Underreporting of Purchases in the U.S. Consumer Expenditure Survey

Stephanie Eckman
Fellow
Survey Research Division
RTI International
Washington DC 20005
United States
seckman@rti.org

Acknowledgements: The research reported here was supported by an ASA-NSF-BLS Fellowship. The author thanks Laura Erhard, Parvati Krishnamurty, Adam Safir, Lucilla Tan, and Erica Yu for discussion on the ideas behind this research, help with data processing, and comments on draft versions of the paper. Ruben Bach, Ashley Amaya and Jill Dever also provided comments.

Forthcoming in *Journal of Survey Statistics and Methodology*

Abstract

Motivated misreporting occurs when respondents give incorrect responses to survey questions to shorten the interview; studies have detected this behavior across many modes, topics, and countries. This paper tests whether motivated misreporting affects responses in a large survey of household purchases, the U. S. Consumer Expenditure Interview Survey. The data from this survey inform the calculation of the official measure of inflation, among other uses. Using a parallel web survey and multiple imputation, this paper estimates the size of the misreporting effect without experimentally manipulating questions in the survey itself. Results suggest that household purchases are underreported by approximately 5 percentage points in three sections of the first wave of the survey. The approach used here, involving a web survey built to mimic the expenditure survey, could be applied in other large surveys where budget or logistical constraints prevent experimentation.

Statement of Significance

Household purchases may be underreported in the first wave of the U. S. Consumer Expenditure Interview Survey. The approach used here, involving a web survey built to mimic the expenditure survey, could be applied in other large surveys where budget or logistical constraints prevent experimentation.

1 Introduction

Researchers are increasingly aware of the shortcuts that respondents take to make surveys less burdensome: when the structure of the questions allows respondents to learn which responses trigger additional questions, they can avoid giving those responses. This phenomenon, called motivated misreporting, has been detected in surveys across several modes, countries, and topics, suggesting it is a wide-spread phenomenon (Duan et al. 2007; Kreuter et al. 2011; Eckman et al. 2014; Tourangeau et al. 2015). In previous studies of motivated misreporting, the most vulnerable questions concern clothing purchases or other routine household purchases (Kreuter et al. 2011; Eckman et al. 2014; Bach and Eckman, 2018). The Consumer Expenditure Interview Survey (CE) asks hundreds of such questions. The collected data are used in the calculation of the Consumer Price Index (CPI) and other important economic analyses. The survey's reliance on filter questions has led researchers to worry about underreporting of purchases in the CE (Bosley et al. 1999; Shields and To 2005; Yan and Copeland 2010; McBride 2013; Clark-Fobia et al. 2018). The National Research Council (2013, pp. 84-85) mentioned the risk of motivated misreporting in its review of the CE:

> It seems likely that respondents learn quickly...that the interview will last longer if they answer "yes" to these [filter] questions.... Fifty percent of field representatives said [this] happened frequently or very frequently.

This study investigates the amount of the measurement error introduced by motivated misreporting in the first wave of the CE. Using a web survey which experimentally varied question order, and data from the CE itself, the study imputes the responses that would be obtained in a hypothetical version of the CE. The results in this paper will be of interest to users of CE data and to many other surveys that use filter questions. In addition, the approach taken in this study could be applied to other large ongoing surveys that cannot launch full-scale experiments of alternative methods but nonetheless want to estimate and understand measurement error in their current questionnaire.

2 Relevant Literature

This study builds on the literature on motivated misreporting. Because experimentation in the CE was not possible, the study relies on multiple imputation to estimate the hypothetical responses in a different version of the CE instrument. The two subsections below summarize the relevant research on motivated misreporting and multiple imputation.

Motivated Misreporting

The literature on motivated misreporting demonstrates that some respondents will give incorrect answers in surveys to reduce the length or burden of the interview. Motivated misreporting occurs in three types of questions: filter, looping, and screening questions (Tourangeau et al. 2015). Both filter and looping questions are present in the CE, although this study focuses on filter questions.

Filter questions can be asked in two formats: grouped and interleafed. The interleafed format asks the follow-ups after each filter. The grouped format asks all filter questions at the start of the section and then asks all relevant follow-up items. Although the questions asked in each format are the same, the order is different, which affects the experience of the respondent. The interleafed format allows a respondent to learn the consequences of a "yes" answer and adjust her reporting to reduce the burden of the survey. The grouped format collects all responses to the filter questions before the respondent realizes that the follow-up questions are coming. Numerous experimental studies have found that filters in the grouped format collect more "yes" responses than those in the interleafed format (see for example: Kreuter et al. 2011, Eckman et al. 2014, Eckman and Kreuter 2018, Bach et al. 2019). These studies suggest that respondents underreport in the interleafed format to reduce the burden and length of the survey, and comparison of filter responses to administrative records supports this explanation (Eckman et al. 2014).

Another way to order the filter and follow-up questions would be to ask all filters (in all sections) first and then all the follow ups: the grouped-overall format. Previous research, associated with the study behind Eckman et al. (2014) but not reported there, suggested that the grouped-overall format is quite awkward for both interviewers and respondents. It requires respondents to first think about vacations, then about clothing, then about furniture, and later to return to each topic to answer the follow up questions: "thinking back to the vacation that you said a household member took in the last three months..." This format is seldom used in surveys.

Looping questions are similar to filter questions. They ask a series of follow-up questions about each event a respondent has experienced: degrees, jobs, purchases. These questions can also be asked in two formats. In the how-many format, the respondent first reports how many events she has experienced and then answers the follow-up questions about each one. This format is similar to the grouped format: the respondent does not realize that the follow-up questions are coming before reporting the number of events. In the go-again format, the respondent answers the follow-up questions about one event and then is asked if he has another event of that type to report. The go-again format is similar to the interleafed format, because the respondent experiences the follow-up questions before deciding to report another event. The how-many format collects more event reports than the go-again format (Eckman and Kreuter 2018).

Motivated misreporting has been detected across several topics, from mental health (Duan et al. 2007; Kessler et al. 1998) to employment (Eckman et al. 2014) and in several countries (Kreuter et al. 2019). The effect also exists in all tested modes. In a small face-to-face study ($n = 304$) in Maryland, Bach et al. (2019) found underreporting by 3.3 percentage points ($t = 1.84; p = 0.066$) in the interleafed format. A U.S. web survey found a larger format effect, 11.4 percentage points ($p < 0.001$; $n = 1,215$; Kreuter et al. (2019)).

The CE asks filter questions in the interleafed format in many sections, leading to the concern that purchases are underreported. Previous research has explored motivated misreporting in filter questions in the CE. Bosley et al. (1999) found no evidence of motivated misreporting, although the study involved only 24 participants and looked at responses in the second wave. McBride (2013) investigated whether reports of purchases decreased over the length of the CE interview detecting some evidence for this trend. In cognitive interviews using CE items, respondents preferred the interleafed condition because it allowed them to think about one item at a time, rather than jumping between items as in the grouped format (Clark-Fobia et al. 2018). Other studies have investigated underreporting of purchases in later waves of the CE, finding little evidence that data quality worsens over waves (Shields and To 2005; Yan and Copeland 2010; Bach and Eckman 2019). However, none of these studies provide strong evidence of motivated misreporting in the first wave of CE, a gap in the literature that this paper addresses. Thus, there is a contradiction in the literature: studies robustly find motivated misreporting in many surveys, yet the phenomenon has not been conclusively demonstrated in the CE, a long survey which uses many filter questions in the interleafed format.

Multiple Imputation

To estimate motivated misreporting in the CE, this study imputes the responses that CE respondents would give if the filter and follow-up questions were asked in the grouped format. Imputation is most often used to fill in item missing data in surveys. For example, respondents may skip or refuse a question about their income. To impute these missing values, researchers might use regression imputation, which involves building a model to predict income from the other variables in the survey. On the cases where income is not missing, income is modeled as a function of the available characteristics (for example, age, home ownership, and marital status). The model can then predict income for the cases where it is missing. Other imputation methods, such as hot deck, are also possible.

The concern with all methods of single imputation is that analysis of the imputed data set does not account for the uncertainty in the imputed values. To correctly account for this uncertainty, researchers use multiple imputation, filling in several values for each missing response. The values might be multiple predictions from one model or predictions from several models. The result is multiple complete data sets. The final analyses then account for the uncertainty in the imputed values by capturing the variation across the data sets (Rubin 1987).

More recently, researchers have expanded the use of multiple imputation beyond item missing data. The *cross-survey imputation* of Rendall et al. (2013) involves collecting all variables of interest in one survey (the "impute-from" survey) and a subset of those variables in the other survey (the "impute-to" survey), which is generally larger or higher quality. In the paper's application, shown in Table 1a, the background variables (X_1) and the outcome variable, Y, were observed (shown with "O") in both surveys. However, one crucial explanatory variable, X_2 was observed in only the impute-from survey. The authors imputed X_2 in the impute-to survey (shown with "I") and used the multiply-imputed data in analysis. An important condition of this approach is that all variables of interest are collected in one survey: the variables are *jointly observed*.

Other researchers have taken these ideas further and relaxed the jointly-observed requirement. Several studies have used multiple imputation to predict how respondents would have responded in an alternative mode (Powers et al. 2005; Christensen et al. 2006; Peytchev 2012; Kolenikov and Kennedy 2014; Park et al. 2016). The approach is shown in Table 1b: respondents in Source 1 responded in Mode A (Y_{modeA}) and those in Source 2 responded in Mode

B (Y_{modeB}). The hypothetical responses of Source 2 respondents in Mode A were filled in via multiple imputation. The goal of these studies is to remove the mode effect when combining responses from the two sources. Importantly, in most of these studies, no cases respond in both modes; responses in the two modes are not jointly observed.

Two recent papers used a nonprobability survey to impute a substantive variable missing from a probability data set (Kim et al. 2020; Chen et al. 2020). Researchers collected background variables (*X*) and the outcome of interest (*Y*) with a nonprobability web survey and then imputed *Y* on the probability survey: see Table 1c. In these studies, the *X* and *Y* variables were jointly observed.

The approach in this paper combines elements from the mixed-mode approach (Table 1b) and the nonprobability approach (Table 1c). Section 3 gives details of the nonprobability web survey and the CE, and Section 4 describes the imputation approach in detail.

Table 1: Three Uses of Imputation

Data Source	X1	X2	Y
Impute-to Survey	O	I	O
Impute-from Survey	O	O	O

(a) Rendall et al. Approach

Data Source	X	Y_{modeA}	Y_{modeB}
Source 1	O	O	
Source 2	O	I	O

(b) Mixed-Mode Approach

Data Source	X	Y
Probability	O	I
Nonprobability	O	O

(c) Nonprobability Approach

X = background variables; Y = variable of interest
O = observed data; I = imputed data

3 Data

Data from a web survey designed to mimic portions of the CE support the imputation of counterfactual responses in the CE: how CE respondents would have answered filter questions about purchases if they were asked in the grouped rather than the interleafed format. The two subsections below describe the web and CE data sets used in the imputation.

Web Survey

The web survey was conducted September 27–October 9, 2019, with members of the Lightspeed opt-in panel; 2,198 completed the survey. The median response time was 10.7 minutes (mean 13.7). The panel did not report the number of cases invited to the survey, so no response rate or participation rate is available. The web survey's questionnaire consisted of six sections and was modeled after the CE questionnaire. It contained three sections of background questions: Section 1 asked about demographics, Section 2 about housing unit characteristics, and Section 6 about household income. See the Open Science Framework (https://osf.io/a5vpe/) for question wording. These variables are strongly correlated with purchases and thus are needed for the imputation models described below. The order of these sections matched the order in which they are asked in the CE and did not vary across respondents.

Sections 3, 4, and 5 of the questionnaire contained questions from three purchase sections of the CE: utilities, clothing, and non-health insurances. The 16 filter questions in these sections asked about recent household purchases of (or payments for) items such as electricity, shoes, and life insurance. Each filter question in these sections, if answered with a "yes," triggered follow-up questions. In the utilities section, the follow-ups asked about the amount paid, the billing period, and whether any portion was a deductible business expense. In the clothing section, the follow-ups asked about the cost, the month of the purchase, and for whom the purchase was made. In the insurance section, the follow-ups again asked for the amount, the frequency, and whether any portion of the premiums was paid by the household. The wording of all questions is available at the link in the previous paragraph. Table 2 shows the number of filters and follow-ups in these three sections. These sections were chosen in consultation with CE staff to reflect the diversity of topics included in the CE and because they do not depend on answers in other sections. Questions based on the clothing section have been used in several previous studies (Kreuter et al. 2011; Eckman et al. 2014).

Table 2: Number and Characteristics of Questions in Filter Question Sections, by Section

Section	# Filters	# Follow-Ups[a]	Go-Again Loop
Utilities	5	4	No
Clothing	6	5	Yes
Insurance	5	5	Yes

[a] per filter question

The web survey contained two manipulations. The first randomly assigned respondents to receive the filters in the interleafed format (51% of respondents) or the grouped format (49% of respondents). The grouped format was grouped within each section: respondents answered all filters and follow-ups in one section before moving on to the next. Because of this awkwardness of the grouped-overall format, it was not used in this study. If the grouped-overall format collects more reports of purchases than even the grouped (within-section) format, and these additional purchases are correct, then the version of the grouped format implemented in this study may underestimate underreporting in the CE. The second manipulation randomized the order of the three purchase sections. Between 16% and 17% of respondents answered the sections in each of the six possible orderings. Both randomizations, format and section order, were successful. The mean household size does not differ significantly between the two formats ($F(1,3129) = 28.36; p < 0.001$). The distribution of 21 of the 22 categorical variables shown in Table 4 also did not differ at the 5% level of significance between the formats or the section orders. Only household receipt of self-employment income showed significant deviation between the two formats ($\chi^2(2) = 7.97$; $p = 0.019$), which is not unexpected with multiple tests of significance.

In the clothing and insurance sections, the CE uses a go-again loop to collect information about additional purchases of a given item. The web survey included go-again loops in these sections as well (Table 2). However, although the CE allows respondents to report many purchases of each item, the web survey was limited to only two purchases, to limit the complexity of the web survey programming. In preparing the data for analysis, all purchases after the first for each item in both the CE and web data were discarded. That is, if the respondent reported having two life insurance policies, the filter for that item is still coded as "yes" and not two.

Throughout the questionnaire, explicit *don't know* and *prefer not to say* responses were available. In the CE, these response options are not explicitly offered, but respondents can tell the interviewer they do not know an answer or do not wish to answer a question. In the web mode, this approach of allowing such responses without making the options explicit does not work. Even when explicit response options are offered, online panel participants tend to answer all questions (Hillygus et al. 2014; Kaplan and Edgar 2018). To try to make the web survey participants' behavior similar to the CE participants', the survey included explicit *don't know* and *refused* options on most questions. The survey also displayed text at the beginning calling respondents' attention to these options, following Kaplan and Edgar (2018). Web respondents could also simply skip past most questions without answering. Across the 35,168 filters (2,198 respondents ×16 items), 3.9% of responses were "don't know" responses and 1.3% were "prefer not to say." In addition, 0.53% of filters were simply skipped. To match how the CE works (described below), all filter responses other than "yes" were recoded to "no" for analysis.

Consumer Expenditure Interview Survey

The CE is conducted every month in the United States by Census Bureau interviewers. Selected housing units remain in the sample for four waves, but only data from the first wave are used in this study, because the mechanism of response is likely different in later waves when respondents know that each reported purchase leads to follow-ups (see Bach and Eckman 2019, for a discussion of motivated misreporting in later waves of the CE). During the interview, a household informant reports on purchases by all household members in the previous 3 months.

The instrument contains 23 sections. Several collect background information about the household and its members. The majority ask about purchases using filter and looping questions. There are sections on spending on trips and vacations, appliances, home furnishings, vehicle expenses, and more. The Bureau of Labor Statistics prefers that the survey be administered in person, but respondents can choose to do a telephone interview. About 75% of first wave interviews are conducted in person. In 2018, the median length in the first wave was 75 minutes (Hubener et al. 2019).

To match the web survey as closely as possible, the analysis uses data from September and October 2019. In these months, 932 respondents completed the first wave of the survey, 55% of all eligible cases. (The CE does not calculate response rates for each month and wave.) The data used in this study are available to researchers outside of the Bureau of Labor Statistics only via the onsite visiting researcher program. The data do not include any imputations or editing.

In the CE, interviewers can administer the sections in any order but usually stick to the default order. Data about the order of administration are captured only in the instrument trace files, which were not available for this study. The imputation models discussed below assume that all CE respondents received the six sections used in the web surveys in the default order: demographics, housing unit characteristics, utilities, clothing, non-health insurance, income.

The filter questions in the CE are asked in the interleafed format, but in an unusual way that the self-administered web survey was not entirely able to replicate. For example, during the insurance section, interviewers read a list of the insurance types of interest: life, homeowner's, renter's, car, and other. The respondent should mention which of the insurance types she has. For each type reported, the follow-up questions are asked. If the respondent does not mention a given item, such as renter's insurance, then that item simply does not appear in the data set for that respondent. That is, "no" responses to the filter questions are not recorded in the CE data. To make the CE match the web survey, "no" responses were filled in for items that were missing in the raw data. With this edit and the recoding of missing responses to filter question in the web survey to "no," the two surveys should be aligned. Thirty-one cases (3.33%) reported no purchases in the 16 items used in this study, which is a close match for a same percentage in the web survey (3.37%).

If the household purchased more than one of a given item, the respondent first answers the follow-up questions about one purchase and then can indicate additional purchases and answer follow-ups about each one. As mentioned above, each item was coded as purchased (1) or not (0), even if a respondent reported more than one purchase of a given item. In the CE, respondents reported two or more purchases to 5.9% of the clothing items and 2.7% of the insurance items. In the web, these percentages were 10.2% and 7.0%. These results are in line with generally higher reporting in these two sections by the web respondents, as shown in Section 5. Collapsing multiple purchases makes the analysis in this paper similar to other studies of motivated misreporting.

4 Methods

The ideal design to test for motivated misreporting in the CE would randomly assign respondents to the interleafed or grouped formats, while holding constant other study characteristics. The difference in the number of reported purchases between the formats would be the measure of motivated misreporting. Although this approach has worked well in earlier studies, experimentation is not feasible in the CE because of budget constraints and the importance of the CE data. However, the importance of the CE also means that it is crucial to know whether motivated misreporting is taking place.

Instead, this study uses multiple imputation to estimate the extent of motivated misreporting in Wave 1 of the CE. Table 3 illustrates the approach. The CE is conducted in the interleafed format and collected background variables (X) as well as purchase data ($Y^{interleafed}$). The web survey collected the same background variables and asked the purchase questions in both formats. However, no cases responded in both the interleafed and grouped format: the variables X, $Y^{interleafed}$, and $Y^{grouped}$ are never jointly observed. The web survey did not administer the questions in both formats to the same respondents because those answering the purchase questions twice would likely respond differently the second time or even break off. Just as in the multimode imputation studies discussed in Section 4, joint observation of the $Y^{interleafed}$ and $Y^{grouped}$ is not possible. Responses by the CE respondents to the filter questions in the grouped format are imputed from the observed cells.

Table 3: Design for Imputation of CE Responses in Grouped Format

Data Source	X	$Y^{interleafed}$	$Y^{grouped}$
Consumer Expenditure	O	O	I
Web Survey	O	O	
	O		O

O = observed data; I = imputed data

Twelve imputation models are used to impute the responses in the grouped format, marked "I" in Table 3. Each model predicts multiple responses (yes or no) to the filter questions about household purchases. The following subsections compare the responses of the CE and web respondents, describe the imputation models, and detail the analysis approach.

Comparison of CE and Web Survey Respondents

There are many differences between the web and CE surveys. The CE is an interviewer-administered survey: interviewers can provide motivation to complete the interview and to provide higher quality answers. The CE contains many more sections and items than the web survey. The CE respondents are recruited from a probability sample of the U.S. household population; the web respondents are members of the Lightspeed panel who take surveys in exchange for payment. The extent to which the models can accurately predict the responses of CE respondents in the grouped format depends on how similar the respondents to the two surveys are. If the web respondents are very unlike the CE respondents, in ways that influence purchasing behavior, then the web survey data cannot be used to impute the purchases of the CE respondents.

About half of the length of the web survey was devoted to collecting background information about each case. The variables fall into three categories: demographics, housing unit characteristics, and household income. The full list of background variables is given in Table 4. These variables are the predictors in the imputation models. Although there may be additional variables that would also be related to household purchases, the imputation models can only use variables collected in the CE.

Table 4 compares the CE and web survey respondents on the background variables. Missing responses (don't know, refused, and simply skipped) are combined and shown as a category. (The seven "HU has" variables are exceptions: each has only two response categories: "yes" and "no": missing responses in the CE are not permitted. To match this behavior, all missing responses to these questions in the web survey were recoded as "no.") Statistical comparisons between the two surveys require an assumption about the selection process in the nonprobability web survey. Following guidance from the American Association for Public Opinion Research (AAPOR 2016), the comparisons in Table 4 assume that the web respondents are a simple random sample from the Lightspeed panel; no weights, clustering, or stratification are used in the calculation of standard errors. Seventeen of the 23 variables show significant differences, after Bonferroni correction for multiple testing and adjustment for the geographic clustering of the CE cases. However, there is no concern about overlap: the characteristics of the CE respondents are well represented among the web respondents, indicating that the weak version of the overlap (or common support) assumption needed for the approach in this paper is met (Cunningham 2021, Section 5.1.4). To formally test this assumption, the propensity to be in the CE data set versus the web data set was predicted from the variables shown in Table 4 using a logistic model (Cunningham 2021, Section 5.4). The predicted propensities demonstrate that there is sufficient overlap between the two data sets: propensities range from 0.000148 to 0.979 in the web data set and from 0.0460 to 0.996 in the CE data set.

Imputation models can control for differences between the web and CE respondents in the variables in Table 4. More concerning are the differences that exist after controlling for these variables. The imputation approach described below rests on the assumption that there are no relevant differences between the CE and web survey respondents after adjusting for these characteristics. This assumption, called the conditional independence or unconfoundedness assumption in the causal inference literature, is common yet fundamentally untestable (Angrist and Pischke 2009, Section 3.2.1; and Wooldridge 2010, Section 21.3).

Table 4: Comparison of Demographics of CE and Web Respondents

		Mean/Proportion (SE)		Prob. of F Test[a]
Variable	Category	CE	Web	
HH members	NA[b]	2.43 (0.0084)	2.73 (0.0075)	< 0.0001
HU has pool	No	0.95 (0.0081)	0.80 (0.0085)	< 0.0001
	Yes	0.053 (0.0081)	0.20 (0.0085)	
HU has off-street parking	No	0.52 (0.020)	0.33 (0.010)	< 0.0001
	Yes	0.48 (0.020)	0.67 (0.010)	
HU has porch	No	0.47 (0.018)	0.21 (0.0087)	< 0.0001
	Yes	0.53 (0.018)	0.79 (0.0087)	
HU has apartment	No	0.99 (0.0028)	0.90 (0.0064)	< 0.0001
	Yes	0.01 (0.0028)	0.10 (0.0064)	
HU has central air	No	0.57 (0.022)	0.30 (0.0098)	< 0.0001
	Yes	0.43 (0.022)	0.70 (0.0098)	
HU has window air	No	0.88 (0.014)	0.69 (0.0099)	< 0.0001
	Yes	0.12 (0.014)	0.31 (0.0099)	
HU has solar panels	No	0.98 (0.0050)	0.93 (0.0053)	< 0.0001
	Yes	0.018 (0.0050)	0.065 (0.0053)	
HU single family	No	0.36 (0.018)	0.37 (0.010)	0.0007
	Yes	0.64 (0.018)	0.61 (0.010)	
	Missing	0 (NA)	0.016 (0.0027)	
No. bedrooms	1	0.12 (0.011)	0.096 (0.0063)	0.001
	2	0.25 (0.016)	0.24 (0.0091)	
	3	0.39 (0.016)	0.40 (0.011)	
	4	0.17 (0.012)	0.18 (0.0082)	
	5+	0.057 (0.0087)	0.053 (0.0048)	
	Missing	0.0054 (0.0023)	0.029 (0.0036)	

No. bathrooms	1	0.44 (0.020)	0.39 (0.010)	< 0.0001
	2	0.43 (0.020)	0.43 (0.010)	
	3+	0.12 (0.012)	0.14 (0.0074)	
	Missing	0.0043 (0.0021)	0.037 (0.004)	
R owns home	No	0.37 (0.016)	0.44 (0.011)	< 0.0001
	Yes	0.63 (0.016)	0.54 (0.011)	
	Missing	0 (NA)	0.023 (0.0032)	
R age	18-34	0.12 (0.012)	0.51 (0.011)	< 0.0001
	35-49	0.22 (0.013)	0.16 (0.0079)	
	50-64	0.24 (0.015)	0.16 (0.0077)	
	65+	0.38 (0.015)	0.15 (0.0076)	
	Missing	0.032 (0.0061)	0.20 (0.003)	
R White	No	0.22 (0.021)	0.19 (0.0083)	0.15
	Yes	0.77 (0.022)	0.80 (0.0086)	
	Missing	0.010 (0.0034)	0.017 (0.0028)	
R Black	No	0.89 (0.012)	0.85 (0.0077)	0.024
	Yes	0.10 (0.012)	0.14 (0.0073)	
	Missing	0.010 (0.0034)	0.017 (0.0028)	
R Asian	No	0.93 (0.016)	0.94 (0.0051)	0.24
	Yes	0.059 (0.016)	0.044 (0.0044)	
	Missing	0.010 (0.0034)	0.017 (0.0028)	
R Hispanic	No	0.88 (0.016)	0.90 (0.0063)	0.22
	Yes	0.11 (0.015)	0.085 (0.0059)	
	Missing	0.011 (0.0035)	0.011 (0.0022)	
R married	No	0.50 (0.018)	0.52 (0.011)	0.037
	Yes	0.50 (0.018)	0.47 (0.011)	
	Missing	0.0043 (0.0026)	0.015 (0.0026)	
R never married	No	0.79 (0.015)	0.58 (0.011)	< 0.0001
	Yes	0.20 (0.014)	0.41 (0.011)	
	Missing	0.0043 (0.0026)	0.015 (0.0026)	
R education	No HS degree	0.085 (0.0012)	0.043 (0.0043)	< 0.0001
	HS degree	0.24 (0.015)	0.21 (0.0087)	

	Some college	0.22 (0.013)	0.21 (0.0087)	
	Associate's degree	0.11 (0.0010)	0.013 (0.0071)	
	Bachelor's degree	0.21 (0.015)	0.27 (0.0095)	
	Master's or higher	0.12 (0.011)	0.13 (0.0071)	
	Missing	0.17 (0.0045)	0.0091 (0.002)	
HH wage income	No	0.30 (0.015)	0.30 (0.0097)	0.0009
	Yes	0.67 (0.016)	0.63 (0.0103)	
	Missing	0.027 (0.0085)	0.073 (0.0056)	
HH has self-employment income	No	0.87 (0.013)	0.81 (0.0083)	0.0104
	Yes	0.10 (0.011)	0.14 (0.0073)	
	Missing	0.027 (0.0085)	0.049 (0.0046)	
HH income	< \$30k	0.37 (0.018)	0.17 (0.0081)	< 0.0001
	\$30k-\$50k	0.086 (0.0087)	0.13 (0.0072)	
	\$50k-\$70k	0.058 (0.0073)	0.098 (0.0063)	
	\$70k-\$90k	0.083 (0.0090)	0.085 (0.0060)	
	\$90k-\$120k	0.066 (0.0081)	0.067 (0.0053)	
	≥ \$120k	0.17 (0.012)	0.10 (0.0065)	
	Missing	0.17 (0.016)	0.34 (0.0010)	

[a] *F* statistic for test of hypothesis that means/proportions not equal

[b] Continuous variable; all others categorical

F tests control for geographic clustering of CE cases

The household income variable was collected slightly differently in the two surveys. The CE collects income at the household member level. It first asks for the amount earned by each member. If the respondent does not give an answer, it asks for income in ranges: less than \$5,000; [\$5,000, \$10,000); [\$10,000, \$15,000); [\$15,000, \$20,000); [\$20,000, \$30,000); [\$30,000, \$40,000); [\$40,000, \$50,000); [\$50,000, \$70,000); [\$70,000, \$90,000); [\$90,000, \$120,000); \$120,000 and greater. The web survey used a different approach. Income is among the most sensitive questions asked in surveys (Tourangeau and Yan 2007). Without the motivation provided by an interviewer and the backing of an official government survey, the web survey was unlikely to be able to collect income for each household member. For these reasons, the web survey asked for household income in ranges (the same ranges given above). To make the income data collected in the two surveys comparable, the CE income data were aggregated to the household level. For those households where amounts were reported for each member (68% of households), household income is the sum of the reported amounts. For those households where ranges were reported for all members (5.9% of households), the minimum and maximum income was calculated for each member from the reported range. These two numbers were summed across all household members to capture the minimum and maximum possible household income. The household income was set to the mean of these two numbers. For those households where a mix of ranges and amounts were reported (8.8% of households), income for those members reported in ranges was set to the midpoint of the range and income was then summed across all household members. For the remaining CE households (17.1%), income was set to missing. For all three types of households, the resulting household income was then collapsed into larger ranges. The web survey respondents were more likely not to answer the income question (Table 4). The wage income and self-employment income indicators were also asked at the person level in the CE and at the household level in the web survey. When aggregating the CE data to the household level, a household was marked as having wage income (or self-employment) income if any member was reported to have that income type. These indicators differ significantly in the two sets of respondents but not meaningfully.

Figure 1 compares the percentage of CE respondents who answered with “yes” to each item’s filter question with the percentage among web respondents in the interleafed format. Thirteen of the 16 items show significantly different purchase rates (with Bonferroni correction and clustering adjustment). However, the correlation between the percentages is 0.93, indicating that the reported purchase rates are similar. Web respondents are more likely to be insured and less likely to have expenses for water and sewage and garbage and recycling. Although there are differences between the two groups of respondents, the broad overlap between them supports the imputation approach used below.

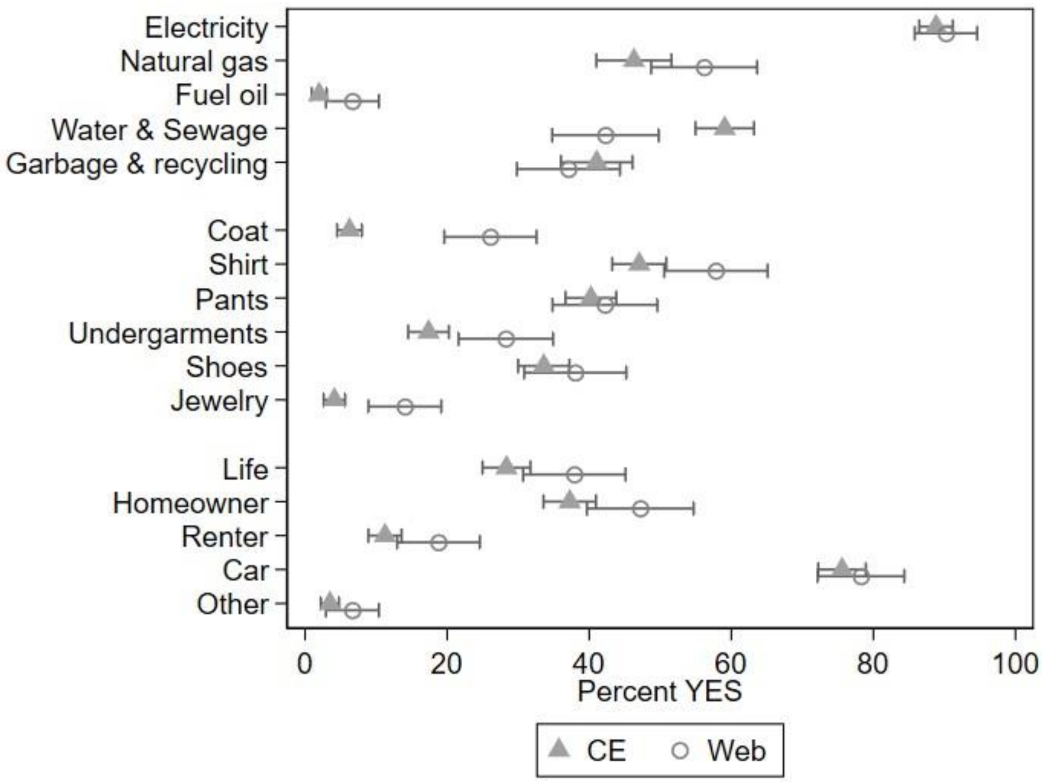

Figure 1: Comparison of Web and CE Respondents on Reported Purchases
NOTE: Only web survey respondents in interleafed format included

Imputation Models

The other factor influencing the accuracy of the imputed values is how well the models explain purchases of the 16 items. Twelve logistic imputation models were fit at the case-item level on all data from the CE and web surveys. In each model, the dependent variable is the filter response, “yes” or “no.” The independent variables in the models are the format (grouped or interleafed), the survey (CE or web), the item (1 through 16), the order of the sections (1 through 6; fixed for CE respondents), and the background variables shown in Table 4. Unfortunately, no information about the CE interviewers was available for inclusion in the models. The models do not account for the clustering of observations into primary sampling units in the CE or into respondents in the web survey. However, analysis of the imputed data does, as discussed in the next subsection.

Table 5 shows the interactions included in each model. Together these models cover all interactions likely to affect purchases. The full set of all interactions could not be fit in one model, because of sparse cells, so the models differ in the interactions they include. In each model in the table, all variables indicated with an “a” in a given column were interacted with each variable with a “b” in that column. When interactions were included in a model, main effects were also included. Variables indicated with an “x” were included without any interactions. Model 1 interacts the grouped (versus interleafed) indicator with each of the background variables to capture how respondent and household characteristics impact the difference between the two formats. Model 2 interacts the survey indicator (CE versus web) with the background variables, because the demographic makeup of the two surveys differs (Table 4). Model 3 contains no interactions. Model 4 interacts the grouped (versus interleafed) indicator with the item indicator, because the two formats may behave differently with some items. Model 5 includes the interaction of the survey indicator and the item, because the mode effect (interpreted broadly to include representation and measurement differences) may differ by item. And Model 6 interacts the item indicator with each of the background variables. Models 7–12 are identical to Models 1–6 but exclude household income because of the challenges in constructing comparable household income measures. As indicated in Table 5, all models include main effects for all variables. Table 6 shows that the models fit the data reasonably and about equally well. The best fitting model is Model 5, according to both the AIC and BIC measures.

Table 5: Specification of Logistic Imputation Models

Variable	Model											
	1	2	3	4	5	6	7	8	9	10	11	12
Grouped (vs. Interleafed)	a	x	x	a	x	x	a	x	x	a	x	x
CE (vs. Web)	x	a	x	x	a	x	x	a	x	x	a	x
Item (1-16)	x	x	x	b	b	a	x	x	x	b	b	a
Section Order	x	x	x	x	x	x	x	x	x	x	x	x
HH members*	b	b	x	x	x	b	b	b	x	x	x	b
HU has pool	b	b	x	x	x	b	b	b	x	x	x	b
HU has off-street parking	b	b	x	x	x	b	b	b	x	x	x	b
HU has apartment	b	b	x	x	x	b	b	b	x	x	x	b
HU has central air	b	b	x	x	x	b	b	b	x	x	x	b
HU has window air	b	b	x	x	x	b	b	b	x	x	x	b
HU has solar panels	b	b	x	x	x	b	b	b	x	x	x	b
HU single family	b	b	x	x	x	b	b	b	x	x	x	b
HU bedrooms	b	b	x	x	x	b	b	b	x	x	x	b
HU bathrooms	b	b	x	x	x	b	b	b	x	x	x	b
HH has wage income	b	b	x	x	x	b	b	b	x	x	x	b
HH has self-employment income	b	b	x	x	x	b	b	b	x	x	x	b
HH income	b	b	x	x	x	b						
R owns home	b	b	x	x	x	b	b	b	x	x	x	b
R age	b	b	x	x	x	b	b	b	x	x	x	b
R white	b	b	x	x	x	b	b	b	x	x	x	b
R African-American	b	b	x	x	x	b	b	b	x	x	x	b
R Asian	b	b	x	x	x	b	b	b	x	x	x	b
R married	b	b	x	x	x	b	b	b	x	x	x	b
R never married	b	b	x	x	x	b	b	b	x	x	x	b
R Hispanic	b	b	x	x	x	b	b	b	x	x	x	b
R education	b	b	x	x	x	b	b	b	x	x	x	b

NOTE: Case base for each model is 50,080 filter questions. “x” indicates variables included without any interactions. In each column, every variable shown with “a” is interacted with every variable shown with “b.”

* Continuous variable; all others categorical

Table 6: Measures of Fit of Imputation Models

Fit Measure	Model											
	1	2	3	4	5	6	7	8	9	10	11	12
AUC	0.798	0.798	0.796	0.798	0.801	0.798	0.797	0.797	0.795	0.8	0.801	.08
AIC	52446.6	52428.8	52569.7	52320.6	51991.7	52365.5	52518.5	52495.4	52631.3	52381.1	52052.1	52482.3
BIC	53434.6	53390.3	53151.9	53035.1	52706.2	53212.4	53400.7	53351.1	53160.6	53042.7	52713.1	53222.2
n	50080	50080	50080	50080	50080	50080	50080	50080	50080	50080	50080	50080

AUC = Area Under Receiver Operating Curve
AIC = Akaike's information criterion
BIC = Bayesian information criterion

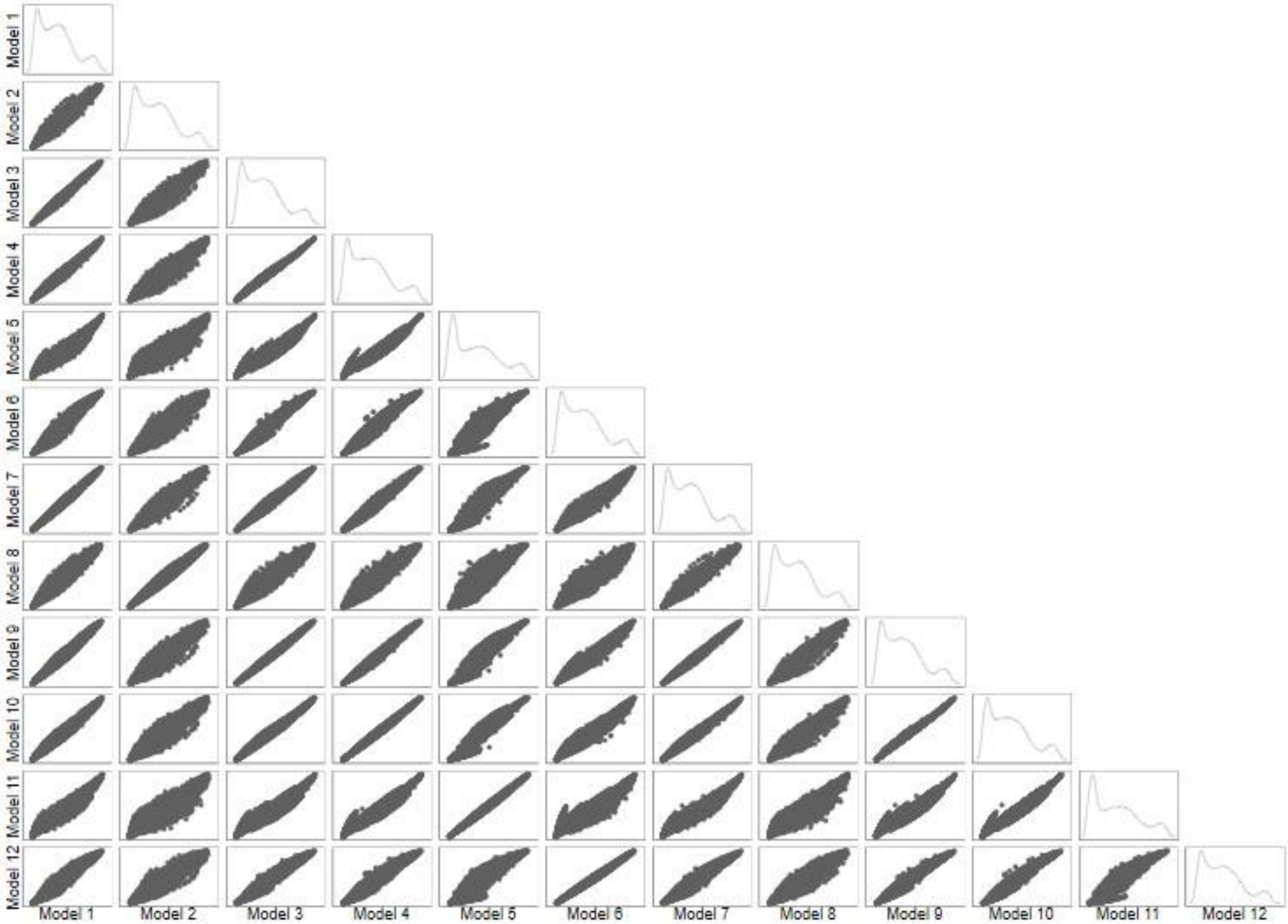

Figure 2: Comparison of Predicted Probabilities (0-100) from Imputation Models
NOTE: Only predictions for CE case-items in the grouped format shown (n=14,912 in each subplot)

The scatterplot matrix in Figure 2 compares the predicted probabilities from each model for the CE cases in the grouped format. The probabilities estimate how likely each case is to purchase each item. The horizontal axis of every subplot ranges from 0 to 100. The main diagonal of the matrix shows the density of the predicted probabilities from each model in a kernel density plot. The vertical axis of the subplots below the diagonal also ranges from 0 to 100. In the subplots below the diagonal, the wider the distribution of points, the more those two models disagree about the probability that a given case would report a purchase of a given item in the (hypothetical) grouped format. The pairwise correlations between the models' predication are all greater than 95%. The smallest correlation is between models 2 and 11.

Estimation of Motivated Misreporting Effect

Twenty responses to each filter question were imputed from each model representing how CE respondents would respond in the grouped format. The main outcome of interest is the motivated misreporting effect in the CE (*MM*), which is the difference between the percentage of filters answered with "yes" in the observed interleafed format $p^{interleafed}$ and the percentage of filters answered with "yes" in the imputed grouped format $p^{grouped}$:

$$MM_m = p^{interleafed} - p_m^{grouped} \quad (1)$$

The *m* subscript refers to the imputation model. Each model gives a different estimate of $p^{grouped}$ and a different motivated misreporting effect. The standard error on ${p^{grouped}}_m$ and on MM_m should account for the uncertainty in the imputations. If the format effect in Wave 1 of the CE is similar to that observed in previous studies, MM_m will be negative, indicating underreporting in the interleafed format. Using a specification curve (Simonsohn et al. 2019), the results section compares the estimates of MM_m across the 12 imputation models to understand how robust they are to model specification. Using 12 models, which vary in their predictions, and making multiple imputations from each one, captures the uncertainty both within and between models. The results section reports estimates from an analysis that combines the imputations from all models (240 imputations in total) to produce a thirteenth estimate of *MM*.

All models and analyses were run in Stata 15.1 (StataCorp LP 2017) and are unweighted. Analyses of the imputed data account for the clustering of the case-item observations within primary sampling units in the CE and within respondents in the web survey.

5 Results

The outcome of interest is the percent of filters answered with "yes" in the interleafed format minus the percent answered with "yes" in the grouped format (Equation 1). In the web survey, the grouped format collects more "yes" responses to the filter questions than the interleafed format does. Purchases are underreported by 6.1 percentage points in the interleafed format relative to the grouped format, across all sections and items (standard error 0.80; see the left side of Table 7). The motivated misreporting effect occurs in each of the three sections and varies from 9.4 percentage points in the clothing section to 3.0 percentage points in the insurance section. These results regarding the motivated misreporting effect are as expected from previous research (Kreuter et al. 2011; Eckman et al. 2014; Tourangeau et al. 2015; Bach and Eckman

2018). The order of the section also significantly impacted the probability of responding “yes” to the filter questions. The second and third sections garnered 3.9% and 5.5% fewer “yes” responses than the first section (all differences significant at 5% level). Because the order of the section was randomized in the web survey, these are true order effects and not topic effects.

In the CE, the 12 imputation models each provide a different set of responses to the filter questions in the grouped format and a different estimate of the motivated misreporting effect in Equation (1). Figure 3 shows the estimated motivated misreporting effects from each model, overall and for the three sections: utilities, clothing, and insurance. This specification curve captures the sensitivity of the results to the different models. In the overall subplot, 10 of the 12 models report a significant effect in the expected direction: the percentage of filter questions answered with a “yes” in the interleafed format is lower than it would be if the questions were asked in the grouped format. In the utilities section, the results have no clear sign, and half the models predict no significant difference in the grouped and interleafed format. In the CE, the order of the section is not recorded but most respondents answer the utility section before the clothing or insurance section. The results shown in the upper right subplot of Figure 3 may be the result of an order effect more than a topic effect. In the clothing section, all models report a significant motivated misreporting effect, and the size of the effect is largest in this section. Models 5, the best-fitting model according to Table 6, reports significant and negative effects overall and in each section.

The right side of Table 7 shows the percentage of filters answered with “yes” in each format, overall and by section in the CE, after imputation. The reported imputed estimates come from the combination of all 12 models and 20 imputation from each one, for 240 total imputations. In the interleafed format, 33.8% of the filter responses were “yes.” In the grouped format, 38.7% were imputed “yes.” The difference between the two formats is −4.9 percentage points (standard error −2.15), slightly smaller than the difference in the web survey (−6.1, standard error 0.80).

6 Discussion

Because experimental manipulation of the questions in the CE was not possible, this paper has used a different approach. Members of a nonprobability panel were randomly assigned to answer CE filter questions about household purchases in the grouped or interleafed format. Multiple imputation was used to understand how CE respondents would have answered in the grouped format. Most models show that respondents underreport purchases by approximately five percentage points to avoid the follow-up questions.

Table 7: Motivated Misreporting Effects in Web and CE

Section	Web			CE		
	Interleafed % Yes	Grouped % Yes	Difference % Points	Interleafed % Yes[a]	Grouped % Yes[ab]	Difference % Points[a]
Overall	36.8	42.9	-6.1*	33.8	38.7	-4.9*
	(0.53)	(0.60)	(-0.80)	(0.73)	(2.10)	(-2.15)
Utilities	40.5	45.7	-5.2*	47.4	46.6	0.82
	(0.75)	(0.84)	(-1.1)	(1.12)	(3.51)	(3.67)
Clothing	32.7	42.1	-9.4*	24.5	34.8	-10.2*
	(0.80)	(1.0)	(1.3)	(0.95)	(3.63)	(-3.71)
Insurance	37.9	40.9	-3.0*	31.3	35.7	-4.4*
	(0.66)	(0.74)	(-1.0)	(0.81)	(1.88)	(-1.97)

Standard errors in parentheses

* Difference significant at 5% level

[a] Standard errors adjust for clustering of observations

[b] Imputed from all models; standard errors adjust for multiple imputation

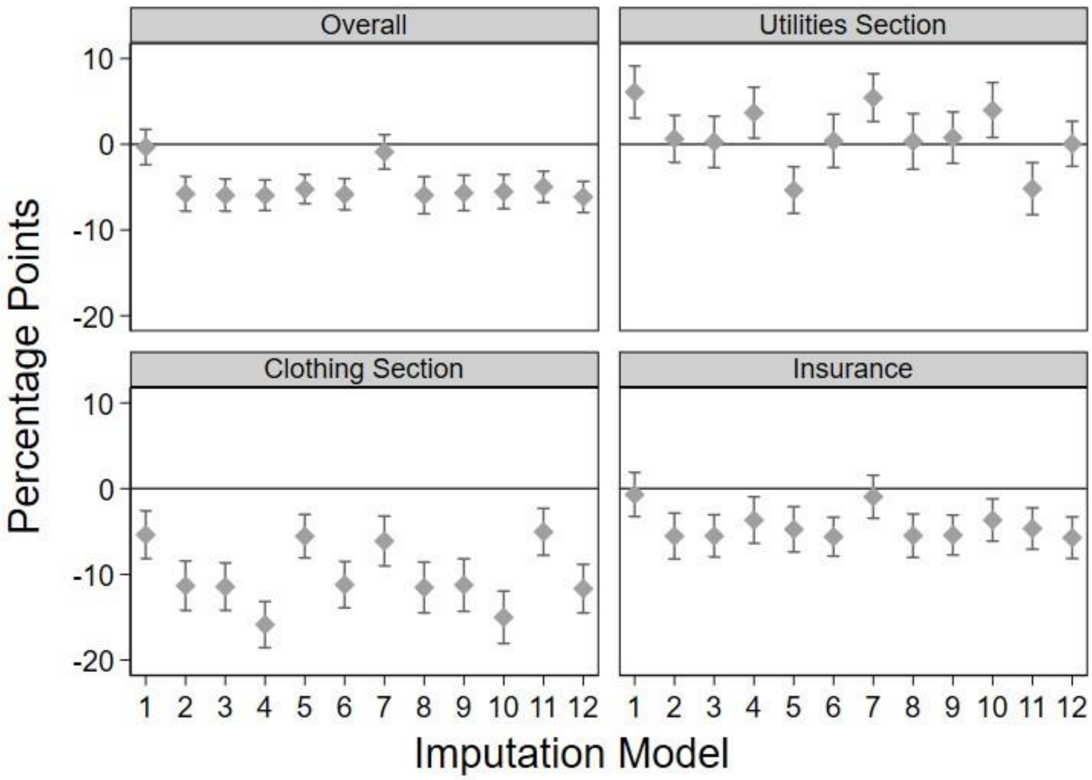

Figure 3: Estimates of Motivated Misreporting Effects across Imputation Models
Difference in "Yes" Percentages: Interleafed – Grouped (see Equation 1)
95% confidence intervals shown, adjusted for multiple imputation and clustering of observations

It is not possible to say what the effects of this motivated misreporting might be on the CPI. The CE data contribute to the weights used to construct the market basket for the CPI. Underreporting may affect the assigned weights. If the motivated misreporting effect is larger toward the end of the questionnaire (which this paper has not tested), and the rate of price increases on goods in those sections is higher or lower than the overall rate of price increases, then bias in the CPI is possible. The ultimate effect of motivated misreporting on the CPI is too complex to speculate on in this paper. However, the CE data are used for many other purposes by policy researchers and academics. Some of these other studies may be more vulnerable to bias because of motivated misreporting. Researchers should carefully think through how underreporting of purchases in the CE could bias their analyses.

To reduce the motivated misreporting effect in the CE, the survey could do several things. First, the CE could switch to a grouped by section or even a grouped-overall approach. In discussions, researchers at the Bureau of Labor Statistics are hesitant to make this change. Many respondents find the grouped approaches difficult because they require jumping from one item to the next in the filter questions and then going back to each purchased item: respondents prefer to think about items one at a time (Clark-Fobia et al. 2018). Another option would be to interview more respondents but ask fewer questions of each one to reduce the burden and length of the survey. Imputation could fill in the missing responses (Gonzalez and Eltinge 2008). This approach would likely increase data collection costs, however, and thus is also not ideal. Future research should focus on identifying techniques to minimize motivated misreporting in the interleafed format, which would benefit the CE and many other important surveys.

The results in this paper depend strongly on the imputation models. The models explain purchases rather well but not perfectly, as shown in Table 5. The models control for the observed differences between the web and CE respondents (Table 4 and Figure 1) but cannot control for unobserved differences. In addition, this study relies on a nonprobability survey; a more representative set of web respondents may result if different imputations and different conclusions. Income data were collected differently in the CE and the web survey. And no interviewer characteristics were available in the CE: controlling for the clustering of cases by interviewer may alter the results.

Nevertheless, these results are the strongest evidence to date that motivated misreporting is taking place in the first wave of the CE. Furthermore, the approach used here is one that other large surveys may find useful. When experimental manipulation in a given survey is not possible, because of budget or practical constraints, a parallel web survey offers a way to estimate how questionnaire changes would affect the collected data.

References

American Association for Public Opinion Research (AAPOR) (2016). AAPOR Guidance on Reporting Precision for Nonprobability Samples. https://www.aapor.org/getattachment/Education-Resources/For-Researchers/AAPOR_Guidance_Nonprob_Precision_ 042216.pdf.aspx Accessed 2020-11-20.

Angrist, J. and Pischke, J. (2009). *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton University Press.

Bach, R. L. and Eckman, S. (2018). Motivated Misreporting in Web Panels. *Journal of Survey Statistics and Methodology*, 6(3):418–430.

Bach, R. L. and Eckman, S. (2019). Rotation Group Bias in Reporting of Household Purchases in the U.S. Consumer Expenditure Survey. *Economics Letters*, 187.

Bach, R. L., Eckman, S., and Daikeler, J. (2019). Misreporting Among Reluctant Respondents. *Journal of Survey Statistics and Methodology*, 8(3):566–588. https://doi.org/10.1093/jssam/smz013

Bosley, J., Dashen, M., and Fox, J. (1999). When Should we Ask Follow-up Questions About Items in List? In *Proceedings of the Section on Survey Research Methods, American Statistical Association*. https://www.bls.gov/osmr/research-papers/1999/pdf/st990220.pdf.

Chen, S., Yang, S., and Kim, J. K. (2020). Nonparametric Mass Imputation for Data Integration. *Journal of Survey Statistics and Methodology*, pages 1–24.

Christensen, K. B., Feveille, H., and Kreiner, S. (2006). Adjusting for Mode of Administration Effect in Surveys using Mailed Questionnaire and Telephone Interview Data. Technical Report Research Report, Department of Biostatistics, University of Copenhagen, Denmark.

Clark-Fobia, A., Kephart, K., and Nelson, D. V. (2018). A Qualitative Study on the Effects of Grouped versus Interleafed Filter Questions. *Survey Practice*, 11(2).

Cunningham, S. (2021). Causal Inference: The Mixtape. Yale University Press.

Duan, N., Alegria, M., Canino, G., McGuire, T. G., and Takeuchi, D. (2007). Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats. *Health Services Research*, (42):890–907.

Eckman, S. and Kreuter, F. (2018). Misreporting to Looping Questions in Surveys: Recall, Motivation and Burden. *Survey Research Methods*, 12(1):59–74.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Presser, S., and Tourangeau, R. (2014). Assessing the Mechanisms of Misreporting to Filter Questions. *Public Opinion Quarterly*, 78(3):721–733.

Gonzalez, J. and Eltinge, J. (2008). Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, page 2081–2088. https://www.bls.gov/cex/cesrvymethsgonzale1.pdf.

Hillygus, D. S., Jackson, N., and Young, M. (2014). Professional Respondents in Nonprobability Online Panels. In *Online Panel Research*, pages 219–237. John Wiley & Sons, Ltd.

Hubener, E., Knappenberger, C., Lee, Y., and Tan, L. (2019). The 2018 CE Data Quality Profile. Technical report. https://www.bls.gov/cex/dqreport2018.pdf.

Kaplan, R. L. and Edgar, J. (2018). Priming Confidentiality Concerns: How Reminders of Privacy Affect Response Rates and Data Quality in Online Data Collection.

Kessler, R. C., Wittchen, H.-U., Abelson, J. A., McGonagle, K., Schwarz, N., Kendler, K. S., Knäuper, B., and Zhao, S. (1998). Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS). *International Journal of Methods in Psychiatric Research*, (7):33–55.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2020). Combining Non-probability and Probability Survey Samples Through Mass Imputation. https://arxiv.org/abs/1812.10694v3.

Kolenikov, S. and Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. *Journal of Survey Statistics and Methodology*, 2(2):126–158.

Kreuter, F., Eckman, S., and Tourangeau, R. (2019). Salience of Survey Burden and Its Effects on Response Behavior to Skip Questions. Experimental Results from Telephone and Web-Surveys. In Beatty, P., Collins, D., Kaye, L., Padilla, J., Willis, G., and Wilmot, A., editors, *Advances in Questionnaire Design, Development, Evaluation and Testing*. Wiley.

Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. (2011). The Effects of Asking Filter Questions in Interleafed versus Grouped Format. *Sociological Methods and Research*, 40(88):88–104.

McBride, B. (2013). Examining Changes in Filter Question (FQ) Reporting in the Consumer Expenditure Quarterly Interview Survey. In Proceedings of the Survey Research Methods Section of the American Statistical Association, pages 4304–4316.

National Research Council (2013). *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. The National Academies Press, Washington, D.C.

Park, S., Kim, J. K., and Park, S. (2016). An Imputation Approach for Handling MixedMode Surveys. *Annals of Applied Statistics*, 10(2):1063–1085.

Peytchev, A. (2012). Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quarterly*, 76(2):214–237.

Powers, J. R., Mishra, G., and Young, A. F. (2005). Differences in mail and telephone responses to self-rated health: use of multiple imputation in correcting for response bias. *Australian and New Zealand Journal of Public Health*, 29(2):149–154.

Rendall, M. S., Ghosh-Dastidar, B., Weden, M. M., Baker, E. H., and Nazarov, Z. (2013). Multiple Imputation for Combined-survey Estimation With Incomplete Regressors in One but Not Both Surveys. *Sociological Methods and Research*, 42(4):483–530.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.

Shields, J. and To, N. (2005). Learning To Say No: Conditioned Underreporting in an Expenditure Survey. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*. http://www.asasrms.org/Proceedings/y2005/files/JSM2005-000432.pdf.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2019). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. https://ssrn.com/abstract= 2694998 Accessed 2020-11-18.

StataCorp LP (2017). *Stata statistical software: Release 15*. StataCorp, College Station, TX.

Tourangeau, R., Kreuter, F., and Eckman, S. (2015). Motivated Misreporting: Shaping Answers to Reduce Survey Burden. In Engel, U., editor, *Survey Measurements. Techniques, Data Quality and Sources of Error*, pages 24–41. Campus, Frankfurt/New York.

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

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data. The MIT Press.

Yan, T. and Copeland, K. (2010). Panel Conditioning in the Consumer Expenditure Quarterly Interview Survey. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*. http://www.asasrms.org/Proceedings/y2010/Files/307812_59394.pdf.