Is the Public Sweet on Sugary Beverages? Social Desirability Bias and Sweetened Beverage Taxes

Melissa A. Knox[1]
Vanessa M. Oddo[2]
Lina Pinero Walkinshaw[3]
Jessica Jones-Smith[3]

September 15, 2019

This version: April 2, 2020

[1] Corresponding author. Department of Economics, University of Washington, Box 353330, Seattle, WA 98195. knoxm@uw.edu

[2] Department of Kinesiology and Nutrition, University of Illinois at Chicago

[3] Department of Health Services, University of Washington.

Abstract

Social desirability bias has been documented in self-reported diet as well as in voting behavior, but not in regards to sweetened beverage consumption or sweetened beverage taxes. We find evidence that respondents in a mixed-mode opinion survey exhibit social desirability bias in both reported sweetened beverage consumption and beliefs about the health and economic benefits of sweetened beverage taxes. We do so in a study of 1,704 adults residing in Seattle, Minneapolis, and the D.C. metro area. Phone respondents in our survey under-report sweetened beverage consumption by 0.63 beverages per week relative to web respondents (average web respondent consumption is 3.55 beverages per week). They also over-report their beliefs about the positive health and economic impacts of sweetened beverage taxes by 0.54 points in an 18-point index (average web respondent index score is 2.79). These differences are measured after we control for selection into survey mode by using matching methods, and we interpret them as occurring due to social desirability bias. In contrast to these findings, there is no modal difference in respondents' stated approval of sweetened beverage taxes, and so we conclude that this question is not subject to social desirability bias.

1. Introduction

The recent introduction of sugar sweetened beverage taxes as health-promoting policies has created a demand for information about the impacts of these taxes on the consumption of sugar sweetened beverages (henceforth shortened to sweetened beverages). A flurry of recent papers have found that these taxes do, in fact, raise beverage prices and reduce purchases in affected areas, a necessary intermediate step in reducing consumption (see Teng *et al.*(2019) for a survey). However, with the potential for consumers to cross borders in order to purchase sweetened beverages outside of the taxed zone (Cawley *et al.*, 2019), inferring sweetened beverage consumption from beverage purchases inside the taxed area may overestimate the consumption reduction due to taxes.

To address this issue, researchers have also used self-reported sweetened beverage consumption, measured either with bespoke surveys or large-scale surveys such as the National Health and Nutrition Examination Survey (NHANES), to evaluate the price responsiveness of consumers to sweetened beverage taxes (Falbe *et al.*, 2016; Zhong *et al.*, 2018; M. M. Lee *et al.*, 2019). These data may also be important for detecting the extent to which consumers avoid these taxes by shopping outside the taxed region. Accurate reports of sweetened beverage consumption are additionally useful for estimating the distributional impacts of sweetened beverage taxes, and optimally designing future tax policies (O'Donoghue and Rabin, 2006).

However, recent public attention around the health effects of sugar sweetened beverage consumption, as well as the growing use of sweetened beverage taxes to curb their purchase, suggests that social norms regarding the acceptability of consuming these beverages may be undergoing a shift (Tamir *et al.*, 2018). As such, self-reported sweetened beverage consumption is increasingly prone to social desirability bias, a form of bias that arises from under-reporting opinions, habits, or behaviors that are in contrast with prevailing social norms. In particular, a study by Klesges and coauthors finds that pre-teen girls' self-reported sweetened beverage consumption is negatively associated with an index of compliance with social norms (Klesges *et al.*, 2004). Their findings suggest that girls who are more compliant report about 10% fewer beverages consumed than the average, and are consistent with previous studies that have found that women tend to understate their daily caloric intake by about 1% due to social desirability bias (Herbert et al. 1997).

In addition to being found in self-reported dietary intake, social desirability bias has previously been documented in surveys about such sensitive topics as voting behavior, weight, and sexual practices (Schläpfer, Roschewitz and Hanley, 2004; Burkill *et al.*, 2016; Jones *et al.*, 2016; Burke and Carman, 2017). This bias appears to be meaningful in a policy sense, but to varying degrees. For example, Schläpfer and co-authors find that survey respondents overstate their willingness to pay for public goods by 10 to 20 times (2004). On the other hand, Burkill and co-authors only find that about 10% of their survey questions about sexual attitudes and behavior are subject to social desirability bias (2016).

The methodology used to identify social desirability bias is also quite varied. Identification requires benchmarking survey responses against a "true" value for the variable, which is likely to depend on respondent characteristics. Some studies take the approach of modeling true values using known information about the relationship between respondent characteristics and these true values. This is the approach taken by Burke and Carman in detecting social desirability bias in reported weight, and is also similar to the method used by Klesge (Klesges *et al.*, 2004; Burke and Carman, 2017). An alternative approach for detecting social desirability bias is to use mixed mode surveys that collect some responses on the web and some over the phone or in person and comparing responses across modes. In these studies, the web responses are taken to be the true response and the phone or in-person responses taken to be the responses contaminated by social desirability bias. Because there is frequently self-selection into survey mode based on the same characteristics that influence survey response, these studies must first correct for this confounding before estimating bias. They do so either by randomizing respondents into mode, or by using propensity scores or other matching methods to remove the influence of selection on observable characteristics on responses (Duffy *et al.*, 2005; Kreuter, Presser and Tourangeau, 2008; Vannieuwenhuyze, Loosveldt and Molenberghs, 2010). Once selection is accounted for in one of these ways, the remaining modal difference in survey responses is interpreted as social desirability bias.

This study follows the second strategy to explore the existence and extent of social desirability bias in self-reported sweetened beverage consumption, attitudes regarding the health and economic effects of sweetened beverage taxes, and overall approval of such taxes. We do so using a mixed mode (phone and web) survey of adult respondents in four U.S. cities. We use matching methods to account for selection into mode on observed demographic characteristics, and find that modal response differences remain for sweetened beverage consumption and attitudes toward these taxes, but not for overall tax approval. These modal response differences are consistent with social desirability bias, and we interpret them as evidence of bias in the first two outcomes. Our findings contribute to the literature on sweetened beverage tax impacts and extend the findings of Klesges (2004), by showing evidence of social desirability bias in reported sweetened beverage consumption among adults of both sexes and across race, ethnicity, and income lines. Our study is also the first to find evidence of bias in self-reported views on the potential health and economic benefits of sweetened beverage taxes.

The existence of social desirability bias in mixed mode surveys measuring dietary intake of sweetened beverages raises questions about the reliability of any self-reported measures related to sweetened beverages. If phone respondents feel social pressure to alter their responses when speaking to another person on the phone, they almost certainly feel a similar pressure when responding to surveys in-person, and we cannot reject the possibility that they also feel this pressure when responding on the web. Even surveys that report beverage purchases based on products or receipts scanned in the home may be called into question, if the social undesirability of sweetened beverage consumption is strong enough to cause households to neglect to record some purchases.

The paper proceeds as follows: In Section 2, we formally model modal effects on survey responses as a combination of effects from selection by different respondent types into different survey modes and effects due to the mode itself, which we interpret as social desirability bias. In Section 3, we describe our survey, and define our key variables. Section 4 describes the methods we use, Section 5 describes our results and sensitivity analysis, and Section 6 discusses our results, study limitations, and conclusions.

2. Mixed Mode Surveys

2.1 Selection Effects and Social Desirability Bias

Mixed mode surveys (i.e. surveys utilizing multiple modes including in-person, phone, and internet) are valuable because they collect data from a broader range of respondents, while being cheaper and faster than phone or in-person only surveys, so more responses can be collected (Vannieuwenhuyze and Loosveldt, no date). Many evaluations of sweetened beverage taxes have had to enroll and collect data from participants on an accelerated schedule, often racing against the clock of the impending tax implementation. For this reason, some of these studies, including our own (Oddo *et al.*, 2019) have employed mixed mode surveys

However, the advantage of mixed mode surveys– reaching different segments of the population through different modes– may also create difficulty in interpreting their results. Web survey respondents are often found to be a selected, non-representative group, although there is increasingly selection bias in phone respondents as technology use shifts away from landlines (Dal Grande et al. 2016, Schonlau 2009). Web survey populations are frequently selected on income and age, two characteristics that have been found to influence sweetened beverage consumption (Bleich *et al.*, 2009), and so conclusions about sweetened beverage consumption-related behavior based on web-only or mixed-mode surveys may suffer from selection bias.

This selection into mode also makes it problematic to assess whether differences in responses across modes are due to selection or social desirability bias, by simply comparing responses across survey modes. Instead, social desirability bias is confounded with selection bias in the measured modal difference (Grewenig *et al.*, 2018).

A handful of recent studies have tried to tease out social desirability bias (or other) from selection bias by randomly assigning respondents into phone and web modes. Generally, these studies find reduced selection by mode, but still find that there are differences in responses by mode for questions about sensitive subjects including health behaviors, sexual practices, and study habits, although not for less sensitive topics. The direction of these effects overwhelmingly suggest that phone and in-person responses are subject to social desirability bias and not other measurement effects, such as non-response (Parks, Pardi and Bradizza, 2006; Woo, Kim and Couper, 2015; H. Lee *et al.*, 2019). Most multi-modal surveys are not able to follow random assignment, however, and so researchers have explored propensity score matching as a means to creating a sample of survey respondents that can be matched across modes on their observable characteristics (Vannieuwenhuyze and Loosveldt, no date; Duffy *et al.*, 2005). These methods can be used both for measuring social desirability bias and for correcting for selection bias, when inferring population values from mixed mode survey responses.

2.2 Formally Decomposing Mode Effects in Mixed Mode Surveys

To formally model the effect of survey mode on response in a two-mode survey, without randomization into mode, we modify the analysis of Vannieuwenhuyze and co-authors (2010) as follows. The available response modes (A) are denoted a and b, and respondents select into response mode according to a subset of their individual characteristics (G), which we will also refer to as a respondent's type. Therefore, G can also take on two values, G_a and G_b, that represent the types that select into mode a and mode b, respectively. An individual's survey responses also depend on their personal characteristics, X, of which G is a subset, and their response is f(X).

The observed mode effect in a two-mode survey is the difference between the average response in mode *a* of those who are the type to respond in mode *a*, and the average response in mode *b* of those who are the type to respond in mode *b*:

$$D = f(X|A = a, G = G_a) - f(X|A = b, G = G_b)$$

This difference has its origin in two sources, as described above. First, the respondents in each mode have, on average, different personal characteristics, X, and so we would expect f(X) to be different for these two populations. This is the effect of selection into mode on the difference measured in D, even if we were able to observe their responses in the same mode. Second, the mode itself may affect the response, so even if we were able to observe the same person's responses in both modes, then we would find that those responses to be different, on average, even though the Xs are the same. Equation 1 shows that *D* can be decomposed into variation from these two sources. It defines the mode effect as the sum of a measurement effect (*M*) and a selection effect (*S*).

Our goal in this paper is to separate out these two contributions to the mode effect and identify the measurement effect alone. Although there are potentially other sources of *M*, our claim, as outlined in this paper, is that *M* can be interpreted as social desirability bias, and this occurs because respondents answering survey questions on the phone feel pressured by the presence of the interviewer to alter their responses, to be consistent with social norms. We further claim that an individual's web survey response is less likely to be affected by this pressure, and web responses represent something close to the true values for the measured outcomes.[4] We formally define *M* in Equation 2 to be consistent with these claims, by defining it to be the difference between the average phone response for phone respondents and the average web response for these same respondents.

Since we are treating the web mode as the mode with no social desirability bias, we will henceforth treat the web response as the "true" value in the population, conditional on individual characteristics. We will define the measurement effect and the selection effect according to this assumption, and call the web mode *b* and the phone mode *a*. First, the measurement effect is defined as the difference between how a type G_a person answers in mode *a* (phone mode) and how a person of the same type would have answered in mode *b* (web mode), if we were able to observe this counterfactual. This definition is shown in equation 2.

The other component of the mode effect given in equation 1 is *S*, the selection effect, given in equation 3. This is the difference between the response of a type *a* responder in mode *a* and how a type *b* responder would have responded in mode *a*. In true random assignment of respondents into mode, S would be zero because type a responders and type *b* responders would be randomly assigned to modes and so this difference would be zero, on average. In this case, any mode effect measured would be entirely due to the measurement effect or social desirability bias.

[4] It is possible that web responses also suffer from social desirability bias. If they do, then our results would be an underestimate of the extent of social desirability bias present. Since we are unable to detect social desirability bias in web responses with our study, we focus on the simplest case of no social desirability bias on the web.

$$D = f(X|A = a, G = G_a) - f(X|A = b, G = b) = MM_a\big(f(x)\big) + S(f(X)) \quad (1)$$

Where

$$M = M_a\big(f(X)\big) = f(X|A = a, G = G_a) - f(X|A = b, G = G_a) \quad (2)$$

And

$$S = S_a\big(f(X)\big) = f(X|A = a, G = G_a) - f(X|A = a, G = G_b) \quad (3)$$

But we could alternatively use the definitions:

$$M = M_b\big(f(X)\big) = f(X|A = b, G = G_b) - f(X|A = a, G = G_b) \quad (4)$$

And

$$S = S_b\big(f(X)\big) = f(X|A = b, G = G_b) - f(X|A = b, G = G_a) \quad (5)$$

Where we would define the measurement effect as the difference between how a web responder “type” would respond on the web versus how they would respond on the phone, and the selection effect as the difference between how web responder types respond on the web and how phone responder types would respond on the web.

In either set of definitions, we are unable to observe the counterfactual terms. In other words, we cannot observe how people of one mode type respond in the other mode, and so cannot directly measure either M or S. Instead, we use linear regression and matching methods to try to reduce S down to zero, so that we can interpret any remaining mode effect as the measurement effect or social desirability bias.

3. Data Collection and Sample

3.1 Survey Design

In 2017, the City of Seattle City passed an ordinance imposing a tax on distributing sweetened beverages in Seattle, which went into effect on January 1, 2018. In order to better understand norms and attitudes around sweetened beverage taxes, we designed a survey to examine the public’s perceptions about the tax itself, self-reported consumption of sweetened beverages, and views on the possible health and economic impacts of the tax, both in Seattle and a demographically similar comparison area (Minneapolis, MN and the combined region of Rockville and Bethesda, MD and Arlington, VA, henceforth referred to as D.C. metro). Eligible participants in Seattle and the comparison area included adults (aged 18 years and older) who answered the screener questions on household income and race/ethnicity, and who spoke or read English or Spanish or read Vietnamese.

Seattle participants (N=851) were recruited prior to the implementation of the tax (October – December 2017). In the comparison area (N=863), participants were recruited between December 2017 and January 2018. In Seattle, we asked participants about the tax that was about to be implemented. In the comparison area, participants were asked about sweetened beverage taxes more generally. The survey was administered online and via the telephone, with the assistance of a professional survey research firm, Ironwood Insights, LLC.

3.2 Variable Definitions

Details of the questionnaire have been described by Oddo et al (2019). Briefly, we asked participants to report their beliefs around the tax and its economic and health impacts, using a 4-category Likert scale. The response options included strongly approve (strongly agree), somewhat approve (somewhat agree), somewhat disapprove (somewhat disagree), and strongly disapprove (strongly disagree). In addition, for some questions, participants were read two statements and asked to indicate if the first or second statement was "much closer" or "somewhat closer" to their own attitudes. Participants could also respond that they “don’t know.” We collapsed these responses into 3-category variables (e.g. approve, disapprove, or don’t know). Individuals who refused to provide a response were excluded from the analysis.

Demographic characteristics were collected among all participants. Participants reported their education level (some high school, completed high school, some college or vocational training, completed college or university, or completed graduate or professional degree), gender (male, female, self-identify), age (18–30 years old, 31–40 years old, 41–50 years old, 51–64 years old, ≥ 65 years), annual household income (<$30,000, $30,000–$59,999, $60,000–$89,999, $90,000–$120,000, >$120,000), marital status (married, widowed/divorced/separated, single, living with partner) and political party identification (Democrat, Republican, Independent, Other).

Race and ethnicity were asked as separate questions. Individuals were then categorized as: people who are non-Hispanic white, people who are non-Hispanic Black, people who are non-Hispanic Asian, people who are non-Hispanic of “other” races, and people who are Hispanic. We categorized Native Hawaiian and Pacific Islanders, American Indian and Alaska Natives, and those reporting two or more races as non-Hispanic of an “other” race. We defined low-income as household income below < 260% (federal poverty level [FPL]) and high-income as ≥ 260% FPL based on their self-reported total annual household income and given household size.

Participants were asked about their consumption of sweetened beverages during the prior 30 days, using a modified version of the NHANES Dietary Screener Questionnaire (none or < 1 per week, 1 per week, 2–6 per week, 1 per day, ≥ 2 day, don’t know). Total weekly consumption was then calculated by assigning the median value of each consumption category to the respondents.

Our survey also included a number of questions gauging respondents’ perceptions of the healthfulness of sweetened beverages, and the potential health and economic consequences of sweetened beverage taxes, as well as whether or not the respondent approved of the tax. We include the binary tax approval variable as one of the outcome variables in our study, following Oddo *et al.*, (2019). We also follow this previous work and combine responses to the questions that specifically address the health and economic effects of taxes into a summary score, in order to better capture overall perceptions around the possible health and economic impacts of the tax, and to reduce problems arising from multiple inference. This score, which we will refer to as the tax impacts score, is comprised of responses to nine questions about participants’ attitudes toward the impacts of sweetened beverage taxes on: child well-being, public health, cross-border shopping, small businesses, the economy, job loss, family finances, vulnerable populations, and on autonomy over beverage choice. A participant received a − 1 if they perceived that the impact of the tax would be negative (e.g., tax will not improve child well-being) and a

1 if they believed that the impact of the tax would be positive (e.g., tax will improve child well-being). If they responded that they “don’t know,” they received a 0 for that question. Scores ranged from − 9 to + 9 (making the full scale 18 points), with a higher score interpreted to mean that the impacts of the sweetened beverage tax in Seattle or sweetened beverage taxes more generally were perceived as more positive. The exact questions used in forming the tax impact score are listed in Supplemental Table 1 of the paper.

3.3 Descriptive Results

Table 1 shows the differences in both the demographic characteristics and responses to key survey questions by response mode. Even with our attempts to balance respondent characteristics across modes, some differences in mode of response remained. The top panel shows that web responders are more likely to be in the high-income category (with family income versus ≥ 260% of the FPL for their household size) and are younger (< 50 years old). Among non-Hispanic Blacks and among non-Hispanic Asians, respondents were more likely to respond on the web (versus the phone). On the contrary, among Hispanics, a higher prevalence responded to the survey via phone (versus web). There were not differences in mode of response among non-Hispanic Whites. Web respondents were also less likely to report that they identify themselves as Democrats and more likely to report that they identify themselves as Republicans or Other, but equally likely to report that they are politically independent, when compared to phone responders.

Table 1: Summary Statistics for Survey Sample by Response Mode

	Mean All (N=1,714)	Mean Phone (N=703)	Mean Web (N=1,011)	Difference	P-Value
Weekly Sweetened Beverage Consumption	3.11	2.49	3.55	-1.07	0.00
Approves of Tax	59%	60%	57%	3%	0.24
Tax Impact Score	2.37	2.79	2.08	0.72	0.00
Lives in Seattle	50%	60%	43%	17%	0.00
High Income (≥ 260% FPL)	53%	49%	56%	-7%	0.00
Non-Hispanic White	71%	73%	70%	3%	0.13
Non-Hispanic Black	9%	8%	10%	-2%	0.10
Non-Hispanic Asian	9%	4%	12%	-8%	0.00
Hispanic	11%	13%	9%	3%	0.03
50 or Younger	54%	38%	65%	27%	0.00
Some College or Below	40%	42%	38%	4%	0.14
Completed College or Above	60%	58%	62%	4%	0.14
Married or Partnered	49%	48%	51%	3%	0.18
Democrat	48%	51%	46%	6%	0.02
Independent	28%	28%	28%	0%	0.83
Republican or Other	24%	21%	26%	-5%	0.02

The outcome variables that we analyze for social desirability bias in this study are the first three variables in Table 1. We find that reported weekly sweetened beverage consumption (2.49 drinks per week versus 3.55) and the tax impact score (2.79 mean impact score versus 2.08) significantly differ between phone and web respondents, but that the modal difference in tax approval is small and not significant (60% approval on the phone versus 57% on the web). Since we expect actual sweetened beverage consumption to vary by race/ethnicity and income, and have reason to believe that tax approval and beliefs around the impact of the tax would also vary by these characteristics, these raw differences are not surprising, but they are also consistent with phone responses being skewed toward social norms. Given the attention to the potential for sweetened beverage taxes to be regressive, it is perhaps not as likely that stating approval for the tax would succumb to social desirability bias, as suggested by the lack of modal differences to responses to this question.

To explore the ways in which social desirability bias and selection bias might both be contributing to the modal differences in Table 1, Table 2 shows the three outcome variables of interest stratified by income (< 260% of the FPL versus ≥ 260% FPL), both for the full sample and for each of the modes separately. We choose to analyze by income and mode here (rather than by some other demographic characteristic and mode), because our study was designed to be adequately powered for this comparison.

Full Sample. This is a replication of the averages by mode presented in Table 1. As we show in Table 1, there are significant differences in two of these three outcomes by mode, with phone respondents reporting lower sweetened beverage consumption (2.49 beverages per week instead of 3.55), and more positive tax impact scores (2.79 out of 18 instead of 2.08 out of 18).

Low-Income Sample. The modal difference in reported sweetened beverage for the low-income group is almost twice as high as that for the full sample, with all incomes combined. There is no significant difference in tax approval across modes for the low-income group, and the difference is similar to that found for the full sample. The tax impact score, on the other hand, shows larger modal differences (1.18 points for the low-income group versus 0.64 points for the full sample).

High-Income Sample. The high-income sample shows approximately the same modal difference as the low-income group, although on average, they report their consumption to be lower overall. The average modal difference in tax approval and the tax impact score is smaller than for the low-income group, however, and neither difference is statistically significant.

This decomposition shows that there are large differences in responses between modes that are consistent with social desirability bias, and that those differences are roughly similar within high- and low-income groups. The end result is that looking only at the difference in responses across income group or mode will not give the full picture of either social desirability bias or the effect of characteristics such as income on the outcomes in our study. Because selection into mode occurs based on a number of characteristics (i.e. more than just income), what this decomposition ultimately shows is that we will need to use selection correction techniques that are able to account for a range of respondent characteristics. We explain those techniques in the next section.

Table 2: Mean Outcome Variables in Survey by Mode and Income Group

	Phone Mean (N=703)	Web Mean (N=1,101)	Difference	P-Value
Weekly Sweetened Beverage Consumption	2.49	3.55	-1.07	0.00
Agrees with Tax	0.60	0.57	0.03	0.24
Tax Impact Score	2.79	2.08	0.72	0.00
Low-Income Respondents (N=805)	Phone Mean (N=359)	Web Mean (N=446)	Difference	P-Value
Weekly Sweetened Beverage Consumption	2.82	3.93	-1.11	0.00
Agrees with Tax	0.59	0.54	0.05	0.16
Tax Impact Score	2.69	1.50	1.18	0.00
High-Income Respondents (N=909)	Phone Mean (N=344)	Web Mean (N=565)	Difference	P-Value
Weekly Sweetened Beverage Consumption	2.13	3.25	-1.12	0.00
Agrees with Tax	0.62	0.60	0.02	0.65
Tax Impact Score	2.91	2.53	0.38	0.25

4. Methods

We attempt to control for selection into mode and reduce selection bias into our results using two techniques. First, we employ linear regression to control for the effect of covariates such as income, age, race/ethnicity, political affiliation, and any other characteristic we think might contribute both to the response mode and the outcomes._This is the simplest way of measuring measurement effects if we believe that controlling for observable differences between respondents in both modes is sufficient for eliminating selection bias, as discussed in Section 2. This method also has the advantage of showing the relationship between survey responses and respondent characteristics, which may also be of interest to policy makers. Estimating treatment effects using linear regression can create bias, however, even in the absence of unobserved selection since linear regression both assumes a functional form for the relationship between covariates and treatment status, and it extrapolates treatment status into statistical space, where there may not be common support if the distribution of covariates differs across survey modes (Imbens, 2015).

We look at covariate balance in Column 1 of Table 3 by displaying the standardized difference in means and the ratio of the variance for our covariates of interest before any balancing. The difference in means is the mean for phone respondents subtracted from the mean for web respondents. This quantity is then standardized by dividing by a combination of the standard deviations in both populations. There are no test statistics to guide the choice of a meaningful standardized difference in means, but values above 0.10 are typically considered large. The variance ratio shows the similarity of the variance of the covariates between the two groups, and provides us with additional information about the distribution of the covariates in both samples. Variance ratios closer to 1 imply samples that are more closely matched in the second moment of their distributions (Austin, 2009). The unmatched sample in Column 1 appears to be mostly unbalanced in sweetened beverage consumption, Seattle residence, identifying as non-Hispanic Asian, and age. By design, it is relatively balanced across income and race/ethnicity, with non-Hispanic Asians being the exception. Due to the lack of balance in some covariates shown in Table 3, as well as the limitations of trying to predict treatment with only a linear combination of covariates, we employ covariate matching methods in estimating the effect of phone survey mode on responses. The procedure employed in matching methods can be described using the notation presented in Section 2, and in Imbens (2015). Matching allows us to create a pseudo-counterfactual group that stands in for the second term in Equation 2 above.

This is achieved by finding an individual or group of individuals in the web response group who match each individual in the phone response group. The match's response to key outcome variables is treated as representative of how the phone responders would have responded, had they been given the survey on the web. Additionally, we report the average effect of treatment in the entire sample population, which means that we are performing the same procedure in reverse – starting with web respondents, finding phone respondents who match with them on observable characteristics, and measuring the gap between the web respondent's response and matched phone respondent(s) response (similar to the measurement effect as defined in Equation 4 above).

With some assumptions, as described in Imbens (2015), we can use the stand-in for the counterfactual for each individual to produce the average treatment effect of the phone mode on survey response

$$\tau = E[f(X|A=a) - f(X|A=b)] \qquad (6)$$

The main assumption is that of unconfoundedness between mode and response, conditional on observable characteristics. In other words, in order for equation 6 to measure the true effect of mode on response net of selection into mode, we must assume that there are no characteristics of individuals that influence their mode and their response. This is a strong assumption, and one that is a common source of debate in the causal effects literature. In our case, it essentially comes down to the assumption that there are no intrinsic differences between phone and web respondents with similar demographic characteristics and that it is only the act of responding on either the phone or the web that creates differences in their survey responses. Another way of stating this assumption is that the only differences between the response "types" Ga and Gb are observable and that these differences go away when we compare people with similar characteristics who respond in different modes. Aside from observing the actual counterfactual, which is impossible, the only way we can be certain that unconfoundedness holds is if we randomly assign respondents into mode. Without randomization in the current study, then, we can only claim that our results are supportive evidence for the existence of measurement effects consistent with social desirability bias. As we will discuss further in the limitations of our paper, although we cannot reject intrinsic differences between at least some phone and web respondents, we also have no reason to believe that these intrinsic differences would in any way be correlated with increased consumption of sweetened beverages.

To perform matching, we employ three different methods, and the procedure each uses to find a match is briefly described here. The first is propensity score matching (PSM), which is commonly used in the economics and health literature, but only matches treated and control groups on a single metric formed from a combination of covariates (Rosenbaum and Rubin, 1984). The second is nearest neighbor matching (NN Match) on all covariates with replacement and using the Mahalanobis distance, as suggested by Abadie and Imbens (Abadie *et al.*, 2004; Abadie and Imbens, 2006). This method allows closer matching on all covariates and reduces the likelihood that they will be matched on only a few covariates dominating the propensity score. Finally, we estimate treatment effects using inverse probability weighted regression adjustment (IPWRA). This technique estimates the probability of treatment based on covariates, and then uses the inverse of this probability to weight regression coefficients in a linear regression of the outcome on treatment and other covariates. This method is considered “doubly robust” in that it will give estimate the treatment effect correctly even if the model predicting either treatment or the outcome is incorrectly specified (Cattaneo, 2010; StataCorp, 2019).

Table 3 shows the covariate balance achieved with each of the three methods. While there are no test statistics available to compare covariate balance, the IPWRA method appears to have the smallest standardized differences across modes, while the superiority of PSM over NN Match varies by covariate. For this reason, and the doubly robust property of IPWRA, we conclude that IPWRA produces the closest estimate of the “true” treatment effect.

Table 3: Covariate Balance before Matching and Using Three Matching Methods

	1		2		3		4	
	Unbalanced Covariates		Propensity Score Matching		Nearest Neighbor Matching		Inverse Probability Weighted Regression Adjusted	
	Standardized Difference in Means	Variance Ratio	Standardized Difference in Means	Variance Ratio	Standardized Difference in Means	Variance Ratio	Standardized Difference in Means	Variance Ratio
Weekly Sweetened Beverage Consumption	-0.27	0.75	0.10	1.33	-0.18	0.76	0.09	1.38
Lives in Seattle	0.36	0.98	-0.02	1.00	0.07	1.00	0.00	1.00
High Income (≥ 260% FPL)	-0.16	1.02	0.02	1.00	-0.01	1.00	-0.01	1.00
Non-Hispanic White	0.05	0.95	0.08	0.91	0.05	0.95	0.06	0.94
Non-Hispanic Black	-0.05	0.86	-0.04	0.87	-0.02	0.93	-0.06	0.83
Non-Hispanic Asian	-0.30	0.37	-0.11	0.69	-0.05	0.85	-0.06	0.83
Hispanic	0.09	1.25	0.05	1.13	0.02	1.05	0.02	1.06
50 or Younger	-0.54	1.03	-0.04	1.00	-0.07	1.01	-0.01	1.00
Completed College or Above	-0.09	1.04	0.00	1.00	0.03	0.98	0.00	1.00
Married or Partnered	-0.07	1.00	-0.03	1.00	0.00	1.00	-0.02	1.00
Democrat	0.12	1.00	-0.04	0.99	0.00	1.00	-0.02	1.00
Independent	-0.02	0.98	0.02	1.02	-0.01	0.99	0.03	1.03

Note: Standardized difference in means is phone mean response minus web mean response, divided by the standard deviation of the outcomes. Variance ratio is the ratio of the variances in each group. For columns 2-4, the full set of matching variables includes all variables shown, interactions between all variables including squared terms, and interactions between all interactions, including quadrupled terms. We also employ matching on all variables except consumption for the consumption outcome, but these results are not shown. Results are not materially changed in that specification.

5. Results

5.1 Linear Regression Results

The results of a linear regression of our three outcomes on survey mode and relevant covariates are shown in Table 4. We find that phone respondents report 0.78 fewer sweetened beverages consumed per week, even once we control for income, race and ethnicity, age and education, covariates that affect selection into response mode but also affect consumption. Phone respondents also report an impact score that is 0.54 points higher (on an 18-point scale) than web respondents, once we control for respondent characteristics. The effect of phone mode on tax agreement is both small (3%) and not statistically significant. These values are consistent with the raw differences in mean response by mode shown in Table 1, although for weekly consumption and the tax impact score, the absolute value of the difference between modes has declined by about 25% (0.78 versus 1.07 beverages per week and 0.54 versus 0.72 for the impact score). This is consistent with some, but not all, of the modal differences being explained by income, race/ethnicity, age, and other characteristics. While some race/ethnicity variables explain sweetened beverage consumption, only education and identifying a Democrat explain both consumption and impact score in this linear model (Table 4).

Table 4: Results from OLS Regression of Outcome Variables on Response Mode with Demographic Control Variables

	1	2	3
	Weekly Sweetened Beverage Consumption	Agrees with Tax	Impact Score
Phone Response	-0.78***	0.03	0.54**
	[0.21]	[0.03]	[0.24]
Weekly Sweetened Beverage Consumption		-0.02***	-0.17***
		[0.00]	[0.03]
Lives in Seattle	-0.70***	0.01	0.07
	[0.21]	[0.03]	[0.23]
High Income (≥260% FPL)	-0.32	-0.02	-0.01
	[0.21]	[0.03]	[0.24]
Non-Hispanic White	0.21	0.05	0.04
	[0.34]	[0.04]	[0.40]
Non-Hispanic Black	0.97*	-0.04	-0.42
	[0.50]	[0.06]	[0.49]
Non-Hispanic Asian	-0.54	-0.05	-0.37
	[0.44]	[0.06]	[0.53]
Hispanic	0.79**	0.02	-0.09
	[0.36]	[0.04]	[0.40]
50 or Younger	0.99***	0.10***	0.26
	[0.21]	[0.03]	[0.24]
Completed College or Above	-0.80***	0.13***	1.29***
	[0.23]	[0.03]	[0.25]
Married or Partnered	0.23	0.04	0.53**
	[0.21]	[0.02]	[0.23]
Democrat	-0.45*	0.14***	1.60***
	[0.27]	[0.03]	[0.28]
Independent	-0.28	0.01	-0.13
	[0.30]	[0.04]	[0.32]
Observations	1,702	1,609	1,702
R-squared	0.07	0.07	0.10

Note: High income is family income ≥ 260% of the Federal Poverty Line. Linear regression with robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

5.2 Matching Results

The average treatment effect (ATE) estimated using all three matching methods and for all three outcomes is shown in Table 5. The ATE is the average difference in phone mode responses and web mode responses when responders in all modes are compared to one or many responders in the other mode that are matched to them based on one of the three matching models explained above. Column 1 shows the estimated ATE for our three outcomes using Propensity Score Matching (PSM), column 2 shows the ATE estimated using Nearest Neighbor Matching (NNM), and column 3 shoes the ATE estimated using Inverse Probability Weighted and Regression Adjusted Matching (IPWRA). All three models produce estimated ATEs that are consistent with the regression results shown in Table 4, but we focus on the IPWRA results since this method is doubly robust to model misspecification. In the IPWRA columns, we see that responding in the phone mode, which we assume is due to social desirability bias, is responsible for 0.63 of the 1.07 fewer beverages per week reported by phone respondents relative to web respondents in Table 1. This is 59% of total difference, implying the selection accounted for the other 41% of the observed modal difference. The IPWRA results in column 3 also show that social desirability bias leads respondents to overestimate their opinion of the benefits of the tax by 0.55 points (on an 18-point scale). This is 76% of the original difference, implying that selection accounted for 24% of the modal difference in scores in the raw data. Consistent with all of our other comparisons, we find that there is no social desirability bias in responses to the tax approval question.

Table 5: Estimated Effect of Phone Survey Mode on Survey Responses

	Propensity Score Match	Nearest Neighbor Match	Inverse Probability Weighted and Regression Adjusted
Weekly Sweetened Beverage Consumption	-0.77*	-0.90***	-0.63**
	[0.44]	[0.26]	[0.25]
Approves of Tax	0.03	0.02	0.04
	[0.03]	[0.03]	[0.03]
Impact Score	0.65*	0.58**	0.55**
	[0.35]	[0.29]	[0.27]

Note: Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

5.3 Sensitivity of Results

We tested the sensitivity of our results both to different ways of defining income and consumption of sweetened beverages, and to the exclusion of respondents from outside of Seattle. First, since our analysis involves turning a categorical outcome (income) into a dichotomous measure of high- and low-income, we re-run our matching results using something closer to actual family income as a control variable. Panel A of Table 6 shows the results of the same matching procedure shown in Table 5, but with the midpoint of the respondent's reported income range as the income variable (see Supplemental Table 1 for income categories).

Second, we somewhat reverse this procedure to test for our results' sensitivity to the way we convert sweetened beverage consumption to a continuous variable. Panel B of Table 6 shows the estimated effect of phone response on our outcomes using the same matching methods as above, but with consumption measured as dichotomous variable that is defined as one when consumption is greater than the median reported consumption of one sweetened beverage per week.

Third, we re-run the original specification from Table 5, but only include responses by Seattle residents. From Table 1 above, we see that Seattle has considerably lower reported sweetened beverage consumption, a result that suggests that Seattle residents may have different attitudes about these beverages or may experience different social norms around them. Additionally, the survey was conducted right before a sweetened beverage tax was implemented in Seattle, but this was not the case for the other areas in the study. As such, Seattle residents may have been exposed to more media surrounding the health impacts of sweetened beverages or may have had more opportunities to consider their opinions of sweetened beverage taxes. Overall, we want to consider the possibility that Seattle residents will somehow exhibit different levels of social desirability bias than respondents in the comparison areas. Panel C of Table 6 shows the estimated effect of phone mode on our outcomes.[5] Finally, there are 59 individuals in our survey who reported that their household income was below 260% of the federal poverty line for a household of their size when they were initially screened into our survey, but then subsequently reported a household income that was inconsistent with this initial categorization. Similarly, another 14 respondents reported incomes above 260% of the federal poverty line in the screener, but reported inconsistently low incomes thereafter. Therefore, we also present the results when we exclude those 73 individuals (N =1,641) as a sensitivity test. These results are shown in Panel D of Table 6.

[5] Summary statistics and linear regression results for only Seattle respondents are given in Supplemental Tables 2 and 3, respectively.

Table 6: Estimated Effect of Phone Survey Mode on Survey Responses using Alternative Specifications

	Propensity Score Match	Nearest Neighbor Match	Inverse Probability Weighted and Regression Adjusted
Panel A – Redefine Income			
Weekly Sweetened Beverage Consumption	-0.22	-1.08***	-0.60**
	[0.40]	[0.27]	[0.26]
Agrees with Tax	0.03	0.04	0.04
	[0.03]	[0.03]	[0.03]
Impact Score	0.65*	0.83***	0.54**
	[0.35]	[0.29]	[0.27]
Panel B – Redefine Consumption			
Weekly Sweetened Beverage Consumption	-0.12***	-0.13***	-0.12***
	[0.04]	[0.03]	[0.03]
Agrees with Tax	0.04	0.02	0.03
	[0.04]	[0.03]	[0.03]
Impact Score	0.51	0.64**	0.43
	[0.39]	[0.28]	[0.27]
Panel C – Seattle Only			
Weekly Sweetened Beverage Consumption	-0.67*	-0.47	-0.37
	[0.35]	[0.30]	[0.31]
Agrees with Tax	0.04	0.02	0.08**
	[0.04]	[0.04]	[0.03]
Impact Score	0.55	0.91**	0.93***
	[0.41]	[0.41]	[0.35]
Panel D – Remove Inconsistent Incomes			
Weekly SSB Consumption	-1.01**	-0.91***	-0.68***
	[0.47]	[0.26]	[0.26]
Agrees with Tax	0.07**	0.02	0.05
	[0.03]	[0.03]	[0.03]
Impact Score	0.66**	0.52*	0.58**
	[0.28]	[0.29]	[0.28]

Note: Robust standard errors in brackets. *** p<0.01, ** p<0.05, * p<0.1

Overall, the results in Table 6 are similar to those in Table 5. In Panel A, we see that increasing the variation in income results in similar coefficients and statistical significance to Table 5's results, with the exception of the treatment effect on consumption that was estimated using propensity score matching. In Panel B, we redefine consumption to be a binary outcome (high versus low). This affects not only the treatment effect on consumption, but also on the other outcomes, since consumption is used as a control variable in those columns. The consumption outcomes are not directly comparable between Tables 5 and 6, although Table 6 does show a negative and significant treatment effect, as we would expect from Table 5. The tax impact score outcome is interesting in that the coefficients are close to those found in Table 5, but with much lower (or no) statistical significance. This suggests that we are creating additional noise by reducing the variation in consumption as a control variable, but this likely does not invalidate our main findings.

In Panel C, we find that the estimated coefficients on consumption are smaller and less significant in the Seattle-only sample. The coefficients are likely reduced because consumption is lower in Seattle (reported average consumption is 2.5 beverages per week in Seattle versus 3.1 per week in the full sample), and so we would expect a bias that reduces reported consumption by the same percentage to produce a larger reduction in the full sample than in Seattle only, in absolute terms. We also note that the measured bias for the tax impact score is larger in Seattle only sample, but Seattle has slightly more positive attitudes toward the tax (2.68 in Seattle versus 2.11 in the comparison area), so a similar mechanism could be at work here.

In Panel D, we see that removing the individuals with inconsistently reported incomes does not affect our results, aside from reducing our standard errors in some cases. Without these 73 observations, we also find that the effect of the phone mode on reported tax agreement is statistically significantly different from zero and positive in two of the three specifications.

6. Discussion and Conclusion

6.1 Discussion

The objective of this study was to investigate the existence of social desirability bias in self-reported sweetened beverage consumption, attitudes regarding the health and economic effects of sweetened beverage taxes, and overall approval of such taxes. The preceding section demonstrates evidence of social desirability bias in two of the three outcomes – sweetened beverage consumption and attitudes toward the tax, but not tax approval – by showing that modal differences in outcomes remain even when we use matching methods to control for selection bias. Social desirability bias has been previously documented in both the political and health literature, as respondents to phone and in-person surveys have been found to give more of what may be considered socially acceptable responses to sensitive questions compared to respondents to web surveys. Suggestive evidence that reporting reduced sweetened beverage consumption is socially desirable was found by Klesges and coauthors (2004), but they did not find actual bias in self-reported consumption of sweetened beverages. No prior literature has found social desirability bias regarding questions around sweetened beverage taxes, but these findings appear consistent with current media attention around the role of sweetened beverages in diabetes and other negative health outcomes.

If our interpretation is correct, mixed-mode surveys and surveys conducted exclusively over the phone or in-person may be under-reporting sweetened beverage consumption in the population. While we do not attempt to draw inferences about population-level consumption in this study, and so therefore cannot infer anything about under-reporting in the population, we do find that responding on the phone leads to under-reporting consumption by 0.63 beverages per week in our sample. Therefore, by this estimate, true consumption in our phone sample is approximately 3.12 beverages per week instead of 2.49. Since phone respondents make up 41% of our total sample, we expect that the stated average of 3.11 beverages per week found across both modes in our survey actually under-estimates sweetened beverage consumption in our sample by 25% (relative to what they would have reported if all respondents took a web-based survey).

We can compare our findings to Klesges and co-authors, who found that sweetened beverage consumption could be under-reported by up to 10% due to social desirability bias. Translating both sets of results into daily consumption, we find that people are under-estimating their consumption by about 0.09 beverages per day, while they found that under-reporting might be up to 0.3 beverages per day. Notably, their sample was made up of pre-teen African-American girls, while ours covers adults of both sexes and includes other race and ethnicity groups. Average reported consumption in our sample is much lower, as well (about 3 beverages per week instead of 3 beverages per day in their survey). The effect of survey mode on perceptions of public support for the tax among our sample is less clear than the impact on consumption. None of our specifications found differences in overall tax support by survey mode, although we do find that the combined impression of the health and economic benefits of these taxes are lower for web respondents, suggesting that phone respondents may overstate how beneficial they truly believe the tax to be. Our estimates vary by estimation method, but phone respondents are overstating their positive attitude to sweetened beverage taxes by approximately 0.54 points on an 18-point scale (relative to web respondents). This implies that the average impact score in our phone population is 2.24, not 2.79, and that if all respondents were given a web-based survey, our average impact score would be 2.14, not 2.37. Thus, we are overestimating the true value of this score by at least 11%.

6.2 Limitations

In this study, we are fundamentally trying to measure the difference in responses by the same respondent across modes, a task that is impossible since we only have access to the respondent's response in one mode. We attempt to address this by creating a phone and web sample that look similar to each other in observable characteristics, and then applying matching techniques to further correct when the samples still show selection bias. Any remaining modal differences in survey responses, then, can potentially be interpreted as the causal impact of mode on responses. In the case of sensitive topics, we may believe that this causal impact operates through the mechanism of social desirability bias.

Another interpretation of these remaining differences, however, is that there are unobservable differences between respondents in the two modes, making it impossible for us to either match respondents properly or observe that we have failed to match them. In this case, we cannot exclude the interpretation that the remaining modal difference in responses arises from respondents in one mode being of a completely different type than respondents in the other mode. Given almost ubiquitous access to the internet among urban Americans in 2017, it may be implausible to think that web respondents are somehow fundamentally different from phone respondents. This claim is supported by recent research into web and mixed-mode surveys that find that web respondents in high income countries are increasingly, although still not perfectly, representative of the population as web access continues to grow (Leenheer, 2013; Eckman, 2016; Grewenig *et al.*, 2018). On the other hand, our web respondents are recruited from survey panels, groups of people who have agreed to be contacted by survey firms. There is some evidence that these panelists are more active than other survey respondents: they are more politically active, are earlier adopters of technology, and even eat out more

(Duffy *et al.*, 2005). However, there is less evidence that the use of these panels leads to biased results, once researchers control for sociodemographic characteristics (Campbell, Venn and Anderson, 2018) Relatedly, one could also argue that, given current use of caller id and voicemail to screen calls and overall low response rates in phone surveys, the respondents who agree to answer a survey on the phone may have become the population outliers (Duffy *et al.*, 2005).

Ultimately, while we cannot definitively claim that there is no confounding between response mode and response that cannot be controlled for using observable characteristics, the fact remains that there are meaningful differences in response by mode, and none of the observable characteristics of web respondents (richer, whiter, more technologically savvy) are associated with higher sweetened beverage consumption (Bleich *et al.*, 2009). For this reason, we claim that our results are evidence in support of social desirability bias, but that more work remains to be done. Other researchers have reduced selection bias in multi-mode surveys by randomizing respondents into survey mode, and this procedure could be used in future work on this topic to further investigate the extent to which modal differences remain after randomization. These remaining differences, if they existed, could be interpreted as causal with more confidence than our own.

The other limitations of our study relate to the way in which data was collected. We do not collect information on body weight, which has been shown to be related to consumption. We also do not know the characteristics of the phone survey interviewers, nor do we know the ways in which these characteristics may or may not interact with social desirability bias. All interviewers worked with a script, which was available in English and Spanish, and were trained, but it is possible that interviewer demographics (or perceived demographics) affected the response to phone interviews in ways that might affect our findings (West and Blom, 2016). However, we have no reason to believe that this effect would work systematically in one direction, or that it could explain any of our findings. We also include in our limitations the fact that dietary intake is difficult to measure, even when responses are not susceptible to social desirability bias. Our survey employed a screening question, adapted from the NHANES, which is likely to be less accurate than a question that involves a 24-hour recall of consumption (the "gold standard" for dietary consumption). As such, we do not claim that the average reported beverages per week represent a true population mean.

6.3 Policy and Research Recommendations

Our results suggest using caution to interpret self-reported dietary intake of sweetened beverages in surveys that interview respondents over the phone or in-person, but they also call all self-reported measures of sweetened beverage consumption into question. Survey respondents likely feel greater pressure to bias their responses toward socially desirable values when responding on the phone or in-person, but this is no guarantee that they do not also feel social pressure when responding on the web or when reporting their purchases through scanning receipts or products. Social desirability bias has previously been detected in web surveys, for example when survey respondents are directly asked about sensitive voting behaviors (Brown-Iannuzzi, Najle and Gervais, 2019). If self-reported sweetened beverage consumption is prone to under-reporting regardless of survey mode, then researchers will be seriously hampered in their ability to measure the impacts of sweetened beverage taxes on consumption, understand linkages between sweetened beverage consumption and population health, and optimally design sweetened beverage tax policies.

Policy makers, therefore, should consider that consumers may consume more sweetened beverages than they actually report; in particular, they should be wary of solely relying on self-reported measures of intake when evaluating the effectiveness of these policies. Relatedly, policy makers should consider strengthening their public messaging regarding the health and economic benefits of sweetened beverage taxes, even if they believe that attitudes are generally positive. Without a pro-tax messaging campaign, that informs the public about the positive health and economic effects of these taxes, the taxes may eventually lose public support. In fact, recent successful efforts to block U.S. municipalities from enacting future beverage taxes by banning the taxes at the state level have relied heavily on informational campaigns that focused on the negative economic effects of the taxes (Daniels, 2019; White, 2019). These campaigns, often funded by the beverage industry, may ultimately shift social norms in the direction of more favorable attitudes toward sweetened beverages, with unpredictable effects on public health.

For researchers, we recommend that future surveys of self-reported sweetened beverages either be conducted on the web, where there is less likely to be social desirability bias in the self-reports, or via a mixed-mode survey that is explicitly designed to oversample populations across characteristics that influence both consumption and response mode. In our data, we find that income, age, and Hispanic ethnicity are the three demographic characteristics that affect both phone responses and reported consumption, although we caution that our data are only relevant for Seattle and our four comparison cities. This oversampling will allow researchers to further explore the existence of social desirability bias in self-reported consumption using either linear regression or one of the matching techniques explored here. We also see a need for future researchers to design surveys that explore potential social desirability bias in web reports of sweetened beverage consumption, a possibility that we cannot speak to with our study design. Designs that employ direct and indirect measures of sweetened beverage consumption, such as that used by Brown-Iannuzzi, Najle and Gervais (2019) in their study of voting behavior, may prove useful here.

6.4 Conclusion

This paper looks for evidence of social desirability bias in self-reported measures of sweetened beverage consumption, support for sweetened beverage taxes, and attitudes toward the health and economic benefits of these taxes by comparing survey responses given over the phone to those given over the web. We use intentional oversampling of populations less likely to complete the survey in each mode and established matching methods to create samples balanced on demographic characteristics to isolate the effect of survey mode on response, net of the effect of selection into survey mode. We find evidence of substantial underreporting of sweetened beverage consumption in phone respondents compared to web respondents, which we attribute to social desirability bias. There was no evidence of social desirability bias in reporting approval of a sweetened beverage tax, although there was evidence of a small effect of social desirability on report of positive or negative health and economic impacts of the tax.

Acknowledgments

The authors would like to thank James Krieger, Brian E. Saelens, Nadine Chan, and Mary Podrabsky, who contributed to the design of the survey used in this study.

Funding

This study was primarily funded by the City of Seattle, with additional support provided by Kaiser Foundation Health Plan of Washington and Arcora – The Foundation of Delta Dental of Washington. Partial support for this research also came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, P2C HD042828, to the Center for Studies in Demography & Ecology at the University of Washington. The City of Seattle provided input on the evaluation components and methods. The funding bodies did not influence analyses, interpretation or the writing of this manuscript.

Availability of data and materials

The de-identified dataset used during the current study is available from the corresponding author upon reasonable request and approval by the City of Seattle.

Ethics approval and consent to participate

The University of Washington School of Public Health Institutional Review Board determined that this survey was exempt under category 2. Thus, this research is exempt from the federal human subjects regulations and need for consent did not apply. Note that information was obtained in such a manner that subjects could not be identified by the research team.

References

Abadie, A. *et al.* (2004) *Implementing matching estimators for average treatment effects in Stata*, *The Stata Journal*. Available at: https://journals.sagepub.com/doi/pdf/10.1177/1536867X0400400307 (Accessed: 22 August 2019).

Abadie, A. and Imbens, G. W. (2006) *LARGE SAMPLE PROPERTIES OF MATCHING ESTIMATORS FOR AVERAGE TREATMENT EFFECTS*, *Econometrica*. Available at: http://emlab.berkeley.edu/users/imbens/ (Accessed: 22 August 2019).

Austin, P. C. (2009) 'Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples.', *Statistics in medicine*. Wiley-Blackwell, 28(25), pp. 3083–107. doi: 10.1002/sim.3697.

Bleich, S. N. *et al.* (2009) 'Increasing consumption of sugar-sweetened beverages among US adults: 1988–1994 to 1999–2004', *The American Journal of Clinical Nutrition*. Narnia, 89(1), pp. 372–381. doi: 10.3945/ajcn.2008.26883.

Brown-Iannuzzi, J. L., Najle, M. B. and Gervais, W. M. (2019) 'The Illusion of Political Tolerance: Social Desirability and Self-Reported Voting Preferences', *Social Psychological and Personality Science*. SAGE PublicationsSage CA: Los Angeles, CA, 10(3), pp. 364–373. doi: 10.1177/1948550618760147.

Burke, M. A. and Carman, K. G. (2017) 'You can be too thin (but not too tall): Social desirability bias in self-reports of weight and height', *Economics & Human Biology*. North-Holland, 27, pp. 198–222. doi: 10.1016/J.EHB.2017.06.002.

Burkill, S. *et al.* (2016) 'Using the Web to Collect Data on Sensitive Behaviours: A Study Looking at Mode Effects on the British National Survey of Sexual Attitudes and Lifestyles', *PLOS ONE*. Edited by M. A. Cardoso, 11(2), p. e0147983. doi: 10.1371/journal.pone.0147983.

Campbell, R. M., Venn, T. J. and Anderson, N. M. (2018) 'Cost and performance tradeoffs between mail and internet survey modes in a nonmarket valuation study', *Journal of Environmental Management*. Academic Press, 210, pp. 316–327. doi: 10.1016/J.JENVMAN.2018.01.034.

Cattaneo, M. D. (2010) 'Efficient semiparametric estimation of multi-valued treatment effects under ignorability', *Journal of Econometrics*. North-Holland, 155(2), pp. 138–154. doi: 10.1016/J.JECONOM.2009.09.023.

Cawley, J. *et al.* (2019) 'The Economics of Taxes on Sugar-Sweetened Beverages: A Review of the Effects on Prices, Sales, Cross-Border Shopping, and Consumption', *Annual Review of Nutrition*, 39(1), pp. 317–338. doi: 10.1146/annurev-nutr-082018-124603.

Daniels, J. (2019) *California soda tax bill shelved, in reprieve for beverage industry*, *cnbc.com*. Available at: https://www.cnbc.com/2019/04/22/california-soda-tax-bill-shelved-in-reprieve-for-beverage-industry.html (Accessed: 6 March 2020).

Duffy, B. *et al.* (2005) 'Comparing data from online and face-to-face surveys', *International Journal of Market Research*, 47(6). Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.376.621&rep=rep1&type=pdf (Accessed: 13 June 2019).

Eckman, S. (2016) 'Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?', *Social Science Computer Review*. SAGE PublicationsSage CA: Los Angeles, CA, 34(1), pp. 41–58.

doi: 10.1177/0894439315572985.

Falbe, J. *et al.* (2016) ‘Impact of the Berkeley excise tax on sugar-sweetened beverage consumption’, *American Journal of Public Health*, 106(10), pp. 1865–1871. doi: 10.2105/AJPH.2016.303362.

Grewenig, E. *et al.* (2018) *Can Online Surveys Represent the Entire Population?* Available at: www.iza.org (Accessed: 24 August 2019).

Imbens, G. W. (2015) ‘Matching Methods in Practice: Three Examples’, *Journal of Human Resources*. University of Wisconsin Press, 50(2), pp. 373–419. doi: 10.3368/jhr.50.2.373.

Jones, M. K. *et al.* (2016) ‘A Comparison of Web and Telephone Responses From a National HIV and AIDS Survey.’, *JMIR public health and surveillance*. JMIR Publications Inc., 2(2), p. e37. doi: 10.2196/publichealth.5184.

Klesges, L. M. *et al.* (2004) ‘Social desirability bias in self-reported dietary, physical activity and weight concerns measures in 8- to 10-year-old African-American girls: results from the Girls health Enrichment Multisite Studies (GEMS)’, *Preventive Medicine*. Academic Press, 38, pp. 78–87. doi: 10.1016/J.YPMED.2003.07.003.

Kreuter, F., Presser, S. and Tourangeau, R. (2008) ‘Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity’, *Public Opinion Quarterly*. Narnia, 72(5), pp. 847–865. doi: 10.1093/poq/nfn063.

Lee, H. *et al.* (2019) ‘Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era’, *Social Science Computer Review*. SAGE PublicationsSage CA: Los Angeles, CA, 37(2), pp. 234–247. doi: 10.1177/0894439318756867.

Lee, M. M. *et al.* (2019) ‘Sugar-Sweetened Beverage Consumption 3 Years After the Berkeley , California , Sugar-Sweetened Beverage Tax’, *American journal of public health*, pp. 1–3. doi: 10.2105/AJPH.2019.304971.

Leenheer, J. (2013) ‘Does it pay off to include non-internet households in an internet panel?’, *International Journal of Internet Science*, 8(1), pp. 17–29.

O’Donoghue, T. and Rabin, M. (2006) ‘Optimal sin taxes’, *Journal of Public Economics*, 90(10–11), pp. 1825–1849. doi: 10.1016/j.jpubeco.2006.03.001.

Oddo, V. M. *et al.* (2019) ‘Perceptions of the possible health and economic impacts of Seattle’s sugary beverage tax’, *BMC Public Health*. BioMed Central, 19(1), p. 910. doi: 10.1186/s12889-019-7133-2.

Parks, K. A., Pardi, A. M. and Bradizza, C. M. (2006) ‘Collecting data on alcohol use and alcohol-related victimization: a comparison of telephone and Web-based survey methods.’, *Journal of Studies on Alcohol*, 67(2), pp. 318–323. doi: 10.15288/jsa.2006.67.318.

Rosenbaum, P. R. and Rubin, D. B. (1984) ‘Reducing Bias in Observational Studies Using Subclassification on the Propensity Score’, *Journal of the American Statistical Association*. Taylor & Francis, Ltd.American Statistical Association, 79(387), p. 516. doi: 10.2307/2288398.

Schläpfer, F., Roschewitz, A. and Hanley, N. (2004) ‘Validation of stated preferences for public goods: a comparison of contingent valuation survey response and voting behaviour’, *Ecological Economics*. Elsevier, 51(1–2), pp. 1–16. doi: 10.1016/J.ECOLECON.2004.04.006.

StataCorp (2019) *STATA TREATMENT-EFFECTS REFERENCE MANUAL: RELEASE 16*. College Station, TX:

Stata Press. Available at: https://www.stata.com/manuals/te.pdf (Accessed: 22 August 2019).

Tamir, O. *et al.* (2018) 'Taxation of sugar sweetened beverages and unhealthy foods: a qualitative study of key opinion leaders' views', *Israel Journal of Health Policy Research*. BioMed Central, 7(1), p. 43. doi: 10.1186/s13584-018-0240-1.

Teng, A. M. *et al.* (2019) 'Impact of sugar-sweetened beverage taxes on purchases and dietary intake: Systematic review and meta-analysis', *Obesity Reviews*, (January), pp. 1–18. doi: 10.1111/obr.12868.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2010) 'A Method for Evaluating Mode Effects in Mixed-Mode Surveys', *Public Opinion Quarterly*, 74(5), pp. 1027–1045. doi: 10.1093/poq/nfq059.

Vannieuwenhuyze, J. T. A. and Loosveldt, G. (no date) 'Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects', *Sociological Methods & Research*, 42(1), pp. 82–104. doi: 10.1177/0049124112464868.

West, B. T. and Blom, A. G. (2016) 'Explaining Interviewer Effects: A Research Synthesis', *Journal of Survey Statistics and Methodology*. Oxford Academic, 5(2), p. smw024. doi: 10.1093/jssam/smw024.

White, J. B. (2019) *Is Big Soda winning the soft drink wars?*, *Politico.com*. Available at: https://www.politico.com/agenda/story/2019/08/13/soda-tax-california-public-health-000940 (Accessed: 6 March 2020).

Woo, Y., Kim, S. and Couper, M. P. (2015) 'Comparing a Cell Phone Survey and a Web Survey of University Students', *Social Science Computer Review*. SAGE PublicationsSage CA: Los Angeles, CA, 33(3), pp. 399–410. doi: 10.1177/0894439314544876.

Zhong, Y. *et al.* (2018) 'The Short-Term Impacts of the Philadelphia Beverage Tax on Beverage Consumption', *American journal of preventive medicine*. Elsevier.