT'S CODING RULES:

- Voter (1,786) = QA8:1(1,786)
- Non-voter(1,678) = QA8:5(373) + QA7:5(1,093) + QA3:5(212)
- Missing ( 39) = QA8:8(39)

► CODING RULES OF THE PRESENT RESEARCH NOTE:

- Voter (1,786) = QA8:1(1,786)
- Not-found(1,093) = QA7:5(1,093)
- Non-voter( 585) = QA8:5(373) + QA3:5(212)
- Missing ( 39) = QA8:8(39)

► VALIDATION OF WHETHER RESPONDENTS IN THE 1974 ANES VOTED IN THE 1974 ELECTION:



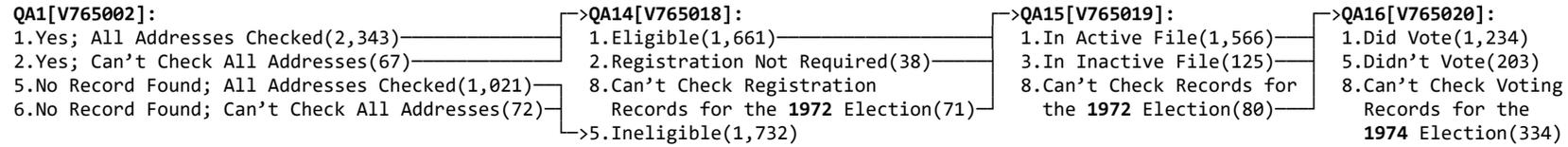
► PRESSER & TRAUOGOTT'S CODING RULES:

- Voter (1,155) = QA12:1(1,155)
- Non-voter(2,080) = QA12:5(1,594)
- Missing ( 268) = QA12:8(268)

► CODING RULES OF THE PRESENT RESEARCH NOTE:

Voter (1,155) = QA12:1(1,155)  
 Not-found(1,093) = QA1 :5(1,021) + QA1:6(72)  
 Non-voter( 987) = QA12:5(486) + QA10:5(1,594) - NotFound(1,093)  
 Missing ( 268) = QA12:8(268)

► **VALIDATION OF WHETHER RESPONDENTS IN THE 1972 ANES VOTED IN THE 1972 ELECTION:**



► **PRESSER & TRAUOGOTT'S CODING RULES:**

Voter (1,234) = QA16:1(1,234)  
 Non-voter(1,935) = QA16:5(203) + QA14.5(1,732)  
 Missing ( 334) = QA16:8(334)

► **CODING RULES OF THE PRESENT RESEARCH NOTE:**

Voter (1,234) = QA16:1(1,234)  
 Not-found(1,093) = QA1 :5(1,021) + QA1:6(72)  
 Non-voter( 987) = QA12:5(486) + QA10:5(1,594) - NotFound(1,093)  
 Missing ( 334) = QA16:8(334)

■ 1978 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1978 ANES:

- Q2. Was there a voter registration record available for the person whose name is on this label?
  - Q6. Was this person registered, or otherwise eligible, to vote in the November 7, 1978 general election?
  - Q9. Does the record indicate that this person voted in the November 7, 1978 general election?
- [[Http://www.electionstudies.org/studypages/1978post/1978post\\_qnaire.pdf](http://www.electionstudies.org/studypages/1978post/1978post_qnaire.pdf)]  
[[Http://www.electionstudies.org/studypages/1978post/nes1978.pdf](http://www.electionstudies.org/studypages/1978post/nes1978.pdf)]

► VALIDATION OF WHETHER RESPONDENTS IN THE 1978 ANES VOTED IN THE 1978 ELECTION:

<b>Q2[V781400]:</b>	<b>Q6[V781408]:</b>	<b>Q9[V781409]:</b>
1.Yes(1,465)	1.Yes(1,412)	1.Voted(960)
3.Registration Not Required(0)	2.No(53)	2.Didn't Vote(420)
5.No Record Available(797)		8.Can't Check Record for This Election(32)
7.No Name Obtained(35)		
9.Said Being Registered at Another Address but Didn't Give Enough Information to Check Registration Record at That Address(7)		

► CODING RULES OF THE PRESENT RESEARCH NOTE:

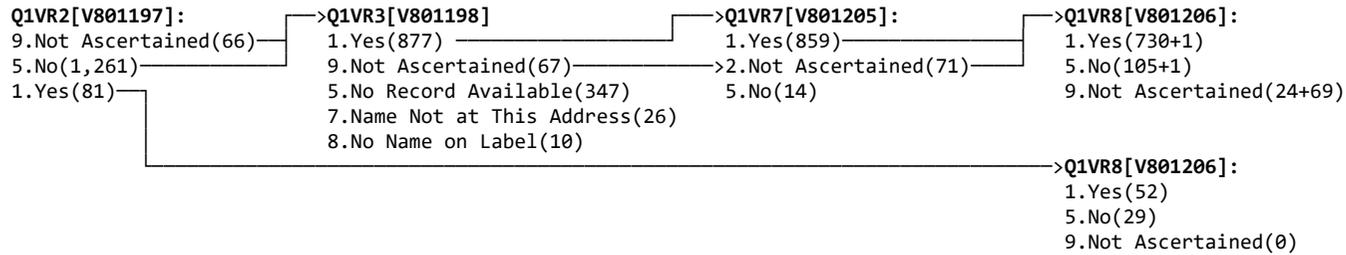
- Voter (960) = Q9:1(960)
- Not-found(797) = Q2:5(797)
- Non-voter(473) = Q6:2(53) + Q9:2(420)
- Missing ( 74) = Q2:7(35) + Q2:9(7)  
+ Q9:8(32)

■ 1980 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1980 ANES:

- Q1VR2. Do election regulations allow election day registration of voters in this county (or municipality)? V801197
  - Q1VR3. Is there a voter registration record available for the person whose name appears on this label at the label address? V801198
  - Q1VR7. Was this person eligible to vote in the November 4, 1980 general election? V801205
  - Q1VR8. Dose the record indicate that this person voted in November 4, 1980 general election? V801206
- [The questionnaire of the 1980 validation exercise is not available on the ANES website.]  
[\[Http://www.electionstudies.org/studypages/1980prepost/nes1980.pdf\]](http://www.electionstudies.org/studypages/1980prepost/nes1980.pdf)

► VALIDATION OF WHETHER RESPONDENTS IN THE 1980 ANES POST-ELECTION SURVEY VOTED IN THE 1980 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

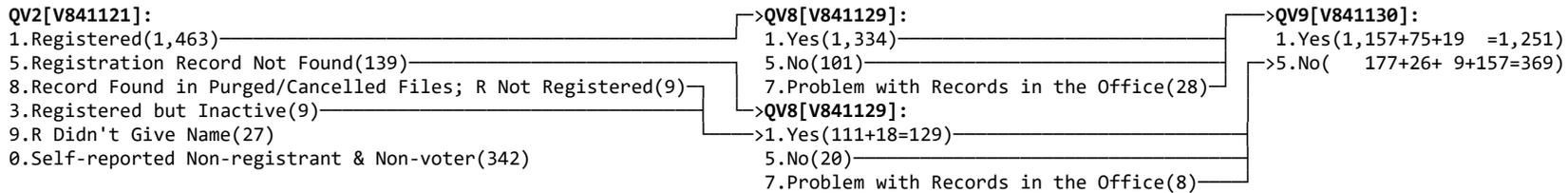
- Voter (783) = Q1VR8:1(783)
- Not-found(347) = Q1VR3:5(347)
- Non-voter(149) = Q1VR7:5(14) + Q1VR8:5(135)
- Missing (129) = {Q1VR3:7,8}(36) + Q1VR8:9(93)

■ 1984 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1984 ANES:

- QV2. Were you able to find a voter registration record for this respondent?
- QV8. Were you able to gain access to the records (or lists) that would allow YOU to determine whether R voted in the November 6, 1984 general election?
- QV9. According to the answers given about determining whether someone voted or not: Did R vote in the November 6, 1984 general election?
- QVS. Validation Summary[V84113]  
[\[Http://www.electionstudies.org/studypages/1984prepost/1984prepost\\_qnaire\\_vv.pdf\]](http://www.electionstudies.org/studypages/1984prepost/1984prepost_qnaire_vv.pdf)  
[\[Http://www.electionstudies.org/studypages/1984prepost/nes1984.pdf\]](http://www.electionstudies.org/studypages/1984prepost/nes1984.pdf)

► VALIDATION OF WHETHER RESPONDENTS IN THE 1984 ANES POST-ELECTION SURVEY VOTED IN THE 1984 ELECTION:



QVS[V84113]:

- {QV2:1 & QV8:1,5,7 & QV9:1}→ 1.Vote Record Indicates That R Voted(1,248)
- {QV2:1 & QV8:5,7 & QV9:1}→15.Some Evidence That R Voted but It Is Impossible to Say If(3)
- {QV2:1,3,8 & QV8:1,7 & QV9:5}→ 2.Vote Record Indicates R didn't Vote(89)
- ↳12.Vote Record Indicates R didn't Vote(115)
- {QV2:1,3,8 & QV8:5 & QV9:5}→13.Found Registration Record but No Voting Records; Can't Make Judgement(9)
- {QV2:1 & QV8:5 & QV9:5}→ 4.Found Registration Record but No Voting Records; Can't Make Judgement(17)
- {QV2:5 & QV8:1,5 & QV9:5}→ 3.Didn't Found Registration Record; All Registration Records Accessible(85)
- {QV2:5 & QV8:7 & QV9:5}→ 6.Didn't Found Registration Record; Not All Registration Records Accessible(6)
- {QV2:5 & QV8:1,5,7 & QV9:5}→14.Didn't Found Registration/Voting Record; All Sources Accessible(48)
- {QV2:9}→ 5.R Didn't Give Name(19)
- ↳11.R Didn't Give Name(8)
- {QV2:0}→10.Didn't Try to Validate; Self-reported Non-registrant & Non-voter(342)

► CODING RULES OF THE PRESENT RESEARCH NOTE:

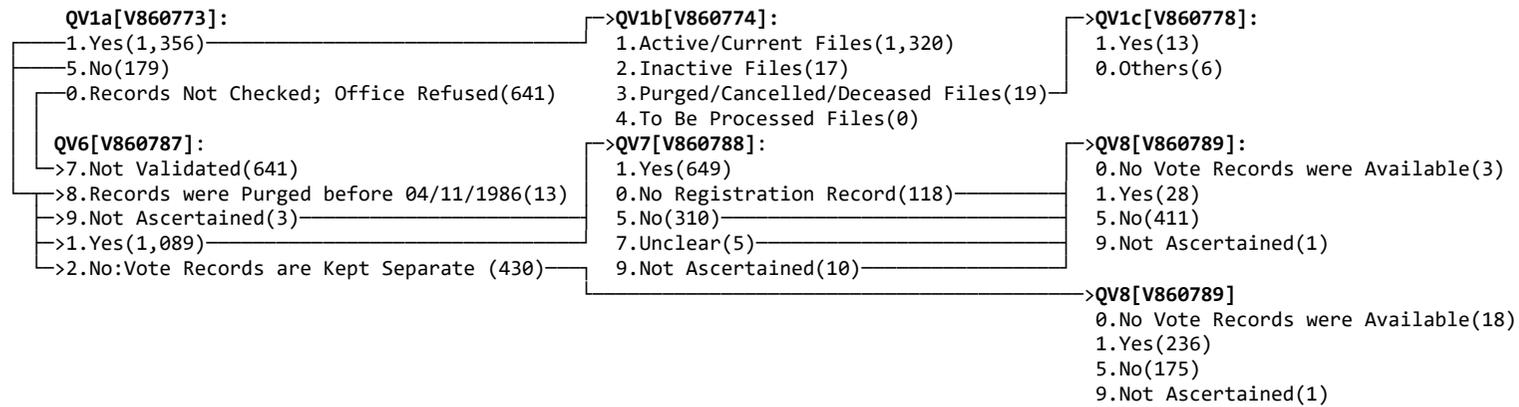
- Voter (1,248) = QVS:1(1,248)
- Not-found( 139) = QVS:3(85) + QVS:6 (6) + QVS:14(48)
- Non-voter( 204) = QVS:2(89) + QVS:12(115)
- Missing ( 398) = QVS:15(3) + QVS:4 (17) + QVS:13(9) + QVS:5(19) + QVS:11(8) + QVS:10(342)

■ 1986 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1986 ANES:

- QV1a. Does the respondent have a registration record in this office?
  - QV1b. Where did you find the record?
  - QV1c. Purge date before 04/11/1986?
  - QV6. Are registration records (or master file) updated with vote information?
  - QV7. Does the registration/master file indicate that R voted?
  - QV8. Do vote records indicate that R voted?
- [[Http://www.electionstudies.org/studypages/1986post/1986post\\_qnaire\\_vv.pdf](http://www.electionstudies.org/studypages/1986post/1986post_qnaire_vv.pdf)]  
 [[Http://www.electionstudies.org/studypages/1986post/nes1986.pdf](http://www.electionstudies.org/studypages/1986post/nes1986.pdf)]

► VALIDATION OF WHETHER RESPONDENTS IN THE 1986 ANES VOTED IN THE 1986 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

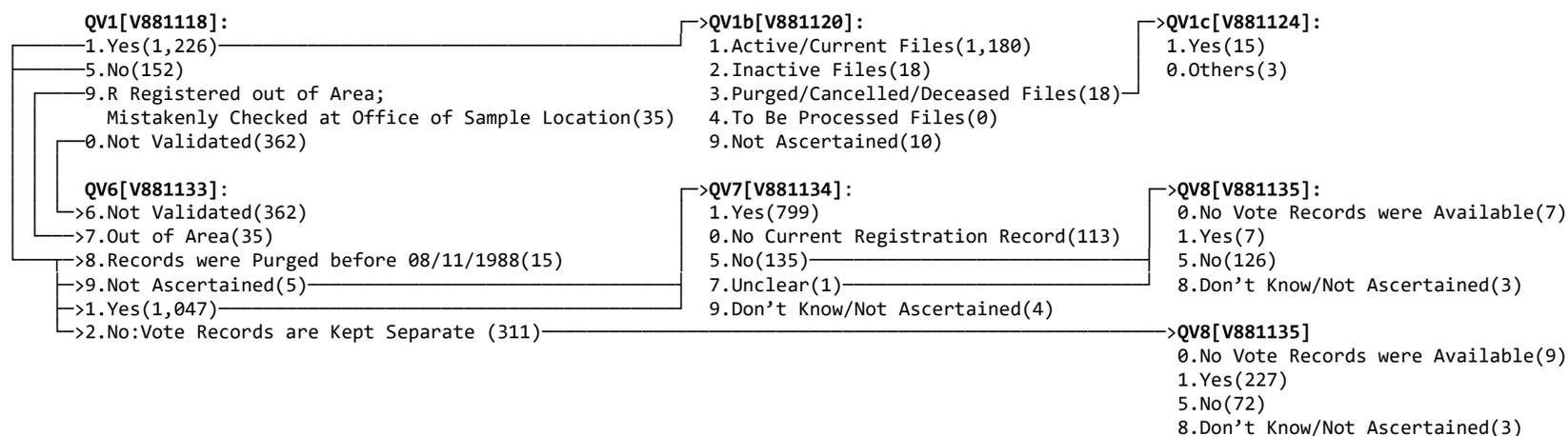
- Voter (913) = QV7:1(649) + QV8:1(28+236)
- Not-found(179) = {QV1a:5 & QV8:0,5,9}(179) + {QV1a:1 & QV7:0 & QV8:0,5,9}(0)
- Non-voter(424) = QV1c:1(13) + {QV1a:1 & QV6:2 & QV8:5}(121) + {QV1a:1 & QV6:1,9 & QV7:5 & QV8:0,5,9}(290)
- Missing (660) = QV1a:0(641) + {QV1a:1 & QV6:2 & QV8:0,8}(13) + {QV1a:1 & QV6:1,9 & QV7:7,9 & QV8:0,5,9}(6)

■ 1988 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1988 ANES:

- QV1. Does the respondent have a registration record in this office? V881118
  - QV1b. Where did you find the record?
  - QV1d. Purge date before 08/11/1988?
  - QV6. Are registration records (or master file) updated with vote information?
  - QV7. Does the registration/master file indicate that R voted in November 1988?
  - QV8. Do vote records indicate that R voted in November 1988 General Election?
- [[Http://www.electionstudies.org/studypages/1988prepost/1988prepost\\_qnaire\\_vv.pdf](http://www.electionstudies.org/studypages/1988prepost/1988prepost_qnaire_vv.pdf)]  
 [[Http://www.electionstudies.org/studypages/1988prepost/nes1988.pdf](http://www.electionstudies.org/studypages/1988prepost/nes1988.pdf)]

► VALIDATION OF WHETHER RESPONDENTS IN THE 1988 ANES POST-ELECTION SURVEY VOTED IN THE 1988 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

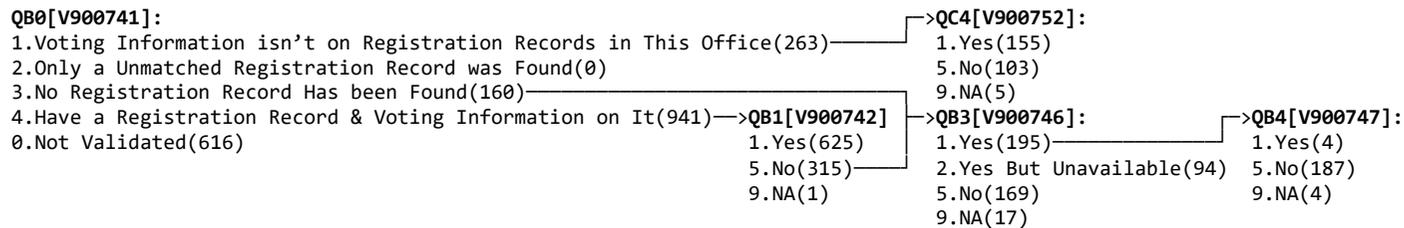
- Voter (1,033) = QV7:1(799) + QV8:1(7+227)
- Not-found( 147) = {QV1a:5 & QV7:0,9}(113) + {QV1a:5 & QV8:0,5,8}(30) + {QV1a:1 & QV7:0}(4)
- Non-voter( 190) = QV1c:1(15) + {QV1a:1 & QV6:2 & QV8:5}(47) + {QV1a:1 & QV6:1,9 & QV7:5 & QV8:0,5,8}(128)
- Missing ( 405) = QV1a:0(362) + {QV1a:1 & QV6:2 & QV8:0,8}(7) + {QV1a:1 & QV6:1,9 & QV7:7 & QV8:0,5,8}(1) + QV1a:9(35)

■ 1990 ANES

► KEY ITEMS IN THE VALIDATION EXERCISE OF THE 1990 ANES:

- QB0. Checkpoint. V900741 V882032
- QB1. Does the registration record indicate that R voted in the November, 1990 general election?
- QB3. Are there any records or files on which voting information for 1990 is recorded, other than three registration records, e.g., ballot application, poll lists, signature lists, (other registration records)?
- QB4. Do any of the voting records indicates that R voted on November 6, 1990?
- QB6. Does the registration record indicate that R voted in the November, 1988 general election?
- QB7. Are there any records or files on which voting information is recorded, other than three registration records, e.g., ballot application, poll lists, signature lists, (other registration records)?
- QB8. Do any of the voting records indicates that R voted in November 1988? V900751 V882041
- QC4. Do any of the voting records indicates that R voted on November 6, 1990?
- QC5. Do any of the voting records indicates that R voted on November 8, 1988?  
[[Http://www.electionstudies.org/studypages/1990post/1990post\\_qnaire\\_vv.pdf](http://www.electionstudies.org/studypages/1990post/1990post_qnaire_vv.pdf)]  
[[Http://www.electionstudies.org/studypages/1990post/nes1990.pdf](http://www.electionstudies.org/studypages/1990post/nes1990.pdf)]

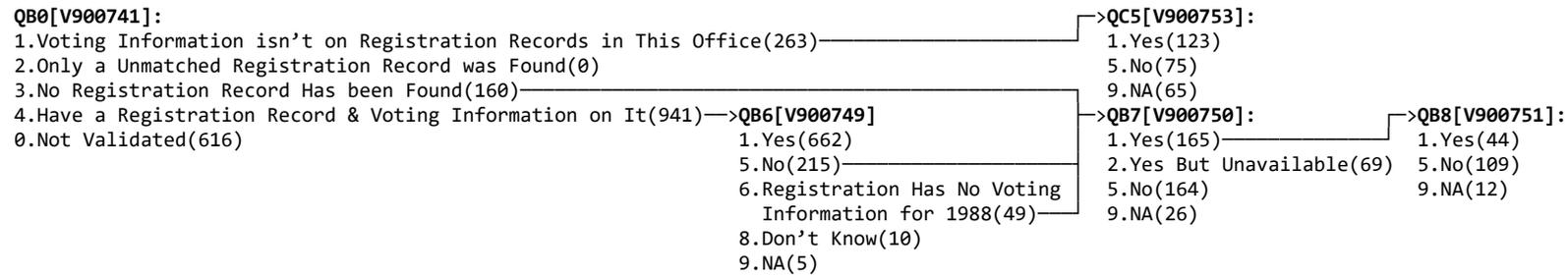
► VALIDATION OF WHETHER RESPONDENTS IN THE 1990 ANES VOTED IN THE 1990 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

- Voter (784) = QB1:1(625) + QB4:1(4) + QC4:1(155)
- Not-found(160) = {QB0:3 & QB3:1 & QB4:5,9}(60) + {QB0:3 & QB3:2,5,9}(100)
- Non-voter(414) = {QB0:1 & QC4:5}(103) + {QB1:5 & QB3:1 & QB4:5,9}(131) + {QB1:5 & QB3:2,5,9}(180)
- Missing (622) = QB0:0(616) + QB1:9(1) + QC4:9(5)

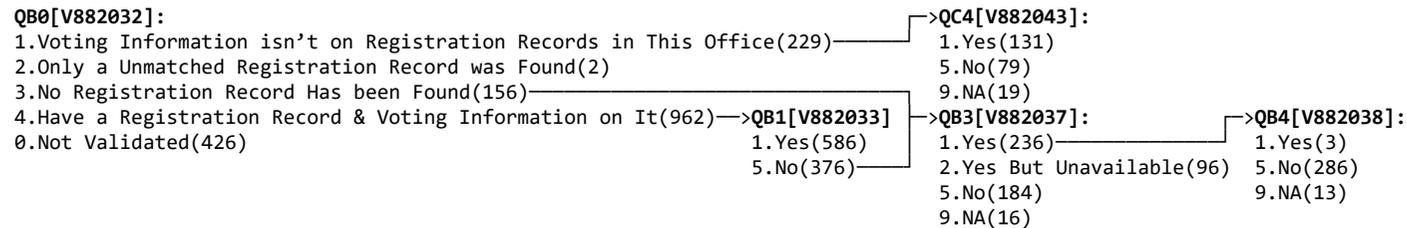
► VALIDATION OF WHETHER RESPONDENTS IN THE 1990 ANES VOTED IN THE 1988 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

Voter (829) = QB6:1(662) + QB8:1(44) + QC5:1(123)  
 Not-found(160) = {QB0:3 & QB7:1 & QB8:5,9}(48) + {QB0:3 & QB7:2,5,9}(112)  
 Non-voter(278) = {QB0:1 & QC5:5}(75) + {QB6:5 & QB7:1 & QB8:5,9}(56) + {QB6:5 & QB7:2,5,9}(147)  
 Missing (713) = QB0:0(616) + {QB6:8,9}(15) + {QB6:6 & QB7:1 & QB7:5,9}(17) + {QB6:6 & QB7:2,5,9}(0) + QC5:9(65)

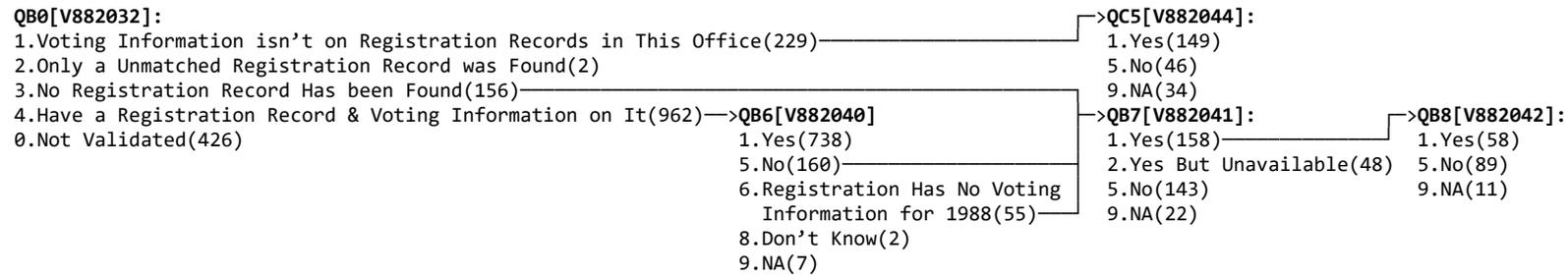
► VALIDATION OF WHETHER RESPONDENTS IN THE 1988 ANES POST-ELECTION SURVEY VOTED IN THE 1990 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

Voter (720) = QB1:1(586) + QB4:1(3) + QC4:1(131)  
 Not-found(156) = {QB0:3 & QB3:1 & QB4:5,9}(82) + {QB0:3 & QB3:2,5,9}(74)  
 Non-voter(452) = {QB0:1 & QC4:5}(79) + {QB1:5 & QB3:1 & QB4:5,9}(151) + {QB1:5 & QB3:2,5,9}(222)  
 Missing (447) = {QB0:0,3}(428) + QC4:9(19)

► RE-VALIDATION OF WHETHER RESPONDENTS IN THE 1988 ANES POST-ELECTION SURVEY VOTED IN THE 1988 ELECTION:



► CODING RULES OF THE PRESENT RESEARCH NOTE:

Voter (945) = QB6:1(738) + QB8:1(58) + QC5:1(149)  
 Not-found(156) = {QB0:3 & QB7:1 & QB8:5,9}(59) + {QB0:3 & QB7:2,5,9}(97)  
 Non-voter(192) = {QB0:1 & QC5:5}(46) + {QB6:5 & QB7:1 & QB8:5,9}(30) + {QB6:5 & QB7:2,5,9}(116)  
 Missing (482) = {QB0:0}(428) + {QB6:8,9}(9) + {QB6:6 & QB7:1 & QB7:5,9}(11) + {QB6:6 & QB7:2,5,9}(0) + QC5:9(34)

#### 4. Simulation

Simulations for estimating how many over-reporters in the 1976 ANES would have voted in the 1972 election under counterfactual scenarios are described as follows. These designs also apply to simulating over-reporters' turnout in the 1974 election.

##### *Simulation A:*

- This simulation manipulates the proportion of NF-over-reporters among overall over-reporters in the 1976 ANES down to a relatively normal level.
- The proportion of NF-over-reporters among overall over-reporters in the 1988 ANES was 49.4% (standard error = 0.04,  $n = 178$ ), which is very close to the average shown in the analysis of Figure 1A (49.9%), so it serves as a good reference for this simulation.
- The simulation is done by repeating the following steps 10,000 times:

Step 1. Draw a value  $P_A$  from a  $t$  distribution with location parameter 49.4%, scale parameter 0.04 and degrees of freedom 177.

Step 2. Generate a weighting variable  $W_{72}$  to adjust the abnormal proportion of NF-over-reporters in the 1976 ANES to  $P_A$ .

Step 3. Weight the data by  $W_{72}$  and calculate the proportion of overall over-reporters in the 1976 ANES who voted in the 1972 election.

##### *Simulation B:*

- This part simulates a scenario where the turnout validation exercises were carried out promptly after the 1972 and 1974 elections.

- The 1988 ANES validated respondents' self-reported turnout twice: once in 1989, a second time in 1991. Of the respondents who reported having voted in the 1988 election and whose registrations were not found by the 1991 validation exercise, 23.2% were in fact confirmed as actual voters by the prompt validation exercise conducted in 1989 (standard error = 0.04,  $n = 82$ ). This result is used for the simulation.

- The simulation is done by repeating the following steps 10,000 times:

Step 1. Draw a value  $P_B$  from a  $t$  distribution with location parameter 23.2%, scale parameter 0.04 and degrees of freedom 81.

Step 2. Draw a value from a Bernoulli distribution with a proportion parameter  $P_B$  for each NF-over-reporter in the 1976 ANES. If the value is equal to 1, recode that NF-over-reporter as an actual voter in the 1972 election.

Step 3. Calculate the proportion of overall over-reporters in the 1976 ANES who voted in the 1972 election.

*Simulation C:*

- The scenarios of Simulation A and B are simulated simultaneously.
- The simulation is done by repeating the following steps 10,000 times:

Step 1. Do the first two steps of Simulation A.

Step 2. Do the first two steps of Simulation B.

Step 3. Weight the data by  $W_{72}$ , and calculate the proportion of overall over-reporters in the 1976 ANES who voted in the 1972 election.

## References

- Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. "Attempts to Improve the Accuracy of Self-Reports of Voting." In *Questions about Questions*, ed. Judith M. Tanur, 138-53. New York: Russell Sage Foundation.
- Ansolabehere, Stephen and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 4(20):437-459.
- Ansolabehere, Stephen, and Eitan Hersh. 2011. "Who Really Votes?" In *Facing the Challenge of Democracy? Explorations in the Analysis of Public Opinion and Political Participation*, eds. Paul M. Sniderman and Benjamin Highton. Princeton, New Jersey: Princeton University Press.
- Belli, Robert F., Michael W. Traugott, and Matthew, N. Beckmann. 2001. "What Leads to Voting Over-Reports? Contrasts of Over-Reporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 4(17):90-108.
- Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia 2011. "The Quality of Government Records and 'Over-estimation' of Registration and Turnout in Surveys: Lessons from the 2008 ANES Panel Study's Registration and Turnout Validation Exercises." in Working Paper no. nes012554. Ann Arbor, Michigan: American National Election Studies. (<http://www.electionstudies.org/resources/papers/nes012554.pdf>)
- Bernstein, Robert, Antia Chadha, and Robert Montjoy. 2001. "Over-Reporting Voting: Why It Happens and Why It Matters." *Public Opinion Quarterly* 1(65):22-44.
- Brenner, Philip S. 2012. "Over-Reporting of Voting Participation as a Function of Identity." *Social Science Journal* 4(49):421-429.
- Cahalan, Don. 1968-1969. "Correlates of Respondent Accuracy in the Denver Validity Survey." *Public Opinion Quarterly* 4(32):607-621.
- Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 4(32):588-606.
- Darmofal, David. 2010. "Re-Examining the Calculus of Voting." *Political Psychology* 31(2): 149-174
- Harbough, William T. 1996. "If People Vote Because They Like to, Then Why Do So Many of Them Lie?" *Public Choice* 89(1/2): 63-76.
- Hill, Kim Q. and Patricia A. Hurley. 1984. "Nonvoters in Voters' Clothing: The Impact Of Voting-Behaviour Misreporting on Voting-Behaviour Research." *Social Science Quarterly* 1(65):199-206.

- Holbrook, Allyson L., and Jon A. Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77(S1): 106-123.
- Kitt, Alice S. and David B. Gleicher. 1950. "Determinants of Voting Behaviour: A Progress Report on the Elmira Election Study." *Public Opinion Quarterly* 14(3): 393-412.
- Larcinese, Valentino. 2007. "Does Political Knowledge Increase Turnout? Evidence from the 1997 British General Election." *Public Choice* 3-4(131):387-477.
- Presser, Stanley and Michael W. Traugott. 1992. "Little White Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting." *Public Opinion Quarterly* 1(56):77-86.
- Presser, Stanley, and Traugott, Michael. 1983. "Correlated Response errors in a panel study of voting." in the Annual Meeting of the American Association of Public opinion Research. Buck Hill Falls, Pennsylvania.
- Presser, Stanley, Michael W. Traugott, and Santa Traugott 1990. "Vote 'Over' Reporting in Surveys." in The International Conference on Measurement Errors. Tucson, Arizona.
- Selb, Peter and Simon Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 1(32):186-196.
- Sigelman, Lee. 1982. "The Nonvoting Voter in Voting Research." *American Journal of Political Science* 1(26):47-56.
- Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Over-Reports Voting?" *American Political Science Review* 2(80):613-624.
- Swaddle, Kevin, and Anthony Heath. 1989. "Official and Reported Turnout in the British General Election of 1987." *British Journal of Political Science* 19(4): 537-551.
- Tittle, Charles R. and Richard J. Hill. 1967. "The Accuracy of Self-Reported Data and Prediction of Political Activity." *Public Opinion Quarterly* 1(31):103-106.
- Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(3): 413-432.
- Traugott, Michael W. and Stanley Presser 1992. "Revalidation of Self-reported Vote." in the Annual Meeting of the American Association for Public Opinion Research. St. Petersburg Beach, Florida.
- Traugott, Michael W., and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behaviour." *Public Opinion Quarterly* 43(3): 359-377.
- Traugott, Santa 1989. "Validating Self-Reported Vote: 1964-1988." in The Annual Meeting of the American Statistical Association. Washington D.C.

Weiss, Carol H. 1968. "Validity of Welfare Mothers' Interview Responses." *Public Opinion Quarterly* 4(32):622-633.

## **A META-ANALYSIS OF SOLUTIONS TO TURNOUT OVER-REPORTING**

---

CHI-LIN TSAI

**Abstract** Turnout over-reporting is a measurement bias that occurs when survey respondents who did not vote report having done so. It threatens the validity of electoral studies, which rely heavily on respondents' self-reports as research data. Over recent decades, survey research has accumulated a wealth of knowledge about solutions to the problem. In this paper, I conduct a meta-analysis of those solutions, aiming to make practical and methodological contributions to the measurement of turnout. This paper presents a solution catalogue, which pollsters will be able to use to tackle turnout over-reporting in the field. In reviewing existing solutions to over-reporting, this paper also indicates directions for further methodological advances in the measurement of turnout.

## **Introduction**

Voting is a fundamental form of political participation in modern democracies. Who votes, why and how are important research topics in political science. Studies rely heavily on sample surveys to collect individual-level data on voter turnout. Respondents' self-reports are, however, error-prone. There is always a noticeable number of respondents 'over-reporting' turnout – they report having voted when they actually abstained. Misremembering may explain why some respondents give false self-reports, but, more importantly, social-desirability bias explains why over-reporting is far more common than under-reporting. Since voting is considered to be the behaviour of a good citizen, non-voters over-report in order to appear in a socially desirable light, while, in this regard, voters have little incentive to under-report (Belli, Traugott, and Beckmann 2001: 479-80; Cahalan 1968: 621; Górecki 2011: 8).

The concern over turnout over-reporting dates back to the 1940s, when pollsters at the time noticed inconsistencies between respondents' self-reports and official records of their turnout (Parry and Crossley 1950). Since then, various solutions have been proposed and tested in the field. Research on this issue has accumulated a wealth of knowledge, but a systematic review is still lacking. This paper fills that gap by conducting a meta-analysis of all the experimental studies on solutions to turnout over-reporting. The purpose of this paper is twofold. First, by compiling a list of solutions to over-reporting, I aim to offer a useful guide for those who intend to measure turnout using opinion polls. Second, by discussing the pros and cons of available solutions, this paper lays the foundations for future advances in the measurement of turnout.

This paper begins by describing the sources of data and methods of analysis. Then, I systematically review each solution to turnout over-reporting – explaining their logic,

evaluating their effects, and discussing potential problems. I also categorise those solutions and assess the performance of each type. The final section poses several questions in need of further investigation, pointing out directions for future research into turnout over-reporting.

## Data and methods

Studies that experimented with solutions to turnout over-reporting were identified through three sources (Table 1). First, through a citation indexing service, *web of Science*, I found 13 peer-reviewed journal articles for this meta-analysis.<sup>1</sup> Second, it is well-known that the American National Election Study (ANES) has frequently made efforts to tackling turnout over-reporting. I checked all the ANES surveys from 1948 to 2013 and found 14 relevant surveys for analysis (two of them overlap with articles identified through the first source). Third, when reviewing those 27 studies, I found seven additional studies: three peer-reviewed articles, three conference papers and one technical report. Taken together, there are 32 studies – 77 experiments in total – for analysis. I summarised these studies in a file that is freely available online (<https://goo.gl/hTrESH>).<sup>2</sup>

---

<sup>1</sup> The search rule is “TOPIC: (((overreport OR over-report OR overreporting OR over-reporting OR misreport OR mis-report OR misreporting) OR (social desirability)) AND (vote OR voter OR voting OR turnout OR election))”.

<sup>2</sup> When summarising these studies, I infer unreported information based on reported information. For example, some studies did not report the sample size, but if they reported the turnout rate estimate, say 70%, and the standard error, say 0.025, then I infer the sample size by  $n = 0.7(1-0.7)/0.025^2$ , given that  $SE^2 = p(1-p)/n$ .

**Table 1. Experimental studies of solutions to turnout over-reporting**

Study	Web of Knowledge	ANES archive	Other sources
Abelson, Loftus, and Greenwald (1992)		×	
ANES Merged Methods Comparison Project (1982)		×	
ANES Panel Study (2004)		×	
ANES Pilot Study (1989)		×	
ANES Time Series Study (1984)		×	
ANES Time Series Study (1996)		×	
ANES Time Series Study (1998)		×	
ANES Time Series Study (2000)		×	
ANES Time Series Study (2008)		×	
ANES Time Series Study (2014)		×	
ANES Time Series Study (2004)		×	
Belli, et al. (1999)	×		
Belli, Moore, and VanHoewyk (2006)	×		
Belli, Traugott, and Rosenstone (1994)		×	
Duff, et al. (2007)	×	×	
Hanmer et al. (2014)	×		
Holbrook and Krosnick (2010a)	×		
Holbrook and Krosnick (2010b)	×		
Holbrook and Krosnick (2013)	×	×	
Keeter, et al. (2002)			×
Kritzing, Schwarzer, and Zeglovits (2012)			×
Locander, Sudman, and Bradburn (1976)			×
Mircea and Gheorghită (2011)			×
Mircea, Comşa, and Camil Postelnicu (2013)	×		
Persson and Solevid (2014)	×		
Presser (1990)	×		
Rogers (1976)			×
Stocké (2007)	×		
U.S. Bureau of the Census (1973)			×
Wu (2006)			×
Waismel-Mano and Sarid (2011)	×		
Zeglovits and Kritzing (2014)	×		

Ideally, studies of solutions to over-reporting should verify respondents' answers against reliable external records of turnout, so as to confirm whether the solutions really increased the proportion of respondents whose answers matched the records. Such records are, however, not always accessible. Even if they are, not all studies can afford to conduct turnout validation exercises. In the 32 studies for analysis, only eight were able to do so.

Furthermore, the use of validation data does not guarantee better assessment of solutions to over-reporting. For example, despite all their efforts, [Mircea and Gheorghiuță \(2011\)](#) were only able to complete validation exercises for 65% of the respondents in their sample. In other words, an assessment based on their validation data has to exclude 35% of respondents (more than 1,400 respondents), and so the results are less precise and less accurate, if respondents who are excluded are systematically different from those who are included.

Due to the difficulties in turnout validation, most studies have made assessments based on respondents' self-reported turnout (e.g. [Belli, Moore, and VanHoewyk 2006](#)). A solution to over-reporting is considered effective if it produces a higher proportion of respondents who report not having voted (or equivalently, if it yields a smaller turnout rate estimate). The use of this criterion is usually justified by two presumptions. First, given an effective random-assignment procedure, any difference between control and experimental groups' turnout rate estimates must be attributed to the experimental solution to over-reporting. Second, it is rare for voters to report not having voted ([Selb and Munzert 2013: 191](#)), so if the experimental group's turnout rate estimate is lower than the control group's, it is highly likely that the solution successfully reduces turnout over-reporting. In this meta-analysis, I use this widely accepted criterion in assessing solutions to turnout over-reporting. I define the effect size of a solution as the ratio of the proportions of self-reported non-voters between control and experimental groups (RPSN, also known as relative risk):

$$RPSN = \frac{\% \text{ Self-reported Nonvoters in Treatment}}{\% \text{ Self-reported Nonvoters in Control}} \quad \ln(RPSN) \sim N(\mu, \sigma^2)$$

This measure is widely used in epidemiology and has a straightforward interpretation. For example, an RPSN of 1.5 means that the proportion of self-reported

non-voters in the experimental group is 50% higher than in the control group (or equivalently, the respondents in the experimental group are 1.5 times as likely as their counterparts in the control group to report that they did not vote). A solution is effective against turnout over-reporting if and only if its RPSN is statistically significantly higher than one. Testing whether  $RPSN > 1$  is mathematically equivalent to testing whether  $\ln(RPSN) > 0$  in a Z-test, since the natural logarithm of RPSN has a sampling distribution approximating to a normal distribution (Sribney and Wiggins 1999).

RPSN has an advantage over other measures of effect size – it takes account of the fact that the higher the actual turnout rate, the fewer the potential over-reporters, and thus the smaller the difference that a solution to turnout over-reporting can make. Suppose that we use a solution to measure the turnout rates of two electorate groups. Group A consists of fewer actual voters than does Group B. In Group A, the solution helps to reduce the self-reported turnout from 50% to 25%, and in Group B, the reduction is from 80% to 70%. In terms of percentage difference, the reduction in Group B (10 points) is smaller than in Group A (25 points), but that does not mean the solution is less effective in Group B. Instead, it is because Group B consists of fewer potential over-reporters (i.e. fewer actual non-voters) for the solution to influence. RPSN reflects this fact, resulting in a value of 1.5 for both groups. In contrast, if the solution also makes a 25-point reduction for Group B (from 80% to 55%), the RPSN increases from 1.5 to 2.25, reflecting the fact that the solution drives a higher proportion of potential over-reporters in Group B to report nonvoting.

I calculate the RPSN of each solution to turnout over-reporting.<sup>3</sup> Some solutions have been tested more than once by different studies, and some results are more precise

---

<sup>3</sup> Some studies counted non-respondents (e.g. don't know, don't remember) among self-reported non-voters, while other studies did not. To standardise measurement, I exclude non-respondents from the

than others, due to different research designs, sample sizes etc. I summarise the results using random-effects meta-analysis (see [Borenstein, Hedges, Higgins, and Rothstein 2009: 69–76](#); also see the online supplement for more details). The meta-analysis can take the precision of each RPSN into account and yield a weighted average RPSN. Considering the differences in the backdrops of the 32 studies, I use random-effects meta-analysis, since it does not assume that the underlying true RPSN is the same in all studies (whereas fixed-effects meta-analysis does).

Throughout this paper, I use the following notation to report the results of each experiment:

For a single experiment: *REC#*:  $RPSN=1.00$ ,  $SE_{Ln}=0.0001$ ,  $P=0.001$

For multiple experiments: *META#*:  $RPSN=1.00$ ,  $SE_{Ln}=0.0001$ ,  $P=0.001$

*REC#* and *META#* are indices for finding the details about an experiment from the online supplementary file. *RPSN* represents the effect size of the solution.  $SE_{Ln}$  is the standard error of the *logarithm* of the RPSN. *P* stands for the p-value, indicating whether a solution makes a statistically significant difference to the proportion of self-reported non-voters. Unless specified otherwise, I report one-tailed p-values for the null hypothesis  $H_0: RPSN \leq 1$  and the alternative hypothesis  $H_a: RPSN > 1$ .

## Solutions to turnout over-reporting

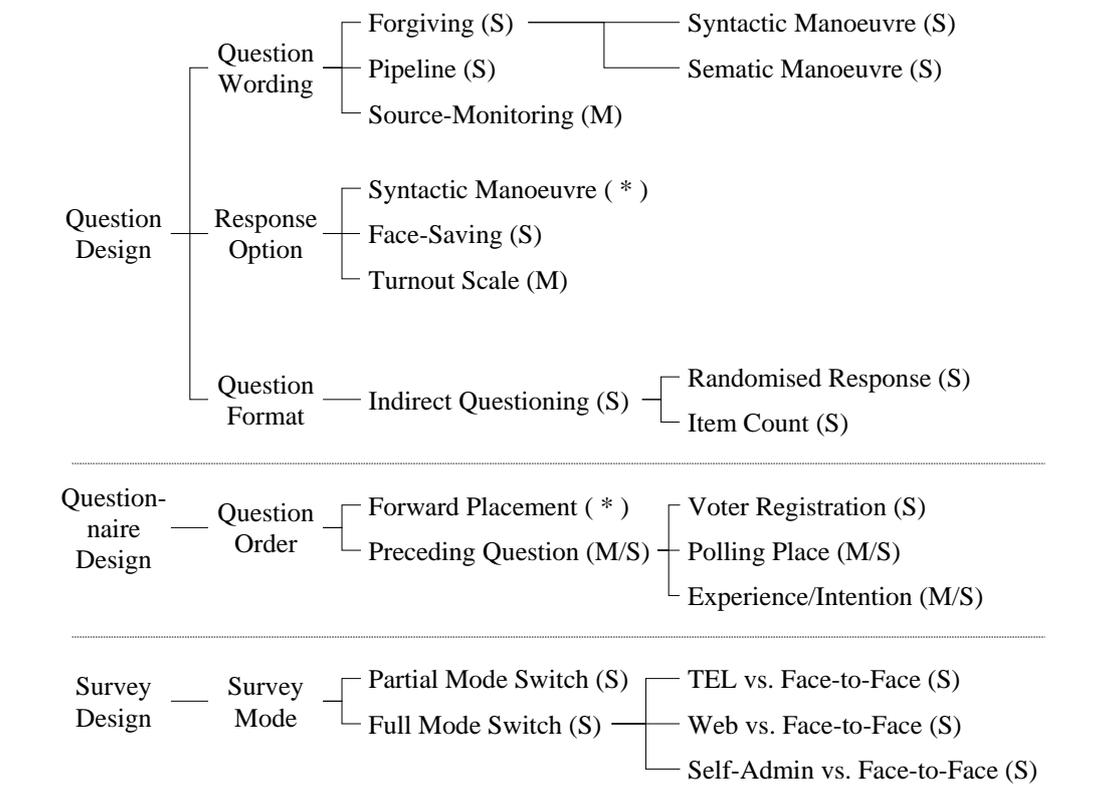
In order to reduce turnout over-reporting, survey methodologists have applied innovative ideas to different parts of survey design (Figure 1). Many of these ideas are aimed at a better question design, so they involve refining the wording, options or

---

RPSN calculation, as long as there is enough information for me to do so. Moreover, if the research designs of the original studies used survey weighting, so does my RPSN calculation.

format of the turnout question. Others improve questionnaire or survey designs to produce positive order or mode effects on turnout measurement. In this section, I review, evaluate and discuss each of these ideas.

**Figure 1. Catalogue of solutions to turnout over-reporting**



NOTES: (M) and (S) indicate solutions that were originally devised to tackle memory failure and social desirability, respectively. (\*) means that original studies were not explicit about which problem to address.

### 1. Question wording

American pollsters in the early 1940s recognised the problem of turnout over-reporting, but until the 1950s, even ANES – the most iconic election survey in world – did not use any special question design to tackle the problem. The turnout question in the 1948 ANES was: “In this election, about half the people voted and half of them didn't. Did you vote?” This wording adheres to the general principles of question design – it is concise, neutral and tries to prevent respondents from thinking of either voting or

nonvoting as a desirable answer. However, this wording does not take any extra precautions against turnout over-reporting.

Four years later, a change in the question wording marked the rise of concerns over turnout over-reporting. In order to address a major cause of the problem – social desirability bias – the 1952 ANES abandoned the principle of neutrality and rephrased the question as follows:

*In talking to people about the election, we find that a lot of people weren't able to vote because they weren't registered, or they were sick or they just didn't have time. How about you, did you vote this time?*

Mentioning nothing about voting, this wording concentrates on delivering a message that nonvoting is prevalent and justifiable. This design is usually called 'forgiving wording'. It aims to make respondents feel less embarrassed at admitting to nonvoting.

Though the 1952 ANES did not experiment with its forgiving wording, an analysis of two later studies – Keeter, Zukin, Andolina, and Jenkins (2002) and Persson and Solevid (2014) – suggests that the use of certain forms of forgiving wording can persuade significantly more respondents to report nonvoting than can the use of neutral wording (*META1: RPSN=1.22, SE<sub>Ln</sub>=0.1084, P=0.035*). This result to some extent justifies ANES' decision to substitute forgiving wording for neutral wording. However, those two studies did not measure turnout in a conventional way. Keeter et al. measured several types of political participation in one question; Persson and Solevid measured the frequency of voting (see *REC131* and *139* for more details). Both are very different from what election surveys (including ANES) usually do, so more investigations into the impact of forgiving wording on turnout measurement are needed.

Forgiving wording has become a standard component of the turnout question of ANES since 1952, and it is now very popular among election surveys around the world. In order to improve this popular wording further, [Abelson, Loftus, and Greenwald \(1992\)](#) tested a hypothesis that, after giving three excuses for nonvoting, it is syntactically more natural to ask respondents whether they ‘missed out’ on voting, rather than whether they voted:

*In talking to people about elections we often find that a lot of people missed out on voting because they weren't registered, they were sick, or they just didn't have time. How about you – did you miss out on voting in the 1988 election for President?*

Their experiment results, however, do not support the hypothesis. Respondents who saw the ‘miss-out’ wording were actually less likely to admit nonvoting than those seeing the classic version of forgiving wording, though the difference was not statistically significant ( $REC27: RPSN=0.96, SE_{Ln}=0.1388, P=0.371, H_a: RPSN$ ).

In addition to syntactic structure, there is also an attempt to improve semantic property of forgiving wording. Forgiving wording conveys an impression that nonvoting is prevalent. ANES usually tells respondents, “*A lot of people were not able to vote.*” Similarly, the 1963 British Election Study (BES, [Butler and Stokes 1979](#)) said, “*A great many people weren't able to vote.*” The 2002 European Social Survey said, “*Some people don't vote nowadays.*” All these sentences are vague about the prevalence of nonvoting, so different respondents may understand forgiving wording in different ways. To investigate this problem, [Mircea and Gheorghiuță \(2011\)](#) experimentally compared a vague determiner – ‘many’ – with two clear determiners – ‘half’ and ‘one out of two’:

*For different reasons, [many people / around half of the people / around one out of two people] were not able to vote at [elections], while others did...*

Their findings were mixed. The *'half'* wording persuaded significantly more respondents to admit to nonvoting than did the *'many'* wording (*REC140: PRSN=1.30, SE<sub>Ln</sub>=0.1108, P=0.009*), but the *'one-out-of-two'* wording did not (*REC141: RPSN=0.99, SE<sub>Ln</sub>=0.1189, P=0.534*). These findings, on the one hand, do not fully support the hypothesis that clear determiners are better than vague determiners. On the other hand, these findings show that a simple amendment such as replacing *'many'* with *'half'* is enough to improve the measurement of turnout, which implies that there is still room for improvement in the way forgiving wording is phrased.

Mircea and Gheorghiuță went some of the way towards refining the part of the wording about the prevalence of nonvoting, but no study has dealt with the justification of nonvoting. Forgiving wording usually comprises three excuses for nonvoting – “weren’t registered”, “were sick” and “didn’t have time”. Do these excuses have different effects on different respondents in different contexts? Are there more persuasive excuses than those three? Can listing more excuses in the wording amplify the effect against turnout over-reporting? All these are awaiting investigation.

Instead of refining forgiving wording itself, [Hanmer, Banks, and White \(2014\)](#) employed the pipeline technique as an additional layer of protection against turnout over-reporting. Their question began with forgiving wording and then used a statement (i.e. a pipeline) to lead respondents to believe that researchers had independent means of checking the truth of their responses. The rationale behind this statement is to replace the pressure to be one who conforms to the social norm of turnout with more powerful pressure to not be perceived as a liar.

*In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. By looking at public records kept by election officials, we can get an accurate report of who actually voted in November, and in previous elections. Of course, these public records do not say who you voted for. Part of our study will involve checking these records against the survey reports...*

This question wording is called an 'actual' pipeline, as the study did validate respondents' answers about turnout after interviews. If the validation exercise is not viable, Hanmer, Banks, and White suggest replacing the actual pipeline with a 'subtle' pipeline, so as to avoid deceiving respondents. Their subtle pipeline wording was:

*In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. We also sometimes find that people who say they voted actually did not vote.*

In Hanmer, Banks, and White' experiments, respondents who were presented with a subtle or actual pipeline were 1.07 and 1.06 times, respectively, as likely as those seeing only forgiving wording to report nonvoting (*REC4 & 5: RPSN=1.07 & 1.06, SE<sub>Ln</sub>=0.0765 & 0.0767, P=0.193 & 0.211*). More precisely, based on validation data, the researchers reported that the proportions of over-reporters in the two pipeline conditions were 19% and 31%, respectively, lower than the proportion of over-reporters in the control condition, but only the effect of the actual pipeline was statistically significant (*REC4 & 5: Ratio of Over-Reporting Rates=0.81 & 0.69, SE<sub>Ln</sub>=0.1719 & 0.1720, P=0.108 & 0.017*).<sup>4</sup> The subtle pipeline did not significantly

---

<sup>4</sup> Here I also report the results based on validation data, because the data not have the problems discussed in the previous section. All the respondents' turnout records are obtained, so selection bias should not be a cause for concern. The sample sizes for analysis of the over-reporting rate are fairly large – almost 800 for each experiment – so estimates should be quite precise. Note that Hanmer, Banks, and White define the over-reporting rate as the proportion of over-reporters among validated non-voters. In contrast, The calculation of self-reported turnout (and RPSN) is based on the entire sample. The effect of the actual pipeline is significant in terms of the reduction in the over-reporting rate but insignificant in terms

reduce over-reporting, probably because its wording was too soft or too equivocal to make respondents aware that being a liar represents a greater norm violation than does being a non-voter. If this is the case, pipeline users – especially those who are unable to conduct turnout validation – are in a dilemma over how to design an effective pipeline without deceiving respondents.

While both forgiving and pipeline wording aim to reduce turnout over-reporting from the social desirability aspect, there are also wording designs that focus on the memory aspect. First, a clear definition of which election is under investigation is essential to prevent confusion and misremembering. The wording of turnout questions changes over time in response to this need. The 1948 and 1952 ANES simply said, “*In this election...*” and “*...did you vote this time*”. The 1972–1976 ANES gave a slightly clearer definition – “*...in the elections this fall*”. Since the 1978 ANES, more precise time descriptions, e.g. “*...in the elections this November*”, has been consistently included in the wording. Similar amendments can be spotted in other surveys too. For example, since 2001, the turnout questions of BES have always define elections of interest by election dates, e.g. “*...the general election on June 7th*” or “*...in the European Elections on the 22nd May*”.<sup>5</sup> All these amendments went straight into the field without experimental testing, either beforehand or afterwards, so whether and how

---

of the reduction in self-reported turnout. This is mainly because the effect is ‘diluted’ when the analysis is based on the entire sample, rather than based only on validated non-voters.

<sup>5</sup> It is common for post-election surveys to ask respondents to report their turnout in more than one election. When the target election was held not long ago before the survey fieldwork, pollsters usually pay more attention to addressing social desirability bias than memory issues, so question wording almost always involves a certain type of forgiving wording, but defines the election of interest concisely, just like the examples in this paragraph. In contrast, when measuring the turnout in an election held some time ago before the survey fieldwork, the focus of attention shifts to refreshing respondents’ memories, so pollsters tend to define the election of interest as clear as possible, but forgiving wording is sometimes omitted. For example, in order to measure the turnout in the 1959 U.K. general election, the 1963 BES used almost every single word of the question to tackle misremembering, but completely abandoned tackling social desirability bias: “*Now think of the last general election, the one four years ago in the autumn of 1959, when the Conservatives were led by Macmillan and Labour by Gaitskell. Do you remember for certain whether you voted then?*”

these amendments affected respondents' self-reported turnout remains unclear. Considering that "early voting" is becoming popular, it is particularly important to know whether the use of the election date in the definition counterproductively confuses those who cast their ballots before Election Day.

A clear definition of which election is under investigation may prevent respondents from retrieving memories of wrong elections, but it may not prevent respondents from recollecting wrong memories of the right election. Actual voting and thinking about voting involve the same cognitive operations associated with whom to vote for and for what reasons. Non-voters – particularly those who usually vote and who had thought about voting in the election under investigation – may confuse memories of voting intention with memories of actual voting.<sup>6</sup> In order to avoid such confusion, [Belli, Traugott, and Rosenstone \(1994\)](#) employed a 'source-monitoring' technique to assist respondents in monitoring correct sources for their memories. They phrased the question by (1) making respondents aware of the possibility of memory confusion, (2) giving a precise definition of the election of interest, and (3) instructing respondents to recollect events that would differentiate voting from nonvoting. Note that the first two sentences of this source-monitoring wording look similar to Hanmer, Banks, and White's subtle pipeline, but they are based on different theories for different aims – one tackles memory failures, the other tackles social desirability bias.

*In talking about elections, we sometimes find that people who thought about voting actually did not vote. Also, people who usually vote may have trouble saying for sure whether they voted in a particular election. In a moment, I'm going to ask you whether*

---

<sup>6</sup> This confusion is not due to misunderstanding the turnout question. Respondents are actually clear about which election is under investigation, and aware that the question is asking about voting behaviour, rather than voting intention. However, they monitor inadequate sources of memories, and so are misled into believing they have voted, when actually they have not.

*you voted on [day of the week, month, day], which was [day(s)/week(s)] ago. Before you answer, think of a number of different things that will likely come to mind if you actually did vote this past Election Day; things like whether you walked, drove, or were driven by another person to your polling place [PAUSE], what the weather was like on the way [PAUSE], the time of day that was [PAUSE], and people you went with, saw, or met while there [PAUSE]. After thinking about it, you may realize that you did not vote in this particular election.*

Belli, Moore, and VanHoewyk (2006) experimented with this source-monitoring wording in three monthly surveys from December 1998 to February 1999, measuring turnout for the U.S. midterm election on 3rd November 1998.<sup>7</sup> These experiments, overall, suggest that using source-monitoring and forgiving wording together can improve turnout measurement significantly more than using forgiving wording alone can do (*META2: RPSN=1.22, SE<sub>Ln</sub>=0.0699, P=0.002*). Furthermore, analysis of each monthly survey separately revealed that the effect of the source-monitoring wording was statistically significant at the conventional significance level only in the February survey (*REC145: RPSN=1.38, SE<sub>Ln</sub>=0.1221, P=0.004*). These findings suggest that source-monitoring wording may not be very useful for surveys that can complete fieldwork within a few weeks after Election Day. However, on the other hand, the significant RPSN of the February survey also suggests that memory failure does not only occur years or decades after an election. Even if the election of interest was held only few months before the survey fieldwork, memory failure can still be a threat to the validity of respondents' self-reported turnout.

---

<sup>7</sup> Belli, Traugott, and Rosenstone (1994) and Belli, Traugott, Young, and McGonagle (1999) are among the first and the second to examine source-monitoring wording, but in their experiments, the effects of source-monitoring wording were confounded with the effects of another solution – face-saving options – so I review these two studies in the next section, when discussing response options.

## 2. Response options

While the wording of turnout questions has changed dramatically since the 1940s, the primitive response options – i.e. “Yes, I voted” and “No, I did not vote” – remain prevalent among election surveys nowadays. In this section, I review three types of innovative response options. Although two of them have proven effective against turnout over-reporting, neither has really taken the place of the yes-no options in most election surveys.

First, [Abelson, Loftus, and Greenwald \(1992\)](#) amended the response options for the turnout question to read “Yes, I missed out” and “No, I voted”, in order to fit with the aforementioned ‘miss-out’ wording, and avoid over-reporting resulting from acquiescence.<sup>8</sup> However, neither this ‘miss-out’ options nor the ‘miss-out’ wording yielded a significant effect on self-reported turnout (*REC27: RPSN=0.96, SE<sub>Ln</sub>=0.1388, p =0.629*).

The second type of response option is closely related to source-monitoring wording – both instruct respondents to avoid memory confusion. Considering that non-voters who recollect correct memories may still be reluctant to admit nonvoting due to social desirability, [Belli, Traugott, and Rosenstone’s \(1994\)](#) redesigned response options as a precaution:

*# I did not vote in the [Month Day] election*

*# I thought about voting this time but didn't*

---

<sup>8</sup> A potential drawback of miss-out wording and options is that, when respondents answer ‘Yes’, some of them may actually mean “Yes, I voted”, rather than “Yes, I missed out”. To avoid misunderstandings, [the 1989 ANES pilot study](#) used a follow-up question to double-check respondents’ answers, and found that all the respondents who said ‘Yes’ meant that they missed out on voting, and those who said ‘No’ meant that they voted (see *REC120*).

*# I usually vote but didn't this time*

*# I am sure I voted in the [Month Day] election*

The first and last options are conventional yes-no options with a clearer definition of the election under investigation. The second and third options are so-called ‘face-saving options’ (Belli, Traugott, Young, and McGonagle 1999). They are a repeat of the first two sentences of source-monitoring wording, but serve a different purpose. Source-monitoring wording alerts respondents to the possibility of memory confusion; the face-saving options mainly aim to be two socially acceptable excuses for nonvoting for countering social desirability bias (Duff, Hanmer, Fark, and White 2007).

Six studies have conducted 14 experiments on face-saving options (ANES 2004a; 2004b; Belli, Moore, and VanHoewyk 2006; Duff et al. 2007; Kritzinger, Schwarzer, and Zeglovits 2012; Zeglovits and Kritzinger 2014). My meta-analysis of these experiments suggests that face-saving options can significantly raise the proportion of self-reported non-voters by 40% (*META3: RPSN=1.40, SE<sub>Ln</sub>=0.0502, P<0.001*).<sup>9</sup>

The third type of response option was inspired by a pre-election survey question that measures respondents’ intentions to vote in a forthcoming election. Instead of forcing respondents to choose between “Yes, I voted” and “No, I didn’t”, Kritzinger, Schwarzer, and Zeglovits (2012) designed a 4-point scale to take account of response uncertainty resulting from blurred memories about turnout:

*# I am sure I did not vote in the [Election] in [Month Year]*

---

<sup>9</sup> Belli and his colleagues (Belli et al. 1999; Belli, Moore, and VanHoewyk 2006; Belli, Traugott, and Rosenstone 1994) and Waismel-Manor and Sarid (2011) experimented on a joint design – source-monitoring wording plus face-saving options – in eight surveys. My meta-analysis suggests that this joint design can significantly increase the proportion of self-reported non-voters (*META4: RPSN=1.27, SE<sub>Ln</sub>=0.0715, P<0.001*).

# I am not sure if I voted but I presumably did not

# I am not sure if I voted but I presumably did

# I am sure that I voted in the [Election] in [Month Year]

These turnout-scale options require respondents not only to answer whether they voted, but also to assess how sure they are about their answers. Like source-monitoring wording, this design aims to address memory failures, but it works in a relatively passive manner. Rather than try to refresh respondents' memories, turnout-scale options evade the problem by allowing respondents to report uncertainty over their answers, but this design produces another problem for data analysis – how to categorise two middle options?<sup>10</sup> [Kritzinger, Schwarzer, and Zeglovits \(2012\)](#), and [Zeglovits and Kritzinger \(2014\)](#) grouped the first two options together as non-voters, and grouped the last two options together as voters. My meta-analysis of their findings from eight surveys yields an *RPSN* of 1.33, suggesting that the turnout scale can produce a significantly better turnout rate estimate than can the yes-no options (*META5: SE<sub>Ln</sub>*=0.0770, *P*<0.001). I also reanalyse their data using two different coding rules. The rule that counts only those who chose the last option as voters yields an *RPSN* of 1.52 (*META5-1: SE<sub>Ln</sub>*=0.0687, *P*<0.001); the other rule that treats the two middle options as missing values yields an *RPSN* of 1.20 (*META5-2: SE<sub>Ln</sub>*=0.0804, *p*=0.012). Different coding rules lead to different effect sizes. Further research needs to explore an optimal coding rule for the use of turnout-scale options.<sup>11</sup>

---

<sup>10</sup> Statistical methods for ordinal variable analysis can save the trouble of categorisation, but it is sometimes inevitable to have a cut-off point for categorisation, particularly when the research aim is to produce a simple estimate of turnout rate.

<sup>11</sup> There is a fourth type of response option designed in reaction to the increasingly common 'early voting'. Because the turnout question defines the election of interest by Election Day, voters who casted ballots before Election Day may misunderstand the wording of the turnout question, and may simply give the answer 'No', though what they mean is "I voted not on but before Election Day". To address this problem, [Holbrook and Krosnick \(2013\)](#) and the [2008 ANES Time Series Study](#) combined the

### 3. Questioning format

Solutions that have been reviewed thus far aim to improve the direct question about turnout. No matter how well-designed the wording and options are, respondents are still required to explicitly report whether they voted or not. Such a direct-questioning technique is notorious for its susceptibility to social desirability bias, since respondents are required to reveal what they may prefer to conceal. A more fundamental solution is to change the question format. Two indirect-questioning techniques – randomised-response and item-count – have been tested in the field. These techniques allow respondents to ‘encrypt’ their truthful answers, and thus weaken the motives for over-reporting. Data collected by these techniques can be decrypted at the aggregate level, so that researchers can still estimate turnout rates.

A simple example of a randomised-response technique is to instruct respondents to flip a coin secretly; if their coins come up heads, they answer the question about whether they voted (yes/no); if the coins come up tails, they simply say ‘No’ (Warner 1971). In this procedure, giving a negative response does not necessarily indicate nonvoting, so respondents can admit nonvoting by simply saying no, and need not worry about being revealed.

Despite being theoretically sound, the randomised-response technique, as a solution to turnout over-reporting, is very disappointing in practice. Holbrook and Krosnick (2010a) and Locander, Sudman, and Bradburn (1976) reported eight turnout estimates based on randomised-response techniques, but none of them was closer to official turnout rates than were estimates based on the direct-questioning technique.

---

options of the turnout question and the options a follow-up question about voting methods (see REC22 and 77). That design aims to tackle turnout under-reporting, and so is beyond the scope of this paper.

Some estimates were even higher than 100% (see *REC13, 14, 16, 17, 18, 20, 21 and 118*).<sup>12</sup> Those nonsensical estimates indicated that many respondents did not follow the instructions properly. The procedure of the randomised-response technique has been criticised for being too ‘mind-boggling’ for respondents to follow ([Droitcour et al. 1991: 188](#)). The findings from the two turnout studies appear to be consistent with the criticism, and cast more doubts on the practicability of the randomised-response technique.

The item-count technique, by contrast, is less mind-boggling. This technique also involves assigning respondents to different groups at random, but it is the researcher who conducts the randomisation, so the technique can be deployed without even making respondents aware of it. Specifically, the technique presents a group of respondents with a list of items and asked ‘*how many*’ items they would answer in the affirmative, and presents the rest of respondents with the same list of items plus a key item (e.g. “I voted in the election held on...”), and asked the same question ([Miller 1984](#)). Respondents are not required to point out *which* items they answer in the affirmative and which in the negative, so they can admit to nonvoting by just excluding turnout from their item counts, and need not worry about being revealed.

The item-count technique is much more successful in the field than the randomised-response technique. [Holbrook and Krosnick’s \(2010b\)](#) and [Mircea and Postelnicu’s \(2013\)](#) applied the item-count technique to turnout measurement and, overall, yielded a 14% increase in the proportion of self-reported non-voters, in

---

<sup>12</sup> The RPSNs of experiments that produced nonsensical estimates are negative. Because the logarithm of a negative value is not defined, significance tests and meta-analysis of those RPSN are inapplicable. An alternative approach is to examine the *differences* between the treatment and control groups’ turnout rate estimates. A meta-analysis based on this approach shows that the randomised-response estimate was significantly higher than its direct-questioning counterpart by 25.4 percentage points ( $SE=0.0572$ ,  $P<0.001$ ;  $H_a:RPSN<1$ ) – i.e. the randomised-response technique in those two studies did not improve but significantly worsened the estimates of turnout rates.

comparison to the use of the direct-questioning technique (*META7*:  $RPSN=1.14$ ,  $SE_{Ln}=0.0779$ ,  $P=0.041$ ). The strength of the item-count technique, however, comes at a price – the potential loss of statistical efficiency due to the nature of indirect-questioning. In those two experimental studies, the standard errors of item-count estimates were, on average, 0.07. Had those estimates been obtained by a direct-questioning measure, the average standard error would have been down to 0.02 – i.e. the uncertainty about turnout estimates would have been smaller.<sup>13</sup> Advanced methods for more efficient use of the item-count technique are available (e.g. [Aronow, Coppock, Crawford, and Green 2015](#); [Corstange 2009](#); [Droitcour et al. 1991](#); [Glynn 2013](#); [Blair and Imai 2012](#)), but none has been applied to turnout measurement.

Additionally, it is interesting to make a comparison between the item-count and pipeline techniques, since they represent two exactly opposite approaches to eliciting truthful self-reports. Based on the assumption that confidentiality promotes candour, the item-count technique persuades respondents not to over-report by keeping their self-reported turnout strictly confidential. In sharp contrast, the pipeline technique dissuades respondents from over-reporting by convincing them that nothing is confidential and lying is detectable (by, for example, checking survey responses against official turnout records). Both techniques have proven effective against turnout over-reporting. Further research needs to assess which one is more effective.

#### 4. Question order

Shifting the focus from question design to questionnaire design, another category of solutions to over-reporting has paid more attention to the interaction between the turnout question and other questions on the questionnaire. For example, the 1972

---

<sup>13</sup> I calculated counterfactual standard errors by  $[p(1-p)/n]^{1/2}$ .

Current Population Survey (U.S. Bureau of the Census 1973: 8) moved the turnout question from the front to the back of the “presumably less sensitive” question on voter registration, and that yielded a 6% increase in the proportion of self-reported non-voters (REC33). In contrast, the 1984 ANES moved the turnout to near the beginning of the questionnaire, in order to avoid any undesirable order effect; however, 43.6% of non-voters still over-reported turnout (REC167) (ANES 1984; Presser, Traugott, and Traugott 1990: 3). More recently, Hanmer, Banks, and White (2014) found that their actual pipeline question about turnout in an election had a desirable, though insignificant, order effect on a latter question about the turnout in another election. (REC7:  $RPSN=1.18$ ,  $SE_{Ln}=0.1126$ ,  $P=0.083$ ; also see REC6 for the subtle pipeline, which did not exert such an effect on the latter turnout question:  $RPSN=0.98$ ,  $SE_{Ln}=0.1272$ ,  $P=0.575$ ).

These three studies, however, are not adequate to explore order effects on turnout over-reporting. The experiment results of the 1972 Current Population Survey were confounded by the nonresponse rate, which was twice as high in the experimental group as in the control group (U.S. Bureau of the Census 1973: 8). The 1984 ANES did not really experiment on the relocation of the turnout question (i.e. no control group). Hanmer, Banks, and White’s (2014) experiments were mainly designed for testing pipeline wording rather than question order.

Presser (1990) conducted two comparatively well-designed experiments to examine order effects on turnout over-reporting. In his first experiment, the turnout question was preceded by a question about the location of the polling place. Arguably, the polling place is “information that must be known in order to vote, but is unlikely to be known by non-voters” (Presser 1990: 558). The polling-place question is supposed

to help respondents recall memories that would differentiate voting from nonvoting, and also to worry non-voters about being perceived as liars, if they intend to over-report but then find themselves do not know the polling place. In this regard, Presser's design seems to be an early application of the source-monitoring and pipeline techniques.

*[Pre-question] Do you happen to know the location of the polling place where people in your neighbourhood go to vote on Election Day in November? Where is that?*

*[Target] In talking to people about elections, we often find that a lot of people were not able to vote in a particular election because they weren't registered, they were sick, or they just didn't have time. How about you, did you vote in the Presidential elections that were held last November?*

In addition to Presser, the 2004 Taiwan's Election and Democratisation Study (Wu 2006) also experimented with the polling-place question, but neither of them found a significant increase in the proportion of self-reported non-voters ( $META8: RPSN=1.05, SE_{Ln}=0.0147, P=0.239$ ). I examine these studies further and identify two problems of the polling-place question. First, a large proportion of self-reported non-voters correctly pointed out the polling place (65.6% and 80.1% in Presser's and Wu's studies, respectively). It seems that the polling place is not something "unlikely to be known by non-voters", so asking about it before turnout cannot assist non-voters in monitoring adequate memory sources. Second, around a quarter of the respondents who could not point out the polling place said they had voted (26.8% and 23.0% in Presser's and Wu's studies, respectively). This finding suggests that the polling-place question is not an adequate pipeline to worry over-reporters about being perceived as liars.<sup>14</sup> Respondents in the two studies might not recognise the polling-place question as a

---

<sup>14</sup> Absentee voting was not a satisfactory explanation of why those self-reported voters did not know the polling place, since there is no absentee voting in Taiwan.

means for verifying their self-reported turnout. Merely placing two questions consecutively in the questionnaire may not be enough to create an order effect. Manifesting the purpose of the pre-question may be essential as well.

In Presser's second experiment, the turnout question was preceded by a question about turnout habits:

*[Pre-question] Thinking back over all the elections since you were first eligible to vote, would you say you have voted most of the time, some of the time, or rarely?*

*[Target] In talking to people about elections, we often find that a lot of people were not able to vote in a particular election because they weren't registered, they were sick, or they just didn't have time. How about you, did you vote in the Presidential elections that were held last November?*

The underlying idea is identical to that of the face-saving response option, “*I usually vote but didn't this time*”, which makes a socially acceptable excuse for nonvoting.<sup>15</sup> However, unlike the face-saving option, Pressers' face-saving question failed to persuade more respondents to admit nonvoting; instead, it decreased the proportion of self-reported non-voters by 11%, though it was not statistically significant ( $REC2: RPSN=0.89, SE_{Ln}=0.1034, P=0.136, H_a: RPSN<1$ ). Presser (1990: 592) conjectures that a single face-saving question may be insufficient to counter social desirability bias. Even worse, he suspects that the use of a face-saving question may counterproductively increase the impulse to over-report for those who want to appear consistent after reporting they usually voted. Furthermore, I also consider respondents' unawareness of the purpose of the face-saving question to be another plausible explanation for the

---

<sup>15</sup> For the sake of convenience, I dub Pressers' second pre-question a ‘face-saving’ question, but this question should be helpful to prevent memory confusion too, because it requires respondents to recall their experience of voting in other elections. The other face-saving questions discussed later in this section also share this feature.

failure of Presser's second experiment. [Abelson, Loftus, and Greenwald \(1992\)](#) and [Holbrook and Krosnick \(2013\)](#) provide some support for this explanation. Both of them attempted to reduce turnout over-reporting by certain forms of face-saving questions but, unlike Presser, they made extra efforts to manifest the purpose of pre-questions, and achieved better results.

Holbrook and Krosnick converted Belli and his colleagues' face-saving response options into two face-saving questions (the first one is conceptually the same as Presser's). Their design used a preamble to tell respondents *implicitly* that, before the turnout question, there were two pre-questions allowing them to make socially acceptable excuses for nonvoting, hence no need to over-report.

*In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, they didn't have time, or something else happened to prevent them from voting.*

*[Preamble] And sometimes, people who USUALLY vote or who PLANNED to vote forget that something unusual happened on Election Day this year that prevented them from voting this time. So please think carefully for a minute about the election held on November 7, and past elections in which you may have voted and answer the following questions about your voting behaviour.*

*[Pre-question 1] During the past six years, did you usually vote in national, state, and local elections, or did you usually NOT vote?*

*[Pre-question 2] During the months leading up to the election held on November 7, did you ever plan to vote in that election, or didn't you plan to do that?*

*[Target] In the election held on November 7, did you definitely vote in person on Election Day, definitely mail in a completed absentee ballot before Election*

*Day, definitely not vote, or are you not completely sure whether you voted in that election?*

Compared to the conventional turnout question (forgiving wording with yes-no options), this design substantively and significantly increased the proportion of self-reported non-voters in the experiment by 46% (*REC22: RPSN=1.46, SE<sub>Ln</sub>=0.1092, P<0.001*).<sup>16</sup> Apparently, this is a much more successful design than Presser's.

As for Abelson, Loftus, and Greenwald's study, the target question was about the turnout for the 1988 U.S. senatorial primary election. Given the fact that many non-voters in the primary election voted in the subsequent Presidential election, the researchers placed a question before the target one to allow the non-voters in the primary election to report their participation in the Presidential election, so that respondents could satisfy a minimal norm of good citizenship and feel less social desirability pressure when answering the target question. Most importantly, the purpose of the pre-question was implicit in the preamble of the question: "no one should be expected to vote all the time".

*There were two elections last fall—the primary election on September 20<sup>th</sup> and the Presidential election on November 8<sup>th</sup>.*

*[Preamble] Most people aren't able to get to vote in every election. How about you —*

---

<sup>16</sup> Holbrook and Krosnick's (2013) design aims to address two drawbacks of face-saving response options. First, the options – "did not vote", "thought about voting this time but didn't" and "usually vote but didn't this time" – are not mutually exclusive, so Holbrook and Krosnick split them into three questions. Second, the voting option – "I am sure I voted in the [Month Day] election" – neglects 'early voting', so Holbrook and Krosnick drew a distinction between "voted in person" and "voted by mail". The 2008 ANES refined these two options further – "voted in person at a polling place on Election Day", "voted in person at a polling place before Election Day", "voted by mailing a ballot to elections officials before the Election" and "voted in some other way". Then, the 2008 ANES experimented with all these designs (the pre-questions plus additional options), but found virtually no improvement to turnout measurement, compared to a question with simply forgiving wording and face-saving options (*REC77: RPSN≈1.000, SE<sub>Ln</sub>=0.1011, P=0.499*).

*[Pre-question] did you vote in BOTH the September and November elections?*

*[Target] Let me make sure I haven't confused you. Specifically, did you vote in the primary election last September 20<sup>th</sup>?*

This design increased the proportion of self-reported non-voters by 27% (*META9: RPSN=1.27, SE<sub>Ln</sub>=0.1817, P=0.095*; the control group saw neither the preamble nor the pre-question). Compared to Presser's design, this one is more successful – its effect size is larger, in the right direction, and significant at the  $p=0.1$  level, though not at the conventional level.

Abelson, Loftus, and Greenwald also experimented on another face-saving question. That experiment again shows the importance of respondents' awareness about the purpose of the pre-question, though this time the importance is demonstrated in the opposite way. The authors did not give any clue about the purpose of the pre-question, and the experiment resulted in failure: the proportion of self-reported non-voters did not increase but slightly decreased by 3%, though the difference was insignificant. (*REC24: RPSN=0.97, SE<sub>Ln</sub>=0.1133, P=0.384, H<sub>a</sub>: RPSN<1*; the control group saw the forgiving wording and the target question but not the pre-question). Note that the wording before the pre-question was nothing more than conventional forgiving wording. It did not mention anything about the purpose of the pre-question.

*In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. How about you –*

*[Pre-question] Thinking back over the last four national elections, that is, the Presidential elections of 1980 and 1984, and the Congressional elections of 1982 and 1986, did you vote in any of these elections?*

[Target] Did you vote in the 1986 elections for United States Congress last November?

The review in this section shows that designs making extra efforts to manifest the purpose of pre-questions outperformed those not doing so. This finding implies that making the question order meaningful to respondents is crucial in capitalising on order effects to address turnout over-reporting. To confirm this finding, future experiments should divide subjects into at least three groups: one is presented with (1) a preamble that manifests the meaning of the question order, (2) a pre-question, and (3) a turnout question; another group is asked the questions without the preamble; the other is a control group, seeing merely the turnout question. This will allow us to disentangle the role of the preamble from the role of the pre-question in the process of creating order effects, and lead us to a better questionnaire design for reducing over-reporting.

## 5. Survey mode

A higher degree of response privacy spares respondents the embarrassment of admitting to nonvoting. One way to do so is to change the questioning format, such as employing the item-count technique. Another is to change the mode of the survey interview. Stocké (2007) experimented with a partial-mode-switch design. In the experiment, all interviews started with a face-to-face mode. After some questions, respondents were randomly split into two groups: one continued with the interviewer-administered mode; the other switched to a self-administered mode and completed the rest part of the questionnaire in private, including a turnout question. The results are impressive – respondents in the self-administered mode were 2.04 times as likely to report nonvoting as those who had to reveal their answers to an interviewer ( $REC37: RPSN=2.04, SE_{Ln}=0.1759, P=0.000$ ).

As discussed previously, being less cognitively demanding is a plausible reason why past studies found the item-count technique worked much better than the randomised-response technique. Compared to the item-count technique, the partial-mode-switching design places even less cognitive burden on respondents. It does not complicate the turnout question at all – respondents still answer a conventional turnout question – whereas the item-count technique makes the question slightly more complicated due to the special questioning format. Moreover, when switching to a self-administered mode, an interviewer hands the questionnaire to a respondent and maintains sufficient distance from it, so the respondent can literally ‘see’ an enhancement of response privacy. In contrast, the way that the item-count technique ensures response privacy is not always obvious to respondents. These two advantages to some extent explain why the experiment of the partial-mode-switching design resulted in a higher RPSN than the item-count technique. When both solutions are viable, the partial-mode-switching design seems to be a better choice than the item-count technique.

Instead of switching mode during the interview, some studies have compared surveys that were administered by completely different modes (henceforth, full-mode-switch). The literature on mode effects has suggested that the presence of an interviewer compromises response privacy and intensifies social desirability bias (Groves et al. 2004: 157). In this regard, a survey mode that minimises the role of the interviewer is supposed to perform better in turnout measurement, but experiments have suggested otherwise: instead of improving turnout measurement, telephone and web surveys tend to result in smaller proportions of self-reported non-voters than face-to-face surveys (*META10 [TEL vs FTF]: RPSN=0.87,  $SE_{Ln}=0.0557$ ,  $P=0.008$ ,  $H_a$ :*

*RPSN*<1; *REC74* [Web vs FTF]: *RPSN*=0.86, *SE<sub>Ln</sub>*=0.0709, *P*=0.016, *H<sub>a</sub>*: *RPSN*<1) (ANES 1982; 1984; 1996; 2000; 2012; Rogers 1976).<sup>17</sup>

It is worth stressing that those experiments applied different modes to different experimental groups throughout not only interviews but also many other stages of survey administration. Interviewer involvement was not the only factor that could have affected experiment results. Other factors, such as the means, channels or strategies for recruiting sampled persons, also varied from one condition (mode) to another in those experiments. Some experiments even used different sampling designs for different modes. This raises the possibility that web and telephone surveys in those experiments produced worse turnout rate estimates not because of the failure to reduce over-reporting, but because of the problem with sample representativeness.

I examine this issue in two aspects – sampling bias and unit nonresponse bias. In the sampling aspect, I scrutinise the sampling designs of those experiments and find that under-sampling of actual non-voters is not a plausible explanation for why web and telephone surveys produced poorer turnout estimates. In the experiment that compared web and face-to-face surveys, although the samples for the two modes were drawn separately, both were probability samples drawn from the same frame and representative of the same population (ANES 2012: 23–7). Given these similarities,

---

<sup>17</sup> The 1998 ANES also compared face-to-face with telephone modes, but did not find a significant mode effect on turnout measurement (*REC166*: *RPSN*=0.97, *SE<sub>Ln</sub>*=0.0748, *P*=0.326, *H<sub>a</sub>*: *RPSN*<1). My meta-analysis does not include this experiment, because its assignment of respondents was not random. Only those who were unable to be interviewed by phone were assigned to the face-to-face mode. Locander, Seymour, and Norman (1976) conducted mode experiments too. They found that telephone interviews and paper-based self-administered interviews produced better turnout measurements than face-to-face interviews, but the differences were not statistically significant (*REC163* & *164*). The authors also compared a self-administered survey with a telephone survey, and found that the latter persuaded more respondents to admitting to nonvoting than the former did, but the difference was not statistically significant (*REC165*). My meta-analysis cannot include these experiments, because the study did not report enough information for me to calculate *RPSN*. However, this study should not make a big difference to my meta-analysis, because of its small sample size ( $n \leq 80$  in each mode).

there is little ground for believing that the web survey under-sampled non-voters (or equivalently over-sampled voters) any more than the face-to-face survey did. As for the five experiments that compared telephone and face-to-face surveys, two of them did use different sampling methods and frames for different modes (the 1982 ANES Merged Methods Comparison Project, and the 2000 ANES Time Series Study). However, even if I excluded those two experiments from the analysis, the results remain largely unchanged (*META10-1*:  $RPSN=0.85$ ,  $SE_{Ln}=0.0585$ ,  $P=0.003$ ,  $H_a: RPSN<1$ , cf. *META10*). Therefore, sampling bias cannot explain why the web and telephone survey produced a worse turnout estimate.

In the unit nonresponse aspect, there were indeed signs that the web and telephone surveys produced worse turnout rate estimates because of having greater difficulty in recruiting sampled non-voters to participate in interviews. One sign of this is that the response rate of the web survey was substantially lower than that of the face-to-face survey (2% vs 38%; see page 31 of the 2012 ANES Post-Election codebook). Another sign is that many respondents in the experiments were initially assigned to the telephone mode but in the end interviewed face-to-face for various reasons, such as they were hard of hearing or did not have a telephone. These respondents significantly outnumbered those who were assigned to the face-to-face mode but interviewed by phone (14% vs 3% in the 1984 ANES Time Series Study; 4% vs 1% in the 1996 ANES Time Series Study; the other three studies did not release such information for comparison).<sup>18</sup> Given these signs and considering the positive correlation between election and survey participation (Voogt and Saris 2003), under-recruitment of sampled

---

<sup>18</sup> In five experiments that compared telephone and face-to-face surveys, only the 1996 and 2000 ANESs reported the response rate of each mode. According to the reports, the differences in response rates between the modes were less than 3 percentage points (see page 9 of the 1996 ANES codebook and pages 11–12 of the 2000 ANES codebook). However, these reported response rates did not take account of the fact that some respondents were assigned to one mode but interviewed in another mode.

non-voters remains a possible reason for the worse turnout estimates from web and telephone surveys.

Overall, there are two possible explanations for why web and telephone surveys in past experimental studies produced worse turnout estimates than face-to-face surveys. One explanation is the difficulty in recruiting sampled non-voters to participate in interviews; the other is the failure to prevent interviewed non-voters over-reporting their turnout. Regardless of which explanation is true, from a practice perspective, replacing the face-to-face mode with a web or telephone mode is not a good solution to turnout over-reporting.

### **Overall effects of the solutions**

After reviewing each of the solutions separately, I take an overall view, discussing them collectively. Table 2 summarises the results of the experiments. Take source-monitoring wording for example. There are three experiments on this solution (Column *Single-n*); one of them reported that the wording significantly improved turnout measurement (Column *Single-Sig.n*). A meta-analysis of the three experiments shows that the wording significantly increased self-reported non-voters by 22% (Column *Single-RPSN*). Additionally, 11 experiments have examined source-monitoring wording jointly with other solutions, so there are 14 experiments in total that have examined source-monitoring wording (Column *Inclusion-n*); eight of them reported that the wording significantly improved turnout estimates (Column *Inclusion-Sig.n*). A meta-analysis of these 14 experiments yields an RPSN of 1.25 (Column *Inclusion-RPSN*). Note that this effect size does not necessarily result from source-monitoring wording, since the experiments examined multiple solutions simultaneously. Therefore, the following analysis focuses on the column for *Single-RPSN*.

**Table 2. Effects of solutions to turnout over-reporting**

	Treatment	Number of Experiments				RPSN		a
		Single		Inclusion		Single	Inclusion	
		Sig.n	n	Sig.n	n	RPSN(SE <sub>Ln</sub> )	RPSN(SE <sub>Ln</sub> )	
Question Design	Wording	2	9	9	21	1.14 (0.04)***	1.19 (0.04)***	
	└ Forgiving	1	4	1	5	1.17 (0.07)*	1.13 (0.07)*	
	└└ Syntactic Manoeuvre	0	0	0	1	-	0.96 (0.14)	c
	└└ Sematic Manoeuvre	1	2	1	2	1.14 (0.14)	1.14 (0.14)	
	└ Pipeline	0	2	0	2	1.07 (0.05)	1.07 (0.05)	d
	└ Source-Monitoring	1	3	8	14	1.22 (0.07)**	1.25 (0.05)***	
	Option	12	22	18	32	1.37 (0.05)***	1.32 (0.04)***	
	└ Syntactic Manoeuvre	0	0	0	1	-	0.96 (0.14)	c
	└ Face-Saving	9	14	15	23	1.40 (0.05)***	1.34 (0.04)***	
	└ Turnout Scale	3	8	3	8	1.33 (0.08)***	1.33 (0.08)***	e
Format	└ Indirect Questioning	2	13	2	13	0.86 (0.15)	0.86 (0.15)	
	└└ Randomised-Response	2	13	2	13	0.86 (0.15)	0.86 (0.15)	
	└└ Item-Count	0	8	0	8	0.49 (0.16)	0.49 (0.16)	f
	└ Item-Count	2	5	2	5	1.14 (0.08)*	1.14 (0.08)*	
Questionnaire	Order	1	10	2	12	1.03 (0.04)	1.08 (0.05)†	
	└ Forward Placement	-	1	-	1	-	-	g
	└ Preceding Question	1	9	2	11	1.03 (0.04)	1.08 (0.05)†	
	└└ Voter Registration	-	1	-	1	-	-	g
	└└ Pipeline	0	2	0	2	1.08 (0.10)	1.08 (0.10)	
	└└ Polling Place	0	2	0	2	1.05 (0.06)	1.05 (0.06)	
	└ Experience/Intention	1	4	2	6	1.02 (0.09)	1.10 (0.09)	
Survey	Mode	1	9	1	9	0.94 (0.08)	0.94 (0.08)	
	└ Partial Mode Switch	1	1	1	1	2.04 (0.18)***	2.04 (0.18)***	
	└ Full Mode Switch	0	8	0	8	0.87 (0.04)	0.87 (0.04)	
	└└ TEL. vs. FTF.	0	6	0	6	0.87 (0.06)	0.87 (0.06)	h
	└└ Web vs. FTF.	0	1	0	1	0.86 (0.07)	0.86 (0.07)	
└ Self-Admin. vs. FTF.	-	1	-	1	-	-	g	

**NOTES:** (a) The columns for ‘Single’ summarise experiments that examined only one treatment (the one in the first column). The columns for ‘Inclusion’ summarise experiments that simultaneously examined multiple treatments (including the one in the first column). (b) ‘n’ means the number of experiments examining the treatment in the first column. ‘Sig.n’ means the number of experiments yielding statistical significant results (P-value<0.05). ‘SE<sub>Ln</sub>’ means the standard error of log-RPSN. (c) The treatment has never been experimented on alone. (d) Validation data showed that the pipeline technique significantly reduced over-reporting. (e) Estimates of the treatment ‘Uncertainty Scale’ are based on coding rules that count the first two options as nonvoting and the last two options as voting. (f) Estimates of the treatment ‘randomised-response’ are based on four out of eight experiments, since the other four experiments yielded unreasonable turnout estimates that make statistical calculation infeasible. (g) The experiments did not report enough information to compute RPSN. These experiments are not used in other calculations either. (h) Estimates of the treatment ‘TEL vs F2F’ are based on five out of six experiments, since one experiment did not report enough information to compute RPSN. \*\*\*P-value (1-tailed)<0.001, \*\*<0.01, \*<0.05, †<0.1. H<sub>0</sub>: RPSN≤1, H<sub>a</sub>: RPSN>1.

Most efforts to address turnout over-reporting have looked at the question design. A majority of the experiments have refined and tested the wording, options or format of the turnout question. Attempts to change the response option have been very successful, leading an overall 37% rise in the number of self-reported non-voters.

Attempts to rephrase the question wording have also succeeded in increasing the self-reported non-voters by 14%. In contrast, attempts to change questioning formats have adversely reduced self-reported non-voters by more than 14%. However, that is entirely due to the dismal failure of the randomised-response technique. Experiments on the item-count technique have been successful, increasing self-reported non-voters by 14%.<sup>19</sup>

The number of experiments on questionnaire and survey designs is relatively small, and the results in general are not very encouraging. Attempts to manipulate order effects have not significantly raised numbers of self-reported non-voters (RPSN=1.03). Attempts to manipulate mode effects have even resulted in an undesirable tendency to reduce self-reported non-voters (RPSN=0.94). Stocké's partial-mode-switch design is a notable exception. By allowing face-to-face survey respondents to answer the turnout question in private, his design increased self-reported non-voters by 104% – an impressive improvement to turnout measurement. Other attempts to replace face-to-face interviews with telephone or web interviews have worsened turnout measurement, causing a reduction of 13% in the numbers of self-reported non-voters.

Taken together, research on different dimensions of designs has had varying degrees of success in tackling turnout over-reporting. On the question design, research has proven fruitful, offering several effective solutions, including forgiving and source-monitoring wording, face-saving and uncertainty-scale options, and the item-count technique.<sup>20</sup> As for questionnaire design, a few experiments have demonstrated the

---

<sup>19</sup> Table 2 shows that many wording and option experiments tested multiple solutions simultaneously (*Single-n vs Inclusion-n*). This is mainly because some wording and options were designed together and thus experimented on together (e.g. Abelson, Loftus, and Greenwald's miss-out design, and Belli, Traugott, and Rosenstone's source-monitoring wording and face-saving options).

<sup>20</sup> Turnout validation data suggest that pipeline wording is also an effective solution.

possibility of capitalising on question order to reduce over-reporting, but there is still no effective design to maximise that desirable order effect. Among survey designs, only partial-mode switch has proved successful. Other attempts at a mode effect, though unsuccessful, were not in vain. They have at least cautioned against potentially adverse effects of telephone and web surveys on turnout measurement.

## **Discussion and conclusion**

Valid survey measurement of voter turnout is fundamental to electoral studies. A problem that detracts from measurement validity is non-voters' tendency to falsely report having voted (i.e. turnout over-reporting). Over recent decades, survey methodologists have developed and experimented with various solutions to over-reporting. Research to date has accumulated a wealth of knowledge, and so it is worth doing a systematic review. In this paper, I conducted a meta-analysis to discuss all the solutions together, aiming to make practical and methodological contributions to the measurement of turnout.

In the practical aspect, I produced a catalogue of solutions to turnout over-reporting for easy reference. With the catalogue to hand, anyone who intends to measure turnout by survey will quickly grasp some idea of what solutions are available, how they work, and which ones may meet actual needs. For example, when survey fieldwork cannot be scheduled to be completed shortly after Election Day, pollsters have to prepare for over-reporting caused by misremembering. Checking the solution catalogue, they will immediately find several choices (i.e. those marked with 'M' in Figure 1). If the aim is to refresh respondents' memories about turnout, pollsters may use source-monitoring wording or a pre-question. However, if the questionnaire has no space for lengthy wording or an additional question, pollsters may opt for the turnout-

scale option – it is a less cumbersome design and can make self-reported turnout less misleading, though it does not directly tackle the problem of misremembering.

The catalogue is not merely a list of unorganised elements; instead, I categorised solutions to turnout over-reporting according to which parts of a survey they modify to tackle a problem. This makes it convenient to use multiple solutions together and profit from synergy. Consider the above example again. The three aforementioned solutions fall into three different categories – wording, option and order. Pollsters looking at the catalogue will easily notice that there is no need to make a choice between those solutions, as each one can be applied to part of the survey, and the three together may provide better protection against over-reporting.

In the methodological aspect, I identified a number of issues in need of further investigation. First, there are solutions that have seen widespread use but lack a proper examination. Forgiving wording is the most important one of this kind. Nowadays, most surveys use a certain form of forgiving wording to phrase their turnout questions, but the effects remain unclear. Experiments are needed to determine whether forgiving wording really deserves space in a questionnaire and, if yes, how to rephrase it to maximise its effect.

Another widespread practice that needs proper examination is the use of an election date to define the election of interest. A definition with a precise date is supposed to prevent respondents retrieving memories about wrong elections, and hence avoid turnout over-reporting. The problem is whether the definition is over-specific, and consequently confuses those who cast ballots before Election Day. This problem requires urgent attention, as early voting is becoming popular.

The second issue in need of further investigation is how to capitalise on the arrangement of questions to tackle turnout over-reporting. Several studies have preceded their turnout questions with one or two questions designed to refresh respondents' memories of voting or to relive the social desirability concern. My meta-analysis suggests that merely placing questions consecutively on a questionnaire may be insufficient to create an order effect; making the question order meaningful to respondents may be essential as well. More evidence is needed to confirm this finding and guide us towards a better questionnaire design to reduce turnout over-reporting.

The third issue concerns how turnout over-reporting relates to survey administration. Past studies have argued that a survey that minimises the role of interviewers is better for eliciting truthful self-reports from respondents, but my meta-analysis suggests otherwise – turnout measurement by face-to-face surveys was in general better than that by telephone and web surveys. Does this mean that minimising the role of interviewers is not a valid solution to turnout over-reporting? Alternatively, is it possible that telephone and web surveys face other problems in measuring turnout, such as sample representativeness? Further studies are needed to investigate these possibilities.

Finally, yet importantly, almost all the studies proposed and tested one single solution to turnout over-reporting at a time, and most of them proved their proposed solutions workable, but whose is better and under what circumstances? To address this question, future studies should experimentally compare existing solutions. For example, it would be interesting to compare the item-count and pipeline techniques, since they are inspired by opposite rationales, yet both have proven effective. The item-count technique embodies the idea that confidentiality promotes candour, while the pipeline

technique leads respondents to believe that nothing is confidential and thus honesty is the best policy. A careful comparison of these two techniques could gather information, not only for better application of these techniques themselves, but also for the development of better solutions.

A major limitation of this paper is the use of the self-reported turnout as the criterion for assessing solutions to over-reporting. In fact, this is a limitation of most studies in this field, so the choice of assessment criteria is inherently limited for my meta-analysis. There are indeed good reasons to believe that assessments based on self-reported turnout can largely reflect the true effectiveness of solutions to turnout over-reporting. However, if reliable validation data are obtainable, future studies should still use validated turnout as the main criterion for making more precise and accurate assessments.

## Supplementary materials

Supplementary materials are freely available online at: <https://goo.gl/hTrESH>

## References

- Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. "Attempts to Improve the Accuracy of Self-Reports of Voting." In *Questions about Questions*, ed. Judith M. Tanur, 138-53. New York: Russell Sage Foundation.
- Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3(1): 43-66.
- Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Voting Over-Reports? Contrasts of Over-Reporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17(4): 479-498.
- Belli, Robert F., Sean E. Moore, and John VanHoewyk. 2006. "An Experimental Comparison of Question Forms Used to Reduce Vote Over-Reporting." *Electoral Studies* 25(4): 751-759.
- Belli, Robert, F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Over-Reporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63(1): 90-108.
- Belli, Robert, Santa Traugott, and Steven J. Rosenstone. 1994. "Reducing Over-Reporting of Voter Turnout: An Experiment Using a 'Source Monitoring' Framework." ANES Technical Report Series: NES010153.
- Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1): 47-77.
- Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. West Sussex: John Wiley & Sons, Ltd.
- Butler, David, and Donald E. Stokes. 1979. *Political Change in Britain, 1963-1970* (Computer file). ICPSR07250-v3. Conducted by David Butler, Bibliographic Citation: Nuffield College, Oxford, and Donald E. Stokes, University of Michigan. Ann Arbor, MI: Inter-university Consortium for Political and Social Research (producer and distributor).
- Cahalan, Don. 1968. "Correlates of Respondent Accuracy in the Denver Validity Survey." *Public Opinion Quarterly* 32(4): 607-621.

- Campbell, Angus, Gerald Gurin, and Warren Miller. 1999. American National Election Studies: 1952 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Center for Political Studies (producer and distributor).
- Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT." *Political Analysis* 17(1): 45-63.
- 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*. Ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. Hoboken, New Jersey: Wiley-Interscience.
- Duff, Brian, Michael J. Hanmer, Won-Ho Fark, and Ismail K. White. 2007. "Good Excuses: Understanding Who Votes with an Improved Turnout Question." *Public Opinion Quarterly* 71(1): 67-90.
- Glynn, Adam N.. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1): 159-172.
- Górecki, Maciej A. 2011. "Why Bother Lying When You Know So Few Care? Party Contact, Education and Over-reporting Voter Turnout in Different Types of Elections." *Scandinavian Political Studies* 34(3): 250-267.
- Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangean. 2004. *Survey Methodology*, 1<sup>st</sup> Ed. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Hanmer, Michael J., Antoine J. Banks, and Ismail K. White. 2014. "Experiments to Reduce the Over-Reporting of Voting: A Pipeline to the Truth." *Political Analysis* 22(1): 115-129.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010a. "Measuring Voter Turnout by Using the Randomized Response Technique." *Public Opinion Quarterly* 74(2): 328-343.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010b. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74(1): 37-67.
- Holbrook, Allyson L., and Jon A. Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77(S1): 106-123.
- Keeter, Scott, Cliff Zukin, Molly Andolina, and Krista Jenkins. 2002. "Improving the Measurement of Political Participation." The 60th Annual Conference of the Midwest Political Science Association (MPSA), Chicago, Illinois, April 25<sup>th</sup>-28<sup>th</sup>, 2002.

- Kritzinger, Sylvia, Steve Schwarzer, and Eva Zeglovits. 2012. "Reducing Over-Reporting of Voter Turnout in Seven European Countries: Results from a Survey Experiment." The 67th Annual Conference of the American Association for Public Opinion Research, Orlando, Florida, May 17-20<sup>th</sup> 2012.
- Locander, William, Seymour Sudman, and Norman Bradburn. 1976. "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association* 71(354): 269-275.
- Miller, Judith. 1984. "A New Survey Technique for Studying Deviant Behaviour." Ph.D. Dissertation, George Washington University.
- Mircea, Comşa, and Andrei Gheorghişă. 2011. " 'Many', 'Half' or 'One of Two'? Assessing Counter-biasing Technique to Reduce the Self-reported Turnout." The 4<sup>th</sup> Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, July 11<sup>th</sup>-22<sup>nd</sup>, 2011.
- Mircea, Comşa, and Camil Postelnicu. 2013. "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique." *International Journal of Public Opinion Research* 25(2): 153-172.
- Parry, Hugh J., and Helen M. Crossley. 1950. "Validity of Responses to Survey Questions." *Public Opinion Quarterly* 14(1): 61-80.
- Persson, Mikael, and Maria Solevid. 2014. "Measuring Political Participation: Testing Social Desirability Bias in a Web-Survey Experiment." *International Journal of Public Opinion Research* 26(1): 98-112.
- Presser, Stanley, Michael W. Traugott, and Santa Traugott. 1990. "Vote 'Over' Reporting in Surveys: The Records or the Respondents?" The International Conference on Measurement Errors, Tucson, Arizona, November 11<sup>th</sup>-14<sup>th</sup>, 1990.
- Presser, Stanley. 1990. "Can Changes in Context Reduce Vote Over-Reporting in Surveys?" *Public Opinion Quarterly* 54(4): 586-593.
- Rogers, Theresa F. 1976. "Interviews by Telephone and in Person: Quality of Responses and Field Performance." *Public Opinion Quarterly* 40(1): 51-65.
- Selb, Peter, and Simon Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 32(1): 186-196.
- Sribney, William and Vince Wiggins. 2009. "Standard Errors, Confidence Intervals, and Significance Tests for ORs, HRs, IRRs, and RRRs." Accessed 03<sup>rd</sup> June 2016. <https://www.stata.com/support/faqs/statistics/delta-rule/>.
- Stocké, Volker, and Tobias Stark. 2007. "Political Involvement and Memory Failure as Interdependent Determinants of Vote Over-Reporting." *Applied Cognitive Psychology* 21(2): 239-257.

- Stocké, Volker. 2007. "Response Privacy and Elapsed Time Since Election Day as Determinants for Vote Overreporting." *International Journal of Public Opinion Research* 19(2): 237-246.
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1948. The ANES 1948 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1976. The ANES 1972-1976 Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1978. The ANES 1978 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1982. The ANES 1982 Merged Methods Comparison Project (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1984. The ANES 1984 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1989. The ANES 1989 Pilot Election Study (dataset). Ann Arbor, MI: University of Michigan, Center for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1996. The ANES 1996 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 1998. The ANES 1998 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 2000. The ANES 2000 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 2004a. The ANES 2004 Panel Study (dataset). Ann Arbor, MI: University of Michigan, Center for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 2004b. The ANES 2004 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 2008. The ANES 2008 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).

- The American National Election Studies ([www.electionstudies.org](http://www.electionstudies.org)). 2012. The ANES 2012 Time Series Study (dataset). Ann Arbor, MI: University of Michigan, Centre for Political Studies (producer and distributor).
- The European Social Survey. 2014. The ESS Round 1 (2002): Documentation Report. Edition 6.4. Bergen, European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC.
- U.S. Bureau of the Census. 1973. *Voting and Registration in the Election of November 1972*. Current Population Report Series: P.20, No.253.
- Voogt, Robert J. J., and Willem E. Saris. 2003. "To Participate or Not to Participate." *Political Analysis* 11(2): 164-179.
- Waismel-Manor, Israel, and Joseph Sarid. 2011. "Can Over-Reporting in Surveys be Reduced? Evidence from Israel's Municipal Elections." *International Journal of Public Opinion Research* 23(4): 522-529.
- Warner, Stanley L.. 1971. "The Linear Randomised Response Model." *Journal of the American Statistical Association* 66(4): 884-888.
- Wu, Chung-li. 2006. "Vote Misreporting and Survey Context: The Taiwan Case." *Issues and Studies* 42(4): 223-239.
- Zeglovits, Eva, and Sylvia Kritzinger. 2014. "New Attempts to Reduce Over-Reporting of Voter Turnout and Their Effects." *International Journal of Public Opinion Research* 26(2): 224-234.



## **A COMPARISON BETWEEN TWO SOLUTIONS TO TURNOUT OVER-REPORTING: THE PIPELINE AND ITEM-COUNT TECHNIQUES**

---

CHI-LIN TSAI

**Abstract** Over recent decades, survey research has developed many effective techniques for reducing respondents' misreporting of turnout, but relatively little is known about which techniques are better. This study compares two techniques – the pipeline and item-count techniques. These techniques are inspired by two opposite rationales. The item-count technique embodies the idea that confidentiality promotes candour, while the pipeline technique leads respondents to believe that nothing is confidential and thus honesty is the best policy. This comparison is based on survey experiments. The most important finding emerging from the experiments is that the item-count technique can persuade more respondents to admit to nonvoting than the pipeline technique can. This finding suggests that, compared to pressuring respondents to tell the truth, granting them greater response privacy is a more effective approach for eliciting truthful responses to a turnout question.

## Introduction

People's reports about their attitudes and behaviour underlie many areas of scientific research. However, eliciting truthful self-reports is a tricky business. A challenge in electoral studies is to prevent survey respondents from reporting having voted when they actually abstained – i.e. turnout over-reporting. The desire to appear in a socially desirable light has been considered the primary cause of turnout over-reporting (e.g. Belli, Traugott and Backmann 2001: 479–80; Cahalan 1968: 621; Górecki 2011: 8). Since voting is widely regarded as a characteristic of good citizens (Clarke et al. 2004: 274), non-voters over-report turnout in order to satisfy their self-image or create a good impression on others (Stocké 2007: 238–9). Moreover, respondents become more susceptible to this social desirability bias when their memories of turnout fade. In that case, respondents may unconsciously retrieve a false but socially desirable memory or, even worse, they may deliberately give a socially desirable answer without putting effort into recall (Belli et al. 1994: 2; Belli et al. 1999: 91–2; Stocké and Stark 2007: 241).<sup>1</sup>

Survey methodologists have proposed a number of solutions to turnout over-reporting. Most of them fall into two categories. Solutions in the first category target the memory issue. Conducting surveys promptly after Election Day is a straightforward means of avoiding misremembering (Stocké and Stark 2007: 255). More sophisticated means involves using a source-monitoring technique to refresh respondents' memories of turnout (Abelson, Loftus and Greenwald 1992: 140–2; Belli, Traugott and Rosenstone 1994; Kritzinger et al. 2012; Presser 1990: 589–90; Waismel-Manor and

---

<sup>1</sup> Some studies have refuted the relation between the memory issue (or say, interview time) and over-reporting (see Belli, Moore, and Van Hoewyk 2006: 118; Holbrook and Krosnick 2013: 111).

Sarid 2011; Wu 2006) or using turnout-scale response options to take account of the uncertainty of respondents' self-reported turnout (Zeglovits and Kritzinger 2014: 5).<sup>2</sup>

The second category of solutions tackles social desirability bias. Many researchers have applied face-saving techniques to turnout measurement (Abelson, Loftus and Greenwald 1992: 142–6; Belli, Traugott and Rosenstone 1994; Campbell, Gurin and Miller 1952: 100; Mircea and Gheorghîță 2011; Presser 1990: 590–91; Zeglovits and Kritzinger 2014). Some have employed indirect-questioning techniques (e.g. Locander, Sudman and Bradburn 1976: 271; Mircea and Postelnicu 2013), or different modes of data collection (ANES 2014: 21; Rogers 1976; Stocké 2007: 240) in order to grant respondents a high degree of response privacy. Other attempts to counter social desirability bias have involved manipulating the question order (Presser, Traugott and Traugott 1990: 3; U.S. Bureau of the Census 1973: 8) and interviewing proxy respondents (U.S. Bureau of the Census 1986: 10).

The item-count technique (ICT) is an indirect-questioning technique that is increasing in popularity in a wide range of fields, including studies on turnout. ICT is based on the idea that confidentiality promotes candour. It grants respondents greater response privacy, lowers the stakes for truth-telling, and hence demotivates them from over-reporting turnout. Compared to another well-known indirect-questioning technique – the randomised-response, ICT has the advantage of practicability, and it has proven to be a much more effective solution to turnout over-reporting (Holbrook and Krosnick 2010a; b).

---

<sup>2</sup> For more details of solutions mentioned in this section, please refer to Paper 2 – A Meta-Analysis of Solutions to Turnout Over-Reporting.

Another noteworthy solution, namely the pipeline technique (PLT), has also proven effective in tackling turnout over-reporting (Hanmer, Banks and White 2014). PLT does not directly tackle the causes of over-reporting; instead, it utilises the fear of being exposed as a liar, leading respondents to believe that dishonesty is detectable and will be detected. This then demotivates respondents from over-reporting turnout, since lying is usually considered more socially undesirable than nonvoting.

There is a sharp contrast between PLT and ICT. Aiming to reduce the psychological pressure of the voting norm, ICT implicitly conveys the message: “No one will know what you say, so please tell the truth.” In contrast, aiming to increase the pressure of norms of honesty, PLT implicitly conveys the message: “We will eventually find out what you did, so why not tell the truth now?” In other words, ICT creates a comfortable atmosphere that encourages respondents to tell the truth, whereas PLT amplifies respondents’ feelings of unease about lying. The rationales behind these two techniques are polar opposite. They represent two very different approaches to addressing turnout over-reporting.

ICT and PLT have proven effective against turnout over-reporting in separate studies, but it remains unclear which technique performs better. In order to answer this questions, I conduct survey experiments to compare ICT and PLT. The experiment results offer pollsters more information for choosing between these two techniques, and also contribute towards the further development of solutions to turnout over-reporting.

This paper begins with a review of ICT and PLT. Then, it proceeds to describe the research design and report the experiment results. This paper concludes with a discussion of findings.

## Development of the techniques

### 1. Item-count techniques

Holding a confidential conversation to elicit what people might prefer to keep secret is not a new idea. In 1565, Archbishop Charles Borromeo already drew on this idea to invent the confessional to alleviate penitents' unease about disclosing unpleasant personal experiences (Krumpal 2013: 2033). Contemporary survey methodologists re-apply this idea to the measurement of sensitive topics. Various questioning techniques have been developed for keeping survey responses strictly confidential and encouraging respondents to answer sensitive questions truthfully. One such technique is the aggregated-response technique, and ICT is a special case of it.

The aggregated-response technique is essentially an encryption scheme. It instructs respondents to 'aggregate' an interference term with their truthful answers to a sensitive question, and then requires respondents to report encrypted answers. Once encrypted, data can only be deciphered to a certain aggregate level (e.g. the mean of answers to the sensitive question). Each individual respondent's truthful answer remains completely confidential. Even interviewers cannot be sure about it.

A simple encryption scheme is  $Y_i = S_i + R_i$ , where  $S_i$  is respondent  $i$ 's truthful answer to a sensitive question,  $R_i$  is a random number from a known distribution, and  $Y_i$  is the encrypted result. The decryption algorithm is  $\bar{S} = \bar{Y} - \bar{R}$  (Warner 1971: 887). For example, respondent  $i$  may generate  $R_i$  by rolling a fair dice in private, and thus  $\bar{R}$  is 3.5. Boruch and Cecil (1979) and Raghavarao and Federer (1979) replaced random numbers with respondents' answers to other questions, and they devised two different

encryption/decryption schemes.<sup>3</sup> Miller (1984) then improved on these schemes in a way to make them more suitable for practical use. Her design was known as the item-count technique (also the unmatched-count technique or the list experiment).

ICT requires a split-ballot design. Respondents are split into two groups at random. One is presented with a list of  $J$  items (e.g. “I owned a gun” and “I donated money to a charity”). These items are usually of no immediate interest to researchers, so they are referred to as ‘non-key’ items.<sup>4</sup> The other group is presented with a list of the same  $J$  items plus a key item (e.g. “I voted in the election held on...”). Respondents then report *how many* items they would answer in the affirmative, and keep their truthful answers to the key item secret. The encryption scheme is  $Y_i = T_i S_i + \sum_{j=1}^J R_{i,j}$ , where  $T_i$  is a group indicator taking on the value of 1 if respondent  $i$  is in the long-list group, and 0 if he/she is in the short-list group.  $S_i$  and  $R_{i,j}$  are respondent  $i$ ’s unobserved answers to the key and  $j^{\text{th}}$  non-key items, respectively. They take on the value 1 for an affirmative answer and 0 for a negative answer. The decryption algorithm is a difference-in-means estimator:  $\bar{S} = \sum_i T_i Y_i / \sum_i T_i - \sum_i (1 - T_i) Y_i / \sum_i (1 - T_i)$ , which produces an estimate of the proportion of affirmative answers to the key item (e.g. an estimated turnout rate of the target population).

---

<sup>3</sup> Boruch and Cecil’s design requires respondents to answer two questions, including a sensitive question of interest. That design randomly divides respondents into two groups. One group of respondents add their two answers together, whereas those in the second group subtract one answer from the other. The decryption algorithm is  $\bar{S} = (\bar{Y}^{(1)} + \bar{Y}^{(2)})/2$ , where  $\bar{Y}^{(j)}$  is  $j^{\text{th}}$  group’s average response (see Droitcour et al. 1991: 187). Raghavarao and Federer’s design requires at least three random groups. In the simplest case, the first group of respondents answer a sensitive question and an additional question, those in the second group answer the same sensitive question and another additional question, and the rest answer all three questions. Each respondent, regardless of group, reports the sum of his/her answers. The decryption algorithm is  $\bar{S} = (\bar{Y}^{(1)} + \bar{Y}^{(2)} - \bar{Y}^{(3)})/(3 - 2)$ .

<sup>4</sup> Non-key items are sometimes referred to as ‘non-sensitive’ or ‘innocuous’ items, but these terms are not always appropriate, since non-key items are not necessarily non-sensitive, and, in theory, as long as not all non-key items are sensitive, ICT should still be valid.

ICT is becoming increasingly popular in various disciplines to measure sensitive topics including turnout (Holbrook and Krosnick 2010a; Mircea and Postelnicu 2013), since it balances the maintenance of confidentiality with the practicability of implementation (Droitcour et al. 1991: 188). This beneficial property, however, comes at a price – the loss of statistical efficiency. ICT collects much less (though better) information about the key item for estimation than does the direct-questioning technique. Respondents in the long-list group merely provide encrypted information, not to mention those in the short-list group; they do not even provide information about the key item at all. Consequently, estimates based on ICT are often surrounded by a greater deal of uncertainty (i.e. higher standard errors).

To recoup the efficiency loss, Droitcour et al. (1991: 189) propose a double-list version of ICT (hereafter DICT). This technique uses two sets of non-key items (with one sensitive item) to form two ICT questions:  $Q_A$  and  $Q_B$ . Respondents in a random subsample answer the long list of  $Q_A$  and the short list of  $Q_B$ , whereas those in the other subsample answer the short list of  $Q_A$  and the long list of  $Q_B$ . DICT collects information about the key item from all respondents, and so allows for more efficient estimation. Specifically,  $Q_A$  and  $Q_B$  generate two separate estimates of the sensitive item; the average of them produces an estimate with a lower standard error.<sup>5</sup>

## 2. Pipeline techniques

PLT is a ‘by-product’ of sustained but unsuccessful attempts at mind-reading. The history of attempts to read human minds can be traced back to at least the 1880s, when the Italian psychologist Angelo Mosso initiated an instrument-based approach to

---

<sup>5</sup> Let  $\bar{A}$  and  $\bar{b}$  be the first group’s average responses to  $Q_A$  and  $Q_B$ , respectively; let  $\bar{B}$  and  $\bar{a}$  be the second group’s average responses to  $Q_B$  and  $Q_A$ , respectively. The DICT estimate is  $[(\bar{A} - \bar{a}) + (\bar{B} - \bar{b})]/2$ , and its variance is  $[V(\bar{A}) + V(\bar{B}) + V(\bar{a}) + V(\bar{b}) - 2 Cov(\bar{A}, \bar{b}) - 2 Cov(\bar{B}, \bar{a})]/4$ .

emotion research. He studied the influences of emotions on human bodies by using a plethysmograph to measure changes in circulation, respiration etc. during fear. Mosso's work appealed to a contemporary, the Italian criminologist Cesare Lombroso, because of its potential "to penetrate into the most secret recesses of the mind of the criminal" (Lombroso 1911: 254). Based on a belief in the link between deception and emotions and its consequential effects on human bodies, Lombroso used instruments like the hydrosphygmograph to monitor variations in suspects' pulse rate and blood pressure during police interrogation to detect false confessions. He was among the first to apply such an approach to lie detection (Bunn 2012: 67–73; Grubin and Madsen 2007: 359).

In the twentieth century, the belief in the links among deception, emotions and physiological reactions continued to underpin the invention and use of modern lie detectors – polygraphs. Despite their widespread use, there is little evidence that polygraphs are accurate. From a large-scale assessment of polygraphs in security-screening uses, the American National Research Council (2003: 212) concludes that the evidence for the validity of polygraphs is "scanty and scientifically weak", and there is "little basis for the expectation that a polygraph test could have extremely high accuracy".

Though failing to function as a genuine 'pipeline' penetrating human minds, polygraphs, as a bogus pipeline, still have the potential to elicit confessions, "if examinees and the public believe that there is a high likelihood of a deceptive person being detected and that the costs of being judged deceptive are substantial" (American National Research Council 2003: 214). It was this potential that Jones and Sigall (1971) exploited to develop the pipeline technique. The technique involves using a polygraph-like machine that purportedly verifies respondents' answers, and some means of

validating that pipeline in the respondents' eyes. Note that the real purpose of using that machine is not to detect lies, but to convince respondents that lies will be detected. Anything that serves this purpose can be a substitute for the machine.<sup>6</sup> Jones and Sigall called their design the 'bogus' pipeline technique, as their machine was not really able to verify respondents' answers. If researchers do have some independent means of verification, the technique is known as the 'actual' pipeline technique.

From a meta-analysis of the social psychology studies on PLT, [Roese and Jamieson \(1993\)](#) conclude that the technique engenders reliable effects in reducing social desirability responding and shifts self-reports toward veracity. It is worth mentioning that studies reviewed by the meta-analysis were small-sample studies ( $n = 24\text{--}225$ ), and all of them used a polygraph-like machine in experiments. In other words, those studies are rather different from survey research, which is usually based on larger samples (e.g.  $n = 1,068$ ), and, more importantly, it is usually impractical to deploy a polygraph-like machine in survey interviews. The application of PLT to survey research therefore requires further consideration.

The major challenge of using PLT in survey research is how to convince respondents that lies are detectable. A verbal alert seems to be the only means of doing so in most interviews. [Hanmer, Banks and White \(2014\)](#) designed two PLT questions to measure turnout. The first one announced that the study was going to check respondents' answers against records of their turnout in electoral offices (and the study did conduct validation exercises afterwards, so it is an actual pipeline).<sup>7</sup> Though this

---

<sup>6</sup> Jones and Sigall's original idea was to tell respondents that the machine could accurately detect their true attitudes and opinions, and then to instruct the respondents to predict the machine's output. In practical applications, some studies followed the original design, while others asked respondents to respond truthfully rather than predict the machine's output – either way they proved to be effective in eliciting more valid self-reports ([Roese and Jamieson 1993: 371](#)).

<sup>7</sup> The question wording was: "In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. *By looking*

design significantly reduced turnout over-reporting, its practical value is limited. Most surveys cannot make such a definite announcement simply because validation exercises are impossible. Given this constraint, the authors' second PLT question (which the authors called 'the subtle pipeline') simply informed respondents that lies about turnout had been exposed, and mentioned nothing about validation exercises.<sup>8</sup> This design also resulted in less over-reporting, but the effect size did not reach the conventional level of statistical significance. A reasonable explanation is that the pipeline was too subtle to convince respondents. Overall, Hanmer, Banks and White's study shows both the possibility and dilemma of PLT in survey uses.

### 3. ICT vs PLT

Table 1 summarises the above review and makes a comparison between the use of ICT and PLT in turnout measurement. Both techniques aim to elicit valid self-reports, but they come from entirely different traditions of truth-seeking, hence very different approaches to achieving the aim. In the belief that confidentiality reduces the cost of truth-telling and promotes candour, ICT guarantees complete confidentiality for respondents in order to motivate them to be candid about their socially undesirable attitudes and behaviours. Taking advantage of the fact that lying is itself socially undesirable, PLT shifts the emphasis onto the cost of lying. By convincing respondents that lies are detectable, PLT attempts to replace the pressure to be one who conforms to the social norm regarding the issue of interest with pressure not to be a liar (Hanmer,

---

*at public records kept by election officials, we can get an accurate report of who actually voted in November, and in previous elections. Of course, these public records do not say who you voted for. Part of our study will involve checking these records against the survey reports. Which of the following statements best describes you?" (Hanmer, Banks and White 2014: 135).*

<sup>8</sup> The question wording is: "In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. *We also sometimes find that people who say they voted actually did not vote. Which of the following statements best describes you?" (Hanmer, Banks and White 2014: 135).*

[Banks and White 2014: 133](#)).

The two techniques involve redesigning different parts of a survey, and pose different challenges. PLT requires pollsters to validate the pipeline in respondents' eyes, but very often this can only be done verbally through question wording in rather restricted ways. It is even more difficult to design a credible pipeline for panel surveys, since respondents may learn about the idea underlying PLT (particularly the bogus type) from repeated interviews ([Ostrom 1973: 258](#)). In quite a different way, ICT is supposed to be more effective when respondents realise its underlying logic ([Droitcour et al. 1991: 195](#)), so frequent and repeated use of ICT should not be a matter of concern. Nevertheless, compared to PLT, ICT involves more fundamental changes to a survey – a split-ballot design and an aggregated-response format. Studies based on ICT also need a larger sample to achieve the same level of statistical efficiency as studies based on PLT. On the whole, it takes more effort to employ ICT than PLT.

The two techniques generate data in different formats. Take turnout measurement for example – data from PLT record each respondent's self-reported turnout and show who reported having vote and who did not. ICT encrypts such individual-level information, so the data only allow for aggregate-level estimates, such as a class of people's turnout rate. This special data format creates some difficulties in data analysis, but it does not devalue ICT in scientific research. In fact, even if individual-level information is available, researchers seldom analyse each survey respondent individually. Instead, what interests researchers is a group of similar people's attitudes or behaviours, such as how many women voted, or why the highly educated vote more often than the less educated. Data from PLT allow researchers to answer these questions, so do data from ICT ([Blair and Imai 2012](#); [Corstange 2009](#); [Holbrook and](#)

Krosnick 2010a; Imai, Park, and Green 2015).<sup>9</sup> Hence the difference in data formats should not be overstated.

Table 1 also summarises published studies that have experimented on either ICT or PLT in cases of turnout measurement (Hanmer, Banks and White 2014; Holbrook and Krosnick 2010a; Mircea and Postelnicu 2013). Compared to the conventional measurement of turnout, the two PLT questions reduced the proportions of over-reporters by 6.2 percentage points on average, and ICT in five experiments as a whole reduced self-reported turnout by 6.9 percentage points.<sup>10</sup> These experiment results provide evidence for the effectiveness of ICT and PLT in tackling turnout over-reporting. However, with only these studies, it is insufficient to answer the question about which technique is more effective, since those findings were based on different experiments in different surveys that targeted different populations in different elections, and presented control groups with different measurement formats. In order to make a more objective comparison, this study experiments with ICT and PLT in the same survey set-up.

---

<sup>9</sup> For example, a simple method for estimating a group's turnout is to apply a difference-in-means estimator to a sub-sample of that group. For comparison with another group's turnout, just apply the estimator to that second group again. Holbrook and Krosnick (2010a) generalise this method from bivariate to multivariate analysis by using a linear regression to fit the ICT variable ( $Y_i$ ) with the treatment status ( $T_i$ ), a set of independent variables ( $\mathbf{X}_i$ ), and interaction terms between  $T_i$  and each of  $\mathbf{X}_i$ . The coefficient of the interaction between  $T_i$  and education, for example, is an estimate of the relationship between turnout ( $S_i$ ) and education. Blair and Imai (2012) and Corstange (2009) have improved the statistical efficiency of this regression method. Imai, Park, and Green (2015) develop a method for using predicted  $S_i$  as an independent variable to model another dependent variable.

<sup>10</sup> These are arithmetic means of experimental results.

**Table 1. Comparisons between ICT and PLT**

Technique	ICT					PLT	
Approach	Obviating the need of social desirability responding					Replacing one social desirability pressure with another	
Means	Keeping survey responses confidential					Claiming to verify survey responses	
Design	Split-ballot design, aggregated-response format					An alert in question wording	
Difficulty	Data analysis, statistical efficiency					Pipeline credibility, repeated uses	
Measure	Aggregate-level					Individual-level	
Experiment <sup>a</sup>							
Survey mode	Face-to-Face	Web	Web	Web	Tel	Web	Web
Target election <sup>b</sup>	2009 EP	2002 USM	2000 USP	2000 USP	2000 USP	2010 USM	2010 USM
Non-key items / Pipeline type	3 items	3 items	4 items	4 items	4 items	Actual pipeline	Subtle pipeline
Control group question	Forgiving wording with yes-no options					Forgiving wording with face-saving options	
Control group size	1,374	2,018	857	115	175	840	840
Treatment group size	1,374	2,029	3,077	454	353	844	830
Reduction in self-reported turnout <sup>c</sup>	-10.5*	-1.4	-3.1	+0.3	-19.6*	-2.4	-2.6
Reduction in over-reporting <sup>d</sup>	N/A	N/A	N/A	N/A	N/A	-7.6*	-4.8

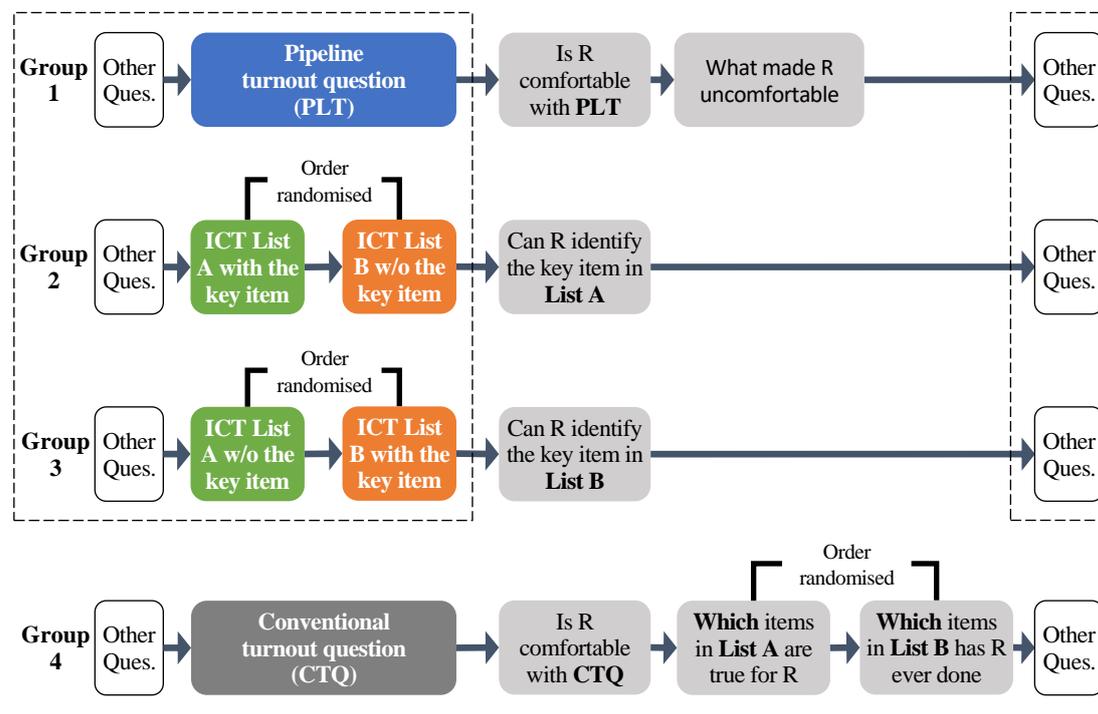
**NOTES:** (a) The first ICT experiment was conducted by [Mircea and Postelnicu \(2013\)](#) during the 2009 Romanian Presidential Election Study. The other ICT experiments were conducted by [Holbrook and Krosnick \(2010a\)](#). Both pipeline experiments were conducted by [Hanmer, Banks, and White \(2014\)](#). (b) ‘EP’, ‘USM’ and ‘USP’ stand for European Parliamentary Elections, U.S. Midterm Elections, and U.S. Presidential Election, respectively. (c) [% self-reported voters in the treatment group] – [% self-reported voters in the in the control group]. (d) [% self-reported voters in the treatment group] – [% self-reported voters in the in the control group].

## Experiment design

I ran two waves of survey experiments through the YouGov Daily Poll in the United Kingdom. Wave 1 fieldwork was carried out immediately after the 2014 European Parliamentary Election (26th May – 12th June 2014). Wave 2 fieldwork was conducted nine months after the election (24th February – 3rd March 2015). The target population for both waves was the adult population of Great Britain. The sample of each wave was drawn independently from YouGov online panel by quota sampling. The sample sizes of Waves 1 and 2 are 18,686 and 7,574, respectively (see Appendix 1 for more details).

As shown in Figure 1, Respondents in Wave 1 were randomly divided into three groups (see dashed rectangles).

**Figure 1. Experiment design**



**NOTE:** Wave 1 did not have Group 4 and did not field meta-questions.

Group 1 answered a PLT turnout question that adopts Hanmer, Banks and White’s (2014) design with necessary adaptation. One modification is to substitute the auxiliary

verb ‘may’ for ‘will’ in the sentence “*Part of our study will involve checking these records against the survey reports.*” Turnout validation is possible and has been carried out several times in the U.K. (e.g. the British Election Study). Plans to conduct validation exercises had been under consideration, but it was uncertain whether there would be sufficient resources for implementation. We therefore decided only to signal the possibility of validation exercises, so as to satisfy the Code of Conduct of the British Market Research Society. Another modification to Hanmer, Banks and White’s design is the use of yes-no options. Since this study aims to assess the effect of PLT *itself* in comparison to ICT, this modification is intended to rule out a synergistic effect from face-saving options.

***PLT:***

In talking to people about elections, we often find that a lot of people were not able to vote because they weren’t registered, they were sick, or they just didn’t have time. By looking at public records kept by election officials, we can get an accurate report of who actually voted in the May 2014 elections to the European Parliament, and in previous elections. Of course, these public records do not say who you voted for. Part of our study may involve checking these records against the survey reports. Which of the following statements best describes you?

- Yes, I voted in the May 2014 elections to the European Parliament.
- No, I did not vote in the May 2014 elections to the European Parliament.

Each of the respondents in Groups 2 and 3 was assigned to two ICT questions, as shown in Figure 1. The two questions contain different non-key items but the same key item – “*I voted in the May 2014 elections to the European Parliament.*” The first question (ICT<sub>A</sub>) adopts [Holbrook and Krosnick’s \(2010a\)](#) non-key items. This design has proven to be effective in a telephone survey that measured the turnout of a highly salient election – the 2000 United States presidential election (see Table 1). I re-assess

this design in the context of a comparatively low salient election.

**ICT A:**

Below are # statements. Please tell us how many of them have you done. You do not have to tell us which you have and have done. Just tell us how many you have done.

- Owned a gun
- Given money to a charity
- Gone to the cinema
- Written a letter to the editor of a newspaper
- Voted in the May 2014 elections to the European Parliament

If you are not sure whether you have done that thing, don't include it in your count. How many have you done? [Response options: 0 to #.]<sup>11</sup>

The second ICT question (ICT<sub>B</sub>) uses another set of non-key items that prompt respondents for their opinions on issues relating to the 2014 European Parliamentary Election. These attitudinal items are more verbose and complicated than the non-key items of ICT<sub>A</sub>, so ICT<sub>B</sub> is presumably a more cognitively demanding question. In this study, ICT<sub>B</sub> serves two purposes. First, it allows for an examination of how different designs of non-key items affect the performance of ICT. Second, and more importantly, ICT<sub>A</sub> and ICT<sub>B</sub> combine to form a DICT, which is expected to produce a more precise estimate of turnout, and thus to reduce the uncertainty of the comparison between ICT and PLT.

**ICT B:**

Below is a list of # statements. Please tell us how many of them are true in your case. You do not have to tell us which are true and which are not. Just tell us how many are true in your case.

- I feel British more than I feel European

---

<sup>11</sup> The questionnaire presented # response options below the question wording for respondents to click, where # = 5 for the long-list groups, and # = 4 for the short-list groups.

- I approve of Britain's membership of the European Union
- I do not want Britain to adopt the Euro as its currency
- I think the EU should take a lot of blame for the global financial crisis
- I voted in the May 2014 elections to the European Parliament

If you are not sure whether a statement applies to you, don't include it in your count. How many of the above statements are true in your case? [Response options: 0 to #.]

The Wave 2 experiment retained all of the Wave 1 design, but incorporated two additional components. First, there were four rather than three groups of respondents in Wave 2. Experimental designs for the first three groups were identical to those in Wave 1. Respondents in Group 4 answered a conventional turnout question (CTQ).

**CTQ:**

In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, they were sick, or they just didn't have time. Which of the following statements best describes you?

- Yes, I voted in the May 2014 elections to the European Parliament.
- No, I did not vote in the May 2014 elections to the European Parliament.

Second, in order to investigate PLT and ICT further, the Wave 2 questionnaires included some meta-questions (see Appendix 2 for wording). Groups 1 and 4 were required to rate how uncomfortable they felt about PLT and CTQ, respectively. Group 4 also answered the non-key items of ICT<sub>A</sub> and ICT<sub>B</sub>, one by one and directly, i.e. they had to indicate *which* items applied to them. As for Groups 2 and 3, I used a meta-question to examine whether they could identify the key item on the list.

## **Randomisation checks**

To assess whether random assignment produced comparable groups of respondents, I compare the distributions of socio-demographic variables across groups in each wave. Appendix 3 provides a detailed analysis. In general, randomisation is effective in both waves, though not perfect. I statistically weight the data to improve comparability across groups further, and I use weighted data in the following analysis.

It is also worth mentioning that, due to unforeseen difficulties, Wave 2 did not collect data from all groups every day. Most interviews of Group 1 were conducted over 24th–25th February 2015, Group 2 over 25th–26th February, Group 3 over 26th–27th February and Group 4 over 1st–2nd March. Nonetheless, the timing difference between groups is small, and the analysis in Appendix 3 suggests that the four groups in Wave 2 are comparable. Therefore, there is little ground for believing that the timing issue undermines the validity of the Wave 2 experiment.

## **Assessment criterion**

Most studies of turnout over-reporting have examined the relative performance of an experimental design over another by comparing aggregate turnout estimates across experiment groups (e.g. [Belli, Moore, and VanHoewyk 2006](#)). The use of this criterion is justified by two presumptions. First, given that turnout under-reporting is extremely rare ([Selb and Munzert 2013, 191](#)), a reduction in turnout estimates must largely result from a reduction in turnout over-reporting. Second, given a successful exercise of randomisation, any difference in turnout estimates across experiment groups must be attributed to the difference between experimental designs. Therefore, a design that produces a smaller turnout estimate is considered a better solution to turnout over-reporting. In this study, I apply this criterion in comparisons between different

questioning techniques.

## Experiment results

### 1. Comparing aggregate turnout estimates

I begin the analysis by comparing different questioning techniques based on aggregate turnout estimates. Figure 2 displays the estimates generated by different techniques in different waves, and Table 2 displays the results of statistical tests for differences between each pair of estimates.<sup>12</sup> The experiment results shows that ICT outperforms PLT, regardless of the timing of fieldwork and the design of non-key items. All ICT estimates are smaller than PLT estimates. In Wave 1, the difference between ICT<sub>A</sub> and PLT estimates is 3.8 [-0.1, 7.8] points with a p-value of 0.058; the difference between ICT<sub>B</sub> and PLT estimates is 4.8 [0.03, 9.6] points with a p-value of 0.049.<sup>13</sup> In Waves 2, the ICT<sub>A</sub> and ICT<sub>B</sub> estimates are smaller than the PLT estimate by 12.4 [5.1, 19.7] and 10.9 [2.1, 19.8] points, respectively, and both differences are statistically significant at conventional levels of statistical significance (p-values=0.001 and 0.016).

The findings above would not be meaningful if both experimental techniques performed worse than the conventional measure. To address this concern, Wave 2 incorporated CTQ into the experiment. The result dispels the concern: all of the ICT and PLT estimates are smaller than the CTQ estimate. The differences between the CTQ and ICT estimates are substantial – at least 14 points – and statistically significant at the 0.01 level at least. The difference between the CTQ and PLT is also notable – 3.3 [-0.4, 6.9] points – though it is not statistically significant at the conventional level (p-

---

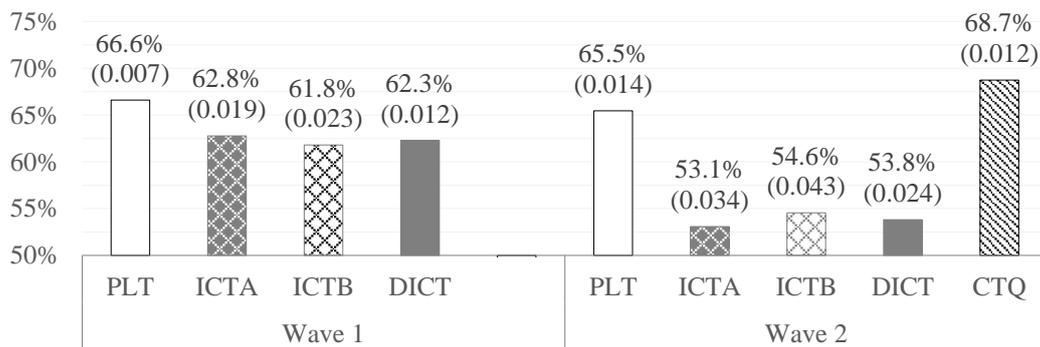
<sup>12</sup> ICT diagnostics in Appendix 4 show no clear violation of the assumptions underlying the technique, so it is reasonable to expect that the ICT design of this study is valid.

<sup>13</sup> Numbers in brackets are the lower and upper bounds of 95% confidence intervals.

value=0.079).

Taken together, the comparison of aggregate turnout estimates shows that ICT yields substantial improvement in turnout measurement and is a more promising questioning technique than PLT (see Appendix 5 for a robust analysis that examines this finding further).

**Figure 2. Turnout estimates by techniques and waves**



NOTE: Numbers in parentheses are standard errors.

**Table 2. Differences in turnout estimates across techniques and waves**

		Wave 1			
		PLT	ICT <sub>A</sub>	ICT <sub>B</sub>	DICT
Wave 2	PLT		-3.8 (0.020) *	-4.8 (0.024) ‡	-4.3 (0.014) **
	ICT <sub>A</sub>	-12.4 (0.037) ***		-1.0 (0.034)	-0.5 (0.017)
	ICT <sub>B</sub>	-10.9 (0.016) *	1.5 (0.061)		0.5 (0.017)
	DICT	-11.6 (0.028) ***	0.7 (0.031)	-0.7 (0.031)	
	CTQ	3.3 (0.019) ‡	15.7 (0.037) ***	14.2 (0.045) **	14.9 (0.027) ***

NOTES: (a) The unit of measurement is a percentage point. (b) Numbers in parentheses are standard errors. (c) Upper-triangle cells show the differences between estimates produced by different techniques in Wave 1. (d) Lower-triangle cells show the differences between estimates produced by different techniques in Wave 2. (e) \*\*\* P-value (two-tailed) < 0.001, \*\* < 0.01, \* < 0.05, ‡ < 0.1

## 2. Investigating the design of PLT

Why does the PLT question yield only a slight improvement in turnout measurement?

PLT aims to substitute the pressure not to be a liar for the pressure to be one who

conforms to the norms of voting. To investigate whether PLT worked as intended in the experiment, I measured respondents' feelings about it. On a scale ranging from 1 (very uncomfortable) to 7 (very comfortable), Group 1's average feeling about the PLT question is 5.9 [5.8, 6.0], whereas Group 4's average feeling about the CTQ is 6.1 [6.0, 6.2]. The difference, though small (0.132 [0.003, 0.263]), is statistically significant ( $p$ -value=0.046). Of 190 respondents who feel uncomfortable about the PLT question (i.e. score<4), 35.3% express "I didn't realise that there are public records for checking" or "I don't like to think of my answers being checked."

The PLT design of this study appears to have exerted extra pressure on respondents, but that extra pressure was not very intense. Had the question wording been phrased in a way that put the respondents under more pressure, PLT might have yielded a more significant improvement in turnout measurement. One possible improvement to the PLT design could be to give a more detailed description of how it is possible to verify survey responses, such as showing respondents what an official record of turnout would look like in a record template (e.g. a table or a list), or even by a real record that lists respondents' turnout histories in elections before the target one.<sup>14</sup> Nevertheless, this might involve the formidable task of preparing each respondent's records in advance, and this is not feasible when records are not publicly accessible.

### 3. Assessing the cost of ICT

In addition to the strength of ICT, I also examine its weaknesses. First, ICT estimates are less precise than PLT estimates. Despite the fact that the sizes of ICT groups (Group 2 plus Group 3) are twice as large as the sizes of the PLT group (Group 1), the standard

---

<sup>14</sup> This would be very similar to [Garber, Green, and Larimer's \(2008: 38\)](#) get-out-the-vote study that showed people a list of their turnout records, in order to convey the message that who votes is publicly accessible, and thus to motivate people to vote.

errors of  $ICT_A$  and  $ICT_B$  estimates are at least 2.4 times as high as the standard errors of PLT estimates (Figure 2).<sup>15</sup> DICT, which combines  $ICT_A$  and  $ICT_B$  in estimation, partly improves precision. The standard errors of DICT estimates are roughly 40% smaller than the standard errors of  $ICT_A$  and  $ICT_B$  estimates, though still 1.7 times as high as the standard errors of PLT estimates.

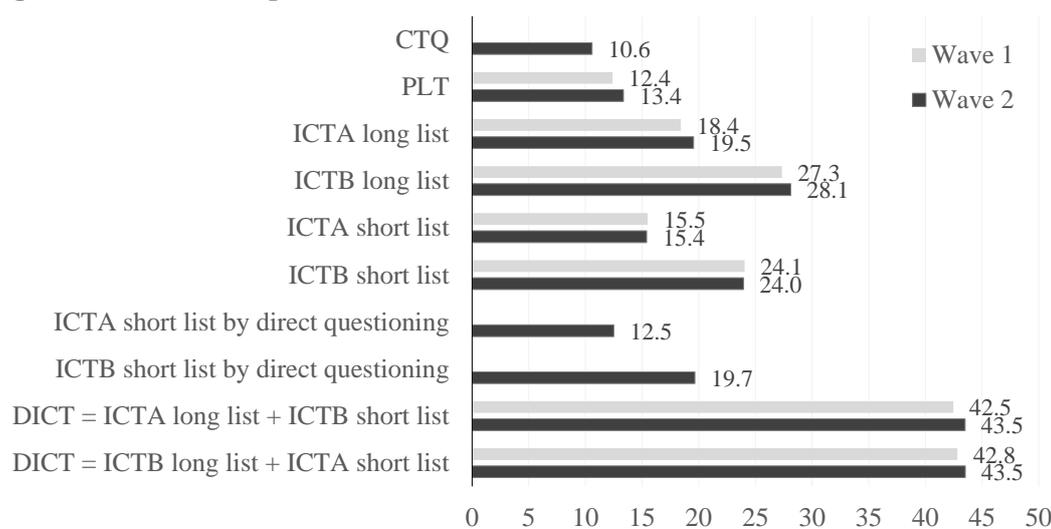
Another weakness of ICT is the long response time. Figure 3 displays the median time that respondents took to answer a question. Given that the differences between waves are small, I focus on Wave 2. The PLT response time is 13.4 seconds, while the ICT response time is 15.4–28.1 seconds, depending on the length of the list and the type of non-key items. Because each item in an ICT list is actually a question, answering an ICT question resembles answering several questions, and so is more time-consuming than answering a PLT question.

Moreover, ICT also requires respondents to put extra effort into adding up their answers about items. To examine this extra cognitive burden, I asked Group ‘4’ to answer each of non-key items directly and separately. It took respondents 12.5 seconds to answer the four  $ICT_A$  non-key items, and 19.7 seconds to answer the four  $ICT_B$  non-key items. In comparison, when respondents (Groups 2 and 3) answered these items in the ICT format, the response time increased by 2.9 and 4.3 seconds, respectively.

Finally, DICT requires each respondent to answer two ICT questions, so it is even more time-consuming. Respondents took more than 42 seconds to answer the DICT design of this study – 3.3 times as much as the PLT response time. Overall, ICT is a more demanding technique than PLT.

---

<sup>15</sup> Regardless of  $ICT_A$ ,  $ICT_B$  or DICT, the computation of ICT estimates always involves Groups 2 and 3, whereas the computation of PLT estimates only involves Group 1.

**Figure 3. Median response time**

**NOTE:** The unit of measurement is a second.

#### 4. Evaluating the impact of non-key items

This study uses two ICT questions to take advantage of the double-list design, and this design also opens up an opportunity to compare two different sets of non-key items. Keeping things simple is a general rule of thumb for survey question designs. Meanwhile, it has been argued that non-key items should be on a similar subject as the key item, in order to prevent making respondents suspicious (Droitcour et al. 1991; Glynn 2013; Kuklinski, Cobb, and Gilens 1997; Tsuchiya, Hirai, and Ono 2007.) However, sometimes it may be difficult to satisfy both rules. ICT<sub>B</sub> represents a situation where non-key items are on a similar subject to the key item (i.e. EU affairs), but more cognitively demanding than those of ICT<sub>A</sub>. Data from the meta-question confirm these differences. ICT<sub>B</sub> does satisfy the rule of subject and valence similarity – only 7.5% [6.2%, 9.0%] of Group3 correctly recognise turnout as the key item of ICT<sub>B</sub>, whereas 44.6% [41.9%, 47.2%] of Group 2 correctly recognise turnout as the key item of ICT<sub>A</sub>. ICT<sub>B</sub> is also more demanding, since its response time is 9 seconds longer than the ICT<sub>A</sub> response time (Figure 3).

Despite these differences, the two ICT questions yield almost identical estimates (62.8% [59.1%, 66.5%] vs 61.8% [57.2%, 66.4%] in Wave 1; 53.1% [46.3%, 59.8%] vs 54.6% [46.1%, 63.0%] in Wave2). The differences in the designs of non-key items have no significant impact on the univariate estimates of the key item. Moreover, the similarity between the ICT<sub>A</sub> and ICT<sub>B</sub> estimates suggests that both ICT designs are valid, and thus justifies the use of DICT, and provides stronger evidence that ICT outperforms PLT.

Although ICT<sub>B</sub> is valid, it has two shortcomings. First, because of the longer response time, ICT<sub>B</sub> is more costly to use than ICT<sub>A</sub>. Second, respondents' answers to the ICT<sub>B</sub> non-key items vary with the question format. When respondents count *how many* of those items apply to them (i.e. the ICT format), the average answer is 2.35 [2.30, 2.41]. In contrast, the average count is 2.22 [2.16, 2.27], when respondents report *which items* apply to them (i.e. answering directly one item after another). The difference (0.13 [0.06, 0.21]) is statistically significant (P-value=0.001). This problem does not occur with ICT<sub>A</sub> – the difference is only 0.02 [-0.03, 0.08] (P-value=0.411).

Why do respondents' answers to the ICT<sub>B</sub> non-key items vary with the question format? Given the complexity of those items, one possible explanation is that respondents miscounted their affirmative answers, when answering those items in the ICT format. This does not necessarily invalidate the estimate of the key item. As long as the short-list and long-list groups miscount in the same way, errors cancel out each other. This is very likely the case for ICT<sub>B</sub>; otherwise, ICT<sub>B</sub> would not have passed the diagnostics and its estimates would not have been so similar to the ICT<sub>A</sub> estimates. Nevertheless, it is still better to avoid such miscounting, since there is no guarantee that the short-list and long-list groups always miscount in the same way; when they

miscount in different ways, estimates of the key item are biased.

Overall, the comparison between ICT<sub>A</sub> and ICT<sub>B</sub> highlights the importance of keeping items simple. Although it is also important to keep items on a similar subject, if compromise is inevitable, simplicity should take priority over subject similarity.

## **5. Examining mechanisms underlying turnout over-reporting**

The experiment results provide a basis for assessing the relative importance of two mechanisms underlying turnout over-reporting. One is ‘impression-management’ – respondents over-report to create a positive impression on others (e.g. interviewers.) This is a controlled mechanism operating only when respondents are aware that their answers to the turnout question are not completely confidential. The second mechanism is ‘self-deception’ – respondents over-report to satisfy their self-image. The confidentiality of survey answers is irrelevant to this mechanism, since it can operate automatically, whenever respondents are faced with a turnout question (Stocké 2007: 238–239).

Different turnout measures have different degrees of effectiveness in preventing these mechanisms being activated. ICT keeps survey responses strictly confidential, so it is pointless to over-report turnout for the purpose of creating a good impression on others. Aggregate turnout estimates based on ICT should be free of impression-management bias. CTQ, by contrast, requires respondents to give a categorical answer about turnout. Non-voters thus have a motive to over-report so as to create a good impression on interviewers, pollsters and any other person who is able to identify their answers. CTQ estimates are therefore prone to impression-management bias. Apart from that, there is no compelling theoretical ground for arguing whether CTQ or ICT is better in preventing self-deception and other kinds of bias (e.g. voters’ over-

representation in the sample). Taken together, the difference between CTQ and ICT estimates is an approximation of the level of over-reporting due to impression-management (Table 3). Bias of this kind in Wave 2 experiment is 14.9 [9.7, 20.2] percentage points (CTQ – DICT = 68.7% – 53.8% according to Figure 2 and Table 2).

**Table 3. Decomposition of turnout estimates**

CTQ-based turnout rate	=	Actual turnout rate	+	Self-deception bias	+	Impression-management bias	+	Other biases	+	Random errors
ICT-based turnout rate	=	Actual turnout rate	+	Self-deception bias	+			Other biases	+	Random errors
Validated turnout rate	=	Actual turnout rate	+					Other biases	+	Random errors

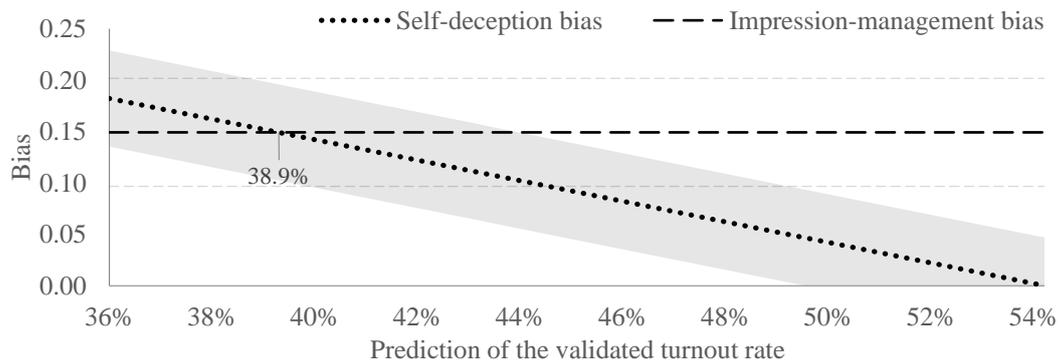
NOTES: The ‘actual turnout rate’ is the entire population’s true turnout rate. The validated turnout rate is the respondents’ turnout rate that is free from over-reporting bias but subject to other biases and random errors.

Assessment of self-deception requires another turnout estimate based on respondents’ official records of turnout (i.e. validation data.) That estimate does not use respondents’ self-reports at all, so it is completely free from any kind of over-reporting bias (though it is still subject to other biases). The difference between an ICT estimate and a validated turnout estimate is therefore an approximation of the level of over-reporting due to self-deception.

The validated turnout estimate is unknown due to lack of validation data, but its range is identifiable – it must range between the actual turnout rate (35.6%) and the ICT estimate (53.8%). Given this range, Figure 4 compares two kinds of over-reporting bias in the Wave 2 experiment. The estimated impression-management bias is 14.9, as mentioned above. Estimates of self-deception bias varies with the prediction of the validated turnout rate. For a prediction smaller than 38.9%, self-deception (dotted line) accounts for more turnout over-reporting than does impression-management (dashed

line). The relative importance of the two mechanisms reverses if the validated turnout rate is larger than 38.9%. (At 38.9%, each mechanism accounts for 50% of over-reporting.) In most of Figure 4, the dotted line lies below the dashed line, suggesting that, in the Wave 2 experiment, impression-management is a more important mechanism underlying turnout over-reporting than self-deception.

**Figure 4. Relative importance of over-reporting mechanisms**



**NOTES:** The grey dashed lines mark a 95% confidence interval for impression-management bias. The grey area is a 95% confidence interval for self-deception bias.

## Conclusion

Survey research has developed various effective techniques for preventing non-voters from reporting having voted, but relatively little is known about which ones are more effective. This study experimentally compares two techniques – pipeline (PLT) and item-count (ICT). These techniques take very different approaches to tackling turnout over-reporting. ICT aims to reduce the cost of truth-telling, whereas PLT seeks to highlight the cost of lying. In order to investigate which approach is more effective, I conduct two waves of survey experiments online in the aftermath of the 2014 European Parliamentary Election in the U.K.

The experiment found that ICT far outperformed PLT as a measure of turnout. Although PLT yielded smaller (and hence presumably better) aggregate turnout

estimates than did the conventional measure, the improvement was only statistically significant at the 0.1 level. In contrast, the improvement with ICT was substantial and statistically significant. These results suggest that ICT is more effective than PLT in reducing turnout over-reporting.

Further investigation found that PLT performed less well in the experiments because the design did not put respondents under enough pressure not to be a liar, and thus failed to divert respondents' attention away from the pressure to over-report. This was a direct consequence of the difficulty in designing an effective pipeline for survey use.

In addition to its strength, the experiment results also showed the weaknesses of ICT. ICT requires a longer response time and a larger sample to achieve the same level of estimation precision as do direct-questioning techniques. Both of these increase the cost of survey administration.

Moreover, by comparing different designs of non-key items, this study showed the importance of keeping items simple. Although it is also desirable to keep the key and non-key items on a similar subject, if compromise is inevitable, simplicity should take priority over subject similarity.

Finally, based on the experiment results, I estimated and compared two major mechanisms underlying turnout over-reporting. I found that impression-management bias accounted for more over-reporting than did self-deception bias in the second wave experiment. If this finding reflects the general situation, then it is impression-management that is a more important mechanism underlying turnout over-reporting, and it should be the prime target to address.

A major limitation of this study is the lack of turnout validation data. This study relies on respondents' self-reported turnout as a proxy measure for turnout over-reporting. It is reasonable to expect this proxy to serve as a good criterion for assessing solutions to over-reporting. However, had I had sufficient resources to conduct validation exercises, I would have drawn conclusions from stronger evidence and with more interesting findings. For example, validation data would make it possible to examine whether some solutions to over-reporting are unexpectedly more likely than others to cause under-reporting.

Gaining a deeper understanding of an existing solution to turnout over-reporting is as important as developing a brand new one. This study has compared ICT and PLT, but there are still other solutions worth further investigation. Continued efforts are therefore needed to gather more information for more efficient use of existing solutions, and also for laying the foundations for future advances in turnout measurement.

## **Supplementary materials**

Supplementary materials are freely available online at: <https://goo.gl/hTrESH>

## **Appendix**

### **1. Experiment administration**

The YouGov Daily Poll in which the experiment was embedded always began with voting intention (i.e. whom to vote for next time) and government approval. There were then a few YouGov tracking questions that might change from day to day but must be asked immediately after government approval. The experiments were appended after those tracking questions. The poll presented experimental questions one at a time, and it did not allow respondents to change their answers to a question once they saw the next one. After the experiments, respondents continued to complete the remaining questions of the Daily Poll.

Table A1 shows the sample sizes of the experiments. None of the AAPOR defined response rates is suitable to calculate and report YouGov quota samples. (For details of YouGov targeted quota sampling, see <https://yougov.co.uk/about/panel-methodology/> and Fieldhouse et al. 2004.) Nonetheless, YouGov reported the number of respondents who dropped out of the interviews. The drop-out rate is higher in Wave 2 than Wave 1; particularly, it is noticeably higher in Group 2 than the other groups in Wave 2. Unfortunately, there is not enough information to explain these differences, but a sensitivity analysis presented in a later section suggests that the findings of this study are quite robust for this drop-out issue.

**Table A1. Sample sizes**

Group	Turnout question	Wave 1 26 May – 12 Jun 2014		Wave 2 24 Feb – 03 Mar 2015	
		Complete	Drop-out	Complete	Drop-out
1	PLT	6,333	113	1,746	91
2	ICT <sub>A</sub> long list and ICT <sub>B</sub> short list	6,150	114	2,016	151
3	ICT <sub>B</sub> long list and ICT <sub>A</sub> short list	6,203	105	1,768	179
4	CTQ	0	0	2,044	142

NOTE: 'Drop-out' means respondents who started the interview but did not complete it.

## 2. Meta-questions

### *Group 1 Meta-question 1:*

Some people may feel a bit uncomfortable when answering certain questions about themselves, while others feel comfortable with that. Here is a question that you answered a moment ago:

[Show the PLT question wording and options]

On a scale below, how comfortable or uncomfortable did you feel when you were answering the question above?

- 1 - Very uncomfortable
- 2
- 3
- 4
- 5
- 6
- 7 - Very comfortable

### *Group 1 Meta-question 2* (only for those who did not tick '7' in the previous question):

Here is a question that you answered a moment ago:

[Show the PLT question wording and options]

What made you feel less than comfortable? Please tick all that apply to you.

- Because I didn't realise that there are public records for checking who actually voted

- Because I don't like to think of my survey answers being checked
- Because voting is a private matter
- Other, please specify \_\_\_\_\_

Group 2 and 3 Meta-question 1 (all Group 2 and 3 respondents):

Here is a question that you answered a moment ago:

[Show the ICTA wording for Group 2 and the ICTB wording for Group 3]

The idea is to keep your answers strictly confidential, especially concerning a key statement that we're particularly interested in. If you had to guess which the key statement is, would you say it is:

- A
- B
- C
- D
- E

**Group 4 Meta-question 1:**

Some people may feel a bit uncomfortable when answering certain questions about themselves, while others feel comfortable with that. Here is a question that you answered a moment ago:

[Show the CTQ wording and options]

On a scale below, how comfortable or uncomfortable did you feel when you were answering the question above?

- 1 - Very uncomfortable
- 2
- 3
- 4
- 5
- 6
- 7 - Very comfortable

**Group 4 Meta-question 2:**

Below is a list of 4 statements. Please tell us which of them you have done. If you are not sure whether you have done that thing, don't tick it. Please tick all that you have done.

- Owned a gun
- Given money to a charity
- Gone to the cinema
- Written a letter to the editor of a newspaper

**Group 4 Meta-question 3:**

Below is a list of 4 statements. Please tell us which of them are true in your case. If you are not sure whether a statement applies to you, don't tick it. Please tick all that are true in your case.

- I feel British more than I feel European
- I approve of Britain's membership of the European Union
- I do not want Britain to adopt the Euro as its currency
- I think the EU should take a lot of blame for the global financial crisis

**3. Random assignment**

Table A2 summarises the results of chi-squared tests for independence between treatment assignment and respondents' characteristics. For most characteristics (16 out of 18 variables in Wave 1 and 20 out of 22 variables in Wave 2), the distributional differences across experiment groups are not statistically significant at the 0.05 level (see column 'Raw'), i.e. randomisation is successful with respect to most characteristics.

It is notable that there are no significant differences in two variables – the attitude toward voting as a civic duty and the self-reported turnout of the 2010 United Kingdom general election (which was recoded from the YouGov's standard variable of voting choice). This indicates the consistency of turnout propensity across groups. Turnout

propensity is the most problematic confounding factor for experimental studies on turnout measures. It is this factor that randomisation aims to rule out. The consistent turnout propensity across groups in the sample of this study provides evidence that randomisation did achieve its aim.

**Table A2. Randomisation checks**

Variable	Category	P-value of Pearson's $\chi^2$ test			
		Wave 1		Wave 2	
		Raw	Weighted	Raw	Weighted
Age	▪18-24 ▪25-39 ▪40-59 ▪≥60	0.019	1.000	0.001	1.000
Gender	▪Male ▪Female	0.185	1.000	0.199	1.000
Social grade	▪A/B ▪C1 ▪C2 ▪D/E	0.116	1.000	0.247	1.000
Newspaper readership	▪Express/Mail ▪Sun/Star ▪Mirror/Record ▪Guardian/Indy ▪Times/FT/Telegraph ▪Other ▪None	0.345	1.000	0.011	1.000
Region of residence	▪North ▪Midlands ▪East ▪London ▪South ▪Wales ▪Scotland	0.813	1.000	0.327	1.000
Party identification	▪Labour ▪Conservative ▪Lib-Dem ▪SNP/PC ▪Other ▪None/do not know	0.153	1.000	0.680	1.000
Ethnicity	▪White British ▪Other white background ▪Chinese ▪Other Asian background ▪Black background ▪Mixed background ▪Other ▪Non-response	0.742	0.430	0.382	0.818
Age finished education	▪≤15 ▪16 ▪17-18 ▪19 ▪≥20 ▪Still at school ▪Non-response	0.420	0.815	0.302	0.149
Highest attained education qualification	▪No formal qualifications ▪Youth training certificate/skill seekers/ recognised trade apprenticeship completed ▪Clerical and commercial ▪City & Guilds certificate ▪City & Guilds certificate advanced ▪ONC ▪CSE grades 2-5 ▪CSE grade 1/GCE O level/GCSE/ School Certificate ▪Scottish Ordinary/Lower Certificate ▪GCE A level or Higher Certificate ▪Scottish Higher Certificate ▪Nursing qualification ▪Teaching qualification (not degree) ▪University diploma ▪University or CNA first degree ▪University or CNA higher degree ▪Other technical, professional or higher qualification ▪Non-response	0.671	0.887	0.783	0.961
Household size	▪1 ▪2 ▪3 ▪4 ▪5 ▪6 ▪≥7 ▪Non-response	0.688	0.541	0.059	0.865
Number of children	▪0 ▪1 ▪2 ▪3 ▪≥4 ▪Non-response	0.539	0.972	0.543	0.694
Household gross income per year	▪<£5000 ▪£5000-9999 ▪£10000-14999 ▪£15000-19999 ▪£20000-24999 ▪£25000-29999 ▪£30000-34999 ▪£35000-39999 ▪£40000-44999	0.523	0.451	0.599	0.394

	▪£45000-49999 ▪£50000-59999 ▪£60000-69999 ▪£70000-99999 ▪£100000-149999 ▪≥£150000 ▪Non-response					
Personal gross income per year	▪<£5000 ▪£5000-9999 ▪£10000-14999 ▪£15000-19999 ▪£20000-24999 ▪£25000-29999 ▪£30000-34999 ▪£35000-39999 ▪£40000-44999 ▪£45000-49999 ▪£50000-59999 ▪£60000-69999 ▪£70000-99999 ▪≥£100000 ▪Non-response	0.144	0.043	0.177	0.394	
Turnout in 2010 U.K. General Election	▪Did not vote ▪Voted ▪Non-response	0.220	0.910	0.162	0.752	
Vote choice in 2010 U.K. General Election	▪Did not vote ▪Conservative ▪Labour ▪Lib-Dem ▪SNP ▪PC ▪BNP ▪Green Party ▪UKIP ▪Other ▪Non-response	0.915	0.866	0.341	0.504	
Voting as a civic duty	▪Strongly disagree ▪Disagree ▪Neither agree or disagree ▪Agree ▪Strongly agree	0.054	0.175	0.183	0.631	
Employment status	▪Working ≥30 hours a week ▪Working 8-29 hours a week ▪Working <8 hours a week ▪Full time student ▪Retired ▪Unemployed ▪Not working ▪Other	0.040	0.223	-	-	
Work type	▪Professional/higher technical work ▪Manager/senior administrator ▪Clerical ▪Sales/Services ▪Foreman/supervisor/other workers ▪Skilled manual work ▪Semi-skilled/unskilled manual work ▪Other ▪Have never worked	0.422	0.344	-	-	
Marital status	▪Married ▪Living as married ▪Separated after being married ▪Divorced ▪Widowed ▪Never married ▪Civil partnership ▪Non-response	-	-	0.111	0.788	
Facebook	▪Use ▪No	-	-	0.270	0.487	
LinkedIn	▪Use ▪No	-	-	0.169	0.138	
Google+	▪Use ▪No	-	-	0.577	0.841	
Twitter	▪Use ▪No	-	-	0.058	0.163	
MySpace	▪Use ▪No	-	-	0.509	0.156	

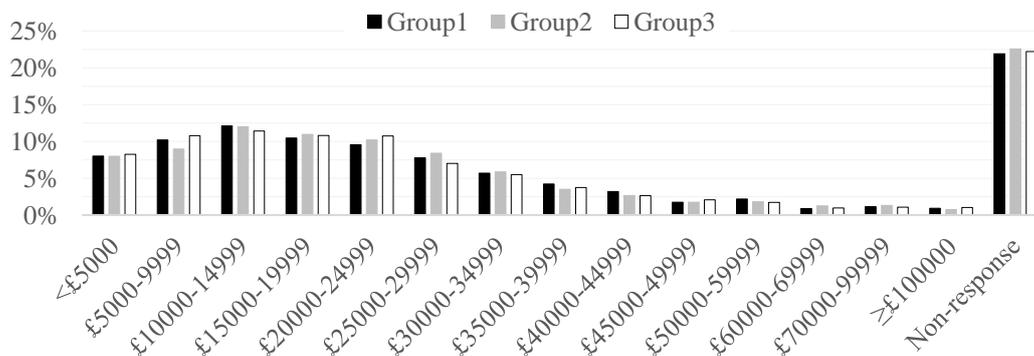
NOTES: (a) Some variables are only available in one wave. (b) I do not change the YouGov variable coding, except to combine categories where expected values are less than 5. (c) The labels ‘Raw’ and ‘Weighted’ refer to p-values calculated for unweighted and weighted data, respectively. (d) To calculate p-values for weighted data, I use Stata commands: *svyset [pw]* and *svy: tabulate*.

Undeniably, the randomisation results are not perfect: the distributions of age in both waves, employment status in Wave 1, and newspaper readership in Wave 2 are statistically different across the groups. Age and newspaper readership are YouGov’s standard weighting variables, so I weight the data to remove the differences in these

two variables. I generate weights for each group in each wave separately by the method of iterative proportional fitting (also known as raking or rim) based on all five standard weighting variables of the YouGov Daily Poll: social grade, newspaper readership, residential region, party identification and the interaction between gender and age.

After weighting, all differences across groups become statistically insignificant at least at the 0.1 level, except for personal income in Wave 1 (see column ‘Weighted’). Further analysis suggests that this income difference should not seriously affect the findings of this study for three reasons. First, as shown in Figure A1, the difference in each income category across groups is fairly small. Even the largest difference is merely 1.8 percentage points (which is in the category of £5,000-9,999 between Groups 2 and 3). Second, the difference lacks a clear pattern. It is difficult to tell which group consists of more high (or low) income respondents. For example, I try to find a pattern by recoding the variable as “<£5,000=£5,000”, “£5,000–9,999=£7,500”, “£10,000–14,999=£12,500” etc., and then regressing the recoded income on the randomisation result. This recoded income does not significantly vary with the groups (p-values>0.30). Third, differences in item non-response rates across the groups are not significant either (p-value>0.43). Taken together, it appears that the income difference across groups is not substantively meaningful, and thus should not undermine the validity of this study.

**Figure A1. Weighted distribution of personal income by groups in Wave 1**



Overall, although randomisation was not perfect, weighting compensates for this issue, and greatly increases the similarity among experiment groups. All analyses in this study are therefore based on weighted data.

#### 4. ICT diagnostics

The validity of ICT rests on three assumptions: treatment randomisation, no design effect and no liar (Imai 2011: 408–409). In the context of this study, the assumption of treatment randomisation means the random assignment of respondents to Group 2 or 3. As discussed in Appendix 3, this assumption holds. The second assumption – no design effect – means that respondents give the same answer to non-key items, no matter whether the key item (turnout) is in the list or not. Suppose that a respondent in Group 3 reports having selected two items in the short list of ICT<sub>A</sub>; if he or she were assigned to Group 2, his or her answer to the long list of ICT<sub>A</sub> must be either two or three (i.e. either selecting two non-key items, or selecting two non-key items plus the key item). Any other answers in this case are in violation of the no-design-effect assumption.

Blair and Imai (2012: 64) developed a statistical test for the design effect. The null hypothesis is “no clear violation of the no-design-effect assumption”. The testing procedure begins by estimating the probabilities of different types of ICT responses. If there is any estimate significantly smaller than 0, the null hypothesis is rejected. The rationale behind this testing is that, by definition, a probability is a real number taking a value between 0 and 1; negative estimates violate this definition, and thus indicate violation of the no-design-effect assumption. Table A2 displays the estimated probabilities of all types of ICT<sub>A</sub> and ICT<sub>B</sub> responses. None of them are negative, so the null hypothesis cannot be rejected (p-values=1.00).

**Table A2. Estimated probabilities of ICT responses**

key-item	Number of non-key item selected	Wave 1		Wave 2	
		ICT <sub>A</sub>	ICT <sub>B</sub>	ICT <sub>A</sub>	ICT <sub>B</sub>
Not select	0	3.6% (0.003)	4.3% (0.003)	3.6% (0.005)	3.8% (0.006)
Not select	1	5.2% (0.006)	7.1% (0.006)	8.2% (0.012)	7.3% (0.011)
Not select	2	24.3% (0.009)	11.8% (0.009)	30.5% (0.017)	12.6% (0.017)
Not select	3	4.0% (0.008)	13.6% (0.010)	4.5% (0.013)	18.9% (0.019)
Not select	4	0.2% (0.003)	1.4% (0.005)	0.1% (0.006)	2.8% (0.011)
Select	0	1.2% (0.005)	0.5% (0.005)	1.6% (0.009)	0.4% (0.008)
Select	1	6.5% (0.007)	7.6% (0.008)	4.0% (0.015)	8.5% (0.014)
Select	2	38.8% (0.009)	20.3% (0.010)	34.9% (0.017)	16.8% (0.019)
Select	3	14.1% (0.006)	27.0% (0.008)	10.5% (0.009)	22.8% (0.015)
Select	4	2.2% (0.002)	6.3% (0.004)	2.1% (0.004)	6.1% (0.007)

NOTE: Numbers in parentheses are standard errors.

The third assumption – no liar – means that respondents do not lie about the turnout item. ICT keeps respondents’ answers about turnout strictly confidential, so respondents usually do not have a motive for lying. One exception is that, when a respondent’s answers to all items in the long list are in the negative, his or her answer to the key item is no longer confidential. In order to cover up the fact of nonvoting, the respondent may still over-report turnout by giving an answer of one. I refer to this issue as the ‘floor effect’, and define it as  $P(Y_i = 1|T_i = 1, Y_i^* = 0)$ , i.e. the probability that a respondent in the long-list group ( $T_i = 1$ ) gives an answer of one ( $Y_i = 1$ ), when the truthful answer is zero ( $Y_i^* = 0$ ). A high number of respondents in the short-list group answering negatively to all non-key items is a sign of the floor effect. In both experiment waves of this study, less than 5% of respondents in each short-list group of ICT<sub>A</sub> and ICT<sub>B</sub> answer negatively to all of non-key items. The ICT design of this study appear to have successfully mitigated the floor effect.

There are two points worth mentioning. First, my definition of floor effect is opposite to Blair and Imai’s (2012: 66) definition:  $P(Y_i = 0|T_i = 1, Y_i^* = 1)$ . This is because I focus on a socially desirable key item, whereas Blair and Imai focused on a socially undesirable key item (e.g. racism). Second, ICT cannot keep respondents’

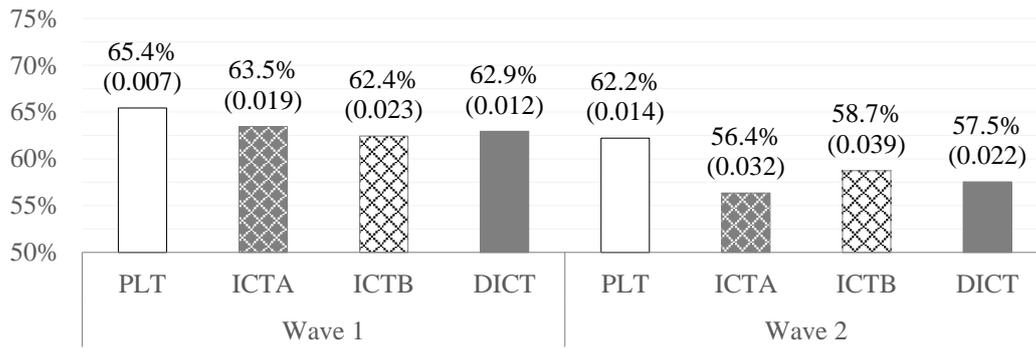
answers about turnout confidential if respondents positively answer all items in the long list. Nonetheless, this should not be a matter of concern for this study, since voters seldom deliberately report not having voted at interview. Overall, the ICT diagnostic results show no clear violation of the three assumptions. It is reasonable to expect that the ICT design of this study is valid.

## **5. Impact of the drop-outs**

The most important finding of this study is that ICT yields lower (and hence better) turnout estimates than estimates generated by PLT. Based on this finding, I argue that ICT can persuade more respondents to admit to nonvoting than PLT can. However, a rival argument is that actual non-voters in the PLT group are more likely to drop out from the experiments than actual voters, whereas actual voters in the ICT groups are more likely to drop out than actual non-voters. Put simply, this rival argument states that ICT yields lower turnout estimates because of the drop-out problem.

To examine this rival argument, I consider an extreme situation where all of the drop-outs in the PLT group were actual non-voters, and all of the drop-outs in the ICT groups were actual voters. Given this situation and the information presented in Table A1, I re-calculate all turnout estimates. Note that this set-up strongly favours the rival argument but, even so, the ICT estimates are still smaller the PLT estimates, as shown in Figure A2. The p-values of two-tailed Z tests for the differences between the PLT and DICT estimates are 0.078 and 0.072 in Waves 1 and 2, respectively. Though larger than the conventional significance level, these p-values are not too far away from it and, most importantly, remember that the set-up of the tests is very much in favour of PLT but against ICT. Therefore, the rival argument is very unlikely to be true.

**Figure A2. Turnout estimates based on the complete and drop-out data**



**NOTES:** (a) Numbers in parentheses are standard errors. (b) Due to a lack of weighting variables in the drop-out data, I cannot generate weights for the drop-outs, so I assign each of the drop-outs a unit weight. (c) I generate a drop-out's turnout answers based on the set-up described in the main text.

## References

- Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. Attempts to Improve the Accuracy of Self-Reports of Voting. In *Questions about Questions*, ed. Judith M. Tanur, 138–153. New York: Russell Sage Foundation.
- ANES. 2014. User's Guide and Codebook for the American National Election Study 2012 Time Series Study. Ann Arbor, Michigan, and Palo Alto, California: The University of Michigan and Stanford University.
- Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Voting Over-Reports? Contrasts of Over-Reporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17(4): 479-498.
- Belli, Robert F., Santa Traugott, and Steven J. Rosenstone. 1994. Reducing Over-Reporting of Voter Turnout: An Experiment Using a 'Source Monitoring' Framework. ANES Technical Report Series: NES010153.
- Belli, Robert F., Sean E. Moore, and John VanHoewyk. 2006. "An Experimental Comparison of Question Forms Used to Reduce Vote Over-Reporting." *Electoral Studies* 25(4): 751-759.
- Belli, Robert, F., Michael W. Traugott, Margaret Young, and Katherine A. McGonagle. 1999. "Reducing Vote Over-Reporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring." *Public Opinion Quarterly* 63(1): 90-108.
- Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1): 47-77.
- Boruch, Robert F., and J.S. Cecil. 1979. *Methods for Assuring Privacy and Confidentiality of Social Research Data*. Philadelphia, Pennsylvania: University of Pennsylvania Press.
- Bunn, Geoffrey C. 2012. *The Truth Machine: A Social History of the Lie Detector*. Baltimore, Maryland: Johns Hopkins University Press.
- Cahalan, Don. 1968. "Correlates of Respondent Accuracy in the Denver Validity Survey." *Public Opinion Quarterly* 32(4): 607-621.
- Campbell, Angus, Gerald Gurin, and Warren Miller. 1952. American National Election Studies, 1952 Time Series Study: Codebook. Ann Arbor, MI: University of Michigan, Center for Political Studies (Producer and Distributor).
- Clarke, Harold D., David Sanders, Marianne C. Stewart, and Paul F. Whiteley 2004. *Political Choice in Britain*. Oxford: Oxford University Press.
- Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT." *Political Analysis* 17(1): 45-63.

- 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*. Ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. Hoboken, New Jersey: Wiley-Interscience.
- Fieldhouse, E., J. Green., G. Evans., H. Schmitt, and C. van der Eijk. 2014. "British Election Study Internet Panel Wave 1: Explanatory Notes." Retrieved from <http://www.britishelectionstudy.com/data-object/2015-bes-internet-panel-wave-1/>. Accessed on 2nd February 2017.
- Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102(1): 33-48.
- Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1): 159-172.
- Górecki, Maciej A. 2011. "Electoral Salience and Vote Over-Reporting: Another Look at the Problem of Validity in Voter Turnout Studies." *International Journal of Public Opinion Research* 23(4): 544-557.
- Grubin, Don, and Lars Madsen. 2007. "Lie Detection and the Polygraph: A Historical Review." *Journal of Forensic Psychiatry and Psychology* 16(2): 357-369.
- Hanmer, Michael J., Antoine J. Banks, and Ismail K. White. 2014. "Experiments to Reduce the Over-Reporting of Voting: A Pipeline to the Truth." *Political Analysis* 22(1): 115-129.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010a. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74(1): 37-67.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010b. "Measuring Voter Turnout by Using the Randomized Response Technique." *Public Opinion Quarterly* 74(2): 328-343.
- Holbrook, Allyson L., and Jon A. Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77(S1): 106-123.
- Imai, Kosuke, Bethany Park, and Kenneth F. Greene. 2015. "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models." *Political Analysis* 23(2): 180-196.
- Imai, Kosuke. 2011. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association* 106(494): 407-416.
- Jones, Edward E. and Harold Sigall. 1971. "The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude." *Psychological Bulletin* 76(5): 349-364.

- Kritzinger, Sylvia, Steve Schwarzer, and Eva Zeglovits. 2012. "Reducing Over-Reporting of Voter Turnout in Seven European Countries: Results from a Survey Experiment." Presented at the 67<sup>th</sup> Annual Conference of the American Association for Public Opinion Research, Orlando, Florida, May 17-20th 2012.
- Krumpal, Ivar. 2013. "Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review." *Quality and Quantity* 47(4): 2025-2047.
- Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. "Racial Attitudes and the 'New South'." *Journal of Politics* 59(2): 323-349.
- Locander, William, Seymour Sudman, and Norman Bradburn. 1976. "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association* 71(354): 269-275.
- Lombroso, Cesare. 1911. *Crime: Its Causes and Remedies*. Translated by Horton P. Henry. *Internet Archive*. Retrieved from <https://archive.org/details/crimeitscausesa00lombgoog>. Accessed on 31st May 2017.
- Miller, Judith. 1984. "A New Survey Technique for Studying Deviant Behaviour." Ph.D. Dissertation, George Washington University.
- Mircea, Comşa, and Andrei Gheorghişă. 2011. "'Many', 'Half' or 'One of Two'? Assessing Counter-biasing Technique to Reduce the Self-Reported Turnout." Presented at the 4<sup>th</sup> Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, July 11<sup>th</sup>-22<sup>nd</sup>, 2011.
- Mircea, Comşa, and Camil Postelnicu. 2013. "Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique." *International Journal of Public Opinion Research* 25(2): 153-172.
- National Research Council. 2003. *The Polygraph and Lie Detection*. Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioural and Social Sciences and Education. Washington, DC: The National Academies Press.
- Ostrom, Thomas M. 1973. "The Bogus Pipeline: A New Ignis Fatuus?" *Psychological Bulletin* 79(4): 252-259.
- Presser, Stanley, Michael W. Traugott, and Santa Traugott. 1990. "Vote 'Over' Reporting in Surveys: The Records or the Respondents?" Presented at the International Conference on Measurement Errors, Tucson, Arizona, November 11<sup>th</sup>-14<sup>th</sup>, 1990.
- Presser, Stanley. 1990. "Can Changes in Context Reduce Vote Over-Reporting in Surveys?" *Public Opinion Quarterly* 54(4): 586-593.
- Raghavarao, Damaraju, and Walter T. Federer. 1979. "Block Total Response as an Alternative to the Randomised Response Method in Surveys." *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1): 40-45.

- Roese, Neal J., and David W. Jamieson. 1993. "Twenty Years of Bogus Pipeline Research: A Critical Review and Meta-Analysis." *Psychological Bulletin* 114(2): 363-375.
- Rogers, Theresa F. 1976. "Interviews by Telephone and in Person: Quality of Responses and Field Performance." *Public Opinion Quarterly* 40(1): 51-65.
- Selb, Peter, and Simon Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 32(1): 186-196.
- Stocké, Volker, and Tobias Stark. 2007. "Political Involvement and Memory Failure as Interdependent Determinants of Vote Over-Reporting." *Applied Cognitive Psychology* 21(2): 239-257.
- Stocké, Volker. 2007. "Response Privacy and Elapsed Time Since Election Day as Determinants for Vote Over-Reporting." *International Journal of Public Opinion Research* 19(2): 237-246.
- Tsuchiya, Takahiro, Yoko Hirai, and Shigeru Ono. 2007. "A study of the Properties of the Item Count Technique." *Public Opinion Quarterly* 71(2): 253-272.
- U.S. Bureau of the Census. 1973. *Current Population Reports, Series P-20, No.253, Voting and Registration in the Election of November 1972*. Washington, D.C.: U.S. Government Printing Office.
- U.S. Bureau of the Census. 1986. *Current Population Reports, Series P-20, No.405, Voting and Registration in the Election of November 1984*. Washington, D.C.: U.S. Government Printing Office.
- Waismel-Manor, Israel, and Joseph Sarid. 2011. "Can Over-Reporting in Surveys be Reduced? Evidence from Israel's Municipal Elections." *International Journal of Public Opinion Research* 23(4): 522-529.
- Warner, Stanley L. 1971. "The Linear Randomised Response Model." *Journal of the American Statistical Association* 66(4): 884-888.
- Wu, Chung-li. 2006. "Vote Over-Reporting and Survey Context: The Taiwan Case." *Issues and Studies* 42(4): 223-239.
- Zeglovits, Eva, and Sylvia Kritzinger. 2014. "New Attempts to Reduce Over-Reporting of Voter Turnout and Their Effects." *International Journal of Public Opinion Research* 26(2): 224-234.

# NOT JUST VALID BUT PRECISE HOW AUXILIARY INFORMATION CAN IMPROVE MODELLING BASED ON THE ITEM-COUNT TECHNIQUE

---

CHI-LIN TSAI

**Abstract** The item-count technique, also known as the unmatched-count technique or the list experiment, is a questioning technique for eliciting respondents' truthful answers to sensitive questions. Statistical inferences based on this technique are more valid but less precise than those based on the direct-questioning technique. This paper introduces a new maximum likelihood estimator for improving the statistical efficiency of the item-count technique. This estimator relies on the item-count technique to make valid estimation of a sensitive issue, and meanwhile the estimator also uses auxiliary information about that sensitive issue to increase the efficiency of estimation. Both Monte Carlo simulations and empirical data analysis provide evidence that the proposed estimator is more efficient than the other two well-known estimators in both univariate and multivariate analysis.

## Introduction

Survey research relies on respondents to provide truthful answers. This is no simple matter when survey questions touch on sensitive issues, and where truthful responses may prompt a degree of social desirability or embarrassment. One technique for revealing real behaviour or attitudes is the item-count technique (also known as unmatched-count technique or the list experiment). The standard design of the item-count technique splits the sample into two groups and presents them with several items (or statements). The control group receives a list of items that are of no interest to the researcher, while the treatment group receives the same list plus the sensitive item that is of interest. Respondents are asked *how many* items they answer in the affirmative. This design provides respondents with higher privacy and thus lower stakes for their truthful answers to the sensitive issue. Researchers cannot be certain who answered the sensitive item in the affirmative, but they can estimate aggregate prevalence.

The idea behind the item-count technique was suggested by Warner (1971: 887) as an alternative to his randomised-response technique. Boruch and Cecil (1979), Raghavarao and Federer (1979) and Miller (1984) continued to develop the idea and finalised the standard design. Several variants have also been devised for improving the applicability of the technique (e.g. Corstange 2009; Droitcour et al. 1991; Trappmann 2014; Hussain, Shah, and Shabbir 2012). To date, the item-count technique has been recognised for its practical superiority (Coutts and Jann 2011) and gained in popularity in a wide range of fields. In political science, it has been used to measure voter turnout (Holbrook and Krosnick 2010; Zeglovits and Kritzingler 2014), electoral fraud (Carkoglu and Aytac 2015; Kiewiet De Jonge 2015; Gonzalez-Ocantos, Kiewiet de Jonge, and Nickerson 2015) and political corruption (Malesky, Gueorguiev, and Jensen 2015).

If confidentiality promotes candour, then the item-count technique should result in less misreporting than the use of direct questioning. This advantage, however, comes at the price of statistical inefficiency. Item-count estimates lack precision, because only the treatment group provides information on the sensitive issue, and that information is not observed directly. Efforts to overcome these drawbacks have focused on question design or analysis methods. [Droitcour et al. \(1991\)](#), for example, propose a double-list experiment, while [Glynn \(2013\)](#) recommends using negatively correlated items to compile lists. [Imai \(2011\)](#) (see also [Blair and Imai 2012](#)), on the other hand, introduce a maximum likelihood estimator as a more efficient method for regression analysis based on the item-count technique.

In this paper, I develop a method that improves both the design of the question and the subsequent analysis. I propose using an auxiliary question to collect information that correlates with the sensitive item. (For example, if the sensitive item measures whether respondents abstained from voting in a past election, a possible auxiliary question is about their intentions to vote in the next election.) I then derive a new maximum likelihood estimator to improve the efficiency of the item-count technique using information contained in that auxiliary variable. Item-count data provide *accurate* information for making inferences about the sensitive item; the auxiliary variable improves inferences by eliciting relatively more *precise* information about the sensitive item.

In this paper, I first review various methods for item-count estimation. I then introduce my proposed method and compare it with other methods. Throughout this paper, examples focus on the issue of reported turnout, but the method has wider applicability.

## Estimators for the standard item-count technique

Consider a standard item-count question where there is one sensitive item and  $J$  reference items (see Appendix 1 for a summary of notation). Let  $T_i$  be the treatment status for respondent  $i$ , where  $i=1, \dots, n$ ;  $T_i=1$ , if the respondent is assigned to the treatment group, and  $T_i=0$ , otherwise. Let  $S_i$  and  $R_{i,j}$  be respondent  $i$ 's potential answers to the sensitive item and the  $j^{\text{th}}$  reference item, respectively, where  $j=1, \dots, J$ . These answers are dichotomous; for example,  $S_i=0$  means that respondent  $i$  answers the sensitive item in the negative;  $R_{i,2}=1$  means the respondent answers the second reference item in the affirmative. By design,  $S_i$  and  $R_{i,j}$  are unobservable; the observed variable is the count of affirmative answers,  $Y_i = T_i S_i + R_i$ , where  $R_i = \sum_{j=1}^J R_{i,j}$ .<sup>1</sup> Given these definitions and assuming treatment randomisation, no design effects, and no liars (Imai 2011: 408–409), any difference between  $Y_i|T_i=0$  and  $Y_i|T_i=1$  is attributed to  $S_i$ . Accordingly, the prevalence of the sensitive item can be estimated as the difference in the mean counts between groups.<sup>2</sup> More formally, with a sample of  $n$  respondents, the estimator is:

$$P(S_i=1) = \frac{\sum_{i=1}^n Y_i T_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n Y_i (1-T_i)}{\sum_{i=1}^n (1-T_i)} \quad \dots \quad \text{Difference-in-Means Estimator} \quad (1)$$

Holbrook and Krosnick's (2010) study of turnout over-reporting provides a concrete example. It has long been established that direct questioning about turnout is prone to over-reporting, i.e. non-voters report having voted. To reduce over-reporting, Holbrook and Krosnick employed the item-count technique in a telephone survey

<sup>1</sup> The definition of  $Y_i$  implies that respondents do not miscount their affirmative answers to items.

<sup>2</sup> The Treatment-Randomisation assumption means that respondents are randomly assigned to either the treatment group or the control group. The No-Design-Effect assumption means that respondents do not change their answers about reference items, depending on their treatment statuses. The No-Liar assumption means that respondents give truthful answers to the sensitive item. However, it does not matter whether respondents give truthful answers to reference items, as long as the No-Design-Effect assumption holds.

( $n=353$ ). Their design included one sensitive item – voting – along with four reference items ( $J=4$ ):

Owned a gun	⋯ $R_{i,1}$
Given money to a charitable organisation	⋯ $R_{i,2}$
Gone to see a movie in a theatre	⋯ $R_{i,3}$
Written a letter to the editor of a newspaper	⋯ $R_{i,4}$
Voted in the Presidential election held on 7th Nov., 2000	⋯ $S_i$

If respondent  $i$  was in the treatment group ( $T_i=1$ ), had seen a movie in a theatre ( $R_{i,3}=1$ ) and voted in the election ( $S_i=1$ ), her answer to the item-count question would be  $Y_i = 1(1) + (0+0+1+0) = 2$ . Had the same respondent been assigned to the control group ( $T_i=0$ ), her answer would have been  $Y_i = \mathbf{0}(1) + (0+0+1+0) = 1$ ; she would not count her turnout in her answer, because turnout did not appear on the control group's list.

Holbrook and Krosnick (2010: 52) reported mean counts for the control and treatment groups of 2.4 and 2.9, respectively. Accordingly, the turnout rate was estimated at  $2.9 - 2.4 = 50\%$  with a standard error of 0.12. The authors also randomly assigned a separate group of respondents to answer a conventional direct question on turnout ( $n=175$ ) and then obtained a higher estimate (72%) with a smaller standard error (0.03). This demonstrates both the usefulness of the item-count technique in reducing over-reporting and its major drawback, i.e. statistical inefficiency.

Holbrook and Krosnick (2010: 53–54) extended the difference-in-means estimator to multivariate analyses. They used ordinary least squares to regress  $Y_i$  on a set of covariates  $\mathbf{X}_i$  and the interaction terms between each of  $\mathbf{X}_i$  and  $T_i$ .<sup>3</sup> Like the

<sup>3</sup>  $\mathbf{X}_i$  is a  $1 \times (1+m)$  vector consisting of a constant term and  $m$  explanatory variables of  $S_i$ . Throughout this paper, a coefficient set is defined as a column vector having the row size equal to the column size of its corresponding covariate set. For example, in the ordinary-least-squares estimator, both  $\delta$  and  $\psi$  are  $(1+m) \times 1$  column vectors.

difference-in-means estimator, the coefficients of interaction terms,  $\delta$ , estimates the prevalence of the sensitive item among the subgroup  $\mathbf{X}_i=\mathbf{x}$ .<sup>4</sup>

$$Y_i = \mathbf{X}_i\boldsymbol{\psi} + T_i\mathbf{X}_i\boldsymbol{\delta} + \epsilon_i \quad \cdots \quad \text{Ordinary-Least-Squares Estimator} \quad (2)$$

$$\mathbf{X}_i\boldsymbol{\psi} = E(Y_i|\mathbf{X}_i, T_i=0)$$

$$\mathbf{X}_i\boldsymbol{\delta} = E(Y_i|\mathbf{X}_i, T_i=1) - E(Y_i|\mathbf{X}_i, T_i=0) = P(S_i=1|\mathbf{X}_i)$$

For example, to examine gender difference in turnout using item-count data, we may fit a model:

$$Y_i = \psi_0 + \psi_1 X_{i,male} + \delta_0 T_i + \delta_1 T_i X_{i,male} + \epsilon_i \quad (3)$$

$\delta$  can be interpreted as though we were interpreting a linear regression  $S_i = \delta_0 + \delta_1 X_{i,male} + \epsilon_i$ . A positive  $\delta_1$  of, say, 0.08 implies that, ceteris paribus, a male's probability of voting is on average 8 percentage points higher than a female's. (In reality, we cannot fit this regression because  $S_i$  is unobserved.)

The ordinary-least-squares estimator is simple but potentially problematic, since a linear functional form is implausible. Modelling the probability of the sensitive item  $P(S_i=1|\mathbf{X}_i)$  on unbounded  $\mathbf{X}_i\boldsymbol{\delta}$  is equivalent to using a linear probability model that might produce nonsensical predictions outside the interval between 0 and 1 (Long 1997: 38-9). The same problem may also arise due to the use of unbounded  $\mathbf{X}_i\boldsymbol{\gamma}$  to model the mean count of reference items  $E(Y_i|\mathbf{X}_i, T_i=0)$ , which is supposed to be bounded between 0 and  $J$ . A solution to this problem is to use appropriate functional forms  $h$  and  $g$  on  $\mathbf{X}_i\boldsymbol{\psi}$  and  $\mathbf{X}_i\boldsymbol{\delta}$ , respectively, and then use nonlinear least squares for estimation (Imai 2011: 409):

<sup>4</sup> If  $\mathbf{X}_i$  merely consists of a constant term, the coefficient of  $T_i$  is identical to the difference-in-means estimate. When there is a categorical covariate in addition to the constant term, the ordinary-least-squares estimates can be replicated by applying the difference-in-means estimator to each sub-group of  $\mathbf{X}_i=\mathbf{x}$ . Beyond one covariate, the ordinary-least-squares estimator is no longer equivalent to the difference-in-means estimator, unless all possible combinations and orders of interaction terms are specified.

$$Y_i = h(\mathbf{X}'_i\boldsymbol{\psi}) + T_i g(\mathbf{X}'_i\boldsymbol{\delta}) + \epsilon_i \quad \dots \quad \text{Nonlinear-Least-Squares Estimator} \quad (4)$$

$$h(\mathbf{X}'_i\boldsymbol{\psi}) = E(Y_i|\mathbf{X}_i, T_i=0)$$

$$g(\mathbf{X}'_i\boldsymbol{\delta}) = P(S_i=1|\mathbf{X}_i)$$

For instance, imposing the logit function on the turnout example yields the model:

$$Y_i = J(1 + e^{-\psi_0 - \psi_1 X_{i,male}})^{-1} + T_i(1 + e^{-\delta_0 - \delta_1 X_{i,male}})^{-1} + \epsilon_i \quad (5)$$

The interpretation of  $\boldsymbol{\delta}$  is analogous to the interpretation of the logistic regression  $P(S_i=1|\mathbf{X}_i) = (1 + e^{-\delta_0 - \delta_1 X_{i,male}})^{-1}$ , as if the sensitive item were directly observed and modelled. A positive  $\delta_1$  of, say, 0.32 means that, ceteris paribus, a male's odds of voting are 1.38 (or  $e^{0.32}$ ) times larger than a female's.

Provided  $h$  and  $g$  are correctly specified, the nonlinear-least-squares estimator is statistically consistent but, in terms of efficiency, there is still room for improvement. Imai (2011) thus derived a maximum likelihood estimator (hereafter IML). He derived IML by first distinguishing four types of respondents, as shown in Table 1.

**Table 1. Classification of respondents by the observed variables  $T_i$  and  $Y_i$**

Type	Treatment status ( $T_i$ )	Item-count answer ( $Y_i$ )	Sensitive item ( $S_i$ )	Reference item count ( $R_i$ )
1	1	0	0	0
2	1	$J + 1$	1	$J$
3	1	$1 \leq Y_i \leq J$	0 or 1	$Y_i - S_i$
4	0	$0 \leq Y_i \leq J$	0 or 1	$Y_i$

Consider the previous turnout example. The first type of respondents were those in the treatment group ( $T_i=1$ ), who reported having not done any item ( $Y_i=0$ ): they neither voted ( $S_i=0$ ) nor engaged in any reference item ( $R_i=0$ ). The second type of respondents were also in the treatment group, but they reported having done all the items ( $Y_i=J+1=5$ ), i.e. they were voters in the election ( $S_i=1$ ) and engaged with all reference items ( $R_i=J=4$ ). The third type of respondents were the rest of the treatment

group, who reported having done some but not all items ( $1 \leq Y_i \leq 4$ ). For instance, if someone of this type gave an answer  $Y_i=2$ , there were two potential outcomes: she voted ( $S_i=1$ ) and did one reference item ( $R_i=2-1=1$ ), or she did not vote ( $S_i=0$ ) but did two reference items ( $R_i=2-0=2$ ). It is impossible to be certain which outcome was true. Finally, the entire control group ( $T_i=0$ ) constituted the fourth type of respondents. Their answers counted only reference items ( $Y_i=R_i$ ), and thus contained no information about whether they voted or not.

Table 1 lays out how the latent variables  $S_i$  and  $R_i$  relate to the observed variables  $T_i$  and  $Y_i$ . Accordingly, the joint probability  $P(T_i, Y_i)$  – which is observed but of no interest to the researcher – can be re-expressed by  $P(R_i, S_i)$  and then factorised into  $P(R_i|S_i)P(S_i)$ , so the prevalence of the sensitive item (i.e. the second factor) can be estimated. Building on this specification and taking  $X_i$  into account, Imai assigned each type of respondents a likelihood function, as shown in Table 2.

**Table 2. Likelihood functions of IML for different types of respondents**

Type	Individual likelihood function ( $L_i^{IML}$ )	Note
1	$h_0(0; X_i\psi_0) \tilde{g}(X_i\delta)$	
2	$h_1(J; X_i\psi_1) g(X_i\delta)$	
3	$h_0(\tilde{y}; X_i\psi_0) \tilde{g}(X_i\delta) + h_1(\tilde{y}-1; X_i\psi_1) g(X_i\delta)$	$1 \leq \tilde{y} \leq J$
4	$h_0(y; X_i\psi_0) \tilde{g}(X_i\delta) + h_1(y; X_i\psi_1) g(X_i\delta)$	$0 \leq y \leq J$

To simplify the notation, let  $h_s(y; X_i\psi_s) = P(R_i=y|S_i=s, X_i=x)$  denote the probability that respondent  $i$  engaged in  $y$  reference item(s), conditional on her unobserved answer to the sensitive item and her status for each covariate.  $g(X_i\delta) = P(S_i=1|X_i)$  denotes the probability that respondent  $i$  engaged in the sensitive item, conditional on her status for each covariate. Conversely,  $\tilde{g}(X_i\delta) = 1-g(X_i\delta) = P(S_i=0|X_i)$ . All these are

probabilities, so  $h_s$  and  $g$  should be functions that restrict predicted values to the unit interval.

As Imai (2011: 411) noted, in order to take account of the correlation between  $S_i$  and  $R_i$ , IML fits  $P(R_i)$  with two sets of coefficients,  $\boldsymbol{\psi}_0$  and  $\boldsymbol{\psi}_1$ . Specifically,  $S_i$  is a right-hand-side variable that models  $P(R_i)$ . Since  $S_i$  is not directly observed, it cannot be simply included in the model. One solution is to estimate two sets of coefficients,  $\boldsymbol{\psi}_1$  for  $S_i=0$  and  $\boldsymbol{\psi}_0$  for  $S_i=1$ . As Equation 6 demonstrates, the difference between the intercepts of the two coefficient sets ( $\psi_{1\alpha}-\psi_{0\alpha}$ ) estimates how  $S_i$  relates to  $R_i$ .

$$\begin{aligned} \begin{cases} P(R_i=y|S_i=0, \mathbf{X}_i=[1,x]) = h_0(y; \psi_{0\alpha}+\psi_{0\beta}X_i) \\ P(R_i=y|S_i=1, \mathbf{X}_i=[1,x]) = h_1(y; \psi_{1\alpha}+\psi_{1\beta}X_i) \end{cases} & \quad (6) \\ \Rightarrow P(R_i=y|S_i=s, \mathbf{X}_i=[1,x]) = h(y; \psi_{0\alpha}+\psi_{0\beta}X_i+(\psi_{1\alpha}-\psi_{0\alpha})S_i+(\psi_{1\beta}-\psi_{0\beta})X_iS_i) & \end{aligned}$$

Though Imai did not mention this explicitly, fitting  $P(R_i)$  with two sets of coefficients also requires the inclusion of the interaction terms between  $S_i$  and each of  $\mathbf{X}_i$ . As can be seen from Equation 6, the specification allows the slopes of  $\mathbf{X}_i$  to vary with  $S_i$ , and the difference between them ( $\psi_{1\beta}-\psi_{0\beta}$ ) is the estimate of interaction effects. This specification is, however, degrees-of-freedom costly, and is often theoretically unnecessary. The full constraint  $\boldsymbol{\psi}_0=\boldsymbol{\psi}_1$  (i.e. no interaction at all), on the contrary, over-simplifies the situation. A more appropriate specification is restricting slopes (e.g. setting  $\psi_{\beta 0}=\psi_{\beta 1}$ ), but letting intercepts vary across  $h_1$  and  $h_0$  to take account of the possible influence of  $S_i$  on  $R_i$  (i.e. allowing  $\psi_{\alpha 0} \neq \psi_{\alpha 1}$ .)

On parameterisation,  $S_i$  as a binary variable is Bernoulli distributed.  $R_i$  is the sum of  $J$  binary variables, so Imai proposed to use a binomial or beta-binomial distribution.<sup>5</sup>

<sup>5</sup> The use of a binomial distribution assumes that  $R_{i,1}, \dots, R_{i,j}$  are identical and independent Bernoulli variables. The use of a beta-binomial distribution allows positive associations among  $R_{i,1}, \dots, R_{i,j}$ . Prentice (1986) extended the beta-binomial distribution to allow moderate negative associations, but its

On estimation, though conventional optimisation methods such as the Newton-Raphson algorithm are usable, Imai simplified the estimation by using the Expectation-Maximisation algorithm and the Bayesian data augmentation algorithm. Among the coefficients in IML,  $\delta$  is of particular interest. If  $g$  is the logit function,  $\delta$  estimates the odds ratio of engaging in the sensitive item, and the interpretation is similar to the discussion of Equation 5.

## Estimation with auxiliary information

IML enhances the estimation of the sensitive issue by exploiting the item-count variables ( $T_i$  and  $Y_i$ ), but does not fully exploit the survey data. Other variables in the data may relate to the sensitive issue and be able to provide additional information. Accordingly, I derive a new maximum likelihood estimator that utilises an auxiliary variable to aid the item-count estimation of the sensitive issue. I begin this section by deriving the estimator (hereafter TML), and then establish the criteria for selecting the auxiliary variable, discuss the statistical properties of TML, and relate it to other estimators.

### 1. Derivation of the likelihood function

In addition to  $T_i$ ,  $Y_i$ , and  $\mathbf{X}_i$ , suppose that the data also contain a Bernoulli variable  $A_i$  collected by the direct-questioning technique from *both* treatment and control groups. I use  $A_i$  as an auxiliary variable and incorporate it into the item-count estimation by modelling the joint probability  $P(R_i, A_i, S_i)$ . The focus remains on  $S_i$ , so I factorise the joint probability into  $P(R_i|A_i, S_i) P(A_i|S_i) P(S_i)$ .<sup>6</sup> Moreover,  $A_i$  is an observed variable;

correlation parameter is difficult to programme. I therefore suggest using the Conway-Maxwell distribution as a computationally feasible alternative; it allows both positive and negative associations among the Bernoulli summands (Kadane 2016).

<sup>6</sup> Alternatively, one may factorise the joint probability into  $P(R_i|A_i, S_i) P(A_i, S_i)$ , and use a bivariate model (e.g. bivariate probit regression) to estimate  $P(A_i, S_i)$ .

$R_i$  and  $S_i$ , though unobserved, are related to the observed  $T_i$  and  $Y_i$ , as elaborated previously. Based on the three observed variables, I classify the respondents into eight types (each type in Table 1 is divided into two sub-types by  $A_i$ ), and assign them likelihood functions as shown in Table 3.

**Table 3. Likelihood functions of TML for different types of respondents**

Type	$T_i$	$Y_i$	$A_i$	Individual likelihood function ( $L_i^{\text{TML}}$ )
1-0	1	0	0 :	$h_0(0; \tilde{X}_i \tilde{\psi}_0) \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$
1-1	1	0	1 :	$h_0(0; \tilde{X}_i \tilde{\psi}_0) k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$
2-0	1	$J+1$	0 :	$h_1(J; \tilde{X}_i \tilde{\psi}_1) \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$
2-1	1	$J+1$	1 :	$h_1(J; \tilde{X}_i \tilde{\psi}_1) k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$
3-0	1	$\tilde{y}$	0 :	$h_0(\tilde{y}; \tilde{X}_i \tilde{\psi}_0) \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + h_1(\tilde{y}-1; \tilde{X}_i \tilde{\psi}_1) \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$
3-1	1	$\tilde{y}$	1 :	$h_0(\tilde{y}; \tilde{X}_i \tilde{\psi}_0) k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + h_1(\tilde{y}-1; \tilde{X}_i \tilde{\psi}_1) k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$
4-0	0	$y$	0 :	$h_0(y; \tilde{X}_i \tilde{\psi}_0) \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + h_1(y; \tilde{X}_i \tilde{\psi}_1) \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$
4-1	0	$y$	1 :	$h_0(y; \tilde{X}_i \tilde{\psi}_0) k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + h_1(y; \tilde{X}_i \tilde{\psi}_1) k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})$

NOTES:  $0 \leq y \leq J$  and  $1 \leq \tilde{y} \leq J$ .  $\tilde{k}_s(\mathbf{X}_i \boldsymbol{\kappa}_s) = 1 - k_s(\mathbf{X}_i \boldsymbol{\kappa}_s)$  and  $\tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) = 1 - g(\mathbf{X}_i \boldsymbol{\delta})$ .  $A_i$  is included in  $h_s$  as a right-hand-side variable, so  $\tilde{X}_i$  is a  $1 \times (1+m+1)$  vector consisting of  $X_i$  and  $A_i$ .

Let  $h_s(y; \tilde{X}_i \tilde{\psi}_s) = P(R_i=y|A_i=a, S_i=s, X_i=x)$  still be the probability that respondent  $i$  engaged with  $y$  reference item(s), but now it is conditional not only on  $S_i$  and  $X_i$  but also on  $A_i$ . Then,  $k_s(\mathbf{X}_i \boldsymbol{\kappa}_s) = P(A_i=1|S_i=s, X_i=x)$  denotes the probability that respondent  $i$  answers the auxiliary question in the affirmative, conditional on  $S_i$  and  $X_i$ .  $g(\mathbf{X}_i \boldsymbol{\delta}) = P(S_i=1|X_i)$  is the focus of attention, denoting the conditional probability that respondent  $i$  engaged in the sensitive item.  $h_s$ ,  $k_s$  and  $g$  are functions restricting predicted probabilities to the unit interval. For the same reasons elaborated in the previous section (particularly Equation 6), TML allows  $P(A_i)$  as well as  $P(R_i)$  to vary with  $S_i$ . On parameterisation, estimation and interpretation, TML is similar to IML. Appendices 2 and 3 provide the classical and expected log-likelihood functions for, respectively, the use of conventional optimisation methods and the Expectation-Maximisation algorithm in TML estimation.

## 2. Selection of the auxiliary variable

There are three criteria for selecting an auxiliary variable: *independence of  $T_i$* , *correlation with  $S_i$* , and *exclusion from  $g$* . The independence criterion ensures no systematic difference in  $A_i$  between the treatment and control groups. The correlation criterion ensures that  $A_i$  contains additional information about  $S_i$  (see next section and Appendix 4 for more discussion). The exclusion criterion ensures that the use of  $A_i$  does not cause omitted-variable bias for the model of  $S_i$ . Remember,  $g$  is an equation that models  $S_i$ , and, by design, a variable in TML can only be either used as  $A_i$  or incorporated into  $g$  as a covariate for modelling  $S_i$ . Any variable that is already included in  $g$  cannot be selected as  $A_i$ ; otherwise, TML will omit that variable from  $g$ , and may cause bias in the model of  $S_i$ .

Note that  $A_i$  in itself is of no immediate interest, except for aiding the estimation of  $S_i$ . As long as the three criteria are met, whether respondents' answers to  $A_i$  are accurate is not a matter of concern. In this regard, a candidate for  $A_i$  is a direct self-report on the sensitive issue. Consider the example where the item-count technique measures whether respondents voted in a *previous* election ( $S_i$  is a retrospective turnout variable). If respondents also answer a conventional turnout question – such as, “Did you vote in [that] election? (yes/no)” – a direct self-report to this question is a candidate for  $A_i$ .

There are other candidates. For example, another one is a variable that reflects responses to a prospective turnout question, such as: “Will you vote in the *next* election? (yes/no)”<sup>7</sup> Since turnout is – at least in part – habitual behaviour (Gerber, Donald, and

<sup>7</sup> If the question is ordinal: “How likely is it that you will vote in the next election? (very likely; likely; unlikely; very unlikely)”, then dichotomise respondents' answers into a dummy variable.

Ron 2003), presumably there is some correlation between prospective and retrospective turnout (correlation with  $S_i$ ). Moreover, future intention cannot be a causal explanation of past behaviour, so it is rare to include respondents' prospective turnout as a covariate in the model of their retrospective turnout (exclusion from  $g$ ). If there is no statistically significant difference between the treatment and control groups' prospective turnout (independence of  $T_i$ ), the variable meets all three criteria, hence it is a suitable candidate for  $A_i$ .

Leaving aside turnout, consider another example about racism. The 1991 American National Race and Politics Survey (Sniderman, Tetlock, and Piazza 1991) used the item-count technique to measure the proportion of white Americans who would be upset by “a black family moving in next door” (i.e.  $S_i$ ).<sup>8</sup> Additionally, the survey also asked white respondents a Likert-scale question: “how do you feel about blacks buying houses in white suburbs? (strongly in favour; somewhat in favour; somewhat opposed; strongly opposed)”. This variable appears to be a good candidate for  $A_i$ . First, there is no statistically significant difference between the treatment and control groups' answers to this house-buying question (independence of  $T_i$ ). Second, its question wording is very similar to the sensitive item, so it is reasonable to expect some correlation between them (correlation with  $S_i$ ). Third, since the two questions are so similar as to be almost tautological to each other, the house-buying variable is arguably not a very meaningful causal explanation of the sensitive item. (When we say: “A black family moving in next door would upset whites, because they oppose blacks buying houses in white suburbs”, we explain almost nothing.) In other words, it is not very meaningful to include the house-buying variable as a covariate in a model of the

<sup>8</sup> These are the data that Kuklinski and his colleagues (1997a; 1997b) used to conduct their influential studies on racial prejudice.

sensitive item. If we do decide to exclude that variable from the covariate set (exclusion from  $g$ ), then it becomes an eligible auxiliary variable.

### 3. Properties of TML

To see how the auxiliary variable aids in item-count estimation, consider two hypothetical cases. First, consider that  $A_i$  is fully independent of  $S_i$  and  $R_i$ . Substantively, this means that respondents' answers to the auxiliary question do not provide additional information about whether they engaged with the sensitive and reference items. In this case, TML is nothing more than a stack of two completely unrelated and separately estimated models – one is IML and the other is a model fitting  $A_i$ . TML estimates of  $\delta$  are no different from their IML counterparts, so there is no improvement in item-count estimation (see Equations A4 and A5 in Appendix 4).

In contrast, assume that  $A_i$  correlates perfectly with  $S_i$  (but researchers are unaware of that). In the turnout example, this means that respondents who voted in the previous election ( $S_i=1$ ) all declare their intention to vote next time ( $A_i=1$ ), whereas those who abstained ( $S_i=0$ ) express no intention to vote at all ( $A_i=0$ ). Note that here  $S_i$  is still unobserved and cannot be fully identified based on  $Y_i$  and  $T_i$  alone. Nonetheless, the observed  $A_i$  in this case contains full information about the unobserved  $S_i$ , so TML can estimate  $\delta$  as if  $S_i$  were directly observed.

For a more intuitive explanation, consider Types 3-0 to 4-1 in Table 3. Since those respondents'  $S_i$  are uncertain, their likelihood functions are specified to exhaust the two possibilities of  $S_i$  – the terms to the right of the plus sign take account of  $S_i=1$  and the left terms take account of  $S_i=0$ . However, when  $A_i=S_i$ , the uncertainty disappears. As illustrated in Table 4, for those whose  $A_i=0$  (Types 3-0 and 4-0), the estimate of  $\tilde{\kappa}_1(\mathbf{X}_i\boldsymbol{\kappa}_1)$  approaches zero, eliminating the possibility of  $S_i=1$  (i.e. the terms to the right

of the plus sign in the likelihood functions). Similarly, for those whose  $A_i=1$  (Types 3-1 and 4-1), the estimate of  $k_0(\mathbf{X}_i\boldsymbol{\kappa}_0)$  approaches zero, eliminating the possibility of  $S_i=0$  (i.e. the terms to the left of the plus sign). The inclusion of  $A_i$  makes  $S_i$  fully identified, and thus greatly improves item-count estimation (see Equations A6 and A7 in Appendix 4).

**Table 4. Likelihood function of TML with a perfect auxiliary variable**

Type	$T_i$	$Y_i$	$A_i$	Individual likelihood function ( $L_i^{\text{TML}}$ )	
1-0	1	0	0	$h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \overbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}^1 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	
1-1	1	0	1	$h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \underbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}_0 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	
2-0	1	$J+1$	0		$h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \overbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}^0 g(\mathbf{X}_i \boldsymbol{\delta})$
2-1	1	$J+1$	1		$h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \underbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}_1 g(\mathbf{X}_i \boldsymbol{\delta})$
3-0	1	$\tilde{y}$	0	$h_0(\tilde{y}; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \overbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}^1 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	+ $h_1(\tilde{y}-1; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \overbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}^0 g(\mathbf{X}_i \boldsymbol{\delta})$
3-1	1	$\tilde{y}$	1	$h_0(\tilde{y}; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \underbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}_0 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	+ $h_1(\tilde{y}-1; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \underbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}_1 g(\mathbf{X}_i \boldsymbol{\delta})$
4-0	0	$y$	0	$h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \overbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}^1 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	+ $h_1(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \overbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}^0 g(\mathbf{X}_i \boldsymbol{\delta})$
4-1	0	$y$	1	$h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \underbrace{k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)}_0 \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})$	+ $h_1(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \underbrace{k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)}_1 g(\mathbf{X}_i \boldsymbol{\delta})$

**NOTES:** The auxiliary variable is assumed to be perfectly and positively correlated with the sensitive item ( $A_i=S_i$ ).  $0 \leq y \leq J$  and  $1 \leq \tilde{y} \leq J$ .

These two cases are for illustrative purposes only. A more realistic case is an  $A_i$  that highly correlates with  $S_i$ . In that case,  $A_i$  does not provide enough information to fully identify  $S_i$ , but  $A_i$  can reduce the uncertainty about  $S_i$ . For instance, people's intentions to vote in the future are often highly correlated with their turnout in the past, but the correlation is imperfect. When respondent  $i$  expresses her intention to vote in the next election ( $A_i=1$ ), that does not necessarily mean that she voted in the previous election ( $S_i=1$ ). However, the closely positive correlation between  $A_i$  and  $S_i$  leads the estimate of  $k_1(\mathbf{X}_i\boldsymbol{\kappa}_1)$  to be larger than that of  $k_0(\mathbf{X}_i\boldsymbol{\kappa}_0)$ , and thus TML will place more

weight on the possibility of  $S_i=1$  than  $S_i=0$ . Reducing the uncertainty about  $S_i$ , TML therefore increases the statistical efficiency of item-count estimation.<sup>9</sup>

#### 4. Relationship with other methods

[Aronow et al. \(2015\)](#) also derived a method to combine an item-count estimate with a direct-questioning estimate of the sensitive item. Their method focuses on a particular type of auxiliary variable – a direct self-report on the sensitive item itself. Moreover, [Aronow et al.](#)'s method is nonparametric, whereas TML is parametric. [Aronow et al.](#)'s method is for univariate analysis only, whereas TML allows for both univariate and multivariate analyses. Most importantly, [Aronow et al.](#)'s method rests on the assumption that a direct self-report on the sensitive item is only prone to over-reporting ( $A_i=1, S_i=0$ ) or under-reporting ( $A_i=0, S_i=1$ ), but not both. This monotonicity assumption restricts the choice of auxiliary variable. In contrast, TML does not rest on that assumption, though it can do if necessary.

Put into different perspectives, TML also links with the other two statistical methods that are less relevant to the item-count technique. One is [Hausman, Abrevaya, and Scott-Morton's \(1998\)](#) method of correction for a misclassified dependent variable. TML can be regarded as an extension of that method, in the case where the auxiliary variable is a direct self-report on the sensitive item. Given respondents' tendencies to misreport sensitive issues, that auxiliary variable falls exactly within the definition of a misclassified variable. From this perspective, it is the item-count variables ( $T_i$  and  $Y_i$ )

<sup>9</sup> It is also possible that  $A_i$  is independent of  $S_i$  but correlates with  $R_i$ . In that case,  $A_i$  is still helpful in the estimation of  $S_i$ . This is because  $S_i$  and  $R_i$  are mutually identified – the more we know about  $R_i$ , the more we know about  $S_i$  (based on the relationship  $Y_i = T_i S_i + R_i$ ). Hence, extracting extra information about  $R_i$  from  $A_i$  improves the estimation of  $S_i$  as well. However,  $R_i$  is the sum of  $J$  items, so devising an  $A_i$  that correlates with  $R_i$  involves dealing with a 1-to- $J$  relationship, which is not as straightforward as devising an  $A_i$  that correlates with a single item,  $S_i$ . Therefore, I do not discuss much about the relationship between  $A_i$  and  $R_i$ .

that provide extra information for TML to correct a misclassified dependent variable ( $A_i$ ) for the purpose of drawing valid inferences about the unobserved true dependent variable ( $S_i$ ). In contrast, from the perspective of the item-count technique, TML still aims to make valid inferences about  $S_i$ , but it is  $A_i$  rather than  $T_i$  or  $Y_i$  that is perceived as the source of extra information (see Appendix 5 for more discussion).

TML is also related to methods for analysing missing data with the aid of non-missing variables. In essence, item-count data are missing data, given that the variable of interest (the sensitive item) is only partially observed. TML can therefore be regarded as estimating the incomplete variable  $S_i$  with the aid of the non-missing variable  $A_i$ . From this perspective, the choice of  $A_i$  is more flexible than discussed in Aronow et al.'s and Hausman, Abrevaya, and Scott-Morton's methods. As Horton and Laird (2001: 23–24) noted, any non-missing variable that is predictive but not explanatory of the key missing variable can improve missing-data analysis to a certain extent. This is echoed by the two criteria for selecting  $A_i$  – correlation with  $S_i$  (i.e. predictive power) and exclusion from  $g$  (i.e. no explanatory power). The difference is that, unlike regular missing data analysis, TML requires  $A_i$  to meet a third criterion – independence of  $T_i$  – in order to maintain treatment randomisation.

## Simulation studies

I assess the performance of TML by Monte Carlo simulations. Two alternative maximum likelihood estimators are included in the simulations for comparison. The first is IML. Imai demonstrated that IML was superior to the nonlinear least-squares estimator of the standard item-count technique. Simulations in this section examine whether TML makes further improvements to IML.

Corstange's (2009) Listit is the second estimator for comparison. Listit works with a non-standard item-count technique in which the control group respondents answer each reference item separately and directly (while the treatment group respondents still answer a standard item-count question). Owing to this non-standard design, Listit has the merit of accommodating non-identical prevalence rates of reference items (Blair and Imai 2012: 57–8).<sup>10</sup> However, since respondents in different groups answer the question in different ways, the non-standard item-count technique runs a higher risk of violating the assumption of no design effect assumption, and that raises concerns about the reliability of Listit (Flavin and Keane 2009). Simulations in this section examine whether TML outperforms Listit. If it does, there is little point in taking the risk of using Listit with the non-standard item-count technique.

## 1. General settings

I consider the case in which an item-count question consists of four reference items ( $J=4$ ) and one sensitive item. Accordingly, I specify the three estimators as follows:

$$\begin{array}{lll}
 \text{TML : } & R_i \mid \mathbf{X}_i, S_i, A_i & \sim \text{Binomial}[4, \text{logit}^{-1}(\tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_s)] \\
 & A_i \mid \mathbf{X}_i, S_i & \sim \text{Binomial}[1, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\kappa}_s)] \\
 & S_i \mid \mathbf{X}_i & \sim \text{Binomial}[1, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\delta})] \\
 \\ 
 \text{IML : } & R_i \mid \mathbf{X}_i, S_i & \sim \text{Binomial}[4, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\psi}_s)] \\
 & S_i \mid \mathbf{X}_i & \sim \text{Binomial}[1, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\delta})] \\
 \\ 
 \text{Listit: } & R_{i,j} \mid \mathbf{X}_i, T_i = 0 & \sim \text{Binomial}[1, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\theta}_j)] \\
 & R_i \mid \mathbf{X}_i, T_i = 1 & \sim \text{Poisson-Binomial}[4, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\theta}_j)] \\
 & S_i \mid \mathbf{X}_i & \sim \text{Binomial}[1, \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\delta})]
 \end{array} \tag{7}$$

<sup>10</sup> Corstange originally assumed that the treatment group's  $R_i$  is binomially distributed with a single parameter  $P_i = P(R_{i,1}=1|\mathbf{X}_i, T_i=1) = P(R_{i,2}=1|\mathbf{X}_i, T_i=1) = \dots = P(R_{i,J}=1|\mathbf{X}_i, T_i=1)$ . This means that reference items should be  $J$  independent and *identical* Bernoulli variables (in terms of their prevalence rates). He approximated the parameter  $P_i$  by  $\sum_{j=1}^J P(R_{i,j}=1|\mathbf{X}_i, T_i=0)/J$ . Blair and Imai pointed out that, in the non-standard item-count technique, because the control group's  $R_{i,j}$  is individually and directly observed, it is more appropriate to assume the treatment group's  $R_i$  follows a Poisson-binomial distribution. Listit can therefore accommodate a situation where  $R_i$  is the sum of  $J$  independent but *non-identical* Bernoulli variables, i.e.  $P(R_{i,1}=1|\mathbf{X}_i, T_i=1) \neq \dots \neq P(R_{i,J}=1|\mathbf{X}_i, T_i=1)$ . IML and TML cannot make such a distributional assumption, since the standard item-count technique does not collect information about each  $R_{i,j}$ .

Considering the complexities of these estimators, I adopt Gelman et al.'s (2008) quasi-Bayesian estimation to ensure numerical stability of optimisation. Specifically, I impose a weakly informative prior – Cauchy(0, 10) – on all parameters, and then estimate them using the Newton-Raphson algorithm. Furthermore, in order to reduce the chance that the algorithm converges a local maximum, each estimation is performed with 11 different sets of starting values.<sup>11</sup>

Simulation data were randomly generated according to Appendix 6. The sample size of each simulated replication is 1,000. Half are the treatment group and half are the control group. Each simulation study is based on 3,000 Monte Carlo replications. Analysis of simulation results focuses on estimates of the sensitive item ( $\delta$ ).

## 2. Simulations without covariates

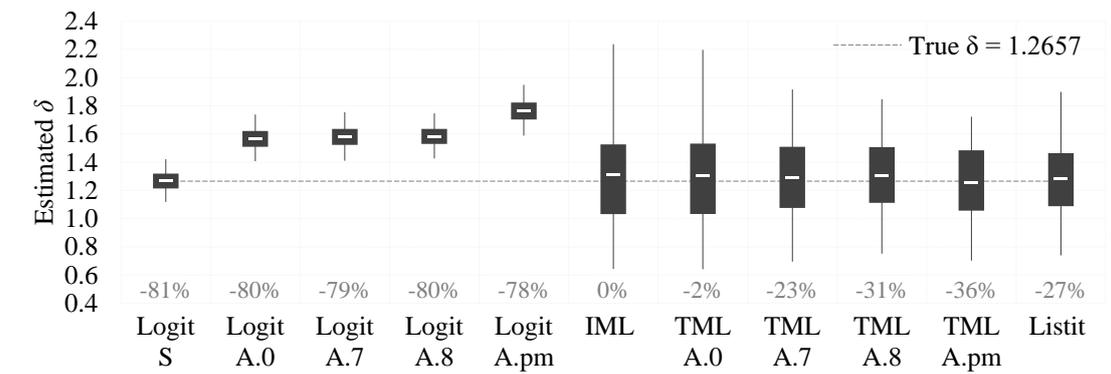
The first simulation study examines the relative performance of TML over other estimators when there are no covariates ( $X_i = 1$  and  $\tilde{X}_i = [1, A_i]$ ). This study also assesses the use of four different auxiliary variables in TML.  $A_{i,0.0}$  contains almost no information about the sensitive item,  $S_i$  – a Pearson's correlation coefficient between them is nearly zero. In contrast,  $A_{i,0.7}$  and  $A_{i,0.8}$  contain different amounts of information about  $S_i$ . (The subscript numbers – 0.7 and 0.8 – denote Pearson's correlation coefficients between  $S_i$  and  $A_i$ .) The fourth auxiliary variable  $A_{i,pm}$  satisfies the positive monotonicity assumption, i.e.  $A_{i,pm} = 1$ , if  $S_i = 1$ , whereas  $A_{i,pm}$  may take

<sup>11</sup> I used Stata 14.2 to conduct simulations. In each simulated replication, Stata automatically determines the first set of starting values by its built-in algorithm, and then randomly generates the rest of 10 sets. Starting values for different estimators are generated independently. Moreover, there is no built-in command in Stata for performing IML and Listit, so I programmed them myself.

a value of 0 or 1, when  $S_i = 0$ .<sup>12</sup> Under this assumption, the specification of TML becomes simpler:  $\text{logit}^{-1}(X_i\kappa_1)$  is fixed at 1, and  $\text{logit}^{-1}(\tilde{\psi}_1) = \text{logit}^{-1}(\psi_1)$ .

Figure 1 presents the simulation results by box plots. Each plot summarises 3,000  $\delta$  estimates from an estimator. The band inside the box is the mean. The greater the gap between the mean and the dashed line (which marks true  $\delta$ ), the larger the bias of the estimator. The length of the whisker is the 95% ‘central range’ (i.e. the range between the 2.5th and 97.5th percentiles). The shorter the whisker, the more efficient is the estimator.<sup>13</sup>

**Figure 1. Simulation results without covariates**



**NOTES:** ‘A.0’, ‘A.7’ and ‘A.8’ denote three dummy variables that have Pearson’s correlation coefficients with the sensitive item of 0.0, 0.7 and 0.8, respectively. ‘A.pm’ denotes a dummy variable that meets the positive monotonicity assumption. ‘Logit-S’ and ‘Logit-A.\*’ denote logistic regressions that use the sensitive item and ‘A\*’, respectively, as the dependent variable. ‘TML-A\*’ denotes a TML that uses ‘A\*’ as the auxiliary variable. Each box plot summarises  $\delta$  estimates from an estimator. The bottom and top of the box are the first and third quartiles. The band inside the box is the mean. The ends of the whisker represent the 2.5th and 97.5th percentiles. The percentage number indicates the relative difference in the 95% central range between an estimator and IML. The quasi-Bayesian estimation is applied to the item-count estimators but not the logistic regressions.

The first estimator is a logistic regression that uses  $S_i$  as the dependent variable.

This is not an item-count estimator, but it serves as a baseline for comparison, showing

<sup>12</sup> If  $S_i$  measures turnout and  $A_{i,pm}$  is direct self-reported turnout, then the positive monotonicity assumption means no under-reporting: voters do not report not having voted.

<sup>13</sup> Another commonly used measure is the root of mean squared error (RMSE). This measure summarises the bias and efficiency of an estimator together. Analysis of simulation results based on RMSE leads to similar conclusions. See the online supplementary materials.

how good the estimation could be, if the sensitive item were observed accurately and analysed directly. The second to fifth estimators are logistic regressions using different auxiliary variables as dependent variables. These are not item-count estimators either. I present them to highlight the differences between the auxiliary variables and the sensitive item.

The remaining six estimators are item-count estimators. None of them is seriously biased, but all of them are much more inefficient than the first logistic regression, as expected. The results of TML-A.0 and IML are remarkably similar. Given that  $A_{i,0.0}$  is uncorrelated with  $S_i$ , it is not surprising that this auxiliary variable cannot provide additional information for TML to enhance estimation efficiency. Nonetheless, TML-A.0 performs as well as IML, suggesting that the use of a non-informative auxiliary variable, though not a help, is at least not a hindrance.

The strength of TML becomes clear when the auxiliary variable correlates with the sensitive item. The stronger the correlation, the more efficient is TML: the 95% central ranges of TML-A.7 and TML-A.8 are 23% and 31% shorter than those of IML, respectively. TML is even more efficient when the monotonicity assumption holds: the 95% central range of TML-A.pm is 36% shorter than that of IML. These results suggest that TML with a properly chosen auxiliary variable can improve the efficiency of item-count estimation, and hence the precision of inferences about the sensitive item.

Furthermore, Listit is also more efficient than IML: its 95% central range is 27% shorter than that of IML. This result is roughly in-between those of TML-A.7 and TML-A.8. (Remember, the simulation settings are in Listit's favour.) Considering that TML does not require the non-standard item-technique whereas Listit does, TML appears to be an attractive alternative to Listit.

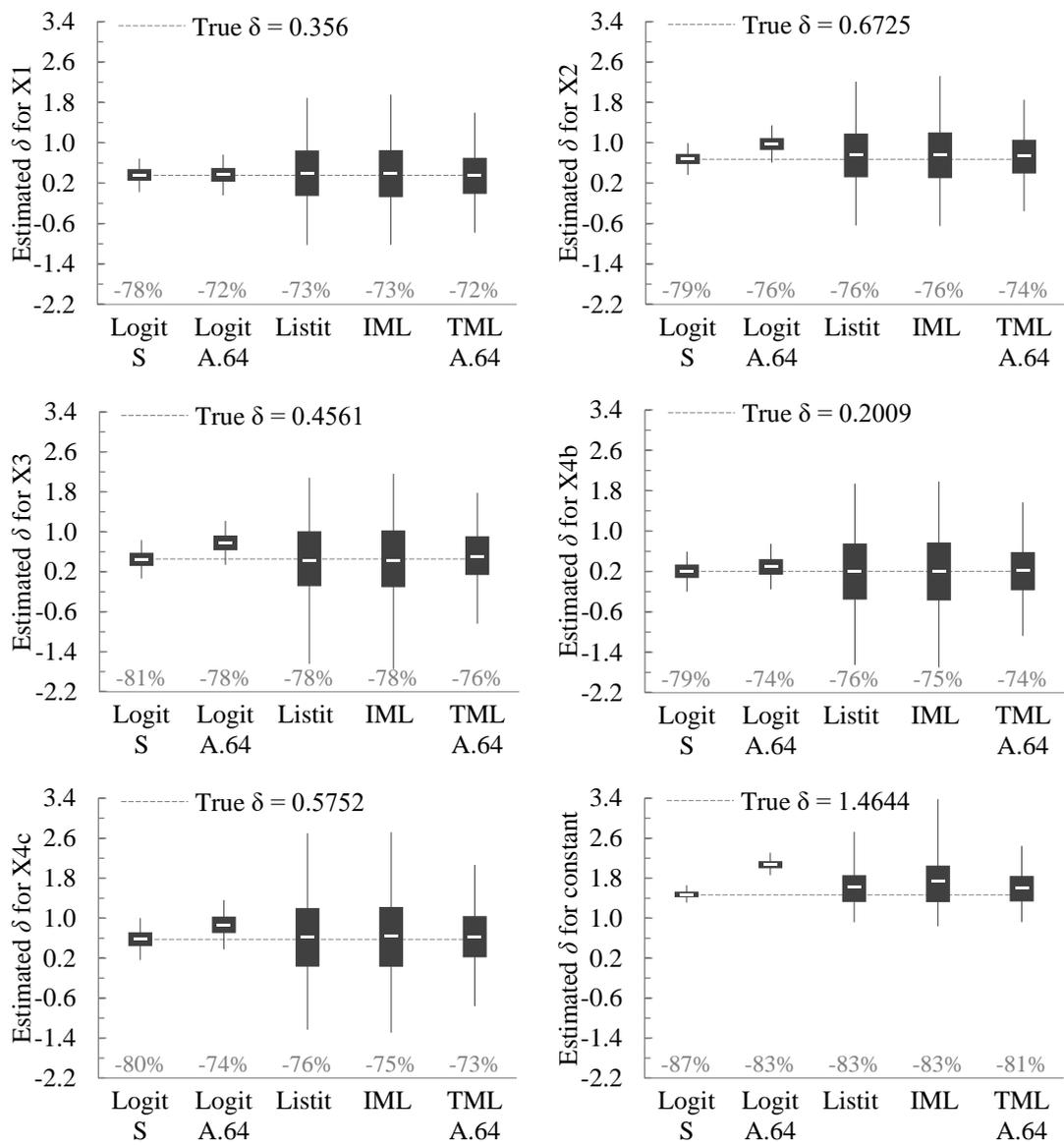
### 3. Simulations with covariates

The second simulation study examines the relative performance of TML over the other estimators in regression analysis. There are five covariates.  $X_{i1}$  is a continuous variable;  $X_{i2}$  is a five-point scale variable;  $X_{i3}$  is a binary variable.  $X_{i4}$  is a 3-category variable converted into two dummies,  $X_{i4b}$  and  $X_{i4c}$ . I do not include interaction terms between  $S_i$  and  $\mathbf{X}_i$  (and  $\tilde{\mathbf{X}}_i$ ) in the equation  $h_s$  of IML and TML for the reasons discussed in a previous section. However, I include interaction items between  $\mathbf{X}_i$  and  $S_i$  in the equation  $k_s$  of TML. In the case of turnout measurement, this specification means that TML does not assume equality between the probabilities of over-reporting and under-reporting.

Figure 2 presents the simulation results in six charts. Each chart summarises the estimates of a regression coefficient from five estimators. The analysis here focuses on TML that uses an auxiliary variable  $A_{i,0.64}$ . (Pearson's correlation coefficient between this variable and the sensitive item is around 0.64.) This TML outperforms both IML and Listit. Overall, its 95% central ranges are 20% – 40% shorter than those of IML, and 16% – 30% shorter than those of Listit (not shown in the charts).<sup>14</sup> These results suggest that TML does represent an improvement in the statistical efficiency of item-count estimation.

<sup>14</sup> Due to space limits, Figure 2 displays  $A_{i,0.64}$  only, but in fact the simulation study also examines other auxiliary variables ( $A_{i,0.05}$ ,  $A_{i,0.74}$ , and  $A_{i,pm}$ ) that have different Pearson's correlation coefficients with the sensitive item. The results show that the stronger the correlation, the more efficient is TML. Moreover, Figure 2 shows that three item-count estimators have noticeable biases for the intercept estimates. These biases diminish when the sample size of each simulated replication increases. See online supplementary materials for all these additional findings.

**Figure 2. Simulation results with covariates**



**NOTES:** ‘A.64’ denotes a dummy variable that has Pearson’s correlation coefficient with the sensitive item of 0.64. ‘Logit-A.64’ and ‘Logit-S’ denote logistic regressions that use A.64 and the sensitive item, respectively, as the dependent variable. ‘TML-A.64’ denotes a TML that uses A.64 as the auxiliary variable. Each chart represents the coefficient for a specific covariate. A box plot in a chart summarises the  $\delta$  estimates of a coefficient from an estimator. The bottom and top of the box are the first and third quartiles. The band inside the box is the mean. The ends of the whisker represent the 2.5th and 97.5th percentiles. The percentage number indicates the relative difference in the 95% central range between an estimator and IML. The quasi-Bayesian approach is applied to the item-count estimators but not the logistic regressions.

### Comparison based on real data

Next, I compare TML and IML using empirical data collected through a YouGov online survey conducted from 25th–27th February 2015 in the United Kingdom. That survey

used the standard item-count technique to measure U.K. respondents' turnout in the 2014 European Parliamentary election:

Below are # statements. Please tell us how many of them have you done. You do not have to tell us which you have and have done. Just tell us how many you have done.

- Owned a gun
- Given money to a charity
- Gone to the cinema
- Written a letter to the editor of a newspaper
- Voted in the May 2014 elections to the European Parliament

If you are not sure whether you have done that thing, don't include it in your count. How many have you done? \_\_\_\_ [Options: 0 to 4 (control group) / 5 (treatment group)]

YouGov drew the sample from its online panel by quota sampling. Sample sizes of the treatment and control groups are 2,016 and 1,768, respectively.

After the item-count question, the survey also asked all respondents about voting choices:

Thinking back to the European Parliamentary Elections on the 22nd May 2014, do you remember which party you voted for then - or perhaps you didn't vote?

- (01) Conservative
- (02) Liberal Democrat
- (03) Labour
- (04) Scottish National Party
- (05) Plaid Cymru
- (06) United Kingdom Independence Party
- (07) Green Party
- (08) British National Party
- (09) Other
- (10) No, I did not vote in the May 2014 elections to the European Parliament
- (11) Can't remember

I dichotomise answers to this question into a variable that takes a value of 1 if a respondent mentioned any of the first nine categories, and 0 otherwise. This dummy

variable is essentially similar to a direct self-reported measure of turnout, so it should correlate with the other measure of turnout, i.e. the sensitive item of the item-count question. However, it is usually not meaningful to model a turnout measure by another turnout measure, so I do not include the dummy variable as a covariate in the model of the sensitive item. Moreover, a t-test shows no significant difference in this dummy variable between the treatment and control groups (two-tailed p-value = 0.187). In short, the dummy variable meets all three selection criteria, so I select it as an auxiliary variable for TML.

I model the sensitive item by seven covariates. Six of them are YouGov's standard variables for sample profiling, so they are presumed to be important in the U.K. context. I recode them as follows: gender (1, male; 0, otherwise), age, social grade (1, AB; 2, C1; 3, C2; 4, DE), residential region (1, England; 0, otherwise), News readership (1, reading certain newspapers; 0, otherwise), and party identification (1, partisan; 0, otherwise). The survey also asked respondents' views on turnout as a civic duty (-2, strongly disagree; -1; 0; 1; 2, strongly agree). I include it in the models as a crucial attitudinal covariate for turnout.

I centre the auxiliary variable and the four dummy covariates to be mean 0, and rescale the three continuous covariates to be mean 0 and standard deviation 0.5. I impose a weakly informative prior – Cauchy(0, 10) – on all parameters, and use quasi-Bayesian estimation. Table 5 presents six models. The standard errors of the  $\delta$  estimates are the focus of analysis. All three TML models yield smaller standard errors than do their IML counterparts, regardless of whether the models have covariates or not, and regardless of whether the models are constrained or not. This result provides further evidence for the statistical efficiency of TML.

**Table 5. Comparison between IML and TML**

	IML Unconstrained	TML Unconstrained	IML Constrained	TML Constrained	IML Unconstrained	TML Unconstrained
	Coef. (SE)	Coef. (SE)	Coef. (SE)	Coef. (SE)	Coef. (SE)	Coef. (SE)
<b><math>\delta</math></b>						
Male			0.235 (0.386)	0.467 (0.236)	0.109(0.374)	0.414(0.246)
Age			0.983 (0.630)	0.539 (0.248)	0.603(0.558)	0.500(0.270)
Social grade			-1.106 (0.421)	-0.615 (0.232)	-1.112(0.400)	-0.632(0.245)
England			-0.155 (0.517)	-0.414 (0.301)	-0.023(0.492)	-0.314(0.305)
News			0.697 (0.407)	0.138 (0.236)	0.545(0.377)	0.107(0.247)
PID			0.036 (0.539)	0.288 (0.306)	0.179(0.419)	0.189(0.317)
Civic duty			3.841 (1.819)	5.926 (0.724)	4.399(1.130)	5.885(0.690)
Intercept	0.341(0.140)	0.140(0.105)	0.067 (0.181)	0.430 (0.159)	0.643(0.316)	0.635(0.170)
<b><math>\kappa_0</math></b>						
Male				0.529 (0.152)		0.523(0.155)
Age				0.282 (0.139)		0.275(0.142)
Social grade				0.133 (0.139)		0.160(0.142)
England				-0.110 (0.219)		-0.109(0.221)
News				0.049 (0.153)		0.060(0.159)
PID				0.753 (0.162)		0.767(0.169)
Civic duty				-0.876 (0.260)		-1.014(0.269)
Intercept		-0.935(0.197)		-1.053 (0.210)		-1.178(0.219)
<b><math>\kappa_1</math></b>						
Male				0.856 (0.280)		0.888(0.257)
Age				0.904 (0.307)		0.880(0.283)
Social grade				-0.175 (0.285)		-0.183(0.257)
England				-0.906 (0.443)		-0.955(0.414)
News				0.328 (0.259)		0.307(0.244)
PID				1.327 (0.273)		1.297(0.244)
Civic duty				0.655 (0.732)		0.989(0.626)
Intercept		5.558(0.371)		2.308 (0.333)		2.071(0.307)
<b><math>\psi_0</math></b>						
Male			0.123 (0.031)	0.084 (0.028)	0.046(0.055)	0.005(0.046)
Age			0.090 (0.036)	0.081 (0.029)	0.027(0.053)	0.029(0.047)
Social grade			-0.210 (0.038)	-0.224 (0.029)	-0.238(0.052)	-0.264(0.046)
England			-0.038 (0.043)	-0.009 (0.038)	-0.098(0.079)	-0.087(0.070)
News			0.011 (0.033)	0.035 (0.029)	0.025(0.052)	0.040(0.046)
PID			-0.001 (0.043)	-0.062 (0.034)	-0.029(0.055)	-0.062(0.047)
Civic duty			0.095 (0.138)	-0.067 (0.051)	-0.182(0.157)	-0.243(0.072)
Auxiliary		0.306(0.074)		0.283 (0.044)		0.269(0.054)
Intercept	0.138(0.028)	0.134(0.048)	0.106 (0.084)	0.066 (0.040)	-0.063(0.108)	-0.031(0.054)
<b><math>\psi_1</math></b>						
Male			0.123 (0.031)	0.084 (0.028)	0.178(0.037)	0.135(0.035)
Age			0.090 (0.036)	0.081 (0.029)	0.150(0.045)	0.109(0.038)
Social grade			-0.210 (0.038)	-0.224 (0.029)	-0.167(0.045)	-0.186(0.039)
England			-0.038 (0.043)	-0.009 (0.038)	-0.004(0.045)	0.034(0.045)
News			0.011 (0.033)	0.035 (0.029)	0.003(0.039)	0.031(0.039)
PID			-0.001 (0.043)	-0.062 (0.034)	0.011(0.048)	-0.048(0.047)
Civic duty			0.095 (0.138)	-0.067 (0.051)	0.282(0.073)	0.147(0.072)
Auxiliary		-5.019(0.364)		0.283 (0.044)		0.281(0.068)
Intercept	-0.051(0.019)	1.755(0.123)	-0.002 (0.067)	-0.006 (0.031)	-0.073(0.032)	-0.080(0.033)
LL	-5063.719	-7326.090	-4966.428	-6804.491	-4981.023	-6822.924

**NOTE:** Constrained models constrain slope coefficients to equality in equations  $\psi_0$  and  $\psi_1$ . ‘SE’ reports a robust standard error. ‘LL’ stands for log-likelihood value. The sample size is 3,784. The analysis is based on unweighted data.

## Limitation

The major limitation of TML is the difficulty of optimisation. As the correlation between the auxiliary variable and the sensitive item increases, TML becomes more efficient, but also more difficult to optimise. Consider an extreme case where  $A_i$  correlates perfectly with  $S_i$ . TML with this auxiliary variable is supposed to be extremely efficient, but actually the optimisation algorithm would never converge because of two problems. First, since  $A_i = S_i$  and both are right-hand-side variables in equation  $h_s$ , the problem of collinearity arises. Second,  $\rho(A_i, S_i) = 1$  means  $k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) = 1$  and  $k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) = 0$ . Then, the so-called ‘separation’ problem arises.<sup>15</sup> These problems may also occur when the correlation between the auxiliary variable and the sensitive item is not perfect but is very strong. Fortunately, whenever these problems occur, they are obvious to researchers, so no one will be misled.

Furthermore, there are solutions to these problems. To address the collinear problem, one can either drop  $A_i$  from equations  $h_0$  and  $h_1$ , or constrain the intercepts of the two equations to equality. (This constraint is equivalent to dropping  $S_i$  from the equations.) As for the separation problem, this is not an uncommon problem. It occurs even in simple logistic regressions. Quasi-Bayesian estimation is a possible solution (Discacciati, Orsini, and Greenland 2015; Firth 1993; Gelman et al. 2008; Heinze and Schemper 2002; Zorn 2005), and the above simulation studies suggest that this solution works well for TML.

<sup>15</sup> This is a limitation of computational re-parameterisation. Suppose that  $k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) = \text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\kappa}_1)$ . As  $\boldsymbol{\kappa}_1$  becomes larger, the inverse logistic function approaches 1 but never quite reaches it. In order to make  $\text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\kappa}_1)$  as close to 1 as possible (and also to make  $\text{logit}^{-1}(\mathbf{X}_i \boldsymbol{\kappa}_0)$  as close to 0 as possible), optimisation algorithm keeps updating the estimates of  $\boldsymbol{\kappa}_1$  towards positive infinity (and updating  $\boldsymbol{\kappa}_0$  towards negative infinity), and thus estimation will never converge.

## **Conclusion**

The item-count technique enhances the accuracy of statistical inferences about sensitive issues at the expense of precision. Most efforts to recoup precision have gone into improving the item-count technique itself. This paper, in contrast, proposes collecting additional information about the sensitive issue, and then using that information to aid item-count estimation. Accordingly, I derived a new maximum likelihood estimator and compared it with two well-known estimators, based on Monte Carlo simulations and real empirical data. The results suggested that the proposed estimator enhances the statistical efficiency of item-count estimation, and hence the precision of inferences about sensitive issues. The main limitation of TML is the difficulty in optimisation, but this limitation is overt, and should not mislead researchers into making invalid inferences.

Although the proposed estimator is derived for the standard item-count technique, the extension to non-standard item-count techniques (e.g. Droitcour et al.'s double-list design and Corstange's Listit) is straightforward. All in all, the proposed estimator holds considerable promise as a means to improve the precision of item-count estimation, and to exploit the potential of the item-count technique. All this estimator needs is a good auxiliary question to collect additional information about the sensitive item. Future studies should therefore treat the auxiliary question as an essential component of the item-count technique: design it together with the item-count question, and use it in the item-count estimation.

## Supplementary materials

Supplementary materials are freely available online at: <https://goo.gl/hTrESH>

## Appendix

### 1. Notation

$n$	Sample size.
$i$	An individual respondent. $i=1, \dots, n$ .
$J$	Number of reference items.
$j$	A reference item. $j=1, \dots, J$ .
$T_i$	Treatment status. $T_i=1$ , if treatment group; $T_i=0$ , otherwise.
$S_i$	Answer to the sensitive item. $S_i=1$ , if affirmative answer; $S_i=0$ , otherwise.
$\tilde{S}_i$	Expectation of $S_i$ (for the EM algorithm).
$R_{i,j}$	Answer to the $j^{\text{th}}$ reference item. $R_{i,j}=1$ , if affirmative answer; $R_{i,j}=0$ , otherwise.
$R_i$	Count of the affirmative answers to the reference items. $0 \leq R_i = \sum_{j=1}^J R_{i,j} \leq J$ .
$Y_i$	Answer to the item-count question ( $T_i S_i + R_i$ ). $0 \leq Y_i \leq J$ , if $T_i=1$ ; $0 \leq Y_i \leq J+1$ , if $T_i=0$ .
$A_i$	Answer to the sensitive item; $A_i=1$ : affirmative; $A_i=0$ : negative.
$\mathbf{X}_i$	An $1 \times (1+m)$ vector consisting of a constant term and $m$ covariates for modelling $S_i$ .
$\tilde{\mathbf{X}}_i$	An $1 \times (1+m+1)$ vector consisting of $\mathbf{X}_i$ and $A_i$ .
$g(\mathbf{X}_i; \boldsymbol{\delta})$	$P(S_i=0 \mathbf{X}_i)$ : the probability of engaging in the sensitive item, given $\mathbf{X}_i$ .
$\tilde{g}(\mathbf{X}_i; \boldsymbol{\delta})$	$1 - g(\mathbf{X}_i; \boldsymbol{\delta})$ .
$k_s(\mathbf{X}_i; \boldsymbol{\kappa}_s)$	$P(A_i=1 S_i=s, \mathbf{X}_i=\mathbf{x})$ : the probability of answering the auxiliary variable in the affirmative, given $S_i$ and $\mathbf{X}_i$ .
$\tilde{k}_s(\mathbf{X}_i; \boldsymbol{\kappa}_s)$	$1 - k_s(\mathbf{X}_i; \boldsymbol{\kappa}_s)$ .
$h_s(y; \mathbf{X}_i; \boldsymbol{\psi}_s)$	$P(R_i=y S_i=s, \mathbf{X}_i=\mathbf{x})$ : the probability of answering $y$ reference item(s) in the affirmative, given $S_i$ and $\mathbf{X}_i$ .
$h_s(y; \tilde{\mathbf{X}}_i; \tilde{\boldsymbol{\psi}}_s)$	$P(R_i=y A_i=a, S_i=s, \mathbf{X}_i=\mathbf{x})$ : the probability of answering $y$ reference item(s) in the affirmative, given $A_i, S_i$ and $\mathbf{X}_i$ .
$L_i^{\text{TML}}$	Individual likelihood function of TML.
$\tilde{L}_i^{\text{TML}}$	Individual likelihood function of TML (for the EM algorithm).
$L_i^{\text{IML}}$	Individual likelihood function of IML.

## 2. Log-likelihood function

Building on Table 3, the log-likelihood function of TML for the entire sample is defined as:

$$\begin{aligned}
 & \ln L^{\text{TML}}(\tilde{\boldsymbol{\psi}}_0, \tilde{\boldsymbol{\psi}}_1, \boldsymbol{\kappa}_0, \boldsymbol{\kappa}_1, \boldsymbol{\delta}; \{Y_i, A_i, T_i, \mathbf{X}_i\}_{i=1}^n) & (A1) \\
 &= \sum_{i \in \text{Type1}} [\ln h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + A_i \ln k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + (1-A_i) \ln \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] \\
 &+ \sum_{i \in \text{Type2}} [\ln h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + A_i \ln k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + (1-A_i) \ln \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \\
 &+ \sum_{i \in \text{Type3} \cup 4} \ln [ h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \times k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{A_i} \times \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{1-A_i} \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + \\
 & \quad h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \times k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{A_i} \times \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{1-A_i} \times g(\mathbf{X}_i \boldsymbol{\delta}) ]
 \end{aligned}$$

## 3. Expectation-maximisation algorithm

Item-count data are incomplete data in the sense that most respondents' answers to the sensitive item are unobserved. In view of this fact, Imai (2011: 410) constructed an EM algorithm to reduce the complexity of IML estimation. This idea applies to TML as well. The EM algorithm for TML is derived as follows.

First of all, treating  $S_i$  as missing data, the E-step of the EM algorithm calculates the expectation of  $S_i$  denoted by  $\ddot{S}_i$ . The calculation is conditional on non-missing data  $Y_i, A_i$  and  $\mathbf{X}_i$ . Although  $T_i$  is also non-missing, it does not correlate with  $S_i$  because of treatment randomisation. Hence the calculation of  $\ddot{S}_i$  does not have to be conditional on  $T_i$ . Following the Bayes' theorem,  $\ddot{S}_i$  can be written as:

$$\begin{aligned}
 \ddot{S}_i &= E(S_i \mid Y_i=y, A_i=a, \mathbf{X}_i=\mathbf{x}) & (A2a) \\
 &= P(S_i=1 \mid Y_i=y, A_i=a, \mathbf{X}_i=\mathbf{x}) \\
 &= \frac{P(S_i=1, Y_i=y, A_i=a \mid \mathbf{X}_i=\mathbf{x})}{P(Y_i=y, A_i=a \mid \mathbf{X}_i=\mathbf{x})}
 \end{aligned}$$

Factorise the numerator:

$$\begin{aligned}
 & P(S_i=1, Y_i=y, A_i=a | \mathbf{X}_i=\mathbf{x}) & (A2b) \\
 = & P(Y_i=y | A_i=a, S_i=1, \mathbf{X}_i=\mathbf{x}) P(A_i=a, S_i=1 | \mathbf{X}_i=\mathbf{x}) \\
 = & P(Y_i=y | A_i=a, S_i=1, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=1, \mathbf{X}_i=\mathbf{x}) P(S_i=1 | \mathbf{X}_i=\mathbf{x})
 \end{aligned}$$

Partition off the denominator according to  $S_i$  and factorise each partition:

$$\begin{aligned}
 & P(Y_i=y, A_i=a | \mathbf{X}_i=\mathbf{x}) & (A2c) \\
 = & P(Y_i=y, A_i=a, S_i=1 | \mathbf{X}_i=\mathbf{x}) + P(Y_i=y, A_i=a, S_i=0 | \mathbf{X}_i=\mathbf{x}) \\
 = & P(Y_i=y | A_i=a, S_i=1, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=1, \mathbf{X}_i=\mathbf{x}) P(S_i=1 | \mathbf{X}_i=\mathbf{x}) \\
 + & P(Y_i=y | A_i=a, S_i=0, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=0, \mathbf{X}_i=\mathbf{x}) P(S_i=0 | \mathbf{X}_i=\mathbf{x})
 \end{aligned}$$

Given the definition  $Y_i = T_i S_i + R_i$ , we know  $Y_i = R_i$ , if  $S_i=0$ , and  $Y_i = T_i + R_i$ , if  $S_i=1$ .

Accordingly, replace  $Y_i$  in  $\ddot{S}_i$  with  $R_i$ :

$$\begin{aligned}
 \ddot{S}_i = & P(R_i=y-t | S_i=1, A_i=a, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=1, \mathbf{X}_i=\mathbf{x}) P(S_i=1 | \mathbf{X}_i=\mathbf{x}) & (A2d) \\
 \div & [ P(R_i=y-t | S_i=1, A_i=a, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=1, \mathbf{X}_i=\mathbf{x}) P(S_i=1 | \mathbf{X}_i=\mathbf{x}) \\
 + & P(R_i=y | S_i=0, A_i=a, \mathbf{X}_i=\mathbf{x}) P(A_i=a | S_i=0, \mathbf{X}_i=\mathbf{x}) P(S_i=0 | \mathbf{X}_i=\mathbf{x}) ]
 \end{aligned}$$

To simplify the notation and connect  $\ddot{S}_i$  back to the likelihood function of TML,  $\ddot{S}_i$  can be rewritten as:

$$\begin{aligned}
 \ddot{S}_i = & \frac{h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})}{h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta}) + h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \check{g}(\mathbf{X}_i \boldsymbol{\delta})} \quad \text{if } A_i=1 & (A2e) \\
 \ddot{S}_i = & \frac{h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta})}{h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) g(\mathbf{X}_i \boldsymbol{\delta}) + h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) \check{g}(\mathbf{X}_i \boldsymbol{\delta})} \quad \text{if } A_i=0
 \end{aligned}$$

Equation A2e applies to respondents whose  $S_i$  is not identifiable, i.e. Types 3 and

4. For Types 1 and 2 respondents, their  $S_i$  are identifiable, so  $\ddot{S}_i=S_i$ :

$$\begin{aligned}
 \ddot{S}_i=1 & \quad \text{if } [T_i, Y_i] = [1, J+1] \quad \text{i.e., Type 1-0 and 1-1 respondents} \\
 \ddot{S}_i=0 & \quad \text{if } [T_i, Y_i] = [1, 0] \quad \text{i.e., Type 2-0 and 2-1 respondents} & (A2f)
 \end{aligned}$$

After it has been generated for all respondents in the sample,  $\ddot{S}_i$  is used as a non-missing version of  $S_i$ . The difference is that, unlike  $S_i$  that is a dummy,  $\ddot{S}_i$  ranges

between 0 and 1, representing the conditional probability that respondent  $i$  engaged with the sensitive item.  $\check{S}_i$  allows us to form a relatively simple log-likelihood function of TML,  $\ln \check{L}^{\text{TML}}$ . The M-step of the EM algorithm maximises  $\ln \check{L}^{\text{TML}}$  with respect to  $\tilde{\boldsymbol{\psi}}_s$ ,  $\boldsymbol{\kappa}_s$ , and  $\boldsymbol{\delta}$ .

$$\begin{aligned}
& \ln \check{L}^{\text{TML}}(\tilde{\boldsymbol{\psi}}_0, \tilde{\boldsymbol{\psi}}_1, \boldsymbol{\kappa}_0, \boldsymbol{\kappa}_1, \boldsymbol{\delta}; \{\check{S}_i, Y_i, A_i, T_i, \mathbf{X}_i\}_{i=1}^n) \tag{A3} \\
&= \sum_{i \in \text{Type1}} [\ln h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + A_i \ln k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + (1-A_i) \ln \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-\check{S}_i) \\
&+ \sum_{i \in \text{Type2}} [\ln h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + A_i \ln k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + (1-A_i) \ln \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \check{S}_i \\
&+ \sum_{i \in \text{Type3} \cup 4} \left\{ \begin{aligned} & \ln [h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \times k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{A_i} \times \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{1-A_i} \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-\check{S}_i) + \\ & \ln [h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \times k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{A_i} \times \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{1-A_i} \times g(\mathbf{X}_i \boldsymbol{\delta})] \check{S}_i \end{aligned} \right\} \\
&= \sum_{i=1}^n \left\{ \begin{aligned} & [\ln h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + A_i \ln k_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + (1-A_i) \ln \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-\check{S}_i) + \\ & [\ln h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + A_i \ln k_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + (1-A_i) \ln \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \check{S}_i \end{aligned} \right\}
\end{aligned}$$

The EM algorithm iterates between the E-step and the M-step. Initially, we can use non-missing data and a set of arbitrarily chosen coefficients  $[\tilde{\boldsymbol{\psi}}_s, \boldsymbol{\kappa}_s, \boldsymbol{\delta}]^{[0]}$  to generate  $\check{S}_i^{[0]}$  and form  $\ln \check{L}^{\text{TML}}$ . (Numbers in superscripted brackets denote the number of iterations.) Maximise  $\ln \check{L}^{\text{TML}}$  to find a new coefficient set  $[\tilde{\boldsymbol{\psi}}_s, \boldsymbol{\kappa}_s, \boldsymbol{\delta}]^{[1]}$ , which is then used to generate a new  $\check{S}_i^{[1]}$ . Substitute  $\check{S}_i^{[1]}$  for  $\check{S}_i^{[0]}$  in  $\ln \check{L}^{\text{TML}}$  to find another set of coefficients  $[\tilde{\boldsymbol{\psi}}_s, \boldsymbol{\kappa}_s, \boldsymbol{\delta}]^{[2]}$ . Repeat these steps until convergence criteria are met.

#### 4. Properties

This appendix discusses the properties of TML using two hypothetical cases. First, suppose that  $A_i$  is completely uncorrelated with  $S_i$  and  $R_i$ , which means:

$$\begin{aligned}
h_s(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_s) &= P(R_i=y|A_i=a, S_i=s, \mathbf{X}_i=\mathbf{x}) \\
&= P(R_i=y|S_i=s, \mathbf{X}_i=\mathbf{x}) \\
&= h_s(y; \mathbf{X}_i \boldsymbol{\psi}_s) \\
k_s(\mathbf{X}_i \boldsymbol{\kappa}_s) &= P(A_i=1|S_i=s, \mathbf{X}_i=\mathbf{x}) \\
&= P(A_i=1|\mathbf{X}_i=\mathbf{x}) \\
&= k(\mathbf{X}_i \boldsymbol{\kappa})
\end{aligned} \tag{A4}$$

Accordingly,  $\ln L^{\text{TML}}$  can be rewritten as:

$$\begin{aligned}
&\ln L^{\text{TML}}(\tilde{\boldsymbol{\psi}}_0, \tilde{\boldsymbol{\psi}}_1, \boldsymbol{\kappa}_0, \boldsymbol{\kappa}_1, \boldsymbol{\delta}; \{Y_i, A_i, T_i, \mathbf{X}_i\}_{i=1}^n) \\
&= \sum_{i \in \text{Type1}} [\ln h_0(0; \mathbf{X}_i \boldsymbol{\psi}_0) + A_i \ln k(\mathbf{X}_i \boldsymbol{\kappa}) + (1-A_i) \ln \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa}) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] \\
&+ \sum_{i \in \text{Type2}} [\ln h_1(J; \mathbf{X}_i \boldsymbol{\psi}_1) + A_i \ln k(\mathbf{X}_i \boldsymbol{\kappa}) + (1-A_i) \ln \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa}) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \\
&+ \sum_{i \in \text{Type3U4}} \ln [h_0(y; \mathbf{X}_i \boldsymbol{\psi}_0) \times k(\mathbf{X}_i \boldsymbol{\kappa})^{A_i} \times \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa})^{1-A_i} \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + \\
&\quad h_1(y-t; \mathbf{X}_i \boldsymbol{\psi}_1) \times k(\mathbf{X}_i \boldsymbol{\kappa})^{A_i} \times \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa})^{1-A_i} \times g(\mathbf{X}_i \boldsymbol{\delta})] \\
&= \left\{ \begin{aligned} &\sum_{i \in \text{Type1}} [\ln h_0(0; \mathbf{X}_i \boldsymbol{\psi}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] \\ &+ \sum_{i \in \text{Type2}} [\ln h_1(J; \mathbf{X}_i \boldsymbol{\psi}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \\ &+ \sum_{i \in \text{Type3U4}} \ln [h_0(y; \mathbf{X}_i \boldsymbol{\psi}_0) \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + h_1(y-t; \mathbf{X}_i \boldsymbol{\psi}_1) \times g(\mathbf{X}_i \boldsymbol{\delta})] \end{aligned} \right\} \\
&+ \left\{ \begin{aligned} &\sum_{i=1}^n [A_i \ln k(\mathbf{X}_i \boldsymbol{\kappa}) + (1-A_i) \ln \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa})] \end{aligned} \right\} \\
&= \ln L^{\text{IML}}(\boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\delta}; \{Y_i, T_i, \mathbf{X}_i\}_{i=1}^n) \quad \dots \text{Model 1 (IML)} \\
&+ \sum_{i=1}^n [A_i \ln k(\mathbf{X}_i \boldsymbol{\kappa}) + (1-A_i) \ln \tilde{k}(\mathbf{X}_i \boldsymbol{\kappa})] \quad \dots \text{Model 2}
\end{aligned} \tag{A5}$$

As a result,  $\ln L^{\text{TML}}$  degenerates into a concatenation of the likelihood functions of two completely unrelated models. That is, we can use IML to estimate  $\boldsymbol{\psi}_s$  and  $\boldsymbol{\delta}$ , and separately use a logistic regression to estimate  $\boldsymbol{\kappa}$ . Theoretically, the sum of the log-likelihoods of the two models will equal the log-likelihood of TML (given that  $k$  is the logit function); the coefficient estimates from the two models will also be the same as the estimates from TML. In other words, TML makes no improvement to IML at all. (One difference is that, unlike  $\boldsymbol{\psi}_s$  in IML,  $\tilde{\boldsymbol{\psi}}_s$  in TML contains an additional coefficient of  $A_i$ , but theoretically that coefficient should be 0 because  $A_i$  is independent of  $R_i$ .)

Second, suppose that  $A_i$  perfectly and positively correlates with  $S_i$ , which means no Types 1-1 or 2-0 respondents in the data:

$$\begin{aligned} k_1(\mathbf{X}_i\boldsymbol{\kappa}_1) &= P(A_i=1|S_i=1, \mathbf{X}_i=\mathbf{x}) = 1 & ; & \quad \check{k}_1(\mathbf{X}_i\boldsymbol{\kappa}_1) = 1 - k_1(\mathbf{X}_i\boldsymbol{\kappa}_1) = 0 \\ k_0(\mathbf{X}_i\boldsymbol{\kappa}_0) &= P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x}) = 0 & ; & \quad \check{k}_0(\mathbf{X}_i\boldsymbol{\kappa}_0) = 1 - k_0(\mathbf{X}_i\boldsymbol{\kappa}_0) = 1 \end{aligned} \quad (\text{A6})$$

Accordingly,  $\ln L^{\text{TML}}$  can be rewritten as:

$$\begin{aligned} & \ln L^{\text{TML}}(\tilde{\boldsymbol{\psi}}_0, \tilde{\boldsymbol{\psi}}_1, \boldsymbol{\kappa}_0, \boldsymbol{\kappa}_1, \boldsymbol{\delta}; \{Y_i, A_i, T_i, \mathbf{X}_i\}_{i=1}^n) & (\text{A7}) \\ &= \sum_{i \in \text{Type 1-0}} [\ln h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + 0 + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] \\ &+ \sum_{i \in \text{Type 2-1}} [\ln h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + 0 + \ln g(\mathbf{X}_i \boldsymbol{\delta})] \\ &+ \sum_{i \in \text{Type 3U4}} \ln [h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \times 0^{A_i} \times 1^{1-A_i} \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + \\ &\quad h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \times 1^{A_i} \times 0^{1-A_i} \times g(\mathbf{X}_i \boldsymbol{\delta})] \\ &= \sum_{i \in \text{Type 1-0}} [\ln h_0(0; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-A_i) \\ &+ \sum_{i \in \text{Type 2-1}} [\ln h_1(J; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] A_i \\ &+ \sum_{i \in \text{Type 3U4}} \{ \ln [h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-A_i) + \\ &\quad \ln [h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) \times g(\mathbf{X}_i \boldsymbol{\delta})] A_i \} \\ &= \sum_{i=1}^n \{ [\ln h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta})] (1-A_i) + \\ &\quad [\ln h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1) + \ln g(\mathbf{X}_i \boldsymbol{\delta})] A_i \} \\ &= \sum_{i=1}^n [(1-A_i) \ln h_0(y; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_0) + A_i \ln h_1(y-t; \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\psi}}_1)] & \dots \text{Model 1} \\ &+ \sum_{i=1}^n [(1-A_i) \ln \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + A_i \ln g(\mathbf{X}_i \boldsymbol{\delta})] & \dots \text{Model 2} \end{aligned}$$

In this case,  $\ln L^{\text{TML}}$  is also a concatenation of the likelihood functions of two models. That is,  $\tilde{\boldsymbol{\psi}}_s$  and  $\boldsymbol{\delta}$  can be estimated in two separate models, and the models can be specified as if their dependent variables,  $R_i$  and  $S_i$ , were directly observed. For example, since  $A_i=S_i$  and given that  $g$  is the logit function, Mode 2 is essentially a logistic regression fitting  $S_i$  to estimate the odds ratio of engaging with the sensitive item (i.e.

$\delta$ ). In other words, all of the uncertainty due to indirect questioning is removed. As the result, TML estimates become much more efficient than IML ones.

This case elaborates how an auxiliary variable can aid in item-count estimation, but it is for demonstration purposes only. There is usually no such perfect auxiliary variable. Even if there was, it might cause computational difficulty (see the discussion of TML limitations).

### 5. Connections

This appendix illustrates the connection between TML and the method of correction for the misclassified dependent variable. According to [Hausman, Abrevaya, and Scott-Morton \(1998: 242\)](#), the log-likelihood function of that method is:

$$\begin{aligned}
 & \sum_{A_i=1} \ln\{ P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x}) + \\
 & \quad [1 - P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x}) - P(A_i=0|S_i=1, \mathbf{X}_i=\mathbf{x})] P(S_i=1|\mathbf{X}_i=\mathbf{x}) \} \\
 & + \sum_{A_i=0} \ln\{ 1 - P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x}) - \\
 & \quad [1 - P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x}) - P(A_i=0|S_i=1, \mathbf{X}_i=\mathbf{x})] P(S_i=1|\mathbf{X}_i=\mathbf{x}) \} \\
 & = \sum_{A_i=1} \ln\{ P(A_i=1|S_i=0, \mathbf{X}_i=\mathbf{x})P(S_i=0|\mathbf{X}_i=\mathbf{x}) + \\
 & \quad P(A_i=1|S_i=1, \mathbf{X}_i=\mathbf{x})P(S_i=1|\mathbf{X}_i=\mathbf{x}) \} \\
 & + \sum_{A_i=0} \ln\{ P(A_i=0|S_i=0, \mathbf{X}_i=\mathbf{x})P(S_i=0|\mathbf{X}_i=\mathbf{x}) + \\
 & \quad P(A_i=0|S_i=1, \mathbf{X}_i=\mathbf{x})P(S_i=1|\mathbf{X}_i=\mathbf{x}) \}
 \end{aligned} \tag{A8a}$$

With the notation of TML, Equation A8a can be rewritten as:

$$\begin{aligned}
 & \sum_{i \in \text{Type}1-1 \cup 2-1 \cup 3-1 \cup 4-1} \ln[k_0(\mathbf{X}_i\boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i\boldsymbol{\delta}) + k_1(\mathbf{X}_i\boldsymbol{\kappa}_1) g(\mathbf{X}_i\boldsymbol{\delta})] \\
 & + \sum_{i \in \text{Type}1-0 \cup 2-0 \cup 3-0 \cup 4-0} \ln[\tilde{k}_0(\mathbf{X}_i\boldsymbol{\kappa}_0) \tilde{g}(\mathbf{X}_i\boldsymbol{\delta}) + \tilde{k}_1(\mathbf{X}_i\boldsymbol{\kappa}_1) g(\mathbf{X}_i\boldsymbol{\delta})]
 \end{aligned} \tag{A8b}$$

or more succinctly:

$$\sum_{i=1}^n \ln [ k_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{A_i} \times \tilde{k}_0(\mathbf{X}_i \boldsymbol{\kappa}_0)^{1-A_i} \times \tilde{g}(\mathbf{X}_i \boldsymbol{\delta}) + k_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{A_i} \times \tilde{k}_1(\mathbf{X}_i \boldsymbol{\kappa}_1)^{1-A_i} \times g(\mathbf{X}_i \boldsymbol{\delta}) ] \quad (\text{A8.3})$$

It can be seen that Hausman, Abrevaya, and Scott-Morton's log-likelihood function is 'embedded' in the log-function of TML. (It is even clear when we compare and contrast Equation A8.3 with Types 3 and 4 respondents' TML log-likelihood functions presented in Equation A1.) Hausman, Abrevaya, and Scott-Morton's log-likelihood function is simpler because their method makes inferences about the latent  $S_i$  from only one observed variable, i.e.  $A_i$ . TML, in contrast, uses three –  $Y_i$ ,  $T_i$  and  $A_i$  – so it can make finer classifications of respondents (eight types in total), and assign different types of respondents with different likelihood functions (In comparison, Hausman, Abrevaya, and Scott-Morton's method only distinguishes those whose  $A_i=1$  from those whose  $A_i=0$ .)

## 6. Simulation data

Simulation data were randomly generated based on ten variables of the 2010 British Election Study (Whiteley and Sanders 2011). Actually, the BES did not collect those variables by the item-count technique, but I used them to create an item-count variable for simulation. I began by recoding these variables in the way described in Table A1, and then analysed them to obtain parameters for data generation. In the analysis, I did not use survey weights, and I dropped respondents who had missing values in any of these variables (list-wise deletion).

**Table A1. Variables for simulations**

	<b>Variable</b>	<b>Label</b>	<b>Response</b>	<b>Recoding</b>
Sensitive item	outcomew	Validated turnout	0) missing 1) voted 2) didn't 3) other	→ missing → 1) voted → 0) didn't → missing
Reference item 1	bq33_1	Experience of seeking help over a crime in home	-1) don't know 1) yes 2) no	→ missing → 1) yes → 0) no
Reference item 2	bq34_1	Experience of seeking NHS treatment	-1) don't know 1) yes 2) no	→ missing → 1) yes → 0) no
Reference item 3	bq35_1	Experience of dealing with risk terrorism	-1) don't know 1) yes 2) no	→ missing → 1) yes → 0) no
Reference item 4	bq36_1	Experience of seeking personal help from MP	-1) don't know 1) yes 2) no	→ missing → 1) yes → 0) no
Auxiliary variable	bq12_1	Direct self-reported turnout	-2) refused -1) don't know 1) voted 2) didn't	→ missing → missing → 1) voted → 0) didn't
Continuous covariate	bq89	Age	-2) refused -1) don't know 18) to 97)	→ missing → missing → copy
Continuous covariate	bq57_5	Voting as a civic duty	-2) refused -1) don't know 1) strongly agree 2) agree 3) neither 4) disagree 5) strongly disagree	→ missing → missing → 5) → 4) → 3) → 2) → 1)
Binary covariate	bq9_1	Party identification	-2) refused -1) don't know 1) none/no 2) Labour 3) Conservatives 4) Lib-Dem 5) SNP 6) Plaid Cymru 7) Green Party 8) UKIP 9) BNP 10) Coalition party 11) other	→ missing → missing → 0) no → 1) yes → 1) yes
Categorical covariate	bq95_1 aq67	Age finish full-time education	-1) don't know 1) 15 or younger 2) 16 3) 17 4) 18 5) 19 or older 6) still at school 7) still at university	→ missing → 1) ≤15 → 2) 16 or 17 → 2) 16 or 17 → 3) ≥ 18 → 3) ≥ 18 → 3) ≥ 18 → 3) ≥ 18

**NOTE:** After list-wise deletion, the sample size for analysis is 2,585.

For simulations without covariates, I generated data by using the following settings. (The Stata code in the online supplementary materials details how I obtained parameters from recoded BES variables.)

$$\begin{aligned}
 R_{i,j} &\sim \text{Bernoulli}(\text{logit}^{-1}(-0.5003)) \\
 R_i &= R_{i,1} + R_{i,2} + R_{i,3} + R_{i,4} \\
 S_i &\sim \text{Bernoulli}(0.78) \quad \therefore \delta = \text{logit}(0.78) = 1.2657 \\
 A_{i,0.7} &\sim \text{Bernoulli}(\text{logit}^{-1}(4.1235S_i - 0.6931)) \\
 A_{i,pm} &= S_i + (1 - S_i)A_{i,0.70} \\
 A_{i,0.0} &\sim \text{Bernoulli}(\text{logit}^{-1}(0.0000S_i - 1.5651)) \\
 A_{i,0.8} &\sim \text{Bernoulli}(\text{logit}^{-1}(5.5000S_i - 1.0285))
 \end{aligned}$$

For simulations with covariates, I generated covariates by:

$$\begin{aligned}
 X_{i1}^* &\sim \text{Truncated Normal}(\mu=51.9231, \sigma=3.0059, \text{min}=18, \text{max}=97) \\
 X_{i1} &= \text{Round}(X_{i1}^*) \\
 X_{i2}^* &\sim \text{Uniform}(0, 1) \\
 X_{i2} &= 1 \quad \text{if } 0.0000 \leq X_{i2}^* < 0.0426 \\
 X_{i2} &= 2 \quad \text{if } 0.0426 \leq X_{i2}^* < 0.1803 \\
 X_{i2} &= 3 \quad \text{if } 0.1803 \leq X_{i2}^* < 0.2716 \\
 X_{i2} &= 4 \quad \text{if } 0.2716 \leq X_{i2}^* < 0.6851 \\
 X_{i2} &= 5 \quad \text{if } 0.6851 \leq X_{i2}^* < 1.0000 \\
 X_{i3} &\sim \text{Bernoulli}(0.7969) \\
 X_{i4}^* &\sim \text{Uniform}(0, 1) \\
 X_{i4b} &= 1 \quad \text{if } 0.2576 \leq X_{i4}^* < 0.6104 \\
 X_{i4b} &= 0 \quad \text{Otherwise} \\
 X_{i4c} &= 1 \quad \text{if } 0.6104 \leq X_{i4}^* < 1.0000 \\
 X_{i4c} &= 0 \quad \text{Otherwise}
 \end{aligned}$$

Then, I centred dummy covariates, rescaled continuous covariates to be mean 0 and standard deviation 0.5, and used these covariates to generate the remaining variables.

$$\check{X}_{i1} = (X_{i1} - \bar{X}_1) / [2 \text{sd}(X_{i1})]$$

$$\check{X}_{i2} = (X_{i2} - \bar{X}_2) / [2 \text{sd}(X_{i2})]$$

$$\check{X}_{i3} = X_{i3} - \bar{X}_3$$

$$\check{X}_{i4b} = X_{i4b} - \bar{X}_{4b}$$

$$\check{X}_{i4c} = X_{i4c} - \bar{X}_{4c}$$

$$R_{i,j}^* = -0.1651\check{X}_{i1} + 0.1029\check{X}_{i2} - 0.0303\check{X}_{i3} + 0.1419\check{X}_{i4b} + 0.1696\check{X}_{i4c} - 0.5033$$

$$R_{i,j} \sim \text{Bernoulli}(\text{logit}^{-1}(R_{i,j}^*))$$

$$R_i = R_{i,1} + R_{i,2} + R_{i,3} + R_{i,4}$$

$$S_i^* = 0.3560\check{X}_{i1} + 0.6725\check{X}_{i2} + 0.4561\check{X}_{i3} + 0.2009\check{X}_{i4b} + 0.5752\check{X}_{i4c} + 1.4644$$

$$S_i \sim \text{Bernoulli}(\text{logit}^{-1}(S_i^*))$$

$$A_i^* = 0.2207\check{X}_{i1} + 0.7911\check{X}_{i2} + 0.7017\check{X}_{i3} + 0.3740\check{X}_{i4b} + 0.9638\check{X}_{i4c} - 0.0823 + (-0.2321\check{X}_{i1} + 0.0766\check{X}_{i2} + 0.2458\check{X}_{i3} - 0.4238\check{X}_{i4b} - 0.7988\check{X}_{i4c} + 3.8891)S_i$$

$$A_i \sim \text{Bernoulli}(\text{logit}^{-1}(A_i^*))$$

## References

- Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behaviour Prevalence." *Journal of Survey Statistics and Methodology* 3(1): 43-66.
- Blair, Graeme, and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1): 47-77.
- Boruch, Robert F., and J.S. Cecil. 1979. *Methods for Assuring Privacy and Confidentiality of Social Research Data*, Philadelphia, Pennsylvania: University of Pennsylvania Press.
- Carkoglu, Ali and S. Erdem Aytac. 2015. "Who Gets Targeted for Vote-Buying? Evidence from an Augmented List Experiment in Turkey." *European Political Science Review* 7(4): 547-566.
- Corstange, Daniel. 2009. "Sensitive Questions, Truthful Answers? Modelling the List Experiment with LISTIT." *Political Analysis* 17(1): 45-63.
- 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(1): 169-193.
- Discacciati, Andrea, Nicola Orsini, and Sander Greenland. 2015. "Approximate Bayesian Logistic Regression via Penalized Likelihood by Data Augmentation." *Stata Journal* 15(3): 712-736.
- 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*. Ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman. Hoboken, New Jersey: Wiley-Interscience.
- Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80(1): 27-38.
- Flavin, Patrick, and Michael Keane. 2009. "How Angry am I? Let Me Count the Ways: Question Format Bias in List Experiments." Presented at the 2008 annual meeting of the American Political Science Association, Boston, Massachusetts, August 28th–31st.
- Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2(4): 1360-1383.

- Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. "Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 47(3): 540-550.
- Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1): 159-172.
- Gonzalez-Ocantos, Ezequiel, Chad P. Kiewiet de Jonge, and David W. Nickerson. 2015. "Legitimacy Buying: The Dynamics of Clientelism in the Face of Legitimacy Challenges." *Comparative Political Studies* 48(9): 1127-1158.
- Gupta, Ramesh C., and Hui Tao. 2010. "A Generalized Correlated Binomial Distribution with Application in Multiple Testing Problems." *Metria* 71(1): 59-77.
- Hausman, Jerry A., Jason Abrevaya, and F.M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-response Setting." *Journal of Econometrics* 87(2): 239-269.
- Heinze, Georg, and Michael Schemper. 2002. "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21(16): 2409-2419.
- 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(1): 37-67.
- Horton, Nicholas J., and Nan M. Laird. 2001. "Maximum Likelihood Analysis of Logistic Regression Models with Incomplete Covariate Data and Auxiliary Information." *Biometrics* 57(1): 34-42.
- Hussain, Zawar, Ejaz Ali Shah, and Javid Shabbir. 2012. "An Alternative Item Count Technique in Sensitive Surveys." *Revista Colombiana de Estadística* 35(1): 39-54.
- Imai, Kosuke. 2011. "Multivariate Regression Analysis for the Item Count Technique." *Journal of the American Statistical Association* 106(494): 407-416.
- Kadane, Joseph B.. 2016. "Sums of Possibly Associated Bernoulli Variables: The Conway-Maxwell-Binomial Distribution." *Bayesian Analysis* 11(2): 403-420.
- Kiewiet de Jonge, Chad P. 2015. "Who Lies about Electoral Gifts? Experimental Evidence from Latin America." *Public Opinion Quarterly* 79(3): 710-739.
- Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997a. "Racial Attitudes and the 'New South'." *Journal of Politics* 59(2): 323-349.
- Kuklinski, James H., Paul M. Sniderman, Kathleen Knight, Thomas Piazza, Philip E. Tetlock, Gordon R. Lawrence, and Barbara Mellers. 1997b. "Racial Prejudice and Attitudes toward Affirmative Action." *American Journal of Political Science* 41(2): 402-419.

- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Newbury Park, California: Sage Publications, Inc.
- Malesky, Edmund J., Dimitar D. Gueorguiev, and Nathan M. Jensen. 2015. "Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment." *American Journal of Political Science* 58(2): 419-439.
- Miller, Judith. 1984. "A New Survey Technique for Studying Deviant Behaviour." Ph.D. Dissertation, George Washington University.
- Prentice, Ross L.. 1986. "Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate Measurement Errors." *Journal of the American Statistical Association* 81(394): 321-327.
- Raghavarao, Damaraju, and Walter T. Federer. 1979. "Block Total Response as an Alternative to the Randomised Response Method in Surveys." *Journal of the Royal Statistical Society – Series B (Methodological)* 41(1): 40-45.
- Sniderman, Paul M., Philip E. Tetlock, and Thomas Piazza. 1991. "Data of 1991 Race and Politics Survey." Survey Research Centre, University of California, Berkeley, CA. Retrieved from [http://sda.berkeley.edu/cgi-bin/hsda?harc\\_sda+natrace](http://sda.berkeley.edu/cgi-bin/hsda?harc_sda+natrace).
- 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(1): 58-77.
- Warner, Stanley L. 1971. "The Linear Randomised Response Model." *Journal of the American Statistical Association* 66(4): 884-888.
- Whiteley, Paul F. and David Sanders. 2011. *British Election Study, 2010: Face-to-Face Survey [computer file]*. Colchester, Essex: UK Data Archive [distributor].
- Zeglovits, Eva, and Sylvia Kritzing. 2014. "New Attempts to Reduce Over-Reporting of Voter Turnout and Their Effects." *International Journal of Public Opinion Research* 26(2): 224-234.
- Zorn, Christopher. 2005. "A Solution to Separation in Binary Response Models." *Political Analysis* 13(2): 157-170.

## CONCLUSION

---

CHI-LIN TSAI

### Summary and discussion of research findings

Valid measurement of voter turnout is plainly a “core business” for election studies. One major problem to overcome is non-voters’ tendency to report having voted, i.e. turnout over-reporting. The main purpose of this doctoral thesis has been to identify effective solutions to that problem. Consisting of four separate but interrelated papers, this thesis has addressed three questions:

- *What solutions are out there?*
- *What works?*
- *Can we do better?*

In Paper 1, I closely re-examined [Presser and Traugott’s \(1992\)](#) study of who over-reports. Their finding – that over-reporters almost never vote – had a significant implication for addressing turnout over-reporting. If their finding were valid, over-reporters should be easily distinguishable from actual voters. Accordingly, it should be possible to use a few behavioural or attitudinal variables to identify over-reporters in surveys, and to correct bias in analyses. However, I revealed that Presser and Traugott’s finding appears to lack generality. This result suggests a gloomy prospect for devising a simple method for distinguishing over-reporters from actual voters, let alone for correcting over-reporting bias after it occurs. Thus, instead of correction, it is prevention that is a more fundamental and promising solution to turnout over-reporting.

Paper 2 addressed the question of “*What solutions are out there?*” In recognising the relative importance of prevention over correction, I conducted a meta-analysis of studies that have experimentally tested methods for preventing turnout over-reporting. Those studies were categorised according to which parts of a survey that they modified to tackle the problem. Different categories of studies have had varying degrees of success in addressing over-reporting. Studies that manipulated the design of the turnout question (i.e. wording, option or format) have been the most fruitful. Studies that focused on questionnaire design demonstrated the possibility of preventing over-reporting by rearranging the position of the turnout question in the questionnaire. However, there is still a lack of an effective design for maximising the order effect. Studies experimenting on survey modes have been the least successful. My meta-analysis of those studies suggests that replacing face-to-face interviews with telephone or web interviews is not an effective solution to turnout over-reporting.

Paper 2 makes both practical and methodological contributions to turnout measurement. I list a catalogue of solutions to over-reporting for easy reference. It has great practical value for those who intend to measure turnout by survey. Anyone with the catalogue can quickly have an overall idea about what solutions are available, how they work, and which ones best suit their needs. Moreover, I have also identified a number of methodological issues in need of further investigation. A particularly important one is the need for comparisons between solutions to turnout over-reporting. Most studies have experimented on only one single solution at a time, so the questions of which solution is better and under what circumstances remain largely unanswered. My meta-analysis has partly answered to this question by examining all the separate studies together. For more complete answers to the question, it is necessary to experimentally compare two or more innovative solutions.

There are many interesting and meaningful comparisons to make. For example, both randomised-response and item-count are indirect-questioning techniques aiming to elicit truthful answers by enhancing response privacy. Which one is better for preventing turnout over-reporting, and why? Holbrook and Krosnick (2010a; b) experimented on both techniques under the same settings in the same surveys, thus allowing for a more objective comparison of techniques.<sup>1</sup> It may be even more important and interesting to compare solutions that are inspired by opposing rationales. Therefore, in Paper 3, I experimentally compared the item-count and pipeline techniques, addressing the question of “*What works?*”

The item-count technique embodies the idea that confidentiality promotes candour, whereas the pipeline technique leads respondents to believe that nothing is confidential and thus honesty is the best policy. The experiment results showed that the item-count technique persuaded more respondents to admit to non-voting than did the pipeline technique, though both performed better than a conventional turnout question. This suggests that creating a comfortable atmosphere for truth-telling is more effective for preventing over-reporting than amplifying feelings of unease about lying. Further analysis indicates that this is mainly due to the relative difficulties in evoking sufficient feelings of unease about lying while not breaching research ethics (at least by the pipeline technique).

While proving its usefulness in preventing turnout over-reporting, the experiment results also clearly demonstrated two drawbacks of the item-count technique: (1)

<sup>1</sup> Holbrook and Krosnick conducted several experiments to compare three types of turnout questions: randomised-response technique, item-count technique and direct-questioning technique (i.e. a standard turnout question). In each of the experiments, respondents were randomly assigned to answer one of three questions. However, Holbrook and Krosnick presented their results in two papers: one compared direct-questioning and randomised-response techniques; the other compared direct-questioning and item-count techniques.

respondents take more time to answer an item-count question than other questions, and (2) inferences based on the item-count technique, though accurate, are less precise than those based on other techniques. Both drawbacks are the price of creating a comfortable atmosphere for truth-telling. Moreover, experiment results suggested that complicated item wording increases the cognitive burden on respondents, and thus can reduce the practicability of the technique. Therefore, it is important to keep the items simple.

Overall, the item-count technique shows greater potential to prevent turnout over-reporting than the pipeline technique. In order to exploit the potential of the item-count technique, Paper 4 addressed the question of “*Can we do better?*” I developed a new statistical estimator, called ‘TML’, for overcoming a major drawback of the item-count technique, i.e. statistical inefficiency. TML takes advantage of the item-count technique to produce valid estimates of turnout, and it also uses an auxiliary variable to increase the efficiency of estimation. Monte Carlo simulations and empirical data analysis showed that TML is more efficient other estimators. TML holds considerable promise as a means to make the item-count technique a better solution to turnout over-reporting.

## **Implications and recommendations for future research**

The research results of these papers have several implications for developing effective solutions to turnout over-reporting. First, *practicability is the number one priority*. In Paper 2, a comparison of two indirect-questioning techniques clearly demonstrated the importance of practicability. Both techniques are inspired by the same rationale, but very different in their practicability. The item-count technique is reasonably practicable, and has repeatedly proven to be an effective solution to turnout over-reporting. In contrast, the randomised-response technique is not very practicable, and no studies

have found it effective against turnout over-reporting. In some experiments, the randomised-response technique even backfired, worsening turnout measurement. Moreover, the experiment results in Paper 3 suggest avoiding over-complicated items, so as not to reduce the effectiveness of the item-count technique. That is also evidence for the importance of practicability.

Second, *awareness promotes effectiveness*. In Paper 2, analysis of order effects found that designs that made extra effort to raise respondents' awareness of the purpose of the question arrangement were more effective against turnout over-reporting than designs not doing so. Then, analysis of mode effects found that [Stocké's \(2007\)](#) mode-switch design greatly reduced turnout over-reporting, and that is a design that clearly manifests itself to respondents. It switched the mode from face-to-face interviewing to self-administered mode *during* the interview, so respondents could easily perceive a higher degree of response privacy being created for truth-telling. Furthermore, studies that manipulated the question design have been more successful in reducing turnout over-reporting than studies that manipulated the questionnaire or survey design. One plausible explanation is that it is easier for respondents to notice changes in the question itself (i.e. wording, option or format) than changes in the question order or survey mode (unless those changes occur during the interview, as in the mode-switch design). These findings all suggest the necessity of raising respondents' awareness of the solution, or at least making respondents 'feel' something different because of the solution.

Third, *speaking softly appears to be more effective than wielding a big stick*. The experiment results in Paper 3 indicate that creating a comfortable atmosphere for truth-telling is more effective against turnout over-reporting than amplifying feelings of unease about lying. Certainly, more research is still needed. However, at this moment, when choosing between approaches to prevent over-reporting, all other things being

equal, researchers should give priority to creating a comfortable atmosphere for truth-telling. Otherwise, researchers should manage to evoke sufficient feelings of unease, while not breaching research ethics. For example, in Paper 3 and [Hanmer, Banks, and White's \(2014\)](#) study, the wording of the pipeline technique told respondents that there are independent means for checking the truth of their responses. We can make this claim more explicit by, for instance, telling respondents which governmental sections are responsible for maintaining turnout records for checking, or by showing respondents what a turnout record look like.

Fourth, *control is better than ignorance*. When none of the solutions is feasible, researchers should still try to assess and control the possible consequences of turnout over-reporting. For example, Paper 2 reviewed some methods for preventing misremembering causing over-reporting, such as the source-monitoring technique. If those methods are unfeasible for practical reasons, we may at least use [Kritzinger, Schwarzer, and Zeglovits' \(2012\)](#) turnout-scale options to let respondents indicate how certain (or uncertain) they are about their answers, so that we have some information to assess and control over-reporting bias.

Additionally, it is recommended that further research is undertaken on the following topics. First, in Paper 2, I listed a catalogue of solutions to turnout over-reporting for easy reference. The catalogue needs constantly updating, in order to maintain its practical value. Second, in Paper 3, I experimentally compared two innovative solutions. More comparisons of this kind are needed, in order to gather more information for the practical use and methodological development of solutions to turnout over-reporting. Third, it is worth trying to synthesise multiple kinds of solutions together, in order to profit from different ideas at the same time. [Holbrook and Krosnick's \(2013\)](#) study is an example. In Appendix 1, I also present a proposal to test

a solution inspired by the pipeline technique and studies on order effects. Fourth, while devising new solutions, future studies should also keep improving existing solutions, just as I did in Paper 4. Fifth, in addition to prevention, correction to turnout over-reporting is difficult but worth trying as a last resort. [Katz and Katz \(2010\)](#) made an attempt. They treated direct self-reported turnout as a misclassified dependent variable, and then used the misclassification model to obtain valid estimates. The estimator that I devised in Paper 4 – TML – is another attempt (if we consider TML to be a model that uses the item-count technique to correct the misclassified self-reported turnout; see Paper 4, Section: “Relationship with other methods”).

Though this thesis has focused on the issue of measuring turnout in post-election surveys, the research results are of wider applicability. In pre-election surveys, respondents’ intentions to turn out are no less important than their reports in post-election surveys. Valid measurement of turnout intention is crucial to the accurate prediction of election outcomes, which is of even greater interest to scholars, pollsters, campaigners and people who care about the outcomes. The findings of this thesis contribute to the measurement of turnout intention, and hence to election predictions. For example, in Appendix 2, I propose two innovative questions about turnout intention for future research to test. One employs the pipeline technique; the other applies the item-sum technique, i.e. a variant of the item-count introduced by [Trappmann et al. \(2014\)](#).

Finally, this thesis makes contributions going beyond electoral studies. Research that involves asking people about sensitive questions can benefit from this thesis. As noted in the introductory chapter, turnout can be a sensitive question. Hence, the findings from the four papers, especially their implications as discussed in this concluding chapter, should be generally applicable to designing, selecting and using

solutions to misreporting on other sensitive issues. Moreover, the solution catalogue listed in Paper 2 is useful for researchers who intend to measure sensitive issues other than turnout. For example, to ask people about their frequency of alcohol abuse, researchers who refer to the catalogue should find several solutions that are immediately applicable, such as the mode-switch design (since it does not require any changes to researchers' original questions or questionnaire designs.) Some solutions in the catalogue, though not directly applicable, may inspire researchers to search for or even innovate new ones. For example, the catalogue can make researchers aware of the item-count technique and inspire them to find an alternative, e.g. the item-sum technique, to measure the frequency of alcohol abuse.<sup>2</sup>

Most importantly, as noted in Paper 3, item count is becoming a very popular technique for eliciting truthful answers on a wide range of sensitive issues. Therefore, TML – which is described in Paper 4 and holds considerable promise as a means of improving item-count techniques – should make significant contributions to a wide range of scientific research. For example, to model women's experiences of abortion, researchers just need to collect information by both item-count and direct-questioning techniques, and then use TML to combine two pieces of information for modelling. Research results based on TML are more accurate and precise than the use of a direct-questioning or item-count technique alone, and so promote a better understanding of a sensitive issue.

All in all, this thesis has gone some way to enhancing our understanding of solutions to turnout over-reporting, and it makes contributions to and beyond electoral

<sup>2</sup> The item-count technique is applicable to measuring a dichotomous sensitive variable, such as "voting vs non-voting". The item-sum technique is applicable to measuring a continuous sensitive variable, such as the frequency of alcohol abuse.

studies. However, as pointed out above, there are still many areas worthy of further exploration. Continued efforts are needed to extend the research results of this thesis and provide more insights into solutions to turnout over-reporting.

## Appendix

### 1. A proposal to experiment with a solution to turnout over-reporting

#### “OH! WHAT A TANGLED WEB”

##### PERSUADING RESPONDENTS NOT TO OVER-REPORT TURNOUT

Individual-level research on voter turnout heavily relies upon survey data but, over the past decades, researchers in this field have been haunted by respondents’ inaccurate responses (Parry and Crossley 1950: 67-9). Turnout over-reporting – i.e., non-voters reporting having voted – is observed in surveys around the world (Selb and Munzert 2014: 191; Swaddle & Heath 1989), and has the potential to distort research on turnout (Silver et al. 1986.) This proposal presents a solution to over-reporting, and an experiment for testing it. The solution involves using a preamble to the turnout item. *The preamble ostensibly explains to respondents the pattern of upcoming questions, but actually aims to remind respondents that once you tell a lie, you need more lies to cover it up.* By making respondents aware of the ‘costs’ of over-reporting, we should reduce their motivation for doing so.

##### THEORETICAL BACKDROP

There are at least two kinds of motivations underlying survey response to sensitive questions: the drive for socially desirable responses, and the drive for honesty (Stocké and Stark 2007: 240). Given that voting is often regarded as the behaviour of a ‘good citizen’ (Clarke et al. 2004: 274), the desire to appear in a socially desirable light has long been identified as an important cause of turnout over-reporting (e.g., Cahalan 1968: 621; Belli et al. 2001: 479-80; Górecki 2011: 8), and methods for mitigating respondents’ sense of social desirability have been proposed to tackle over-reporting (e.g., Abelson et al. 1992; Belli et al. 1994: 4-5; Campbell et al. 1952: 100; Holbrook and Krosnick 2010a, b; Mircea and Gheorghită 2011).

In emphasising the ‘negative’ motivation of social desirability, these studies have overlooked and hence failed to capitalise on the ‘positive’ motivation for truthful reporting. Firstly, respondents may be reluctant to violate the norm of honesty and to risk being revealed as a liar. Secondly, lying is more cognitively demanding than is truth telling (Vrij et al. 2006). Thirdly and most fundamentally, people have an intrinsic aversion to lying and the internal psychic costs that it generates (Gneezy 2005: 338). For several reasons, then, honesty may often be the best policy.

It is not uncommon for social psychologists to elicit more valid self-reports by manipulating people’s motivation for honesty (e.g., Jones and Sigall 1971; Rasinski et al. 2005). Hanmer et al. (2014) applied the pipeline technique – which leads people to believe that the researcher has independent means of checking the truth of responses – to the case of turnout, and found that strengthening motivations for honesty in this way does indeed improve the accuracy of self-reported turnout. Nevertheless, the pipeline technique has its limitations (Nederhof 1985: 273; Aguinis and Henle 2001: 355-356) – e.g., when checking the truth of responses is in fact unfeasible, employing pipeline methods may involve deceiving respondents, which is the last thing researchers want to do. Therefore, I propose an alternative method for motivating respondents to provide accurate reports of turnout behaviour.

##### THE PROPOSED SOLUTION

In most political surveys, the turnout question is invariably followed by a series of questions about

party choice, reasons for voting that way, timing of decision, perhaps method of voting, and so on. Those who over-report turnout at the outset will then face the cognitive and psychic costs of further dishonesty. If they see this coming, they may be persuaded to admit non-voting just to avoid that tangled web. However, respondents are unlikely off their own bats to weigh carefully the costs and gains associated with truthful reporting (Tourangeau et al. 2000: 284). So why not prompt them to do so with a preamble that, while ostensibly a neutral preview of upcoming questions, also alerts respondents to the costs that over-reporting will shortly incur?

**EXPERIMENTAL TREATMENTS**

The table below summarises the experimental conditions (also see the section of ‘WORDING’ and the online demonstration [Http://goo.gl/05LLB3](http://goo.gl/05LLB3)). The core comparison is between Control Group 0, which receives the standard turnout question, and Experimental Group 1, which is alerted immediately before the turnout question to the sequence of follow-up questions facing those who report having voted. One slight concern here is the danger of false negatives: that is, respondents who report non-voting not because they truly abstained but because they want to avoid the extra questions heralded in the prompt. While this is unlikely given the social desirability drive, it can be checked by including a second condition, Group 2, whose preamble alerts respondents to follow-up questions for reported non-voters as well as voters.

Group	Prompt position	Prompt contents	Target N
0	None	N/A	.
1	Before turnout Q	Previews extra Qs for voters	.
2	Before turnout Q	Previews extra Qs for voters + non-voters	.
3	After turnout Q	Previews extra Qs for voters; asks voters to confirm	.
4	After turnout Q	Previews extra Qs for voters + non-voters; asks voters to confirm	.

The experiment can profitably be extended by including two further conditions. These are parallel to Groups 1 and 2 in that the content of prompts is similar; it is the location that changes (see the section of ‘FLOW DIAGRAM.’) In Groups 3 and 4, the prompt follows the standard turnout question and includes a question giving self-reported voters a second chance to admit non-voting. This responds to two points: first, people may not grasp the case for honesty until the implications become clearer; second, respondents tend to answer questions more truthfully when asked to confirm their initial responses (e.g., Clark and Tiffit 1966: 520).

This proposal carries almost no opportunity cost in terms of other questions or proposals – it involves only a small piece of text before or after a question that would be included anyway. These points made, a less powerful experiment could be fielded with smaller group sizes and/or collapsed to include only Groups 0-2.

This experimental design has been piloted in a snowball sample comprised of 129 Taiwanese using Mandarin. The preambles appear to fit together with the standard turnout question well, as the respondents felt no problem with them.

**CONTRIBUTION**

This proposal extends research in that field by focusing on the neglected issue of respondents’ motivation for honesty. Since it involves only a textual preamble or prompt, the proposal is easy to implement. It is also compatible with existing methods (such as using forgiving wording and face-saving options) that are focused more on the social desirability component of respondents’ motivations. Finally, the preamble is so flexible that it can be applied to sensitive questions in a range of other areas (e.g. church attendance,

tax paying), and so this experiment can make a wider contribution to survey methodology.

### WORDING

As will become clear, the exact wording of the preambles depends on what are the follow-up questions in the survey in which this experiment is embedded. The versions below are for illustrative purpose. We are happy to revise these once the otherwise-finalised questionnaire is available.

#### Group 1: Preamble for all eligible electors

*The following questions are about [Election] on [date]. We are interested in things like: Did you vote? If so, which party did you vote for? When did you decide to vote for them? Why did you choose that party? and so on?*

Note: if the turnout question falls within a longer sequence of questions about that election, then the prompt should read “The following questions are also about...”

#### Group 2: Preamble for all eligible electors

*The following questions are about [Election] on [date]. We are interested in things like: Did you vote? If so, which party did you vote for? When did you decide to vote for them? Why did you choose that party? and so on? We also have questions for non-voters, such as: Which party would you have voted for, if you had voted? When did you decide not to vote? Why did you choose to stay at home? and so on.*

Note: if the turnout question falls within a longer sequence of questions about that election, then the prompt should read “The following questions are also about...”

#### Group 3: Prompt and confirmation question for self-reported voters

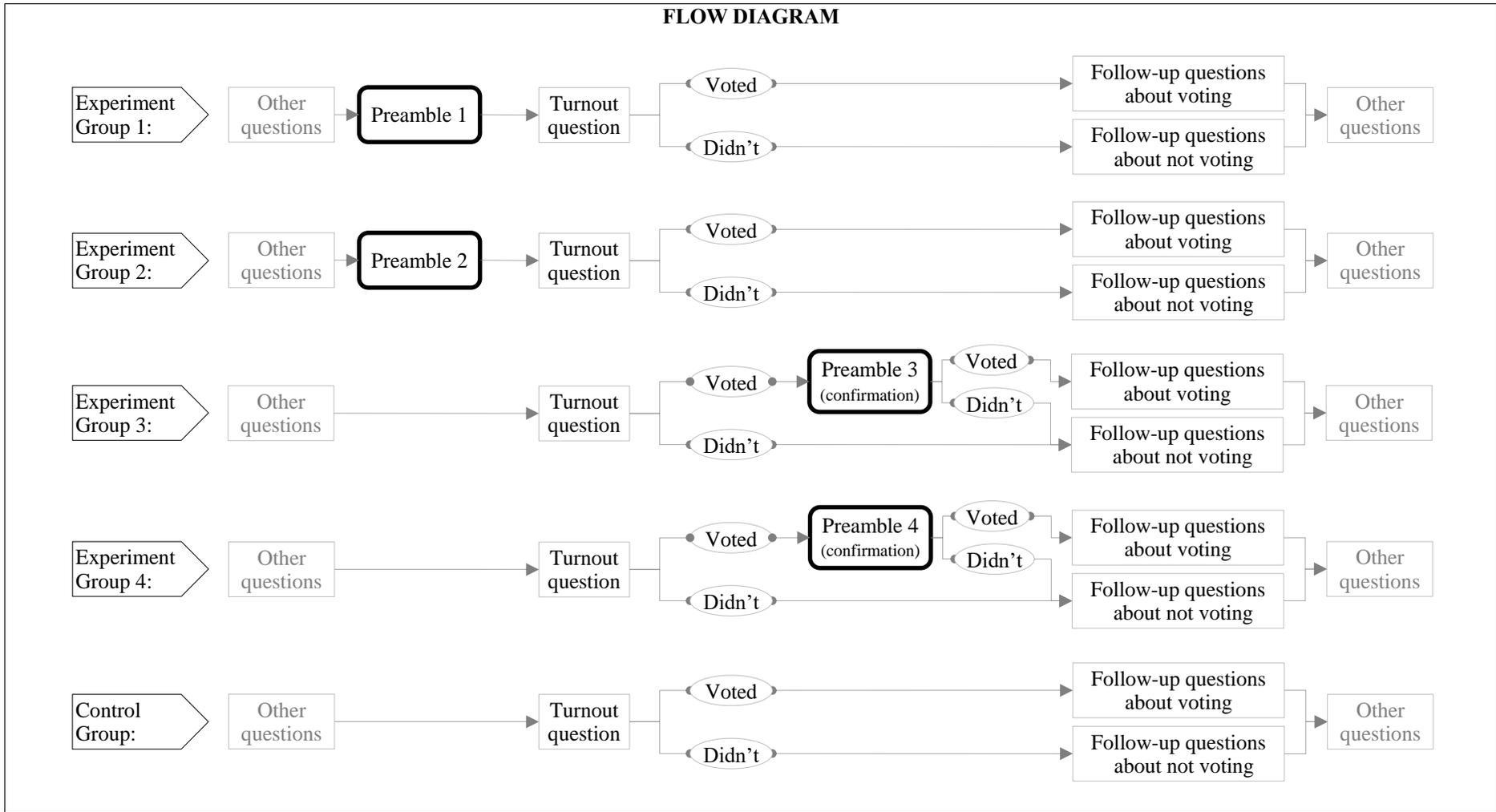
*Let’s talk more about [Election] on [date]. We are interested in things like: Which party did you vote for? When did you decide to vote for them? Why did you choose that party? and so on?*

*Only people who voted in the election should be answering these questions and so, before we move on, can we just double-check whether you voted in [Election] on [date]?*

#### Group 4: Prompt and confirmation question for self-reported voters

*Let’s talk more about [Election] on [date]. We are interested in things like: Which party did you vote for? When did you decide to vote for them? Why did you choose that party? and so on? We also have questions for non-voters, such as: Which party would you have voted for, if you had voted? When did you decide not to vote? Why did you choose to stay at home? and so on.*

*There are different questions for voters and non-voters and so, before we move on, can we just double-check whether you voted in [Election] on [date]?*



## 2. Improvements to the measurement of turnout intention

- The pipeline technique:

*In most countries, government officials retain the records of whether people vote in elections - for example, the UK government keeps such public records in the Clerk of the Crown Office. By looking at these public records, researchers can get an accurate report of who actually voted and who didn't. Of course, these public records do not say who you voted for.*

*According to these public records, we find that a lot of people don't vote these days. By checking survey reports against these public records, we also find that many people said they would vote in a forthcoming election but are later shown by the public records not to have voted. Then there are some people who said they wouldn't vote but actually voted in the end.*

*Please think carefully. If you are entitled to vote, how likely or unlikely is it that you would vote in the next [Election] that will be held in [date]?*

- The item-sum technique (for the treatment group):

*Below are two things. We'd like to ask you to think of a scale from 0 to 4, how likely or unlikely each of these things would be.*

*Below are three things. Please give each of these a score for how likely or unlikely you think it is. Then add the scores of three things together and write the number in the box below. You don't have to tell us the score of each thing -- just tell us the sum of the scores.*

	<i>Very unlikely</i>	<i>Fairly unlikely</i>	<i>Neither</i>	<i>Fairly likely</i>	<i>Very likely</i>
<i>If you have a smartphone, you will get a security software to stop viruses ruining your phone.</i>	0	1	2	3	4
<i>During the next 12 months, there will be times when you don't have enough money to cover your day to day living costs.</i>	0	1	2	3	4
<i>You would vote in the next [Election] that will be held in [date].</i>	0	1	2	3	4

*What is the sum of the scores of the three items?*

## References

- Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. Attempts to Improve the Accuracy of Self-Reports of Voting. In *Questions about Questions*, ed. Judith M. Tanur, 138–153. New York: Russell Sage Foundation.
- Aguinis, Herman, and Christine A. Henle. 2001. “Empirical Assessment of the Ethics of the Bogus Pipeline.” *Journal of Applied Social Psychology* 31(2): 352-375.
- Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. “What Leads to Voting Over-reports? Contrasts of Over-reporters to Validated Voters and Admitted Nonvoters in the American National Election Studies.” *Journal of Official Statistics* 17(4): 479-498.
- Belli, Robert F., Santa Traugott, and Steven J. Rosenstone. 1994. *Reducing Over-Reporting of Voter Turnout: An Experiment Using a ‘Source Monitoring’ Framework*. ANES Technical Report Series: NES010153.
- Cahalan, Don. 1968. “Correlates of Respondent Accuracy in the Denver Validity Survey.” *Public Opinion Quarterly* 32(4): 607-621.
- Campbell, Angus, Gerald Gurin, and Warren Miller. 1952. *American National Election Studies, 1952 Time Series Study: Codebook*. Ann Arbor, MI: University of Michigan, Center for Political Studies (Producer and Distributor).
- Clark, John P., and Larry L. Tiff. 1966. “Polygraph and Interview Validation of Self-Reported Deviant Behaviour.” *American Sociological Review* 31(4): 516-523.
- Clarke, Harold D., David Sanders, Marianne C. Stewart, and Paul F. Whiteley 2004. *Political Choice in Britain*. Oxford: Oxford University Press.
- Gneezy, Uri. 2005. “Deception: The Role of Consequences.” *American Economic Review* 95(1): 384-394.
- Górecki, Maciej A. 2011. “Electoral Salience and Vote Over-reporting: Another Look at the Problem of Validity in Voter Turnout Studies.” *International Journal of Public Opinion Research* 23(4): 544-557.
- Hanmer, Michael J., Antoine J. Banks, and Ismail K. White. 2014. “Experiments to Reduce the Over-Reporting of Voting: A Pipeline to the Truth.” *Political Analysis* 22(1): 115-129.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010a. “Measuring Voter Turnout by Using the Randomized Response Technique.” *Public Opinion Quarterly* 74(2): 328-343.
- Holbrook, Allyson L., and Jon A. Krosnick. 2010b. “Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique.” *Public Opinion Quarterly* 74(1): 37-67.

- Holbrook, Allyson L., and Jon A. Krosnick. 2013. "A New Question Sequence to Measure Voter Turnout in Telephone Surveys: Results of an Experiment in the 2006 ANES Pilot Study." *Public Opinion Quarterly* 77(S1): 106-123.
- Jones, Edward E. and Harold Sigall. 1971. "The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude." *Psychological Bulletin* 76(5): 349-364.
- Katz, Jonathan N., and Gabriel Katz. 2010. "Correcting for Survey Misreports Using Auxiliary Information with an Application to Estimating Turnout." *American Journal of Political Science* 54(3): 815-835.
- Kritzinger, Sylvia, Steve Schwarzer, and Eva Zeglovits. 2012. "Reducing Over-reporting of Voter Turnout in Seven European Countries: Results from a Survey Experiment." The 67th Annual Conference of the American Association for Public Opinion Research, Orlando, Florida, May 17-20th 2012.
- Mircea, Comşa, and Andrei Gheorghişă. 2011. "'Many', 'Half' or 'One of Two'? Assessing Counter-biasing Technique to Reduce the Self-reported Turnout." The 4th Conference of the European Survey Research Association (ESRA), Lausanne, Switzerland, July 11th-22nd 2011.
- Nederhof, Anton J. 1985. "Methods of Coping with Social Desirability Bias: a Review." *European Journal of Social Psychology* 15(3): 263-280.
- Parry, Hugh J., and Helen M. Crossley. 1950. "Validity of Responses to Survey Questions." *Public Opinion Quarterly* 14(1): 61-80.
- Presser, Stanley, and Michael Traugott. 1992. "Little White Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting." *Public Opinion Quarterly* 56(1): 77-86.
- Rasinski, Kenneth A., Penny S. Visser, Maria Zagatsky, and Edith M. Rickett. 2005. "Using Implicit Goal Priming to Improve the Quality of Self-Report Data." *Journal of Experimental Social Psychology* 41(3): 321-327.
- Selb, Peter, and Simon Munzert. 2013. "Voter Overrepresentation, Vote Misreporting, and Turnout Bias in Postelection Surveys." *Electoral Studies* 32(1): 186-196.
- Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Over-reports Voting." *American Political Science Review* 80(2): 613-624.
- Stocké, Volker, and Tobias Stark. 2007. "Political Involvement and Memory Failure as Interdependent Determinants of Vote Over-reporting." *Applied Cognitive Psychology* 21(2): 239-257.
- Stocké, Volker. 2007. "Response Privacy and Elapsed Time Since Election Day as Determinants for Vote Overreporting." *International Journal of Public Opinion Research* 19(2): 237-246.
- Swaddle, Kevin, and Anthony Heath. 1989. "Official and Reported Turnout in the British General Election of 1987." *British Journal of Political Science* 19(4): 537-551.

- Tourangeau, Roger, Lance J. Rips, and Kenneth A. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- 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(1): 58-77.
- Vrij, Aldert, Ronald Fisher, Samantha Mann, and Sharon Leal. 2006. "Detecting Deception by Manipulating Cognitive Load." *Trends in Cognitive Science* 10(4): 141-142.

