The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries

Patrick M. Kuhn[1] and Nick Vivyan[2]

[1] *Associate Professor in Comparative Politics, School of Government and International Affairs, Durham University, Durham, UK. E-mail: p.m.kuhn@durham.ac.uk.*
[2] *Professor of Politics, School of Government and International Affairs, Durham University, Durham, UK. E-mail: nick.vivyan@durham.ac.uk.*

Abstract

To reduce *strategic* misreporting on sensitive topics, survey researchers increasingly use list experiments rather than direct questions. However, the complexity of list experiments may increase *nonstrategic* misreporting. We provide the first empirical assessment of this trade-off between strategic and nonstrategic misreporting. We field list experiments on election turnout in two different countries, collecting measures of respondents' true turnout. We detail and apply a partition validation method which uses true scores to distinguish true and false positives and negatives for list experiments, thus allowing detection of nonstrategic reporting errors. For both list experiments, partition validation reveals nonstrategic misreporting that is: undetected by standard diagnostics or validation; greater than assumed in extant simulation studies; and severe enough that direct turnout questions subject to strategic misreporting exhibit lower overall reporting error. We discuss how our results can inform the choice between list experiment and direct question for other topics and survey contexts.

Keywords: survey experiments, survey design, list experiments, sensitive questions, measurement error, misreporting, satisficing

1 Introduction

How should political scientists elicit sensitive information from survey respondents as to whether they hold attitudes or behave in ways that defy a social norm or formal rule? Direct questions on such topics raise sensitivity concerns among respondents who truly defy the norm or rule ("norm-defiers"), leading them to falsely claim compliance. This direct question *strategic misreporting* (Ahlquist 2018) results in measures of norm-defiance that suffer from false negatives, i.e., defiers wrongly measured as compliers.[1] For example, false negatives arise in turnout studies when respondents who failed to vote in a recent election—and thereby defied the civic norm of electoral participation (Blais and Achen 2019, 476)—claim to have voted—i.e., to have complied with the norm—when asked directly about their turnout (Presser 1990; Belli, Traugott, and Beckmann 2001).

To avoid direct question strategic misreporting, political scientists increasingly use list experiments to ask survey respondents about sensitive topics. By masking individual answers to a sensitive question, list experiments are held to reduce sensitivity concerns among norm-defiers, thereby reducing strategic misreporting and consequent false negative measurements. Recent research, however, suggests the extra cognitive effort demanded by list experiment questions may induce *nonstrategic misreporting* (Ahlquist 2018; Kramon and Weghorst 2019; Riambau and Ostwald 2020): facing a longer, more complex question, respondents may be more likely to either

[1] Our terminology is premised on the goal being to detect norm-defiance, so that a norm-defier is a "positive" case and a norm-complier is a "negative" case.

***Political Analysis* (2022)**
vol. 30: 381–402
DOI: 10.1017/pan.2021.10

Published
16 April 2021

Corresponding author
Nick Vivyan

Edited by
Jeff Gill

satisfice or make mistakes. Crucially, unlike strategic misreporting—which only affects norm-defiers—nonstrategic misreporting plausibly induces reporting errors for both norm-defiers *and* compliers, causing false negatives among the former and false positives among the latter.

How severe is list experiment nonstrategic misreporting in practice? And once we account for realistic levels of such misreporting, does a list experiment on a sensitive topic still reduce overall reporting error compared to a direct question? These questions are critical for researchers deciding between measuring a sensitive variable via direct question or list experiment. Existing research, however, does not answer them directly. Studies of nonstrategic misreporting in list experiments provide circumstantial evidence of its existence via *placebo* tests (Kramon and Weghorst 2019; Riambau and Ostwald 2020) or assume its degree and precise nature in simulations (Ahlquist 2018; Blair, Chou, and Imai 2019), but do not compare list experiment to direct question reporting error. Existing empirical validation studies that do compare list experiment and direct question performance rely on comparison of aggregate prevalence estimates to each other ("comparative prevalence validation") or to a true population benchmark ("population prevalence validation"), neither of which distinguishes true and false positive measurements on the sensitive variable. They may therefore yield similar results whether a list experiment is correcting strategic misreporting among norm-defiers—thereby increasing the true positive rate—or inducing additional nonstrategic misreporting among norm-compliers—thereby increasing the false positive rate (Höglinger and Jann 2018).

In this paper, we provide the first empirical validation of list experiments that distinguishes the increases in true positives (due to reduced strategic misreporting) from the increases in false positives (due to increased nonstrategic misreporting) that they may generate compared to a direct question. We present two new validation studies of list experiments on nonvoting in elections. These were fielded to samples from relatively educated populations in two different contexts: New Zealand and London following their respective 2017 General Elections. Crucially, in both studies, we collect measures of individual respondents' true scores on the sensitive variable, i.e., whether they voted in the election or not, based on official records.

To exploit these true scores in a way that distinguishes true and false positives and negatives, we detail and apply a *partition validation* approach for list experiments. Similar to the approach developed by Höglinger and Jann (2018) for the randomized response technique, it involves partitioning the sample by true score and calculating standard list experiment prevalence estimates within each resulting subsample. We show how the numbers of true positives and false negatives are identified based on the list prevalence estimate among actual norm-defiers, while the numbers of true negatives and false positives are identified based on the list prevalence estimate among actual norm-compliers. By applying this partition validation method, we provide the first empirical assessment of each type of reporting error in list experiments versus direct questions.

We find that, in both the New Zealand and London cases, standard diagnostics and standard validation approaches suggest list experiment measures of nonvoting are unproblematic and probably better than direct measures. However, partition validation based on true scores changes this conclusion. It shows that, while direct questions in both surveys do induce strategic misreporting and consequent false negatives among actual norm-defiers (i.e., nonvoters), neither list experiment successfully reduces false negatives among these respondents. Moreover, both list experiments appear to increase the rate of false positives among actual norm-compliers (i.e., voters), consistent with them inducing additional nonstrategic misreporting compared to the direct question. These false positives are common enough that they imply a rate of list experiment nonstrategic misreporting that is, even under conservative assumptions, double that assumed in extant simulation studies. Taking false positives and negatives together, both list experiments perform significantly worse than direct questions in terms of overall reporting error. In additional

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

analysis, we provide evidence from the London survey that satisficing is an important driver of list experiment nonstrategic misreporting.

Our empirical analysis contributes to the literature on list experiments by providing the clearest evidence to date that list experiments do induce an additional and nontrivial amount of nonstrategic misreporting error in practice, even in relatively educated samples previously thought to be least prone to such behavior (Kramon and Weghorst 2019). This in turn informs applied survey research on sensitive topics by highlighting the need for researchers deciding between a direct question and a list experiment to consider not just the well-known trade-off between strategic misreporting under the direct question and the statistical inefficiency of the list experiment (Blair, Coppock, and Moor 2020), but also the *misreporting trade-off* between direct question strategic misreporting and list experiment nonstrategic misreporting.

To help researchers think through this misreporting trade-off, we develop a simple parameterization of it in the Discussion section. We use our validation results to locate our two studies within the parameter space, then consider the ways in which list experiments in other contexts may plausibly depart from ours and with what consequences for the misreporting trade-off. For the topic of nonvoting or similarly sensitive topics fielded in survey settings like the ones we study, our results suggest that any advantage of list experiment over direct question in terms of reduced strategic misreporting is outweighed in practice by disadvantages in terms of increased nonstrategic misreporting. To be clear, this does not mean that direct questions always outperform list experiments: in other scenarios, where the topic of interest is of sufficiently enhanced sensitivity compared to nonvoting (increasing probability of direct question strategic misreporting among norm-defiers), or where norm-defier prevalence is sufficiently greater than in our cases (increasing the number of respondents “at risk” of direct question strategic misreporting), list experiments will outperform direct questions in terms of expected overall reporting error, provided the amount of list experiment nonstrategic misreporting is similar to that apparent in our surveys. However, we also suggest that researchers surveying respondents in medium- or low-education settings may reasonably expect list experiment nonstrategic misreporting to be more common than we find in the comparatively well-educated setting of New Zealand and London. In such cases, topic sensitivity and norm-defier prevalence will need to be even higher again before list experiments can be expected to outperform direct questions on overall reporting error.

A final contribution of this article is to demonstrate how researchers validating list experiments can use partition validation to fully exploit contexts where true scores on the sensitive variable are available. This is valuable because our results highlight how standard list experiment diagnostics and validation approaches are insufficient to detect list experiment misreporting errors that occur in practice. In particular, standard prevalence validation applied to each of our list experiments suggests unproblematic or superior performance, because the list prevalence estimate of nonvoting is higher than the direct question estimate. Yet partition validation shows how these higher list prevalence estimates are in fact the result of an increase in false positive errors (consistent with an increase nonstrategic misreporting), rather than a reduction in false negative errors (due to reduced strategic misreporting).

2 Misreporting in Direct Questions and List Experiments

To clarify the consequences of direct question strategic misreporting and list experiment nonstrategic misreporting for different types of reporting error, we begin by formally characterizing both processes. Let X_i^* be an indicator capturing the true status of survey respondent $i = \{1, \ldots, N\}$ on the sensitive variable. $X_i^* = 0$ when i complies with the social norm or formal rule of interest, and $X_i^* = 1$ when i defies it. The true prevalence of norm-defiers is thus $\pi = \Pr(X_i^* = 1)$. Let X_i be an indicator capturing respondent i's self-reported status on the sensitive variable, with $X_i = 0$ and $X_i = 1$ indicating reported norm-compliance and -defiance, respectively.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

2.1 Direct Questions and Strategic Misreporting

Direct questions concerning a sensitive topic generally elicit truthful responses from norm-compliers (i.e., $X_i = X_i^* = 0$). However, due to sensitivity concerns (e.g., social desirability bias or fears of the repercussions should a truthful answer be disclosed to third parties), norm-defiers often misreport their true status as $X_i = 0$, thereby generating false negative measurements (Tourangeau and Yan 2007, 863). This is referred to as *strategic misreporting* (Ahlquist 2018). Letting $\theta = \Pr(X_i = 0|X_i^* = 1)$ denote the probability with which norm-defiers strategically misreport for a direct question, the expected proportion of false negatives in the sample is $\theta\pi$.

The direct question estimator of norm-defier prevalence is $\hat{\pi}^{\text{Direct}} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Under the above assumptions, $\mathbb{E}(\hat{\pi}^{\text{Direct}}) = (1-\theta)\pi$. Thus, as θ increases, $\mathbb{E}(\hat{\pi}^{\text{Direct}})$ decreases, inducing downward bias in the norm-defiance prevalence estimator.

2.2 List Experiments and Non-Strategic Misreporting

A list experiment is conventionally assumed to reduce strategic misreporting by asking about the sensitive item indirectly and thereby reducing sensitivity concerns among norm-defiers. A standard design randomly allocates respondents to either a list of J control items (treatment status $T_i = 0$) or a treatment list (treatment status $T_i = 1$) containing the J control items plus the sensitive item. In either case, respondents are asked only to report *how many* of the listed items they affirm, not which items they affirm. Let $Z_{ij}(T)$ be an indicator denoting whether respondent i affirms control item $j = \{1, \dots, J\}$ under treatment status $T = \{0, 1\}$. Define $Y_i(0) = \sum_{j=1}^{J} Z_{ij}(0)$ and $Y_i(1) = \sum_{j=1}^{J} Z_{ij}(1) + X_i$ as respondent i's potential reported item counts under the control and treatment conditions, respectively, and $Y_i = Y_i(T_i)$ as their realized reported item count. Under this design, individual respondents' answers to the sensitive item are masked from the researcher, or anyone else. Yet under the assumption of "no design effects" (i.e., $\sum_{j=1}^{J} Z_{ij}(0) = \sum_{j=1}^{J} Z_{ij}(1)$) and "no liars" (i.e., $X_i(1) = X_i^*$), the researcher can still obtain an unbiased estimate of norm-defier prevalence by taking the difference in means (DiM) of reported item counts for the treatment and control lists, $\hat{\pi}^{\text{List}} = \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)Y_i$, where $N_1 = \sum_{i=1}^{N} T_i$ is the size of the treatment group and $N_0 = N - N_1$ is the size of the control group (Blair and Imai 2012).

Even if list experiment masking eliminates strategic misreporting among norm-defiers, recent research proposes that the added complexity of the list question—which asks respondents to consider multiple items and to sum affirmed items before responding—may lead to increased *nonstrategic misreporting* (Ahlquist 2018; Kramon and Weghorst 2019; Riambau and Ostwald 2020). This occurs when respondents do not properly engage with the list question or make mistakes when answering it. For example, respondents are more likely to satisfice when answering more complex survey questions and do so by devoting less than optimal effort, performing some necessary cognitive steps roughly or skipping them altogether (Krosnick 1991).

Formally, let the indicator S_i^* capture whether respondent i is a list experiment nonstrategic misreporter. $S_i^* = 0$ when respondent i answers the list question via the process conventionally assumed, with potential responses that satisfy the no design effects and no liars assumptions. $S_i^* = 1$ when respondent i answers the list question via an alternative process prone to nonstrategic misreporting errors. Let $\lambda = \Pr(S_i^* = 1)$ denote the probability that a given respondent is a list experiment nonstrategic misreporter. We assume that the expected DiM among nonstrategic misreporters, which we define as $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) = \mathbb{E}(Y_i(1)|S_i^* = 1) - \mathbb{E}(Y_i(0)|S_i^* = 1)$, is not driven by the true rate of norm-defiance among such respondents. Rather, it depends on the decision rule that nonstrategic misreporters use to pick their reported item count, and how the resulting reported item count varies as a function of whether they are asked about the longer treatment list or shorter control list.

Unlike strategic misreporting—which only leads to false negatives—a crucial feature of list experiment nonstrategic misreporting is that it plausibly leads to both false negatives among

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

norm-defiers *and* false positives among norm-compliers. To see this, consider the example "uniform" nonstrategic misreporting scenario hypothesized in Ahlquist (2018) and argued to be plausible in Blair *et al.* (2019). In this scenario, nonstrategic misreporters give an item count that is a random uniform draw from the response options available. This implies expected item counts of $\mathbb{E}(Y_i(0)|S_i^* = 1) = \frac{J}{2}$ for the control list and $\mathbb{E}(Y_i(1)|S_i^* = 1) = \frac{J+1}{2}$ for the treatment list, and an expected DiM among nonstrategic misreporters of $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) = 0.5$. Among nonstrategic misreporters who are actual norm-defiers, this expected DiM is lower than the true norm-defier prevalence of one, leading to false negatives in expectation. Among nonstrategic misreporters who are actual norm-compliers, this expected DiM is higher than the true norm-defier prevalence of zero, leading to false positives in expectation.

This uniform process is just one possible example of a nonstrategic misreporting process. More generally, any such process that generates $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) < 1$ implies false negatives among norm-defiers (as would strategic misreporting), and any nonstrategic misreporting process generating $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) > 0$ implies false positives among norm-compliers (unlike strategic misreporting). Thus, one need not assume a uniform nonstrategic misreporting process to be concerned that a list experiment may generate both false negatives and false positives: a range of nonstrategic misreporting processes can generate both types of error. As will be argued in the following section, this feature of nonstrategic misreporting means that existing list experiment validation approaches can mislead because they do not distinguish false from true positives and negatives.

Before proceeding, we note that list experiment nonstrategic misreporting generally biases the list prevalence estimate, the quantity of interest in much applied research. Ahlquist (2018) and Blair *et al.* (2019) demonstrate the bias in DiM prevalence estimate for two specific types of nonstrategic misreporting process. Importantly, unlike with direct question strategic misreporting, the bias induced by list experiment nonstrategic misreporting can be either negative or positive. To see this, take the case where S_i^* is independent of X_i^*, so that the expected DiM among respondents for whom $S_i^* = 0$ is $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=0}) = \pi$. Then, $\mathbb{E}(\hat{\pi}^{\text{List}}) = (1-\lambda)\pi + \lambda\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})$, which makes clear that for all $\lambda > 0$, the list prevalence estimator will be biased in expectation except in the special case where $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) = \pi$. The bias will be positive when $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) > \pi$ and negative when $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1}) < \pi$. Since both λ and $\mathbb{E}(\hat{\pi}^{\text{List}}_{S^*=1})$ will usually be unobserved, both the magnitude and size of bias will be difficult to gauge in practical applications.

3 Existing Evidence

We argue that existing evidence leaves open important questions about the practical tradeoff between nonstrategic misreporting in list experiments and strategic misreporting in direct questions. We first discuss studies of nonstrategic misreporting in list experiments, before turning to more general list experiment validation studies.

3.1 Existing Evidence on Nonstrategic Misreporting in List Experiments

Two recent studies use *placebo* tests to examine nonstrategic misreporting in list experiments. Riambau and Ostwald (2020) run list experiments where the additional item on the treatment list has a known true sample prevalence of zero, but which yield DiMs that are positive and significant, consistent with nonstrategic misreporting where some respondents condition reported item count on list length. Kramon and Weghorst (2019) compare respondents' item counts for a list of manifestly nonsensitive topics to the counts implied by the same respondents' answers to direct questions on the same topics. The counts should not differ as there should be no direct question strategic misreporting for the list question to ameliorate. But they do differ in 60% of cases (and more so among respondents with lower numeracy and literacy) suggesting that list experiment measures may differ from direct question ones not just by reducing strategic misreporting, but also because their "complexity and difficulty" (p. 4) induces more reporting errors.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

Despite this evidence of list experiment nonstrategic misreporting, key questions remain for applied researchers choosing between list experiments and direct questions. First, Riambau and Ostwald (2020) benchmark list prevalence estimates against true scores only, so cannot gauge the relative magnitude of list versus direct question misreporting error. Second, Kramon and Weghorst (2019) compare list and direct question responses only, so cannot gauge the magnitude of misreporting errors for either relative to the truth. Third, neither study examines measurement of a *sensitive* topic, which is essential to assess whether any increase in nonstrategic misreporting induced by using a list experiment rather than direct question outweighs the reduction in strategic misreporting. The evidence we present below does all of these three things.

Alvarez *et al.* (2019) identify likely survey satisficer respondents via screening questions and show that they respond differently to both direct questions and list experiments than do other respondents. However, in the absence of true scores on the sensitive item, they cannot establish whether these differences arise because satisficers misreport more or because the their true prevalence rate differs, nor whether list experiments induce more or less misreporting than direct questions in each group.

Others examine how problematic nonstrategic misreporting is for list experiment estimators. Building on Ahlquist (2018), Blair *et al.* (2019) suggest a diagnostic test for list experiment measurement error (whether strategic or nonstrategic) that compares maximum likelihood (ML) and nonlinear least squares (NLS) estimates. They show that nonstrategic misreporting does bias list prevalence estimates, but find in simulation studies that the DiM prevalence estimator and NLS estimator are more robust to nonstrategic misreporting than are ML estimators, and that all estimators exhibit only mild biases for the scenarios they consider. They also develop new regression estimators that explicitly model a uniform nonstrategic misreporting process and a "top-biased" process (where nonstrategic misreporters always select the maximum item count available), although they recommend that, due to its "simplicity and robustness" (p. 473), basic DiM should still be used to estimate prevalence alone. These studies offer valuable guidance, but do not speak explicitly to the misreporting trade-off between list experiment and direct question. Moreover, the simulations used must make assumptions about the nature and frequency of nonstrategic misreporting. The validation evidence we present below offers more direct empirical evidence concerning the severity of nonstrategic misreporting error in list experiments in practice.

In sum, existing research suggests that nonstrategic misreporting does occur in practice in list experiments. However, for applied researchers considering whether the use of a list experiment in their survey will reduce overall misreporting on a sensitive topic, further evidence is required on the relative severity of list experiment nonstrategic misreporting and direct question misreporting *in practice*.[2]

3.2 Existing Validation Approaches

The most common approach to list experiment validation in existing studies involves simple comparison of list and direct question prevalence estimates. This approach, which we call *comparative prevalence validation*, invokes a "more is better" assumption (Tourangeau and Yan 2007; Höglinger and Diekmann 2017): the list experiment is judged to improve on the direct question when it estimates greater norm-defier prevalence. By this criterion, the list experiment significantly outperforms a direct question in 63% of 48 comparative validation studies covering a range of sensitive topics (Holbrook and Krosnick 2010).

[2] Blair *et al.* (2020) do examine list experiment versus direct question performance. They focus, however, on a list experiment assumed to eliminate misreporting but which is inefficient versus a direct question which is subject to strategic misreporting but more efficient. We focus on whether a list experiment outperforms a direct question on reporting error even before considering efficiency.

Population prevalence validation additionally compares direct and list experiment prevalence estimates against an observed true population benchmark and makes a “closer is better” assumption: if its prevalence estimate is closer to the population prevalence, the list experiment offers a better measure. For example, Rosenfeld, Imai, and Shapiro (2016) benchmark list and direct prevalence estimates of anti-abortion attitudes against actual population support for anti-abortion measures in a public referendum, and others benchmark election turnout estimates against official population turnout (e.g., Holbrook and Krosnick 2010; Kuhn and Vivyan 2018).

However, both comparative and population prevalence validation approaches may mislead in the presence of nonstrategic misreporting (Höglinger and Diekmann 2017; Höglinger and Jann 2018). Comparative prevalence validation assumes that list experiments only generate higher norm-defier prevalence estimates than direct questions due to reduced strategic misreporting causing reductions in false negative measurements. However, list experiment prevalence estimates may be higher than direct question estimates not because the list question reduces false negatives among norm-defiers, but because nonstrategic misreporting for the list question yields additional false positives among norm-compliers. A similar problem arises with population prevalence validation: compared to a direct question which underestimates population prevalence due to strategic misreporting, a list experiment subject to nonstrategic misreporting may move the norm-defier prevalence estimate closer to the population benchmark by increasing false positives among norm-compliers (Höglinger and Diekmann 2017).[3] To properly assess whether list experiments reduce misreporting compared to direct questions, we need validation approaches that distinguish false from true negative and false from true positive responses (Höglinger and Jann 2018).

4 Validating List Experiments Using True Scores

We add to the above body of evidence by (1) fielding list experiments where we are able to obtain respondents’ true scores on the sensitive variable and (2) exploiting these true scores to distinguish true and false positives and negative measurements. This section sets out how we accomplish (2) given (1).

In a setting where one observes true scores on the sensitive variable, one straightforward extension to the validation approaches discussed above is *sample prevalence validation*. Given respondents’ true scores, we know the true sample prevalence of norm-defiers and can compare list experiment DiM and direct prevalence estimates against this benchmark. Unlike with population prevalence validation, differences in true sample and population prevalence no longer confound the comparison of list and direct estimate performance. However, there remains the problem that, compared to a direct question subject to strategic misreporting, a list experiment subject to nonstrategic misreporting may yield a prevalence estimate closer to the true sample prevalence due to an increase in false positives rather than a reduction in false negatives.

How, then, can we use true scores on the sensitive variable to distinguish false and true positive and negative measurements and thereby properly assess list versus direct question reporting errors? For a direct question, it is straightforward to distinguish reporting errors given access to true scores X_i^* and observed direct question responses: $X_i^{\text{Direct}} < X_i^*$ is a false negative and $X_i^{\text{Direct}} > X_i^*$ is a false positive. Yet distinguishing false and true positives and negatives for a list experiment measure is more challenging. Precisely, because of its masking properties, a list experiment does not yield individual-level measures of the sensitive variable, so these cannot be compared to individuals’ true scores.

[3] True population prevalence may also differ from the true sample prevalence due to sampling or nonresponse biases, such that a prevalence estimate exactly matching true sample prevalence may misleadingly appear inferior when judged against population prevalence.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

However, as Höglinger and Jann (2018) point out in the context of the randomized response technique (which also masks individual responses), one can separately identify the rate of true and false positives and negatives through a process we label *partition validation*. This involves: (a) partitioning the sample based on observed true scores X_i^*; and (b) calculating a DiM prevalence estimate separately for true norm-defier and norm-complier respondents. How does this distinguish error types? First note that, among true norm-defiers, the true prevalence of norm-defiers is, by definition, $\pi_{X^*=1} = 1$, such that only true positives or false negatives are possible. The list DiM prevalence estimate among true norm-defiers, $\hat{\pi}^{\text{List}}_{X^*=1}$, thus gives the rate of true positives in this subsample, while $1 - \hat{\pi}^{\text{List}}_{X^*=1}$ gives the rate of false negatives. Second, among true norm-compliers, the true prevalence of norm-defiers is, by definition, $\pi_{X^*=0} = 0$, and only true negatives or false positives are possible. Thus, the list DiM prevalence estimate among true norm-compliers, $\hat{\pi}^{\text{List}}_{X^*=0}$, gives the false positive rate in this subsample, and $1 - \hat{\pi}^{\text{List}}_{X^*=0}$ gives the true negative rate.

Putting this together, and letting $N_{X^*=0}$ and $N_{X^*=1}$ denote the number of true norm-compliers and -defiers in the sample, respectively, the total implied number of true positives (denoted tp), false positives (fp), true negatives (tn), and false negatives (fn) for the list experiment measure can be computed as follows:

$$tp = N_{X^*=1}\hat{\pi}^{\text{List}}_{X^*=1}, \tag{1}$$

$$fp = N_{X^*=0}\hat{\pi}^{\text{List}}_{X^*=0}, \tag{2}$$

$$tn = N_{X^*=0}\left(1 - \hat{\pi}^{\text{List}}_{X^*=0}\right), \tag{3}$$

$$fn = N_{X^*=1}\left(1 - \hat{\pi}^{\text{List}}_{X^*=1}\right). \tag{4}$$

Based on these quantities, we can compute a confusion matrix—a contingency table of true scores against reported scores (Manning, Raghavan, and Schütze 2009, 307–308)—for the list experiment. Comparing the confusion matrix of the list experiment to that of the direct question can tell us about error mechanisms. If the direct question suffers from *strategic* misreporting, the direct question should generate a nontrivial rate of false negatives among norm-defiers but few false positives among norm-compliers. If the list experiment corrects strategic misreporting and does not induce nonstrategic misreporting, it should generate fewer false negatives than the direct question and no more false positives. If the list experiment fails to correct strategic misreporting, it will generate a nontrivial number of false negatives among norm-defiers, like the direct question. If the list experiment induces additional *nonstrategic* misreporting compared to the direct question, this will generate either or both false negatives among norm-defiers and false positives among norm-compliers, with the mix of false positives and false negatives determined by the precise (unobserved) nonstrategic misreporting process that pertains. Thus, while a nontrivial level of false negatives for the list experiment is an indicator of either strategic or nonstrategic misreporting, a nontrivial level of false positives for the list experiment is a clear indication of some form of nonstrategic misreporting.

We can also summarize and compare the overall rate of reporting errors in the list experiment and direct question in terms of *accuracy*, the fraction of all survey respondents correctly measured on the sensitive variable (Manning *et al.* 2009, 155–156). Höglinger and Jann (2018) focus on this statistic (which they label the "correct classification rate") when validating a randomized response technique.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

5 Data

We apply partition validation to two list experiments on election turnout. We fielded these in New Zealand and in London (UK) and collected measures of respondents' true score on the sensitive variable through inspection of official electoral records.[4] Here, we describe the survey instruments and data collection.

5.1 Survey Instruments

We embedded the New Zealand list experiment in the 2017 New Zealand Election Study (NZES). The NZES collected responses from 3,455 respondents, sampled from the national electoral rolls. Respondents were contacted by mail beginning 4 days after the general election of September 23. Fieldwork continued until early March 2018, although approximately 97% of responses were received within two months of commencement. The London list experiment was fielded via an online YouGov survey of a sample of 3,189 Greater Londoners following the June 8, 2017 UK general election (fieldwork began on June 23 and ended on July 24, 2017). We surveyed Londoners, rather than Britons generally, to make collection of true turnout measures economically feasible (the official records necessary for this must be accessed physically at each local authority office). Further details on sampling and fieldwork for each survey are provided in online Appendix A.

Table 1 presents the list experiment and direct questions used to measure turnout in each survey. In both surveys, all respondents were randomized to either the control or treatment list (the latter being the list question in Table 1 with the item in parentheses included). List response options ranged from zero to four (control group) or five (treatment group), and a "don't know" response was available.

In designing the list experiments, we follow recent practice. Several design choices in particular merit discussion. First, in both the New Zealand and London designs, we include control activities which we expect most respondents to have undertaken ("Discussed the election with...") and which we expect few respondents to have undertaken ("Worked or volunteered for one of the party campaigns" or "Put up a poster for a political party in my window or garden"). This follows Blair and Imai (2012) design advice and is intended to minimize ceiling and floor effects—where respondents affirm or negate all control items, so that their sensitive item response is no longer masked—which may undermine the ability of the list experiment to reduce strategic misreporting.

Second, we included only election-related control items in both list experiments. On the one hand, there is a risk that including election-related control items may prime respondents to become concerned about their general level of political engagement, increasing sensitivity of the turnout item and counteracting any sensitivity-reducing effect of list experiment masking. On the other hand, including control items on a different topic to the sensitive item may draw respondents' attention to that item and enhance its sensitivity (Lax, Phillips, and Stollwerk 2016). Furthermore, the inclusion of low-cost, high-prevalence, election-related activities among list control items may reduce the sensitivity of the turnout item by allowing respondents to indicate that they did at least partake in some election activities, even if they did not vote. On balance, our expectation is that this design choice should reduce the sensitivity of the turnout item and thereby enhance the ability of each list experiment to reduce strategic turnout misreporting compared to the direct question. It should also discourage nonstrategic misreporting for the list experiments, since coherent grouping of question topics reduces cognitive processing costs for respondents, making satisficing or mistakes less likely (Krosnick and Presser 2010).

Third, while the New Zealand control items consist exclusively of "norm-compliant" election-related behaviors, we include two "norm-defiant" behaviors ("avoided watching the leaders debate" and "criticised a politician on social media") among the four London control items.

[4] Replication data and code for this study are available at Kuhn and Vivyan (2020a) and Kuhn and Vivyan (2020b).

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

Table 1. List experiment and direct questions for New Zealand and London surveys.

	New Zealand	London
List experiment	Here is a list of things that some people did, and some people did not do, during the election campaign or on election day. How many of these things did you do? You don't need to tell us which ones you did, justhow many. • Discussed the election with family, friends, or workmates • Saw a news story about the election campaign • Worked or volunteered for one of the party campaigns • (Voted in the election) • Watched the election results coming in on election night How many of these things did you do?	The next question deals with the recent general election on 8th June. Here is a list of four (five) things that some people did and some people did not do during the election campaign or on Election Day. Please say how many of these things you did. Here are the four (five) things: • Discussed the election with family and friends • (Voted in the election) • Criticised a politician on social media • Avoided watching the leaders debates • Put up a poster for a political party in my window or garden How many of these things did you do?
Direct question	Looking at the election results, we can see that a lot of people didn't manage to vote. Did you vote in the election on September 23, did you not manage to vote, or did you choose not to vote? • Cast a vote • Chose not to vote • Didn't manage to vote	Talking with people about the recent general election on 8th June, we have found that a lot of people didn't manage to vote. How about you, did you manage to vote in the general election? • Yes • No • Don't know

Norm-defiant control items were included in a list experiment on turnout with promising population prevalence validation results by Kuhn and Vivyan (2018). They reason that norm-defiant control items signal to respondents that it is recognized that some people do not like or engage with politics, thereby further reducing the potential discomfort of admitting nonvoting. To the extent that this holds, the London list experiment should be more effective than the New Zealand one at reducing strategic turnout misreporting among nonvoters. On the other hand, the inclusion of norm-compliant and -defiant items on a list may confuse respondents, which may result in greater list experiment nonstrategic misreporting in London compared to New Zealand.

Both surveys also include a standard direct turnout question (Table 1, bottom row). In the New Zealand survey, the direct question was asked of all respondents at least 41 questions after the list experiment (itself the second item on the survey). In case exposure to the turnout item in the list question primed list treatment group respondents in any way (Blair and Imai 2012), we subset to direct question responses from list control group respondents in the main validation analysis below.[5]

In the London survey, we have two separate direct measures of turnout based on the same question wording: a baseline (pretreatment) measure from a direct question that YouGov asked of panelists in the days immediately following the election and a measure from a direct question included in our survey for list control group respondents (asked immediately after the list question). For our main validation analysis below, we rely on the latter measure.[6] The "baseline" direct measure of turnout will be used later when testing for the effects of satisficing on list experiment misreporting error.

5.2 True Turnout Measures

Each survey respondents' true turnout in the relevant general election was measured via manual inspection of marked electoral rolls. For New Zealand, true turnout measurements were collected by the NZES team. Of the 3,455 2017 NZES respondents, definitive measures of true turnout were obtained for 3,451 (99.9%): these respondents were successfully located and their turnout status clearly observed on the marked election rolls. Remaining respondents with nondefinitive true turnout measures are treated as missing. For London, we collected definitive true turnout measures for 2,595 respondents (82.4%). The rate of definitive true turnout measurements is lower than for the NZES, because, unlike the NZES, YouGov do not sample directly from the electoral register. The resulting sample may therefore contain respondents (a) who are not on the register or (b) whose name and address details recorded with YouGov contain errors preventing matching to the official register. The lower rate of definitive true turnout measurements in London is a concern for list experiment validation if respondents who do and do not have definitive true turnout scores differ systematically in how they answer list and direct turnout questions. We see little reason for this to be the case, and are reassured by the similarity between the London and New Zealand results below, given the latter sample contains almost no respondents with missing true turnout scores. Online Appendix A gives further details on true turnout measurement.

6 Results

This section examines list experiment versus direct turnout question performance in the New Zealand and London studies. We first summarize results of standard list experiment diagnostics, before presenting the standard information validation results researchers would observe in the absence of sensitive variable true scores. We then present results of partition validation, exploiting the true score measures available in our two studies.

6.1 Standard Diagnostics

For each list experiment, we carried out key diagnostics recommended in the literature (full results reported in online Appendix C). We find no clear indication that either experiment violates key assumptions or yields problematic measures of the sensitive variable.

First, there is no strong evidence of association between treatment assignment and respondent characteristics in either setting. Second, both experiments pass diagnostics for violations of the "no design effects" assumption: we find no negative estimated proportions of "respondent types" (Blair *et al.* 2019, 468–469, 473) and fail to reject the null hypothesis of no design effects in formal significance tests (Blair and Imai 2012, 63–65). Third, following Blair and Imai (2012), we check for potential "ceiling" or "floor" effects, where substantial numbers of respondents either negate or affirm all control items and are therefore incentivized to strategically misreport on the sensitive

[5] Online Appendix F shows that substantive results hold when using the full sample direct turnout measure.

[6] Online Appendix F shows that substantive results hold when using the baseline direct question instead of the control group direct question.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

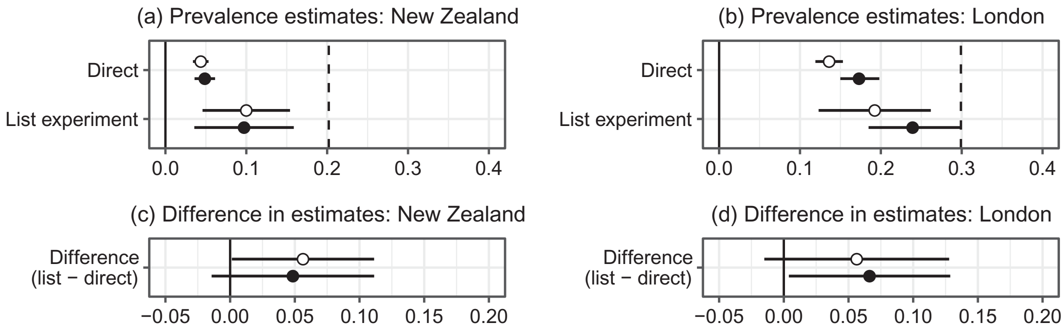

Figure 1. Estimated prevalence versus true population prevalence. Notes: Plots (a) and (b) show direct and list estimates of nonvoting rates for the New Zealand and London surveys, respectively. Dashed vertical lines denote actual population nonvoting rates. Plots (c) and (d) show differences between direct and list estimates. Open and filled circles denote raw and population-weighted estimates, respectively. Horizontal bars indicate 95% confidence intervals.

item in the treatment condition (thus violating the "no liars" assumption) for fear their answer on it could be inferred. Analysis of observed control group item counts suggests little potential for floor effects in New Zealand and ceiling effects in London. Although there is mild potential for ceiling effects in New Zealand—where 6% of the control group affirm all items—and floor effects in London—where 9% of the control group negate all items—this is less severe than in existing published list experiments (e.g., Blair, Imai, and Lyall 2014; Corstange 2018; Kuhn and Vivyan 2018). Fourth, exploiting respondents' answers to list and direct turnout questions, we run the Aronow *et al.* (2015) *placebo* test. This simultaneously tests the no design effect, no liars assumptions, plus two additional assumptions: a "monotonicity" assumption that no false positives occur for the direct question; and a "treatment independence" assumption that list experiment treatment assignment is uncorrelated with direct question response. In both studies, at the 0.05 significance level, we fail to reject the null hypothesis that all four assumptions hold. Finally, the model misspecification test developed in Blair *et al.* (2019, 460) to detect list experiment measurement error—due either to nonstrategic misreporting or to other error processes—yields no significant evidence of such error.

6.2 Standard Information Validation

We now assess list and direct question performance using standard information validation approaches: *comparative* and *population prevalence validation*. Figure 1 compares direct and list prevalence estimates of nonvoting—the sensitive behavior of interest—against each other and against true population prevalence (i.e., true nonvoting prevalence among the eligible electorate in the New Zealand and Greater London populations, according to official records). For comparisons against a population benchmark, we show both raw and population-weighted direct and list prevalence estimates.[7]

In New Zealand and London, both the direct question and list experiment underestimate nonvoting compared to true population prevalence. Yet, consistent with better performance under the more- and closer-is-better assumptions, the list estimate is substantially higher than the direct question estimate and closer to population prevalence. In New Zealand, the raw direct question underestimates nonvoting by 15.8 points, while the raw list experiment only does so by 10.2 points, roughly a one-third reduction in error. A similar reduction in error occurs in London, where the raw direct question underestimates nonvoting by 16.3 points and the raw list experiment only does so

[7] New Zealand, estimates are weighted to the distribution of age, gender, region (Auckland versus non-Auckland), and elector type (general electoral roll versus Māori electoral roll). London, estimates are weighted to the population distribution of age, gender, and educational qualifications using regression adjustment (Rosenfeld *et al.* 2016).

Table 2. Confusion matrices.

(a) New Zealand: direct question			
	Measured		
Actual	**Voter**	**Nonvoter**	***N***
Voter	0.998	0.002	1,617
	[0.996, 1]	[0, 0.004]	
Nonvoter	0.29	0.71	100
	[0.199, 0.378]	[0.622, 0.801]	

(b)New Zealand: list experiment			
	Measured		
Actual	**Voter**	**Nonvoter**	***N***
Voter	0.925	0.075	3,219
	[0.872, 0.977]	[0.023, 0.128]	
Nonvoter	0.393	0.607	186
	[0.06, 0.726]	[0.274, 0.94]	

(c) London: direct question			
	Measured		
Actual	**Voter**	**Nonvoter**	***N***
Voter	0.985	0.015	1,111
	[0.978, 0.992]	[0.008, 0.022]	
Nonvoter	0.268	0.732	164
	[0.199, 0.335]	[0.665, 0.801]	

(d) London: list experiment			
	Measured		
Actual	**Voter**	**Nonvoter**	***N***
Voter	0.921	0.079	2,218
	[0.843, 0.998]	[0.002, 0.157]	
Nonvoter	0.429	0.571	336
	[0.204, 0.652]	[0.348, 0.796]	

Notes: Rows in each table define *actual* turnout status: voter ("negative") or nonvoter ("positive"). Columns define *measured* turnout status. Cells contain row proportions with bootstrap 95% confidence intervals in brackets. Rightmost column gives raw *N* of actual voters and nonvoters in estimation sample. Respondents with nondefinitive true turnout measurements are omitted.

by 10.7 points. In New Zealand, the difference between the raw direct and list prevalence estimates is statistically distinguishable from zero with 95% confidence, but the difference between the weighted estimates is not. In London, the difference between the raw prevalence estimates is not distinguishable from zero, but the difference between the weighted estimates is. Thus, in both surveys, standard information validation indicates that list experiment performs as well as—and probably better than—direct question.

6.3 Partition Validation

Table 2 presents the confusion matrices that result from applying partition validation, exploiting our measure of true scores (i.e., of respondent nonvoting verified using official records).[8] The direct question confusion matrices show that, in both New Zealand (Table 2a) and London (Table 2c), the direct question does appear to suffer from strategic misreporting. Actual voters (norm-compliers) are extremely unlikely to falsely report being a nonvoter—less than 1% and 2% do so in New Zealand and London, respectively. In contrast, actual nonvoters (norm-defiers) falsely report voting much more frequently—29% and 26.8% do so in New Zealand and London, respectively.[9]

Given that the direct turnout questions do indeed suffer from strategic misreporting, do the list experiments reduce overall reporting error? Table 2b and 2d suggests not. First, consider actual nonvoters, who frequently misreport and generate false negatives for the direct question. Rather than reducing false negatives among this group, there is no statistically distinguishable difference between the rate of false negatives recovered by the list experiment and direct question in either New Zealand or London. Point estimates for the difference are actually positive—10.3 points (95% CI: [−24.8, 45.6]) in New Zealand and 16 points (95% CI: [−7.6, 39.6]) in London—indicating more false negatives for the list experiment, if anything. There is thus little evidence that either list experiment alleviates symptoms of direct question strategic misreporting.

[8] Confidence intervals for partition validation are computed via nonparametric bootstrap. All quantities of interest are computed for each given resample.

[9] Consistent with this, online Appendix B shows that in follow-up questions in the London survey actual nonvoters report being less comfortable about directly revealing their turnout than do actual voters.

Second, consider actual voters, of whom the direct question correctly classified all but a tiny proportion in both New Zealand and London. Table 2b and 2d shows an increase in the rate of false positives (the rate of measured nonvoting) among this group when using the list experiment rather than the direct turnout measure. For New Zealand, the estimated increase is 7.3 points and distinguishable from zero (95% CI: [2.1, 12.5]). For London, it is 6.4 points, though not distinguishable from zero (95% CI: [−1.4, 14.2]). The increases in false positives among norm-compliers induced by the list experiments are particularly consequential for overall reporting error, because norm-compliers make up 94% and 86% of the vote-validated New Zealand and London samples, respectively. The increases are also what we would expect to see if list experiments induce nonstrategic misreporting not present for the direct questions.

What do our results imply about the proportion of nonstrategic misreporters for each list experiment? While partition validation alone does not identify this quantity, we compute two implied proportions based on different sets of assumptions. First, we take a conservative approach, assuming that only false positives can be confidently attributed to nonstrategic misreporting (discounting false negatives among norm-defiers as potentially driven by strategic misreporting), and that the response process of nonstrategic misreporters contributes positive measurements only. Under these assumptions, the implied proportion of list experiment nonstrategic misreporters is simply the frequency of false positives among norm-compliers expressed as a proportion of all responses: 0.07 for both New Zealand and London. Second, we take a less conservative approach. This still makes the cautious assumption that only false positives can be confidently attributed to nonstrategic misreporting, but now assumes that these misreporters contribute positive measurements and negative measurements at equal rates (the expected outcome of a uniform response process). Under these assumptions, the implied proportion of nonstrategic misreporters is double the observed proportion of false positives in the sample (in expectation, whatever proportion of cases are false positives, there will be an equivalent proportion where nonstrategic misreporters contribute true negatives). This less conservative approach implies that the proportion of nonstrategic misreporters is 0.14 for both New Zealand and London. All of these implied proportions are substantially higher than the proportion of respondents assumed to be nonstrategic misreporters (0.03) in existing simulation studies (Ahlquist 2018; Blair *et al.* 2019).

Figure 2 summarizes overall direct question and list experiment classification performance as measured by accuracy. Unsurprisingly, given that the point estimate of the rate of false positives among actual voters and of false negatives among actual nonvoters was higher for the list experiment than the direct question in both surveys, the list experiment performs worse than the direct question in terms of overall classification accuracy for both New Zealand and London. The difference in the accuracy of the two measures is 7 points in New Zealand (95% CI: [2.06, 12.66]) and 8 points in London (95% CI: [0.23, 15.23]). In online Appendix E, we show that the list experiment also tends to underperform the direct question according to alternative classification performance measures, including recall of voting and recall of nonvoting. In online Appendix F, we show that both list experiments continue to underperform direct questions in terms of reporting accuracy when we use alternative direct question measures and alternative list estimators.

We emphasize the following key findings from this partition validation. First, the partition validation results are consistent with the notion that list experiments induce nonstrategic misreporting that is largely absent for a direct question. Second, even conservative estimates of the proportion of list experiment nonstrategic misreporters in our data are twice as large as the proportion assumed in existing simulation analyses of list experiment nonstrategic. Third, despite the promising results of standard information validation and standard information diagnostics, once we use true scores to improve validation, both list experiment measures are shown to generate more misreporting error overall than corresponding direct question measures.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

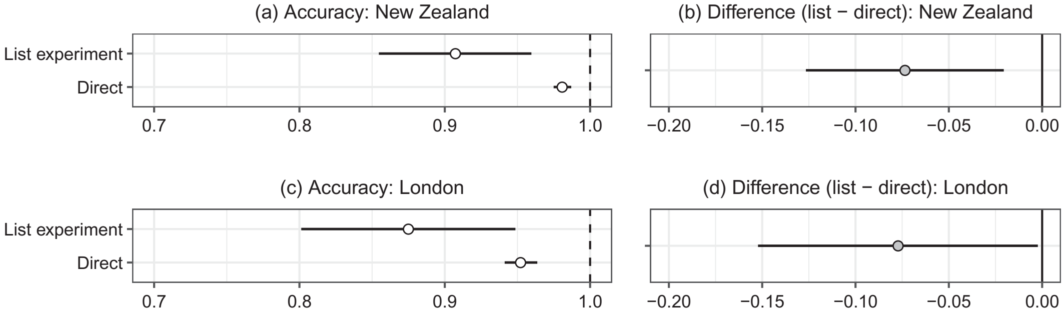

Figure 2. Measurement accuracy of direct and list experiment turnout measures. Notes: Plots (a) and (c) display, for New Zealand and London, respectively, direct and list accuracy. Dashed vertical lines indicate a perfect score. Plots (b) and (d) display differences in list and direct question accuracy. Horizontal lines indicate bootstrapped 95% confidence intervals.

6.4 Evidence of Nonstrategic Misreporting Due to Satisficing

We now provide evidence that satisficing behavior may be an important driver of list experiment nonstrategic misreporting. Recall that satisficing involves respondents answering more complex list experiment questions differently to direct questions because of the shortcuts they adopt to limit time and effort spent on the former. To examine whether such behavior drives list experiment underperformance, we focus on the London survey, where we have the measures necessary to identify respondents who exhibit satisficing-consistent behavior when answering the list experiment.

We identify probable satisficers based on two pieces of information: recall of the listed items and time taken to answer the list experiment question. To measure recall, we rely on a follow-up question which asked respondents to recall the first and last items on the list they had seen two questions earlier. Respondents were presented with open text boxes to record their answers, or could tick “don’t know.” Responses were coded as offering correct recall if they were judged to describe the correct activity using any form of words. We classify a respondent as a probable satisficer if they were unable to correctly recall either the first or last item on the list *and* if they are also in the bottom quartile in terms of time taken to answer the list experiment question. Measured in this way, the proportion of probable list experiment satisficers in the London sample is 0.12, slightly lower than the implied proportion that we computed under the less conservative set of assumptions in the previous subsection.[10]

In Figure 3, we subset the London sample into probable satisficers and nonsatisficers and, for each subgroup, use partition validation to calculate the accuracy of the list measure and of the direct measure based on the “baseline” direct turnout question asked by YouGov of all respondents following the 2017 General Election. Figure 3a shows that the list measure of turnout is clearly less accurate among probable satisficers (right panel) than among probable nonsatisficers (left panel). This would be expected if nonstrategic misreporting due to satisficing drives list experiment inaccuracy.

Figure 3b also shows that the list measure of turnout obtained from probable satisficers is significantly less accurate than the direct measure obtained for the same respondents. Among probable nonsatisficers, the list experiment still does not outperform the direct question in terms of accuracy, but the difference between the two measures is substantially smaller and indistinguishable from zero with 95% confidence. These patterns are consistent with the list question

[10] The proportion of respondents who failed to correctly recall either the first or last list item was 0.39 (0.45 for the first item only; 0.56 for the last item only). In online Appendix G, we show that our findings regarding satisficing and accuracy are robust to different measurement strategies for identifying satisficers.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

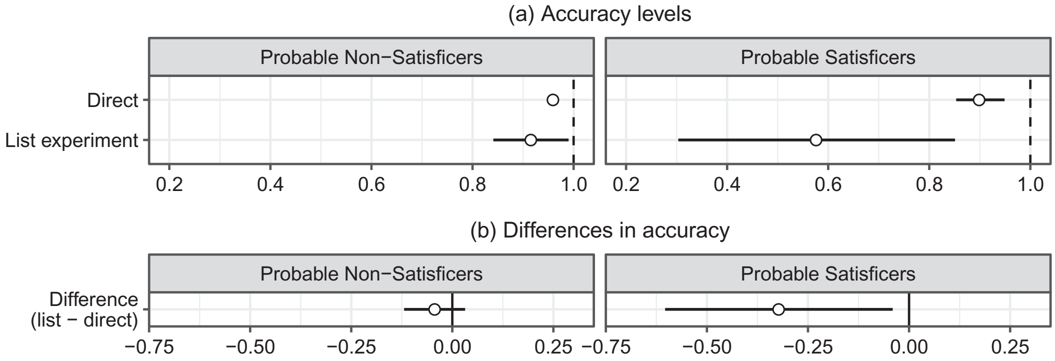

Figure 3. Measurement accuracy by probable list experiment satisficing, London sample. Notes: Plot (a) shows direct and list accuracy among probable nonsatisficers (left panel) and satisficers (right panel) in the London sample. Plot (b) shows differences in list and direct question accuracy. Horizontal lines indicate bootstrapped 95% confidence intervals.

inducing satisficing and nonstrategic misreporting that respondents do not engage in when asked a direct question.

6.5 Sample Prevalence Validation

Would the relative underperformance of the list experiments have been detected if we had simply used the true scores to conduct *sample prevalence validation* (comparing prevalence estimates to overall true sample prevalence), rather than partition validation? To check this, we perform sample prevalence validation in online Appendix D. For New Zealand, sample prevalence validation results partially concur with our partition validation finding of list experiment underperformance: the list prevalence estimate is further than the direct prevalence estimate from the true sample prevalence, although the 95% confidence interval for the difference between the two marginally overlaps with zero.[11] For London, whereas partition validation revealed the list experiment to perform substantially worse than the direct question in terms of reporting error, sample prevalence validation shows the list experiment performing no worse—and probably better—than the direct question. The key reason for this discrepancy is that the false negatives and false positives partially cancel out, and therefore go unnoticed, when assessing overall prevalence estimates.

7 Discussion

The relative magnitude of list experiment nonstrategic misreporting error and direct question strategic misreporting will depend on a number of factors that are likely to vary by context. What, then, are the implications of our validation results for researchers considering list experiments for other topics or in other settings?

To address this question, we begin with a simple parameterization of the general trade-off between direct question strategic misreporting and list experiment nonstrategic misreporting. We do so based on the models developed in Section 2, making the simplifying assumption that nonstrategic misreporter status is independent of norm-complier/defier status and that the expected list DiM among nonstrategic misreporters is 0.5. Based on this, online Appendix H derives an indifference function which, for a given level of true norm-defier prevalence (π) and proportion of list experiment nonstrategic misreporters (λ), gives the proportion of norm-defiers that must strategically misreport for the direct question (θ) such that expected list and direct question accuracy are equalized. Figure 4 plots resulting indifference curves for three levels of true

[11] This result demonstrates how list experiment nonstrategic misreporting can generate overall list prevalence estimates worse than those from a direct question subject to strategic misreporting.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

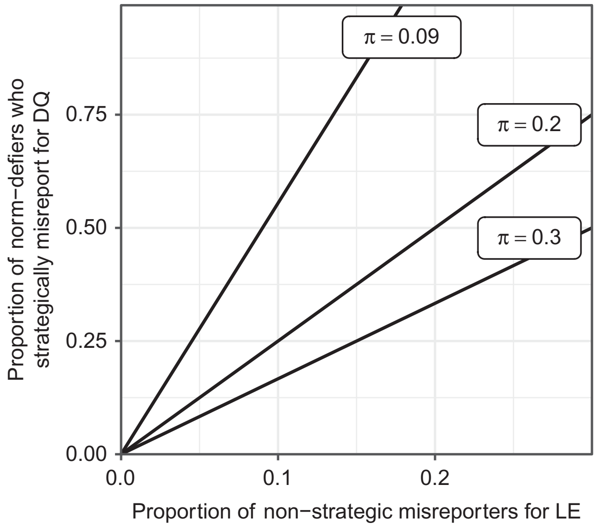

Figure 4. The direct question versus list experiment accuracy trade-off. Notes: Indifference curves showing, for varying proportions of list experiment nonstrategic misreporter (*x*-axis), what proportion of norm-defiers must strategically misreport for the direct question (*y*-axis), such that expected direct question and list experiment accuracy are equalized. The area above (below) a curve indicates superior list experiment (direct question) accuracy. Each curve assumes a different level of true norm-defier prevalence (π). All curves assume nonstrategic misreporter status is independent of norm-defier status and expected list DiM of 0.5 among nonstrategic misreporters.

norm-defier prevalence. The area above (below) a curve indicates superior expected list experiment (direct question) accuracy.

To decide whether a list experiment can be expected to improve on a direct question in terms of reporting error, a researcher planning a study will need to make assumptions about where their case is located in the parameter space depicted in Figure 4. Our validation studies provide an initial benchmark location that can help inform this judgement. Regarding the *x*-axis, the implied proportion of nonstrategic misreporters in our list experiments was 0.07 under conservative assumptions and 0.14 under less conservative assumptions, higher than the proportions previously assumed in simulation studies. Regarding the *y*-axis, the observed proportion of norm-defiers who misreported for a direct turnout question was 0.29 in the New Zealand survey and 0.27 in the London survey. Consider the indifference curve for $\pi = 0.09$, which corresponds to the true proportion of norm-defiers pooling our surveys. In Figure 4, any combination of the aforementioned *x* and *y* values falls below the indifference curve. Thus, if we set parameter values based on our validation results concerning the topic of nonvoting among New Zealand and London survey respondents, we arrive at a region of the parameter space where expected list experiment accuracy is inferior to expected direct question accuracy (consistent with the actual differences in overall accuracy observed in Section 6).

In what ways would we expect studies of other topics in other survey settings to depart from this region of the parameter space, and with what consequences for relative list experiment performance? First, we might plausibly expect a shift upward along the *y*-axis in Figure 4 for some cases of interest. While our evidence clearly indicated direct question strategic misreporting among nonvoters, failing to vote merely violates a social norm in the countries studied. Other researchers may be interested in sensitive attitudes or behaviors that invite serious legal or physical consequences if admitted to. For these topics, such as support for U.S.-led security forces among Afghans (Blair *et al.* 2014) or acceptance of clientelistic payoffs among voters (Corstange 2018), strategic misreporting under the direct question is plausibly higher than we found for nonvoting. In addition, whereas we studied self-complete surveys, direct question strategic misreporting may increase in interviewer-administered surveys, where social desirability concerns

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

may be more acute (Tourangeau and Yan 2007). Whether due to changes in topic or mode of administration, how much higher would direct question strategic misreporting need to be for the list experiment to outperform the direct question in expectation? Given the assumptions of Figure 4, if π were 0.09 and if the proportion of nonstrategic misreporters was similar to that implied by our data under conservative assumptions (i.e., 0.07), then a researcher would need to believe that the proportion of norm-defiers who strategically misreport for the direct question is above 0.39 to reasonably expect the list experiment to outperform the direct question in terms of accuracy. If the proportion of nonstrategic misreporters was similar to that implied by our data under less conservative assumptions (i.e., 0.14), the researcher would want to be confident that the proportion of norm-defiers strategically misreporting for a direct question is more than 0.78.

Second, comparison across indifference curves in Figure 4 makes clear that the misreporting trade-off between list experiment and direct question depends on the true prevalence of norm-defiers for the sensitive topic of interest. For a fixed amount of direct question strategic misreporting and list experiment nonstrategic misreporting, the relative accuracy of the list experiment increases as true norm-defier prevalence increases. The marginal cost of additional list nonstrategic misreporting (in terms of relative accuracy) is also lower when true norm-defier prevalence is higher (indicated by the slopes of the indifference curves). This is because increases in norm-defier prevalence mean more norm-defier survey respondents and a greater proportion of the overall sample susceptible to strategic misreporting incentives if asked a direct question. In the cases we study, norm-defier prevalence among survey respondents is low (0.09 pooling across both surveys), which advantages the direct question over the list experiment. In a case where the level of direct question strategic misreporting and list experiment nonstrategic misreporting were equivalent to those apparent in our studies, but where norm-defier prevalence rose to 0.3, the list experiment would be expected to outperform the direct question in terms of accuracy. Researchers in other settings will therefore need to carefully develop priors about the likely prevalence of norm-defiers among the respondents they expect to survey. For example, studying the sensitive topic of vote-buying, Corstange (2018, 81) cites local observers as estimating that at least half of Lebanese electors have their votes "bought," suggesting a much higher norm-defiance prevalence than in our case, and therefore more favorable conditions for a list experiment on this dimension.

Turning to the third parameter that varies in Figure 4, the proportion of list experiment nonstrategic misreporters, we contend that the cases we study are relatively favorable to the list experiment on this dimension. On the one hand, high-cost face-to-face surveys may encourage better engagement and therefore less nonstrategic misreporting than our self-complete surveys. On the other hand, our New Zealand list experiment was embedded in a reputable national election study, where respondents will plausibly have felt a greater sense of duty to engage with the survey than usual, and the experiment was also only the second question on the survey, minimizing disengagement due to tiring. Moreover, existing placebo studies suggest that reporting error in list experiments is greater among less educated respondents, who are less well equipped to process and answer a more complex list-style survey question (Kramon and Weghorst 2019; Riambau and Ostwald 2020). From this perspective, the rate of nonstrategic misreporting in the list experiments we have studied should be comparatively low, since the New Zealand and London populations from which we sample exhibit high levels of literacy, numeracy, and general education, from a comparative perspective. Even in these cases, our evidence is consistent with 7%—or even 14%, depending on assumptions—of respondents being list experiment nonstrategic misreporters. Researchers planning surveys on sensitive topics in developing countries where education levels are lower—and where list experiments are increasingly used (Kramon and Weghorst 2019)—thus have reason to expect higher levels of list nonstrategic misreporting than we have found, an effective shift rightward in Figure 4. In such cases, the amount of direct question strategic

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

misreporting and/or norm-defier prevalence would need to be considerably higher than in our case for expected list experiment accuracy to reach expected direct question accuracy.

In sum, compared to the cases we have validated, others may present more favorable scenarios for the list experiment in terms of the misreporting trade-off, since they may feature higher rates of direct question strategic misreporting among norm-defiers and higher prevalence of norm-defiers. However, we also suggest that for other cases, the proportion of list experiment nonstrategic misreporters is likely to be similar to or greater than the nontrivial proportions implied in our data. Where this is higher, all else equal, relative list experiment accuracy will be lower than in the cases we study.

The misreporting trade-off is not the only one between direct questions and list experiments. Blair *et al.* (2020) examine a different and better known trade-off between direct question prevalence estimate bias (due to strategic misreporting) and list prevalence estimate inefficiency (due to masking via aggregation of sensitive and control items). They characterize this *bias-variance* trade-off in terms of prevalence estimate root-mean-square error (RMSE) and, for varying sample sizes, show at what level of direct question strategic misreporting the list experiment RMSE is as good as the direct question RMSE. Does our analysis of the misreporting trade-off—as informed by our empirical validation results—have any implications for the choice between direct question and list experiment beyond those that emerge from consideration of the bias-variance trade-off?

A simple example suggests it does. Take the same three norm-defier prevalence levels considered in Figure 4 and assume the proportion of list experiment nonstrategic misreporters is 0.07. Adopting the Blair *et al.* (2020) parameterization of the bias-variance trade-off and assuming a healthy sample size of 5,000 in each case, for the list prevalence estimate to achieve an RMSE as good as the direct question, the proportion of norm-defiers strategically misreporting for the direct question must reach 0.27 (when $\pi = 0.09$), 0.12 (when $\pi = 0.2$), and 0.08 (when $\pi = 0.3$). All three of these threshold rates of direct question strategic misreporting among norm-defiers are lower than those found when the same scenarios are considered in terms of the misreporting trade-off as in Figure 4 (0.39, 0.18, and 0.12, respectively). In other words, in all three scenarios, the level of direct question strategic misreporting required for the list experiment to match the direct question on the bias-variance trade-off is *lower* than that required for the list experiment to match the direct question on the misreporting trade-off.[12] This example shows that, when one takes reporting error as an evaluation criterion and considers levels of nonstrategic misreporting that are conservative given our evidence, the range of settings in which list experiments outperform direct questions is narrower than previously thought. Therefore, researchers choosing between a list experiment and a direct question on any sensitive topic should consider not just the well-known trade-off between bias under the direct question and inefficiency under the list experiment, but also the trade-off between strategic misreporting error for the direct question and nonstrategic misreporting error for the list experiment.

8 Conclusion

This paper has provided the first empirical validation of list experiments that distinguishes between true and false negative and positive measurements on the sensitive variable. In doing so, it provided new evidence of the problem of list experiment nonstrategic misreporting in practice. We examined nonvoting in elections across two different countries with relatively educated populations, exploiting measures of true respondent turnout behavior to perform partition validation in both settings. We found that list experiments induced respondents to nonstrategically misreport more than direct questions and at higher rates than assumed in existing

[12] This will not always be true: for example, when sample size is small, the direct question will outperform the list experiment on prevalence estimate RMSE even for very high levels of strategic misreporting.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

simulation studies. This nonstrategic misreporting was not detected by standard diagnostics or validation approaches, but was still severe enough that both list experiments underperformed simple direct questions (which themselves suffered from strategic misreporting) in terms of overall reporting accuracy.

Our findings highlight the importance of the trade-off between direct question strategic misreporting and list experiment nonstrategic misreporting. We showed how, given a simple parameterization for this general trade-off, our empirical findings can help researchers gauge what level of direct question strategic misreporting and norm-defier prevalence would need to pertain for a list experiment to be accuracy-improving.

Our findings also underline the importance of future research into the reduction of list experiment nonstrategic misreporting. One approach to this involves embedding attention check questions to identify respondents more likely to nonstrategically misreport (Oppenheimer, Meyvis, and Davidenko 2009; Eady 2017; Alvarez *et al.* 2019). However, Alvarez *et al.* (2019) point out that dealing with identified inattentives is not straightforward, since dropping such respondents could lead to selection bias in inferences.

Another recent suggestion is to include a *placebo* statement (i.e., a statement no respondent can truthfully affirm) in the control list, such that the total available items is equalized in the control and treatment list (Riambau and Ostwald 2020). If list experiment nonstrategic misreporters reported item counts are a function of list length, this would yield an expected DiM of zero among nonstrategic misreporters. However, as shown formally in online Appendix I, this approach does not eliminate errors due to list experiment nonstrategic misreporting, since nonstrategic misreporters who are actual norm-defiers still contribute false negatives. It does though, under certain assumptions, at least allow researchers to sign the resulting prevalence estimate bias as negative.

If it is not possible to meaningfully reduce list experiment nonstrategic misreporting, might one generally use list experiments alongside direct questions to estimate bounds for norm-defier prevalence? This may work if: (a) strategic misreporting downward biases the direct prevalence estimator; (b) nonstrategic misreporting upward biases the list experiment estimator. While (a) should hold generally, (b) is problematic, since Section 2 showed that nonstrategic misreporting can up- or downward bias list prevalence estimates. If nonstrategic misreporter and actual norm-defier status are orthogonal, we at least know that list prevalence estimates are upward biased when the expected DiM among nonstrategic misreporters exceeds true norm-defier prevalence. Applied researchers will not be able to fully verify this condition but could check its plausibility by estimating the DiM among probable nonstrategic misreporters (identified using either an attention check or a mixture of list experiment recall questions and response time information). We encourage further research into this potential approach.

Acknowledgments

We are grateful for feedback from Gidon Cohen, Dominik Duell, Mona Morgan-Collins, Kosuke Imai, Benjamin Lauderdale, Mark Pickup, David Redlawsk, Jacob Shapiro, Jack Vowles, and seminar or panel participants at the University of Essex, the North East Research Development Workshop (NERD) at Durham University and University of Newcastle, the London School of Economics and Political Science, the Zurich Center for Comparative and International Studies, EPSA 2018, APSA 2019, the 2017 and 2019 PSA Political Methodology Group, and EPOP 2019. We also thank the editor and anonymous reviewers for their thorough and insightful comments. Moreover, we thank Jack Vowles and the NZES team for facilitating data collection in New Zealand and Samuel Mellish for superb assistance in collecting validated vote data for London. This work was funded by the British Academy/Leverhulme Small Grants scheme (SG162963) and a research support grant from the Faculty of Social Sciences and Health at Durham University.

Data Availability Statement

Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code and can be viewed interactively at Kuhn and Vivyan (2020a) at https://doi.org/10.24433/CO.3695413.v1. A preservation copy of the same code and data can also be accessed via Dataverse at Kuhn and Vivyan (2020b) at https://doi.org/10.7910/DVN/W90Q7B.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017.pan.2021.10.

References

Ahlquist, J. S. 2018. "List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimates." *Political Analysis* 26(1):34–53.

Alvarez, R. M., L. R. Atkeson, I. Levin, and Y. Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27(2):145–162.

Aronow, P., A. Coppock, F. W. Crawford, and D. P. Green. 2015. "Combining List Experiments and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3(1):43–66.

Belli, R. F., M. W. Traugott, and M. N. Beckmann. 2001. "What Leads to Vote Overreports? Contrasts of Overreports to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics* 17(4):479–498.

Blair, G., W. Chou, and K. Imai. 2019. "List Experiments with Measurement Error." *Political Analysis* 27(4):455–480.

Blair, G., A. Coppock, and M. Moor. 2020. "When to Worry About Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114(4):1297–1315.

Blair, G., and K. Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1):47–77.

Blair, G., K. Imai, and J. Lyall. 2014. "Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan." *American Journal of Political Science* 58(4):1043–1063.

Blais, A., and C. H. Achen. 2019. "Civic Duty and Voter Turnout." *Political Behavior* 41(2):473–497.

Corstange, D. 2018. "Clientelism in Competitive and Uncompetitive Elections." *Comparative Political Studies* 51(1):76–104.

Eady, G. 2017. "The Statistical Analysis of Misreporting on Sensitive Survey Questions." *Political Analysis* 25(2):241–259.

Höglinger, M., and A. Diekmann. 2017. "Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT." *Political Analysis* 25(1):131–137.

Höglinger, M., and B. Jann. 2018. "More Is Not Always Better: An Experimental Individual-Level Validation of the Randomized Response Technique and the Crosswise Model." *PLoS One* 13(8):e0201770.

Holbrook, A. L., and J. A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Techniques." *Public Opinion Quarterly* 74(1):37–67.

Kramon, E., and K. Weghorst. 2019. "(Mis)Measuring Sensitive Attitudes with the List Experiment: Solutions to List Experiment Breakdown in Kenya." *Public Opinion Quarterly* 83(1):236–263.

Krosnick, J. A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Journal of Cognitive Psychology* 5(3):213–236.

Krosnick, J. A., and S. Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by J. D. Wright and P. V. Marsdent, 263–314. 2nd edn. San Diego, CA: Elsevier.

Kuhn, P. M., and N. Vivyan. 2018. "Reducing Turnout Misreporting in Online Surveys." *Public Opinion Quarterly* 82(2):300–321.

Kuhn, P. M., and N. Vivyan. 2020a. "Replication Data for: The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." Code Ocean. https://doi.org/10.24433/CO.3695413.v1.

Kuhn, P. M., and N. Vivyan. 2020b. "Replication Data for: The Misreporting Trade-Off Between List Experiments and Direct Questions in Practice: Partition Validation Evidence from Two Countries." Harvard Dataverse, V1. https://doi.org/10.7910/DVN/W90Q7B.

Lax, J., J. Phillips, and A. Stollwerk. 2016. "Are Survey Respondents Lying About Their Support for Same-Sex Marriage?" *Public Opinion Quarterly* 80(2):510–533.

Manning, C. D., P. Raghavan, and H. Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45(4):867–872.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press

Presser, S. 1990. "Can Context Changes Reduce Vote Over-Reporting?" *Public Opinion Quarterly* 54(4):586–593.

Riambau, G., and K. Ostwald. 2020. "Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore." *Political Science Research and Methods* 9:172–179. doi:10.1017/psrm.2020.18.

Rosenfeld, B., K. Imai, and J. N. Shapiro. 2016. "An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions." *American Journal of Political Science* 60(3):783–802.

Tourangeau, R., and T. Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5):859–883.

https://doi.org/10.1017/pan.2021.10 Published online by Cambridge University Press