Year 4 (2020-2021) Project Report

Response Willingness in Consecutive Travel Surveys:
An Investigation Based on the National Household Travel Survey Using a Sample Selection Model

Prepared for **Teaching Old Models New Tricks (TOMNET) Transportation Center**

By

Xinyi Wang
Email: xinyi.wang@gatech.edu

F. Atiyya Shaw
Email: atiyya@gatech.edu

Patricia L. Mokhtarian
Email: patmokh@gatech.edu

Kari Watkins
Email: kari.watkins@ce.gatech.edu

School of Civil and Environmental Engineering
Georgia Institute of Technology
790 Atlantic Drive, Atlanta, GA 30332

May 2022

TECHNICAL REPORT DOCUMENTATION PAGE

<table>
<tr><td colspan="2">1. Report No.
N/A</td><td>2. Government Accession No.
N/A</td><td colspan="2">3. Recipient's Catalog No.
N/A</td></tr>
<tr><td colspan="3" rowspan="2">4. Title and Subtitle
Response Willingness in Consecutive Travel Surveys:
An Investigation Based on the National Household Travel Survey Using a Sample Selection Model</td><td colspan="2">5. Report Date
May 2022</td></tr>
<tr><td colspan="2">6. Performing Organization Code
N/A</td></tr>
<tr><td colspan="3">7. Author(s)
Xinyi Wang, https://orcid.org/0000-0002-3564-9147
F. Atiyya Shaw, https://orcid.org/0000-0001-8717-5118
Kari Watkins, https://orcid.org/0000-0002-3824-2027
Patricia L. Mokhtarian, https://orcid.org/0000-0001-7104-499X</td><td colspan="2">8. Performing Organization Report No.
N/A</td></tr>
<tr><td colspan="3" rowspan="2">9. Performing Organization Name and Address
School of Civil and Environmental Engineering
Georgia Institute of Technology
790 Atlantic Drive, Atlanta, GA 30332</td><td colspan="2">10. Work Unit No. (TRAIS)
N/A</td></tr>
<tr><td colspan="2">11. Contract or Grant No.
69A3551747116</td></tr>
<tr><td colspan="3" rowspan="2">12. Sponsoring Agency Name and Address
U.S. Department of Transportation,
University Transportation Centers Program,
1200 New Jersey Ave, SE, Washington, DC 20590</td><td colspan="2">13. Type of Report and Period Covered
Research Report
(2020 – 2021)</td></tr>
<tr><td colspan="2">14. Sponsoring Agency Code
USDOT OST-R</td></tr>
<tr><td colspan="5">15. Supplementary Notes
N/A</td></tr>
<tr><td colspan="5">16. Abstract
Declining survey response rates have increased the costs of travel survey recruitment. Recruiting respondents based on their expressed willingness to participate in future surveys, obtained from a preceding survey, is a potential solution but may exacerbate sample biases. In this study, we analyze the self-selection biases of survey respondents recruited from the 2017 U.S. National Household Travel Survey (NHTS), who had agreed to be contacted again for follow-up surveys. We apply a probit with sample selection (PSS) model to analyze (1) respondents' willingness to participate in a follow-up survey (the selection model) and (2) their actual response behavior once contacted (the outcome model). Results verify the existence of self-selection biases, which are related to survey burden, sociodemographic characteristics, travel behavior, and item non-response to sensitive variables. We find that age, homeownership, and medical conditions have opposing effects on respondents' willingness to participate and their actual survey participation. The PSS model is then validated using a hold-out sample and applied to the NHTS samples from various geographic regions to predict follow-up survey participation. Effect size indicators for differences between predicted and actual (population) distributions of select sociodemographic and travel-related variables suggest that the resulting samples may be most biased along age and education dimensions. Further, we summarized six model performance measures based on the PSS model structure. Overall, this study provides insight into self-selection biases in respondents recruited from preceding travel surveys. Model results can help researchers better understand and address such biases, while the nuanced application of various model measures lays a foundation for appropriate comparison across sample selection models.</td></tr>
<tr><td colspan="3">17. Key Words
Self-Selection Bias; Nonresponse Bias; Probit with Sample Selection Model; National Household Travel Survey; Model Measurement; Sampling Frame</td><td colspan="2">18. Distribution Statement
No restrictions.</td></tr>
<tr><td>19. Security Classif.(of this report)
Unclassified</td><td>20. Security Classif.(of this page)
Unclassified</td><td>21. No. of Pages
41</td><td colspan="2">22. Price
N/A</td></tr>
</table>

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENTS

This study was funded by a grant from a USDOT Tier 1 University Transportation Center, supported by USDOT through the University Transportation Centers program. The authors would like to thank the TOMNET UTC and USDOT for their support of university-based research in transportation, and especially for the funding provided in support of this project. The authors would like to thank three anonymous reviewers, whose suggestions improved this report.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Declining survey response rates have increased the costs of travel survey recruitment. Recruiting respondents based on their expressed willingness to participate in future surveys, obtained from a preceding survey, is a potential solution but may exacerbate sample biases. In this study, we analyze the self-selection biases of survey respondents recruited from the 2017 U.S. National Household Travel Survey (NHTS), who had agreed to be contacted again for follow-up surveys. We apply a probit with sample selection (PSS) model to analyze (1) respondents' willingness to participate in a follow-up survey (the selection model) and (2) their actual response behavior once contacted (the outcome model). Results verify the existence of self-selection biases, which are related to survey burden, sociodemographic characteristics, travel behavior, and item non-response to sensitive variables. We find that age, homeownership, and medical conditions have opposing effects on respondents' willingness to participate and their actual survey participation. The PSS model is then validated using a hold-out sample and applied to the NHTS samples from various geographic regions to predict follow-up survey participation. Effect size indicators for differences between predicted and actual (population) distributions of select sociodemographic and travel-related variables suggest that the resulting samples may be most biased along age and education dimensions. Further, we summarized six model performance measures based on the PSS model structure. Overall, this study provides insight into self-selection biases in respondents recruited from preceding travel surveys. Model results can help researchers better understand and address such biases, while the nuanced application of various model measures lays a foundation for appropriate comparison across sample selection models.

1. INTRODUCTION

High-quality survey data provide the foundation for research and policymaking across many fields. While novel data sources are actively being examined for use in transport applications, both currently and for the foreseeable future traditional travel surveys will continue to play an irreplaceable role in providing critical data for use in travel demand modeling, regional planning, and policymaking. However, survey response rates are in continuous and significant decline, thus requiring increased efforts toward respondent recruitment. Further necessitating these increased efforts is the fact that low response rates and their accompanying nonresponse biases can threaten the validity of survey data, and thus contingent research findings (National Research Council, 2013).

Survey teams have employed a range of efforts aimed at increasing response rates and improving survey data quality. Among the most common tools are the use of passive datasets such as GPS records (Bohte and Maat, 2009) and targeted marketing data (Shaw et al., 2021), novel survey formats (e.g., interactive surveys; Collins et al., 2012), and targeted sampling frames (e.g., online panels; Circella et al., 2016), to name a few. Another approach, which is the focus of this report, is to *recruit survey respondents who had expressed willingness to be contacted again in a previous survey*; this approach has been shown to produce a significantly higher response rate and lower cost per valid response relative to random sampling (Amarov and Rendtel, 2013; Kim et al., 2019; Circella et al., 2020).

This recruitment method has some similarities to the approach used in panel studies in that both nominally draw respondents from preceding surveys. Accordingly, both approaches are subject to attrition biases. There are some important differences, however. For one thing, respondents to a panel study are normally informed at the outset that participation in the study involves completing multiple surveys (and therefore that agreement to participate signifies agreement to complete multiple surveys), whereas in the present case, the willingness to complete a later survey is an entirely separate decision, not even presented to the respondent at the entrance to the initial study. Other differences reside in the survey purpose, contents, or outcome. Specifically, panel surveys focus on repeated observations on a set of variables for the same sample unit over time (Lavrakas, 2008), which allows the tracking of specific variables or study interests. In contrast, recruiting respondents from a previous survey is not a periodical behavior, and the follow-up survey may have relatively little in common with the initial one. The use of this recruitment method: (1) increases the survey response rates obtained on follow-up surveys; (2) reduces the financial burden for local transportation agencies and researchers; and (3) facilitates the expansion of the variable set of the preceding survey and enables data fusion across datasets (Shaw et al., 2022). In view of the plethora of single cross-section surveys and the challenges of conducting panel studies (notably time and money, among others), using a prior cross-sectional survey to help recruit for the next one is certainly an attractive prospect.

However, in the transportation domain, this recruitment method has not been widely adopted nor carefully examined. A major potential drawback of recruiting respondents based on their willingness expressed in a preceding survey is the non-representativeness that may be inherent in that sample (Couper et al., 2007). Accordingly, the present study is interested in the following questions: (1) Who is more likely to respond to a follow-up survey? (2) How does recruiting

respondents based on their willingness expressed in a preceding travel survey bias the follow-up survey sample? (3) In view of the importance (in sample size, geographic scope, and information value) of the National Household Travel Survey (NHTS), how helpful is it to use the NHTS in particular as the springboard for follow-on survey recruitment? Specifically, what survey sample could we expect if we recruited respondents from the 2017 NHTS respondents in different geographic regions in the U.S.?

To address the questions raised above and bridge the gap in the literature regarding recruiting survey respondents from a preceding travel survey, we do the following:

(1) We analyze the first-stage *self-selection* (willingness to participate in a follow-up survey) and second-stage *non-response* (actual response behavior) biases simultaneously for respondents *recruited from a previous travel survey* (the NHTS), using a probit with sample selection (PSS) model, which could remedy the model coefficient biases. We also propose several standardized PSS model performance measures to enable model comparisons.
(2) We apply the PSS model to a holdout sample to decompose biases (e.g., dataset bias, self-selection bias, non-response bias) accumulated along the way and further analyze the representativeness of the recruited survey respondents by comparing sample and population marginal distributions for various variables.
(3) We predict follow-up survey samples from different geographic regions in the U.S. as another PSS model application example, and check the model's generalizability.

By understanding the dataset biases that can result when respondents are recruited from a preceding survey, researchers/practitioners can better assess the tradeoff between data quality and resource constraints associated with respondent recruitment. Moreover, understanding these biases would allow survey developers to adjust their invited sample – for example, by oversampling underrepresented groups in the follow-up surveys. This work would, therefore, be especially useful for transportation professionals if the NHTS in particular retained the willingness question as a recurring item in future surveys, thereby allowing local agencies and researchers to recruit follow-up respondents from the NHTS sample efficiently. Even outside of the NHTS, the contributions of this report have general findings and implications for researchers using the approach of recruiting respondents from prior cross-sectional surveys.

The rest of this report is organized as follows. We begin with the literature review in Section 2. We describe the data source used in this study in Section 3. Section 4 introduces details of the probit with sample selection (PSS) model and summarizes six modified model performance measures. In Section 5, we present and analyze model results, including both model interpretations and performance measures. In Section 6, we apply the calibrated model to a holdout sample to decompose sample biases and predict follow-up survey participation in diverse geographic regions in the U.S. We close with a summary of findings in Section 7.

2. LITERATURE REVIEW

As mentioned, continuously declining survey response rates make it increasingly difficult for survey developers to obtain high-quality survey data with the same survey budgets as in the past. To enhance response rates, researchers and practitioners have developed and applied many approaches for aiding in the survey recruitment process.

We first summarize a few commonly used recruitment approaches and the accompanying sample biases. The use of *survey incentives* is an effective approach to increase survey response rates; examples of these include lotteries, tokens, and philanthropic donations (Edwards et al., 2002, Young et al., 2020). Coryn et al. (2020) found the lottery to be the most cost-effective incentive format, while Parsons and Manierre (2014) showed that unconditional incentives might exacerbate the overrepresentation of females among survey respondents. Using *different survey modes* (e.g., mail, phone, and web) is another way to increase response rates of specific population groups. For example, web surveys have (at least in the past) been found to generate a much lower response rate than mail surveys in general (Manfreda et al., 2008, Hardigan et al., 2012), but younger generations such as college students are more responsive to web surveys (Shih and Xitao, 2008, Börkan, 2010). However, the sample may retain biases associated with the sampling mode, i.e., a mode effect. In a survey aimed at college students, Carini et al. (2003) found that web survey respondents gave more favorable responses regarding computing and information technology than the paper survey respondents. Survey developers could also obtain higher response rates by carefully selecting the *sampling frame* (Wolf et al., 2005). In recent years, scholars have used commercially-operated online opinion panels, consisting of people who pre-register for survey participation in return for rewards (e.g., cash, vouchers), to reach out to survey respondents and enhance response rates (Neufeld and Mokhtarian, 2012; Miller et al., 2020; Chauhan et al., 2021). Some companies that operate these online opinion panels allow quota sampling within the panelists to ensure a (more) representative sample regarding the selected control variables (usually sociodemographic variables). Still, this does not guarantee the representativeness of other variables. For example, a recent study by this team found that online opinion panel respondents have significantly lower life satisfaction than respondents recruited from other sources, even when controlling for socio-demographics (Wang et al., 2022).

Another approach, as previously detailed in the Introduction, entails *the recruitment of survey respondents who indicated willingness to respond in prior surveys* (e.g., Lin et al., 2011). As with the other recruitment approaches discussed, this method also results in unrepresentative samples. Couper et al. (2007) modeled internet users' willingness to do an online survey and their subsequent follow-up response. They concluded that self-selected samples of internet users are not representative of the population with respect to demographic, financial, and health-related variables. In another example, Germany's Federal Statistical Office developed an access panel (a pool of persons willing to take part in voluntary surveys) from a large-scale household survey. The access panel was then used as the sampling frame for multiple surveys, and was found to be unrepresentative by multiple teams. Specifically, Amarov and Rendtel (2013) explored the survey participation propensity of the access panel and identified self-selection biases existing in multiple variables, including age, household size, and item-nonresponse. An accompanied simulation experiment (Tobias et al., 2013) on the selection process of the access panel emphasizes the importance of constructing proper statistical models for the access panel recruitment to ensure the appropriate usage of this high-response-rate and low-cost recruitment method. Similarly, Adriaan and Jacco (2009) applied bivariate logistic regressions to analyze the selectivity of the nonresponse of an online panel, which was recruited using a three-stage process: participation in a first telephone interview, willingness to be recontacted, and final agreement to participate in the online panel. The authors found selection biases with regards to age, income, and personal computer ownership.

Although transportation studies on this topic are limited, some studies have examined the nonresponse bias in travel surveys, which could inform the analysis of self-selection biases in recruiting survey respondents from a preceding travel survey. Wittwer and Hubrich (2015) reached out to survey nonrespondents with an abbreviated survey, and found that age and household size have significant differences between main survey respondents and nonrespondents. de Haas et al. (2018) used information obtained from a screening survey and found that age, gender, and education influence people's willingness to participate in a household travel survey panel. They also found that willingness to participate in a travel survey could modify model coefficients and slightly improve the fits of mode choice models.

This study aims to address the literature gap by examining the practice of recruiting respondents from the NHTS for a statewide travel survey, and constructing a proper statistical model for the recruitment process in the transportation context. We apply the probit with sample selection (PSS) model for analysis, which remedies the selection biases by allowing correlations between the unobservables in the selection and outcome equations (Heckman, Tobias, & Vytlacil, 2001). The PSS model was proposed by van de Ven and van Praag (1981), which is modified from the Heckman model (Heckman, 1976; originally designed for correcting sample selection biases in linear regressions) to fit binary outcome dependent variables. In the transportation domain, sample selection models have been applied for various purposes, one of the most common of which is to correct for residential self-selection effects (Cao, 2009; Chen, Wu, Chen, Zegras, & Wang, 2017; van Herick & Mokhtarian, 2020). In that context, outcomes are observed for both "selected" and "unselected" groups. In other contexts, including ours, outcomes are only observed for "selected" cases – for us, the cases who self-select into both being willing to respond, and actually responding, to a follow-up survey (Alemi, Circella, Mokhtarian, & Handy, 2019; Sun, Wang, & Wan, 2019). In this study, we select the PSS model structure since it both fits our data structure (see Section 3) and matches the conceptual reasoning (see Section 4.1).

3. DATA DESCRIPTION

The National Household Travel Survey (NHTS) is a repeated cross-sectional travel survey conducted by the Federal Highway Administration, and is widely used by regional planning agencies across the United States. The Georgia subsample of the 2017 NHTS constitutes the survey dataset used for this study. The NHTS typically obtains household, individual, vehicle, and trip information using several survey instruments; these include a recruitment survey, a retrieval survey, travel logs, and a vehicle odometer mileage form. In 2017, for the first time, NHTS allowed states to opt into including a question regarding respondents' willingness to participate in follow-up travel surveys, and Georgia was one of the six states/regions that chose to do so. We segmented NHTS Georgia respondents based on their willingness to participate in a follow-up survey as well as their actual response behavior to the follow-up survey (see *Decisions 1 and 2* in Figure 1)[1]. The follow-up survey, denoted the GDOT survey in Figure 1, is further discussed later in this section.

[1] The NHTS public dataset is available at https://nhts.ornl.gov. Access to the *Decision* variables will be given upon request.

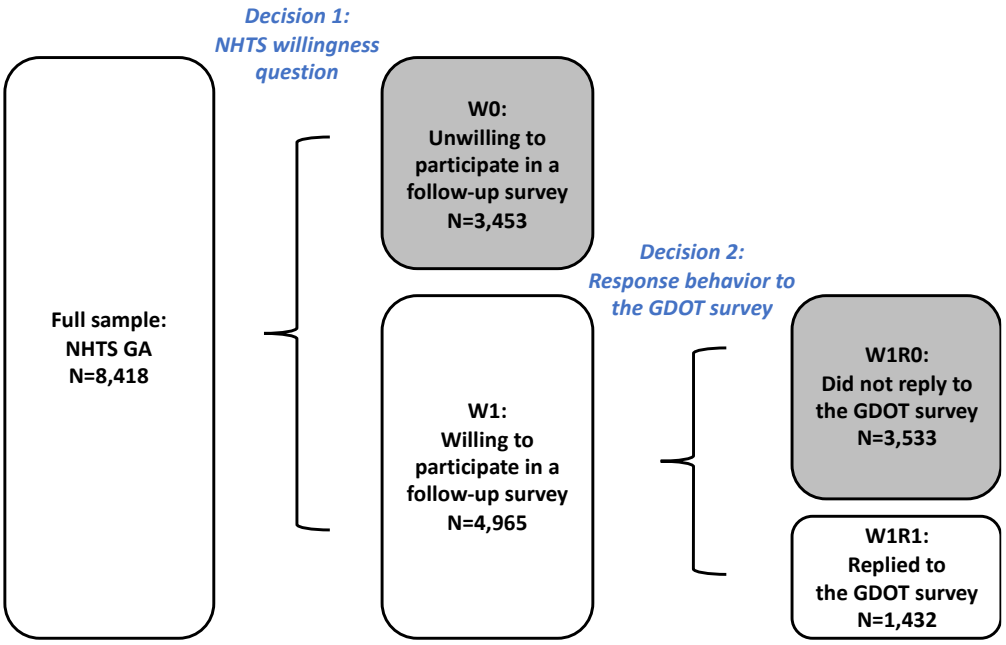


Figure 1 Data sources and structure of analysis

As shown in Figure 1, the first decision was made through the willingness question in the NHTS (i.e., "Would you be willing to participate in a follow-up survey?"). This question is *only* asked of the main household respondent (i.e., the respondent who answered household-related questions in the retrieval survey), and solely of those living in the regions (i.e., states or Metropolitan Planning Organization areas) that specifically requested the inclusion of this question, with Georgia being one of those regions as mentioned before. As such, we used only the main household respondents for analysis purposes, as we did not have additional information regarding other household members' willingness to participate in a follow-up survey. The final working dataset comprised 8,418 respondents, 4,965 of whom indicated a willingness to participate in a follow-up survey (W1), whereas the remaining 3,453 respondents did not want to be contacted again for future surveys (W0).

For the 4,965 NHTS respondents who indicated a willingness to participate in a follow-up survey, their second decision (Figure 1) was made through their actual response to a follow-up survey, the Georgia Department of Transportation-funded Emerging Technologies Survey (GDOT survey, Kim et al., 2019). The GDOT survey is a 15-page attitudinally-rich travel survey with an emphasis on the impacts of emerging technologies on travel behavior. Our research team mailed the GDOT survey to the 4,965 NHTS respondents in September 2017. The respondents could either mail the completed paper survey back using the postage-paid reply envelope we provided, or use the URL we also provided to complete the survey online. Ultimately, 1,432 of the 4,965 NHTS respondents replied to the GDOT survey (W1R1), while the remaining 3,533 did not reply (W1R0). Thus, at this point, we have segmented all 8,418 NHTS Georgia respondents based on the two decisions. We note that for the purpose of this report, the GDOT survey was used only to segment/classify respondents; all respondent data was obtained from the NHTS.

In Table 1, we present descriptive statistics for each segment and the overall sample. In the full sample, the average household size is 2.13, the average age is 55.6 years, and 53% of the sample is female. Overall, participants are highly educated, with 48% of the participants reporting they have a bachelor's degree or higher. Compared to respondents who are unwilling to be contacted (W0), respondents who are willing to be contacted for a follow-up survey (W1) tend to be younger (means of 54.35 versus 57.30 years). On average, the W1 segment conducts more trips on the selected travel day (4.16 versus 3.52 trips) and lives in denser areas (859.07 versus 769.92 housing unit per sq. mi.). Among the respondents willing to be contacted, those who replied to the GDOT survey (W1R1) tend to be older than those who did not reply (W1R0, 59.00 versus 52.46 years). The W1R1 segment conducts more trips (4.47 versus 4.03) on the selected travel days, and they come from less dense areas than other groups.

In the following sections, we separate the final working dataset (N=8,418) into a training set (60%, N=5,051) and a test set (40%, N=3,367) to enable appropriate model evaluation.

Table 1 Descriptive statistics of the working dataset (sample means/shares)

		Full sample:	W0: Unwilling to be contacted	W1: Willing to be contacted	W1R0: Willing but did not reply	W1R1: Willing and did reply
	Sample size	**8,418**	**3,453**	**4,965**	**3,533**	**1,432**
Household sociodemographic	Household size (persons)*	2.13	2.17	2.10	2.13	2.01
	Home ownership (yes)	0.75	0.80	0.71	0.66	0.84
Individual sociodemographic	Age*	55.56	57.30	54.35	52.46	59.00
	Has a medical condition (yes)	0.13	0.13	0.13	0.14	0.12
	Gender (female)	0.58	0.57	0.59	0.60	0.55
	Born in US (yes)	0.93	0.91	0.94	0.93	0.95
	Race: white (yes)	0.73	0.74	0.72	0.69	0.79
	Education†					
	Less than a high school graduate	0.038	0.043	0.035	0.041	0.022
	High school graduate or GED	0.19	0.20	0.17	0.18	0.15
	Some college or associates degree	0.30	0.29	0.30	0.30	0.30
	Bachelor's degree	0.24	0.23	0.24	0.24	0.26
	Graduate degree or professional degree	0.24	0.23	0.25	0.24	0.28
	Worker (yes)	0.54	0.52	0.56	0.59	0.48
Travel-related characteristics	No. of trips recorded in one-day travel diary *	3.90	3.52	4.16	4.03	4.47
	Transit usage frequency*[1]	0.64	0.40	0.81	0.95	0.46
Survey-related characteristics	Household income - missing value	0.035	0.064	0.015	0.016	0.011
	VMD - "I don't know"	0.25	0.32	0.20	0.21	0.17
	VMD - "I prefer not to answer"	0.015	0.025	0.009	0.009	0.008
Land use characteristics	Housing units per sq. mi.*	822.51	769.92	859.07	920.44	707.68

Notes:

[1] 0=Never; 1=Less than once a month; 2=1-3 times a month; 3=1-2 times a week; 4=3-4 times a week; 5=5 or more times a week.

* Treated as continuous variables for modeling; descriptive statistics are sample means.

† Treated as continuous variables for modeling; descriptive statistics are sample shares.

The remaining variables are binary variables. For simplicity, we only show sample shares of one category as indicated in the table.

All descriptive statistics are unweighted. We provide weighted distributions in Table 7, including population distributions based on the 2018 American Community Survey five-year estimates and the full NHTS Georgia sample.

4. METHODOLOGY

4.1 Model Structure and Application

As described in the last section, for this report we model and analyze two consecutive decisions made by the 2017 NHTS Georgia respondents: (1) their willingness to participate in a follow-up survey and (2) their actual response behavior to the follow-up survey. The perspective we take is that the target behavior of interest is the participation in the second survey *by anyone*, and the goal is to obtain consistent estimates of the coefficients of the explanatory variables in the model predicting that behavior. But since we are only able to observe the second decision for NHTS respondents who are willing to participate in a follow-up survey (i.e., respondents who are self-selected, and so received a follow-up survey), modeling the observed response behavior only of this subsample could produce biased (econometrically inconsistent) estimates of those coefficients, relative to their true values in the population at large.

To address the self-selection bias, Heckman (1976) proposed the sample selection model as a corrective method for linear regression models. Given the binary nature of the two decisions in our case (i.e., willing/unwilling to participate, respond/do not respond to the follow-up survey), we apply the analogous corrective method for discrete choice models, the probit with sample selection (PSS) model (van de Ven and van Praag, 1981), to deal with the self-selection bias.

In the PSS model, we have a selection model and an outcome model, which correspond to the willingness and response decisions, respectively. The selection and outcome models are defined as

$$y_i^{S^*} = \boldsymbol{z_i}\boldsymbol{\gamma} + \varepsilon_i^S, \tag{1}$$

$$y_i^{O^*} = \boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i^O, \tag{2}$$

$$y_i^S = \begin{cases} 0, & if\ y_i^{S^*} < 0 \\ 1, & otherwise \end{cases} \tag{3}$$

$$y_i^O = \begin{cases} unobserved, & if\ y_i^S = 0 \\ 0, & if\ y_i^S = 1\ and\ y_i^{O^*} < 0 \\ 1, & if\ y_i^S = 1\ and\ y_i^{O^*} \geq 0, \end{cases} \tag{4}$$

where $y_i^{S^*}$ is the continuous latent variable indicating the tendency for individual i to be willing to participate in a follow-up survey; $y_i^{O^*}$ is the tendency for individual i to respond to the follow-up survey (the GDOT survey); $\boldsymbol{z_i}$ and $\boldsymbol{x_i}$ are vectors of explanatory variables for the selection and outcome models, respectively; $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the corresponding coefficient vectors; and ε_i^S and ε_i^O are error terms that capture the unobserved effects in the two models. As is standard, we assume that the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{5}$$

In the observed choice formulations (Eqs. 3-4), y_i^S is the observed binary selection choice (willing to participate in a follow-up survey = 1, unwilling = 0), and y_i^O is the observed binary outcome choice (responds to the follow-up survey = 1, does not respond = 0). We observe the outcome if and only if the latent selection variable $y_i^{S^*}$ is positive (or y_i^S =1). Finally, we estimate the

parameters $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho}$ using maximum likelihood estimation. The log-likelihood can be written as

$$\ell\left(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho}\right) = \sum_{i:y_i^S=0} ln(\Phi(-\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}})) + \sum_{\substack{i:y_i^S=1,\\ y_i^O=1}} ln(\Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right)) + \sum_{\substack{i:y_i^S=1,\\ y_i^O=0}} ln(\Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, -\boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right)), \tag{6}$$

where $\Phi(\cdot)$ represents the cumulative univariate standard normal distribution function and $\Phi_2(\cdot)$ represents the cumulative bivariate normal distribution function. With this model formulation, we can calculate three sets of probabilities: the marginal probabilities of being willing or not (Eqs. 7-8), joint probabilities of being willing and responding or not responding (Eqs. 9-10), and conditional probabilities of responding or not, given willingness (Eqs. 11-12).

Marginal probabilities:

$$P\left(y_i^S = 0\right) = \Phi(-\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}) \tag{7}$$

$$P\left(y_i^S = 1\right) = \Phi(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}) \tag{8}$$

Joint probabilities:

$$P\left(y_i^S = 1, y_i^O = 0\right) = \Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, -\boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right) \tag{9}$$

$$P\left(y_i^S = 1, y_i^O = 1\right) = \Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right) \tag{10}$$

Conditional probabilities:

$$P\left(\, y_i^O = 0 \mid y_i^S = 1\right) = \Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, -\boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right)/\Phi(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}) \tag{11}$$

$$P\left(\, y_i^O = 1 \mid y_i^S = 1\right) = \Phi_2\left(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}\right)/\Phi(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}) \tag{12}$$

The three sets of probabilities reflect distinct statistical explanations, which should be appropriately used under different model applications. In Table 2, we summarize a few application scenarios and the corresponding probabilities, in the context of a two-stage survey sample recruitment. This study will mainly focus on the first application scenario while lightly touching on the third one in Section 6.2. It is worth mentioning here that, similar to any other model, prediction errors exist in the PSS model applications. We summarize several model performance measures in the next section to help evaluate the quality of the model.

Table 2 Applications of the PSS model in different scenarios

Scenario	Model and probability used in the *prediction*
1. Decomposition of the deviation (i.e., bias) of the follow-up survey sample from the population into its various components (e.g., dataset bias, self-selection bias, prediction errors). This is enabled by comparisons of the predicted sample and population distributions at various stages of the model.	• Use the selection model and the marginal probability of selection $P(y_i^S = 1)$ for the prediction of people who are willing to participate in a follow-up survey. • Use the joint model and joint probability of selection and outcome $P(y_i^S = 1, y_i^O = 1)$ for the final prediction of follow-up survey respondents.
2. Prediction of the response to a second-stage survey following a large-scale first-stage survey (e.g., NHTS) that contains the willingness question. Survey developers conduct a small-scale field test of the second-stage survey to enable the estimation of the PSS model, and then apply the *outcome model* to the remainder of the willing first-stage sample to predict the size and characteristics of the full-scale second-stage sample.	• Use the conditional probability $P(y_i^O = 1 \mid y_i^S = 1)$ to predict the second-stage response of the willing first-stage sample.
3. Prediction of the response to a second-stage survey following a large-scale first-stage survey (e.g., NHTS) that does *not* contain the willingness question. Survey developers do not know the response willingness of the first-stage sample, and adopt a PSS model estimated from other datasets / regions to predict the size and characteristics of the second-stage sample.	• Using a joint model estimated from other datasets, compute the joint probability $P(y_i^S = 1, y_i^O = 1)$ to predict the second-stage response from the full first-stage sample.

4.2 Model Performance Measures

Due to the two-level model structure of the PSS model, the usual discrete choice model performance measures cannot be directly applied, which might explain why PSS models have diverse performance measures in the literature. Accordingly, we aim to address the lack of clarity in the literature surrounding PSS measures by providing a resource for six frequently used categories of model measures, adjusted based on the PSS model structure: the log-likelihood, McFadden's pseudo R-squared, information criteria, correlation, root mean squared error, and success table. Table 3 provides definitions of the six measures, and gives examples of them being applied within the literature. We also demonstrate their use by calculating all of them for the PSS model developed in this report in Section 5.2.

Since both selection and outcome models are binary probit models, we first introduce the log-likelihoods for three models associated with the PSS model: the equally-likely (EL) model, market-share (MS) model, and full model (Eqs. 13-15). Log-likelihoods provide direct measures of the model performance, but they do not allow model comparisons across studies since the values are related to the sample size. McFadden's pseudo R-squared (ρ^2) provides a measure that is derived from the log-likelihoods but is bounded between 0 and 1. A higher ρ^2 means greater information explained by the model (Mokhtarian, 2016). Eqs. 16 and 17 are ρ^2s with EL and MS bases, respectively. Information criteria such as the Akaike information criterion (AIC, Eq. 18) and Bayesian information criterion (BIC, Eq. 19) are also based on log-likelihoods. These criteria

penalize the number of model coefficients to promote parsimony, which could be used for model selection. However, similar to the drawback of log-likelihoods, we do not have a benchmark for such information criteria. The three log-likelihood-associated categories of measures are suitable when the overall PSS model performance is required, such as for Scenarios 1 and 3 in Table 2.

Another model performance measure is the correlation coefficient between predicted probabilities and observed choices. Since the observed choice is a binary variable and the predicted probability is a continuous variable, we apply point-biserial correlation coefficients (Eq. 20), which range between -1 (the wrong outcome is predicted with certainty) and 1 (the correct outcome is predicted with certainty). The closer r_{pb} is to 1, the better the model. Root mean squared error (RMSE) measures the (square root of the) average squared discrepancy between the observed choice (0 or 1) and the predicted probability (Eq. 21). For our model, RMSE ranges between 0 and 1, with smaller RMSE indicating better prediction results. Although the correlation and RMSE measures do not provide an overall measure of the PSS model but only measure separate model performances of the selection and outcome models, they are instrumental under specific application scenarios. For example, in the bias decomposition application (Scenario 1 in Table 2), separate performance measures provide comparable prediction error indicators between selection and outcome models as we decompose biases step by step (see Section 6.1 for more details). Separate model performance measures are also useful when we only need the performance of a single model (e.g., the outcome model performance with known selection results, Scenario 2 in Table 2).

The last model performance measure category is the probability-based success table, which was originally proposed by McFadden (2000). Given the two-level model structure of the PSS model, we could generate a 3×3 matrix based on the observation and model prediction results ($y_i^S = 0$; $y_i^S = 1\ and\ \ y_i^O = 0$; $y_i^S = 1\ \ and\ y_i^O = 1$). Eq. 22 calculates the number of cases in the mn^{th} cell in a success table. Success tables allow both overall model performance measures (i.e., overall prediction accuracy) and alternative-specific measures (i.e., success proportion, success index). Success tables are usually computed for both training and test sets to examine the generalizability of the model.

Table 3 Model performance measures for probit with sample selection models

Measure	Formula	Eq.	PSS examples
Log-likelihood	Equally-likely model: $\ell(\mathbf{0}) = \sum_{i:y_i^s=0} ln\left(\frac{1}{2}\right) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=1}} ln\left(\frac{1}{2}\times\frac{1}{2}\right) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=0}} ln\left(\frac{1}{2}\times\frac{1}{2}\right)$ $= -N_{y_i^s=0}\, ln\, 2 - \left(N_{y_i^s=1,y_i^o=1} + N_{y_i^s=1,y_i^o=0}\right) ln\, 4$	(13)	
	Market share model: $\ell(\boldsymbol{c}) = \sum_{i:y_i^s=0} ln\left(\frac{N_{y_i^s=0}}{N}\right) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=1}} ln\left(\frac{N_{y_i^s=1,y_i^o=1}}{N}\right) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=0}} ln\left(\frac{N_{y_i^s=1,y_i^o=1}}{N}\right)$ $= N_{y_i^s=0} \ln N_{y_i^s=0} + N_{y_i^s=1,y_i^o=1}\, ln\, N_{y_i^s=1,y_i^o=1} + N_{y_i^s=1,y_i^o=0}\, ln\, N_{y_i^s=1,y_i^o=0} - N\, ln\, N$	(14)	Ruiz and Habib (2016), Stavropoulou (2011)
	Full model: $\ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho}) = \sum_{i:y_i^s=0} ln(\Phi(-\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}})) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=1}} ln(\Phi_2(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, \boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho})) + \sum_{\substack{i:y_i^s=1,\\ y_i^o=0}} ln(\Phi_2(\boldsymbol{z_i}\widehat{\boldsymbol{\gamma}}, -\boldsymbol{x_i}\widehat{\boldsymbol{\beta}}; \hat{\rho}))$	(15)	
McFadden's pseudo R-squared[1]	$\rho_{EL}^2 = 1 - \frac{\ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho})}{\ell(\mathbf{0})}$	(16)	Drucker and Khattak (2000)
	$\rho_{MS}^2 = 1 - \frac{\ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho})}{\ell(\boldsymbol{c})}$	(17)	
Information criteria	$AIC = 2k - 2\ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho})$	(18)	Alemi et al. (2019)
	$BIC = \ln(N)\, k - 2\ell(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}, \hat{\rho})$	(19)	
Point-biserial correlation coefficient	$r_{pb} = \frac{m_1 - m_0}{s}\sqrt{\frac{n_1 n_0}{n^2}}$ where m_1 and m_0 are the average probabilities for the binary alternatives; s is the standard deviation of the probabilities for all cases; n_1 and n_0 are the number of cases for each alternative; n is the sum of n_1 and n_0. Selection models use the marginal probability to calculate m_1 and m_0 with the full dataset. Outcome models use the conditional probability to calculate m_1 and m_0 with observed selected samples only.	(20)	Van de Ven and Van Praag (1981) (similar idea)
Root mean squared error (RMSE)	$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{p}_i)^2}{n}}$ where y_i is the observed choice; $\hat{p}_i$ is the predicted probability of that choice (uses marginal probability for the selection model, and conditional probability for the outcome model); n is the number of cases.	(21)	-

Success table	$N_{ab} = \sum_i I_i^a \hat{p}_i^b$ where N_{ab} is the expected number of cases whose observed choice is a and predicted choice is b; I_i^a is an indicator function which equals 1 when the observed choice of case i corresponds to a, and equals 0 otherwise; and $\hat{p}_i^b$ is the predicted probability for case i to choose b.	(22)	-

[1] Note that in this case, "equally likely" means that the two alternatives for each of the two models are equally likely, not that the three possible final combinations ($y_i^s = 0;\ y_i^s = 1, y_i^o = 1; and\ y_i^s = 1, y_i^o = 0$) are equally likely. That is, the respective probabilities for those three events are ½, ¼, ¼, not 1/3, 1/3, 1/3.

5. RESULTS

In this section, we first present the PSS model result (Table 4) and then measure the model performance with the six metrics presented in the previous section (Table 5).

5.1 Model Results

(i) Selection model
The selection model explains respondents' willingness to participate in a follow-up survey. We organized the explanatory variables into three categories: household- and individual-level sociodemographic characteristics, travel-related characteristics, and survey-related characteristics (Table 4).

Among the household-level sociodemographic characteristics tested, we see that respondents from larger households are less willing to participate in a follow-up survey compared to respondents from smaller households; we propose that one reason for this finding may reside in the format of the NHTS. Specifically, NHTS requires all household members five years of age or older to complete the personal section in the retrieval survey and record their travel on the designated travel day. As such, it is more time-consuming and burdensome for larger households to complete the NHTS requirements, which may weaken the motivation of the main household respondent to volunteer for another survey. Furthermore, the log transformation of household size indicates that the impact on survey willingness of a one-person increase in household size becomes weaker (but still negative) as the household size grows. The model also shows that homeowners are less willing to participate in a follow-up survey, perhaps because of the greater value of time for those who own homes. On the one hand, moderate correlations between homeownership and vehicle ownership (0.37), and between homeownership and household income (0.36), suggest that the homeownership variable may be considered a proxy indicator of middle-to-high-income households. Respondents from such households would have higher values of time and thus be less willing to take follow-up surveys. On the other hand, individuals who own homes tend to be at different life stages relative to those who rent, and as such have a higher value of time (e.g., a later career stage with more demands on their time)[2]. Respondents from such households would have higher values of time and thus be less willing to take follow-up surveys.

Among individual-level sociodemographic characteristics, we find that younger people, women, and people who were born in the U.S. are more willing to participate in a follow-up survey. We also find that individuals who have a medical condition restricting them from traveling outside the home are more willing to participate than people who do not have such restrictions. On the one hand, the travel-limited group comprises primarily older individuals who may be retired and thus have more time for doing surveys. The results may also reflect the altruism of the travel-limited group, possibly suggesting that they seek to contribute to society in ways that are accessible to them. On the other hand, their interest and participation in travel-related surveys may also highlight the unmet travel demands of these individuals.

[2] We also investigated whether the presence of children might be a source of time poverty, but the correlation between homeownership and household size was only 0.06.

Among travel-related characteristics tested, the model shows that people who report more trips on the designated travel day are more willing to participate in a follow-up survey, which runs counter to our expectations. Based on the findings regarding household size, we conjectured that having to record more trips would reduce the willingness to participate in a follow-up survey. A resolution of the paradox might reside in the individual's liking for travel. Specifically, travel-liking people might record their travel logs more comprehensively (e.g., walk one block to buy coffee in the middle of the workday, pick up dry cleaning on the way back home), and also be eager to complete a future travel survey[3]. In contrast, those reporting fewer trips might tend to ignore trivial, non-mandatory, short trips or stops because they are not sensitive enough to catch these trips and/or they want to alleviate the burden of completing the travel logs. Alternatively, even without especially liking traveling, heavy travelers may still be interested in the subject precisely because it is such a big part of their lives. Accordingly, they may be more likely than others to express willingness to be surveyed again, whether or not they are too busy traveling to actually respond when the invitation comes. Moreover, frequent transit users are also more willing to participate in a follow-up survey, which might be due to their desire to improve the quality of their travel experience by providing feedback through travel surveys.

Survey-related characteristics constitute a group of variables unique to the selection model: item non-responses. In NHTS, many questions provide choices of "I don't know" and "I prefer not to answer", which allows respondents to protect their privacy for sensitive information (e.g., income) and avoid imprecise estimations (e.g., vehicle-miles driven, VMD). In our model, we combine "I don't know" and "I prefer not to answer" for the household income question and treat both of these responses as indicative of respondents who choose to protect their privacy. The resultant variable is called the household income missing value indicator, and the negative sign of the coefficient implies that respondents who are more protective of their privacy are less willing to participate in a follow-up survey[4]. Regarding VMD, since the variable is self-estimated by NHTS respondents, we believe some respondents who do not care much about their travel might be unclear about their annual VMD. As such, "I don't know" may represent an apathetic attitude toward travel, whereas "I prefer not to answer" reflects a privacy-protective attitude, and accordingly we keep those responses separate for VMD. The model shows that both respondents who are less interested in their travel behavior and respondents who are protective of their privacy regarding travel behavior, are less willing to respond to a follow-up survey.

(ii) Outcome model

The outcome model explains the actual, observed response to the GDOT survey for NHTS respondents who reported being willing to participate in a follow-up survey. The outcome model contains two groups of explanatory variables: household- and individual-level sociodemographic characteristics, and land use characteristics.

[3] Since the NHTS did not measure travel-liking attitudes, we could not test our hypothesis with the presented PSS model. However, to investigate this conjecture we constructed a binary probit model for respondents' willingness to participate in a follow-up survey using the GDOT survey data, which measured respondents' willingness to participate in yet another follow-up survey as well as the travel-liking attitude. Results indicated that the travel-liking attitude positively associated with the willingness to participate at a significance level (p-value) of 0.001.

[4] When we treated the two responses ("I don't know" and "I prefer not to answer") as separate variables, their coefficients were very similar.

Homeownership is the household-level sociodemographic characteristic that was found to be significant in both the selection and outcome models. Interestingly, however, the variable has opposing signs in the two models. Specifically, homeowners were less willing to participate in a follow-up survey than the renters, but among respondents who *are* willing to participate in a follow-up survey, homeowners are *more* likely to respond than renters. One reason for the latter outcome may be that homeowners are more likely to receive the follow-up survey because they move less often, whereas the follow-up survey might not reach renters due to address changes. However, we do not have reliable records of everyone who had moved and thus did not receive the GDOT survey invitation. Another reason might be that homeowners were initially less willing to commit their time to a follow-up survey due to having more household responsibilities, but once opting in, the same commitment to one's responsibilities makes them more likely to follow through.

Age and medical conditions are individual-level sociodemographic characteristics that are significant in both selection and outcome models, albeit also with opposing signs. In general, younger people report being more willing to participate in a follow-up survey compared to older people, while among respondents expressing willingness to participate in a follow-up survey, older people are more likely to actually respond than younger people. Potentially, younger people are less reachable (i.e., more transient) or less able to participate when the time actually comes, even though they may aspire to be helpful. As previously discussed, respondents with travel-restricting medical conditions are more willing to participate in a follow-up survey compared to respondents who do not have such restrictions. However, among people willing to participate in a follow-up survey, medically-restricted respondents are less likely to respond than people who do not have any travel restrictions. It is possible that the medical conditions that restrict travel might also limit these respondents from completing the follow-up survey (e.g., poor eyesight); it is also possible that the medical conditions worsened during the approximately one-year interval between surveys[5]. The outcome model also shows that white, higher-educated people are more likely to respond to the follow-up survey, while workers are less likely to respond to the follow-up survey than non-workers, probably due to time constraints on the part of the worker group.

The land use characteristics are the variable group unique to the outcome model, as they were only found to be significant in this model. We find that people from less dense areas are more likely to respond to the follow-up survey, which could be related to the types of individuals who typically live in lower density areas in Georgia (e.g. older, more likely to be retired)[6].

[5] The 2017 NHTS was conducted between April 2016 and May 2017. The GDOT survey was distributed in September 2017. Accordingly, the interval between the two surveys varies from 4 months to 1.5 years, but we do not know the specific gap for a given individual, since the date of completion of the NHTS survey was not provided with the data.
[6] We checked the correlations of housing density with the home ownership (-0.18), household size (-0.11), age (-0.13), and worker (0.077) variables, but none of them were large enough to cause collinearity concerns.

Table 4 Probit with sample selection model results (N=5,051)

Variables	Coefficient	Std. Error
Selection model: willingness to participate in a follow-up survey		
Household sociodemographic		
Household size (log transformed)	-0.185***	0.0377
Homeowner	-0.178***	0.0469
Individual sociodemographic		
Age	-0.00726***	0.00139
Has a medical condition	0.150*	0.0581
Female	0.111**	0.0369
Born in US	0.194**	0.0694
Travel-related characteristics		
No. of trips	0.0478***	0.00629
Transit usage frequency	0.0579*	0.0230
Survey-related characteristics		
Household income - missing	-0.857***	0.106
VMD - "I don't know"	-0.464***	0.0424
VMD - "I prefer not to answer"	-0.796***	0.140
Constant	0.188*	0.0852
Outcome model: response to the follow-up survey		
Household sociodemographic		
Homeowner	0.417***	0.0606
Individual sociodemographic		
Age	0.0120***	0.00178
Has a medical condition	-0.331***	0.0733
Race: white	0.106*	0.0534
Education	0.0746***	0.0215
Worker	-0.181***	0.0540
Land use characteristics		
Housing units per sq. mi.	-0.0528*	0.0246
Constant	-0.619***	0.129
Error terms correlation		
ρ	-0.574***	0.0964

*** Coefficient is statistically significant at the 0.001 level.
** Coefficient is statistically significant at the 0.01 level.
* Coefficient is statistically significant at the 0.05 level.
Note: Insignificant variables removed from the model include no. of vehicles per driver in the household, no. of children in the household, frequency of walk trips, and usage of delivery services, among others.

(iii) Error terms

The correlation of the error terms in the selection and outcome models is highly significant and sizable (-0.574), which indicates that the self-selection bias (expressed willingness to participate in a follow-up survey) significantly influences whether or not an individual responds to a follow-up survey. Specifically, its negative value signifies that on net, unobserved characteristics that *increase* the reported willingness to participate in a follow-up survey will tend to *decrease* the

tendency to actually do so. Or conversely, unobserved factors that decrease the reported willingness (e.g., a sense of responsibility leading one to count the cost before agreeing to do something) might be the same factors that influence respondents to keep the commitment once they opt in to the follow-up survey. Having already seen this pattern from the three *observed* explanatory variables with opposing signs in the selection and outcome models (i.e., homeownership, age, and medical condition), it is not hard to imagine that it could prevail among *unobserved* variables as well.

5.2 Model Performance Measures

In this section, we apply model performance measures from the six categories proposed in Section 4.2 to our PSS model. Table 5 presents measures from the first five categories including log-likelihood, McFadden's pseudo R-squared, information criteria, correlation, and root mean squared error. The success table is presented in Table 6.

As discussed previously, we cannot compare log-likelihoods and information criteria with models in other studies due to the varying sample sizes, whereas McFadden's pseudo R-squareds are comparable given their 0 to 1 range. In this study, McFadden's pseudo R-squareds are relatively low, which could result from the nature of predicting survey participation. The willingness to participate in a follow-up survey and the actual response also depend on people's mood and time pressure *at the moment*, which are unobserved in our dataset but may explain a large share of the variability in the dependent variables. In the literature, the model fits regarding survey willingness and actual response are similar to ours. For example, Wittwer and Hubrich (2015) developed a binary logistic regression model of survey response behaviors and McFadden's pseudo R-squared was 0.052 (relative to the constant-only model benchmark). Regarding an internet survey, Couper et al. (2007) obtained Cox and Snell pseudo R-squareds of 0.044 and 0.067 for the willingness and response models, respectively[7].

Table 5 Probit with sample selection model measures (N=5,051)

Measure	Formula	Value
Log-likelihood	$\ell(\mathbf{0})$	-5571.517
	$\ell(\boldsymbol{c})$	-5231.426
	$\ell(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, \hat{\rho})$	-4921.783
McFadden's pseudo R-squared	ρ^2_{EL}	0.117
	ρ^2_{MS}	0.059
Information criteria	AIC	9885.567
	BIC	10022.640
Point-biserial correlation coefficient	r_{pb}	r_{pb}(selection model) = 0.274
		r_{pb}(outcome model) = 0.271
Root mean squared error (RMSE)	$RMSE$	$RMSE$ (selection model) = 0.473
		$RMSE$ (outcome model) = 0.439

The last model performance measure is the probability-based success table. As shown in Table 6, the bolded numbers on the diagonal represent the number of correct predictions, while the off-

[7] To enable the comparison between our PSS model and the two single models in Couper et al. (2007), we calculate the Cox and Snell pseudo R-squared with the formula $1 - \left(\frac{L(\boldsymbol{c})}{L(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, \hat{\rho})}\right)^{2/N}$, and the value is 0.115.

diagonal elements are the number of misclassifications. Based on the success table, we calculate overall prediction accuracy (sum of the diagonal elements divided by the total, which is 0.41 for the training set) and the alternative-specific accuracy (i.e., success proportion). Specifically, a *success proportion* is the number of correct predictions of a specific choice divided by the total number of predictions of that choice. For example, 45% of the people who are predicted to be unwilling to participate in a follow-up survey (y_i^S=0) actually do not want to participate in a follow-up survey. We could further normalize success proportions by the corresponding observed shares to obtain *success indices*, which directly compare the performance of the calibrated model with the market-share prediction for each alternative. In general, we expect the success index to be greater than 1, signifying superiority of the final model over the market-share model. Larger success indices indicate more accurate predictions. For example, our model is respectively 1.11, 1.10, and 1.21 times better than the market-share model in predicting the three outcomes. Table 6(b) is the success table based on the test set. Recall that we separated the final working dataset (N=8,418) into a training set (60%, N=5,051) and a test set (40%, N=3,367) to enable appropriate model evaluation. In general, the PSS model has quite similar performances in the training and test sets, which indicates good generalizability of the model to "new" data drawn from the same context.

Table 6 Success table

(a) Training set

	Pred. (y_i^S=0)	**Pred.** (y_i^S=1, y_i^O=0)	**Pred.** (y_i^S=1, y_i^O=1)	**Row total**	**Obs. share**
Obs. (y_i^S=0)	**935.16**	787.90	340.94	2064	0.41
Obs. (y_i^S=1, y_i^O=0)	787.89	**957.84**	356.27	2102	0.42
Obs. (y_i^S=1, y_i^O=1)	340.99	355.92	**188.10**	885	0.18
Column total	2064.04	2101.66	885.31	5051	
Pred. share	0.41	0.42	0.18		
Success prop.	0.45	0.46	0.21		**Acc.= 0.41**
Success index	1.11	1.10	1.21		

(b) Test set

	Pred. (y_i^S=0)	**Pred.** (y_i^S=1, y_i^O=0)	**Pred.** (y_i^S=1, y_i^O=1)	**Row total**	**Obs. share**
Obs. (y_i^S=0)	**630.51**	531.05	227.44	1389	0.41
Obs. (y_i^S=1, y_i^O=0)	536.72	**652.80**	241.47	1431	0.43
Obs. (y_i^S=1, y_i^O=1)	216.26	216.85	**113.89**	547	0.16
Column total	1383.49	1400.70	582.80	3367	
Pred. share	0.41	0.42	0.17		
Success prop.	0.46	0.47	0.20		**Acc.= 0.41**
Success index	1.10	1.10	1.20		

Note: Calculations contain rounding errors.

6. PSS MODEL VALIDATION AND APPLICATION

In this section, we will first apply the PSS model to the hold-out NHTS Georgia sample (the test set) to further validate our model results (Parady, Ory, & Walker, 2021) and retrieve sample biases in the follow-up survey from multiple sources (Scenario 1, Table 2). We will then apply the PSS model to selected states in diverse geographic regions of the US (California, Massachusetts,

Minnesota, North Carolina, and New York) and the full 2017 NHTS national sample, to predict follow-up survey participation and test the transferability of the PSS model (Scenario 3, Table 2).

6.1 Inside Georgia: Breakdown of sample biases

In this section, we apply the PSS model to the test set to predict respondent participation in the follow-up survey, and compare the marginal distributions of several selected variables with the corresponding population[8] distributions derived from the 2018 American Community Survey five-year estimates (https://www.census.gov/programs-surveys/acs). By analyzing the distribution divergence between the follow-up survey respondents and the population, we summarize the potential biases residing in the sampling method, i.e., recruiting respondents from a preceding travel survey. Figure 2 visualizes the five bias sources: dataset bias, household representative bias, self-selection bias, non-response bias, and prediction error. Please see Table 7 for detailed distributions.

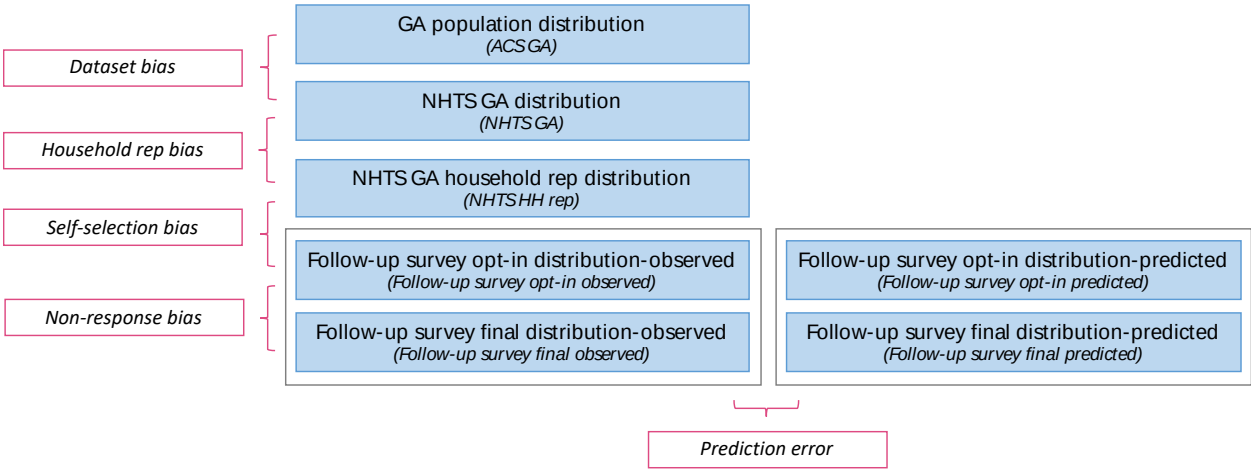


Figure 2 Distribution bias breakdown

The PSS model has demonstrated the existence of self-selection biases through the highly significant and sizable correlation between the error terms in the selection and outcome models. Self-selection bias, however, is not the only source that contributes to the marginal distribution divergence between the follow-up survey respondents and the population (i.e., the bias in the follow-up survey respondents). As shown in Figure 2, the first contribution arises from any coverage, sampling, and non-response biases associated with the dataset of the preceding survey, which is the 2017 NHTS in our case. Since the 2017 NHTS created individual and household weights using the 2015 ACS data as control variables, and since we used the 2018 ACS data to determine the "true" population distribution[9], the dataset bias associated with those control variables is trivial (columns 1 and 2 in Table 7).

[8] Although we refer to these as "population" distributions for convenience and because they presumably closely approximate the true distributions, they are in fact based on samples, and accordingly the ACS data has been weighted by the U.S. Census Bureau to correct for sampling and other biases.

[9] The 2015 (5-year) ACS estimates were the most recent available when the NHTS was administered in 2016-2017. However, the 2018 ACS provided the most recent 5-year estimates when we conducted the analysis. Since the latter involve data from 2014 to 2018, we expect them to provide a good estimate for the middle two years (2016-2017) during which the data for both surveys was collected.

The second contribution to bias comes from the fact that only people who answer the household-related questions in the retrieval survey – i.e., "household representatives (reps)" – are asked the willingness question in the NHTS. The follow-up survey (i.e., the GDOT survey) was therefore delivered only to household representatives and not to any other household members. The household representative filter results in individual-level biases (e.g., age, gender). The household-level variables are not influenced since household weights are the same across household members. Consequently, the marginal distributions of individual-level variables have sizable differences between the 2017 NHTS Georgia sample and the household representative sample (columns 2 and 3 in Table 7). If the household representative filter could be removed (i.e., if the willingness question were asked of all NHTS respondents), we would expect a more representative follow-up survey sample (see Appendix A for details of a scenario that simulates this hypothetical situation, with results that support the conjecture).

The distribution divergence between NHTS household representatives and individuals who are willing to participate in a follow-up survey (opt-in) reflects the self-selection bias (columns 3 and 4 in Table 7). The distribution divergence between the opt-in individuals and individuals who actually complete the follow-up survey reflects a non-response bias (columns 4 and 6), which might result from multiple reasons, such as the opt-in individual being no longer willing or able to do the follow-up survey at the time when it was received, or the follow-up survey not reaching the opt-in individual due to an address change.

The distribution divergence between the observed follow-up survey final respondents and the corresponding PSS predicted results indicates the prediction error (columns 4 versus 5 and columns 6 versus 7 in Table 7).

Table 7 Marginal distributions of selected variables

(a) Individual-level

Column number	1	2	3	4	5	6	7	8	9
Dataset	*ACS GA*[1]	*NHTS GA*[2]	*NHTS HH reps*[2]	*Follow-up survey opt-in observed*[2]	*Follow-up survey opt-in predicted*[2†]	*Follow-up survey final observed*[2]	*Follow-up survey final predicted*[2‡]	*Percent change*[3]	*Effect size*[3]
Age									
18-24	0.13	0.13	0.043	0.046	0.053	0.018	0.025	-0.81	0.43**
25-34	0.18	0.17	0.16	0.18	0.18	0.09	0.11	-0.37	
35-44	0.18	0.19	0.20	0.22	0.20	0.16	0.17	-0.04	
45-54	0.18	0.17	0.20	0.19	0.20	0.22	0.20	0.12	
55-64	0.16	0.17	0.20	0.21	0.19	0.26	0.23	0.42	
65+	0.17	0.17	0.20	0.17	0.17	0.25	0.26	0.52	
Gender									
Male	0.48	0.48	0.41	0.41	0.42	0.45	0.44	-0.08	0.08
Female	0.52	0.52	0.59	0.59	0.58	0.55	0.56	0.07	
Education									
Less than a high school graduate	0.062	0.070	0.051	0.042	0.052	0.019	0.038	-0.38	0.61***
High school graduate or GED	0.36	0.25	0.20	0.18	0.20	0.19	0.17	-0.53	
Some college or associates degree	0.30	0.30	0.31	0.32	0.31	0.27	0.29	-0.02	
Bachelor's degree	0.17	0.21	0.24	0.24	0.23	0.26	0.24	0.36	
Graduate degree or professional degree	0.10	0.17	0.21	0.21	0.21	0.26	0.26	1.53	
Worker	0.59	0.62	0.63	0.62	0.62	0.59	0.56	-0.05	0.06
Hispanic	0.078	0.083	0.075	0.066	0.073	0.058	0.058	-0.26	0.08
Asian/Pacific Islander	0.046	0.039	0.034	0.025	0.026	0.0090	0.017	-0.63	0.14*
Black	0.31	0.31	0.32	0.35	0.33	0.26	0.26	-0.18	0.12*
Native American	0.0090	0.0036	0.0037	0.0017	0.0033	0.0034	0.0028	-0.69	0.07
White	0.62	0.63	0.62	0.60	0.62	0.70	0.70	0.13	0.17*
Commute mode									
Private vehicle	0.94	0.93	0.92	0.90	0.91	0.96	0.93	-0.02	0.16*
Taxi	0.0030	0.0077	0.0050	0.0079	0.0091	0.0017	0.0059	0.97	
Public transit	0.022	0.032	0.041	0.055	0.042	0.017	0.032	0.44	
Walk	0.016	0.018	0.022	0.022	0.021	0.007	0.018	0.13	
Bike	0.0025	0.0065	0.0090	0.013	0.012	0.015	0.0085	2.40	
Other	0.013	0.0074	0.0050	0.005	0.007	0.00	0.0077	-0.41	
Commute time									
0-10 min	0.21	0.21	0.21	0.22	0.21	0.18	0.21	-0.01	0.17*
10-20 min	0.30	0.26	0.27	0.27	0.27	0.25	0.27	-0.11	
20-30 min	0.21	0.20	0.19	0.18	0.18	0.18	0.17	-0.16	

30-60 min	0.23	0.27	0.27	0.27	0.27	0.34	0.28	0.21	
60-90 min	0.033	0.048	0.047	0.052	0.044	0.036	0.050	0.53	
90+ min	0.015	0.026	0.017	0.019	0.020	0.015	0.020	0.28	

(b) Household-level

Column number	1	2	3	4	5	6	7	8	9
Dataset	*ACS GA*[4]	*NHTS GA*[5]	*NHTS HH reps*[5]	*Follow-up survey opt-in observed*[5]	*Follow-up survey opt-in predicted*[5†]	*Follow-up survey final observed*[5]	*Follow-up survey final predicted*[5‡]	*Percent change*[3]	*Effect size*[3]
Household size									
1	0.27	0.28	0.28	0.31	0.32	0.31	0.31	0.14	0.12*
2	0.33	0.33	0.33	0.29	0.31	0.33	0.35	0.05	
3+	0.40	0.39	0.39	0.40	0.37	0.36	0.34	-0.14	
Household income									
Less than $24,999	0.22	0.27	0.27	0.29	0.29	0.22	0.23	0.04	0.08
$25,000 to $49,999	0.23	0.23	0.23	0.23	0.23	0.25	0.22	-0.08	
$50,000 to $74,999	0.18	0.16	0.16	0.14	0.15	0.17	0.17	-0.06	
$75,000 to $99,999	0.12	0.11	0.11	0.11	0.10	0.09	0.12	-0.05	
$100,000 to $149,999	0.13	0.14	0.14	0.13	0.13	0.16	0.15	0.14	
More than $150,000	0.11	0.086	0.086	0.10	0.10	0.12	0.12	0.05	
Vehicle ownership									
0	0.067	0.078	0.078	0.091	0.092	0.040	0.062	-0.07	0.10*
1	0.33	0.35	0.35	0.37	0.37	0.36	0.34	0.03	
2	0.38	0.34	0.34	0.33	0.32	0.36	0.34	-0.10	
3+	0.22	0.23	0.23	0.20	0.22	0.24	0.26	0.14	
Homeowner	0.63	0.62	0.62	0.57	0.58	0.75	0.75	0.20	0.25*
Number of children									
0	0.70	0.68	0.68	0.67	0.69	0.75	0.73	0.04	0.10*
1	0.13	0.13	0.13	0.14	0.13	0.10	0.13	-0.02	
2	0.11	0.12	0.12	0.12	0.12	0.11	0.11	-0.03	
3+	0.060	0.063	0.063	0.06	0.05	0.037	0.038	-0.37	

Notes: For each variable, the sum of category shares might not equal 1 due to rounding errors.

[1] 2018 ACS individual weights are applied.

[2] NHTS individual weights, based on the 2015 ACS individual weights, are applied.

[3] Comparison between the population distribution and follow-up survey predicted distribution (columns 1 and 7).

[4] 2018 ACS household weights are applied.

[5] NHTS household weights are applied.

* Small effect size ($w = 0.10$). ** Medium effect size ($w = 0.30$). *** Large effect size ($w = 0.50$).

† Calculated with $P(y_i^S = 1)$. ‡ Calculated with $P(y_i^S = 1, y_i^O = 1)$.

Beyond the bias breakdown, the sum of all biases and errors shown in Figure 2, which indicates the distribution divergence between the population and the predicted follow-up survey respondents, is of the most concern[10]. A small distribution divergence indicates that the follow-up survey sample is expected to be representative of the population, which is a positive sign that recruiting respondents from a preceding survey is efficient and reasonable. Otherwise, a large divergence indicates that a biased follow-up survey sample is expected, which may call for some sampling remedies to improve its representativeness. Accordingly, in Table 7, we present the percentage change (column 8) and effect size (ES, column 9) between the population (column 1) and the predicted follow-up survey respondents (column 7). The definition of ES (w) is as follows (Cohen, 1977):

$$w = \sqrt{\sum_{i=1}^{m} \frac{\left(P_{prd(i)} - P_{pop(i)}\right)^2}{P_{pop(i)}}}, \tag{23}$$

where m is the number of variable categories; $P_{prd(i)}$ is the predicted proportion of category i in the follow-up survey (Table 7, column 7); $P_{pop(i)}$ is the actual proportion of category i in the population (Table 7, column 1). In general, a smaller ES indicates similar distributions. Cohen (1977) provides references for ES magnitudes: effect sizes of 0.10, 0.30, and 0.50 are considered as small, medium, and large, respectively.

Among the individual-level variables (Table 7a), the distributions of education and age in the follow-up survey samples diverge most widely from the corresponding population distribution. Specifically, the follow-up survey respondents overrepresent highly educated and older groups. In the case of education, we see that the bias begins with the original set of NHTS respondents, and is amplified at the second stage of predicted response to the GDOT survey. The two commute-related variables show that we have a larger share of follow-up survey respondents who use non-private vehicles for commuting compared to the population, which might further contribute to the larger share of long commute times. The effect sizes of the household-level variables have overall smaller magnitudes than those of the individual-level variables (Table 7b). Homeownership has the largest effect size of 0.25. Specifically, the follow-up survey recruits a larger share of homeowners, which might relate to the survey mode (mailing) used for the follow-up survey: homeowners are more likely to receive the survey since they have permanent mailing addresses, while renters might not receive the follow-up survey due to address changes.

In Appendix B, we provide a visualization of selected variables shown in Table 7. The visualization presents the changing trajectories of the marginal distributions from the population to the predicted follow-up survey respondents.

6.2 Outside Georgia: What does the follow-up survey sample look like?

In this section, we test the transferability of the PSS model to different populations, by checking the representativeness of follow-up survey respondents for selected states in diverse geographic

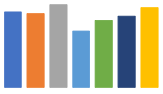
[10] The distribution divergence between the population and the *observed* follow-up survey respondents is of interest in an *ex post* analysis, but here we focus on *ex ante* applications of the PSS such as those in Scenarios 2 and 3 of Table 2. The distribution divergence metrics between the population and the *predicted* follow-up survey respondents could serve as benchmarks in Section 6.2.

regions of the US (west to east: California, Minnesota, North Carolina, New York and Massachusetts) and the full 2017 NHTS national sample. Table 8 presents the effect size by state.

In general, different regions have similar effect sizes for a given variable, which indicates a similar divergence level of the marginal distributions between the follow-up survey respondents and the populations in different regions. In that respect, the results show respectable generalizability of the PSS model across different areas. Nevertheless, the effect sizes do vary by state, which might point to regional differences that are not captured by the current PSS model. Moreover, the variations in effect size are not consistent across variables. For example, New York has the most representative follow-up survey sample regarding gender among the seven regions, but is the least representative on commute mode, household vehicles, and homeownership. Some of these large effect sizes of New York doubtless result from its diverse population composition and different lifestyles (e.g., large share of public transit use) compared to other states. Clearly, a model for Georgia is not seamlessly transferable to New York, but then it appears that a model for many other states would not be transferable to New York, either. Aside from New York, the model for Georgia seems to transfer relatively well to states that are dissimilar to it in many ways, including California and Massachusetts, as well as to the United States as a whole.

Overall, similar to findings in the previous section, the follow-up survey respondents are less representative in terms of age and education among the individual-level variables. Homeownership is the household-level variable that is hardest to represent in the follow-up survey.

Table 8 Effect size by different geographic regions

	GA	US	CA	MN	NC	NY	MA	ES by region[1]
Individual-level								
Age	0.43	0.45	**0.50**	0.41	0.44	0.49	0.48	
Gender	0.08	0.10	0.08	0.13	**0.14**	0.06	0.10	
Education	0.61	0.60	**0.67**	0.46	0.54	0.58	0.65	
Worker	0.06	0.07	0.09	0.08	**0.11**	0.05	0.08	
Hispanic	0.08	0.11	0.14	0.15	0.05	**0.16**	0.15	
Asian/Pacific Islander	0.14	0.15	**0.20**	0.10	0.09	0.19	0.12	
Black	0.12	0.09	0.07	0.10	0.09	**0.18**	0.10	
Native American	0.07	0.08	0.07	**0.13**	0.08	0.07	0.08	

White	0.17	0.18	0.32	0.13	0.12	**0.37**	0.18	
Commute mode	0.16	0.10	0.17	0.22	0.15	**0.33**	0.16	
Commute time	0.17	0.12	0.21	0.15	0.11	0.20	**0.33**	
Household-level								
Household size	0.12	0.13	0.16	**0.16**	0.13	0.03	0.16	
Household income	0.08	0.06	0.05	**0.20**	0.05	0.13	0.11	
Household vehicles	0.10	0.12	0.05	0.09	0.06	**0.43**	0.13	
Homeowner	0.25	0.29	0.33	0.21	0.23	**0.46**	0.30	
No. of children	0.10	0.08	0.09	0.12	0.09	0.09	**0.12**	

Note: Bolded numbers are the maximum effect size by row.
[1] Visualization of the effect size for each state in the same order as presented in the table.

7. CONCLUSION

In this study, we identified and analyzed the self-selection bias existing in follow-up survey respondents who were recruited from a preceding travel survey (the 2017 NHTS). We applied a probit with a sample selection (PSS) model to examine the willingness of NHTS respondents to participate in a follow-up survey, together with their actual response behavior. Overall, as expected, we identified self-selection biases among survey respondents recruited from a preceding household travel survey. Findings suggest that the requirements of the preceding survey influenced respondents' willingness to participate in follow-up surveys. In the particular context of NHTS, respondents from survey-burdensome households (e.g., large households) were less likely to report being willing to respond to a follow-up survey, although individuals reporting more trips were unexpectedly more likely to be willing. Respondents' attitudes towards privacy, and some other travel-related characteristics, were also influential to their willingness to be contacted for a follow-up survey. For example, respondents from specific groups (e.g., travel-restricted people, frequent transit users) were more likely to report being willing to participate in a follow-up survey. By participating in travel surveys, these groups may be seeking to improve the quality of their travel. We also found three explanatory variables with opposing signs between the selection and outcome models, a finding that indicated inconsistencies between people's reported willingness (to participate in a survey) and their actual (response) behaviors. Similarly, the negative error term correlations signified that, on net, unobserved characteristics had impacts on selection that were opposite to their impacts on the outcome.

PSS models do not have model performance measures that are consistently reported in the literature. To address this gap, this paper summarizes six well-known model performance measure categories, adjusted based on the PSS model structure: the log-likelihood, McFadden's pseudo R-squared, information criteria, point-biserial correlation coefficient, root mean squared error, and success table. McFadden's pseudo R-squared bounds the model fit between 0 and 1, which is straightforward for understanding and could be used to compare across different PSS models. The success table provides overall model performance measures as well as performance measures for each alternative, which supplies information important to evaluating the model.

We analyzed the representativeness of the follow-up survey respondents regarding 17 selected variables, including sociodemographic and travel-related variables. We decomposed the divergence of the marginal distributions between the population and the predicted follow-up survey respondents into five components, namely dataset bias, household representative bias, self-selection bias, non-response bias, and prediction error. Results showed that the household rep selection contributed to a large proportion of the distribution divergence of individual-level variables. The effect size for marginal distributions showed that education and age were the two least representative individual-level variables in the follow-up survey, whereas homeownership had the largest effect size among the household-level variables.

We also applied the PSS model to different geographic regions of the U.S., namely California, Massachusetts, Minnesota, North Carolina, and New York. Similar effect sizes across states indicated good generalizability of the PSS model, however education, age, and homeownership were still poorly represented among predicted respondents to the follow-up survey for these other states. New York had less representative predicted follow-up survey respondents compared to other states, presumably a consequence of its diverse population composition and different transportation-related lifestyles.

These results can help survey developers assess the representativeness and cost-effectiveness of the proposed sampling frame (i.e., a pool of previous survey respondents), which in turn will suggest adjustments to the sampling frame that can improve the representativeness of the new sample. Specifically, by using this approach to identify likely biases in the follow-up survey sample, study designers may choose to proactively oversample the predicted-to-be-underrepresented groups when recruiting from other data sources (e.g., online opinion panels). We recommend that large-scale travel surveys like the NHTS retain the willingness question as a recurring item, thereby allowing local agencies and researchers to efficiently recruit follow-up respondents from their sample. In fact, we recommend that the question be asked of *all* survey respondents, not only the main household respondent as was the case here. Recruiting future survey respondents from among *all* willing preceding survey respondents could substantially reduce sampling biases at the outset.

In a companion study (Wang, 2021), we analyze the *consequence* of self-selection biases by assessing their influence on travel behavior models developed on the second-stage sample. We examine and compare two techniques (sample weights and sample selection models) that could remedy the influence of unrepresentative samples recruited from a preceding survey on travel behavior models.

The study also has several caveats. First, the follow-up survey is a personal travel survey instead of a household travel survey. Our results do not speak to a situation in which the follow-up survey aims to obtain answers from all household members. If "household willingness-to-respond" is defined to be "willingness of every household member to respond", we would first of all expect a much *lower willingness rate*, and if follow-through response is required from every household member in order to count, we would secondly expect a much l*ower follow-through rate* among the reported-to-be-willing households. We would further expect more severe *biases* on the part of the willing and responsive households. For example, our results suggest that, in view of the heavier burden, larger households will probably be less likely to express willingness to respond and to actually respond to follow-up surveys. Given these concerns, we imagine that it would be prudent, if at all possible, to allow something less than full household participation to "count", at both stages of the process. Nevertheless, it is not presently clear how best to balance the disadvantages of a smaller and more biased sample when requiring full participation, against the disadvantages of incomplete household information when relaxing that requirement.

Another caveat is that the follow-up survey lags the preceding one by an interval ranging from four to 18 months, during which the address and demographic information of the initial survey respondents may have changed without our knowledge. We encourage future studies to explore the impact of time interval on the actual response to follow-up surveys. Moreover, it can be interesting to study the impact of completion modes (e.g., paper, online) for both preceding and follow-up surveys on the willingness to participate.

REFERENCES

Adriaan, H. & Jacco, D. (2009). Nonresponse in the Recruitment of an Internet Panel Based on Probability Sampling. *Survey Research Methods, 3*(2), 59-72.

Alemi, F., Circella, G., Mokhtarian, P., & Handy, S. (2019). What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transportation Research Part C, 102*, 233-248.

Amarov, B. & Rendtel, U. (2013). The recruitment of the access panel of German official statistics from a large survey in 2006: Empirical results and methodological aspects. *Survey Research Methods, 7*, 103-114.

Bohte, W. & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C, 17*, 285-297.

Börkan, B. (2010). The mode effect in mixed-mode surveys: Mail and web surveys. *Social Science Computer Review, 28*, 371-380.

Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M. & Ouimet, J. A. (2003). College Student Responses to Web and Paper Surveys: Does Mode Matter?. *Research in Higher Education, 44*, 1-19.

Cao, X. (2009). Disentangling the influence of neighborhood type and self-selection on driving behavior: an application of sample selection model. *Transportation, 36*(2), 207-222.

Chauhan, R.S., Bhagat-Conway, M.W., Capasso da Silva, D. et al. (2021). A database of travel-related behaviors and attitudes before, during, and after COVID-19 in the United States. *Scientific Data, 8*, 245.

Chen, F., Wu, J., Chen, X., Zegras, P. C., & Wang, J. (2017). Vehicle kilometers traveled reduction impacts of Transit-Oriented Development: Evidence from Shanghai City. *Transportation Research Part D, 55*, 227-245.

Circella, G., Tiedeman, K., Handy, S., Alemi, F. & Mokhtarian, P. (2016). *What Affects Millennials' Mobility? PART I: Investigating the Environmental Concerns, Lifestyles, Mobility-Related Attitudes and Adoption of Technology of Young Adults in California.* UC Davis: National Center for Sustainable Transportation. Available from the authors and at https://escholarship.org/uc/item/6wm51523.

Circella, G., Lee, Y. & Mokhtarian, P. (2020). *Comparison of Alternative Survey Recruitment/Deployment Methods*. Presentation at The ABCs (Attitudes – Behaviors – Choices) of Future Mobility Webinar, June 12. Webinar records and slides available at https://tomnet-utc.engineering.asu.edu/leadership-webinar-series/

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, Inc.

Collins, A. T., Rose, J. M. & Hess, S. (2012). Interactive stated choice surveys: A study of air travel behaviour. *Transportation, 39*, 55-79.

Coryn, C. L. S., Becho, L. W., Westine, C. D., Mateu, P. F., Abu-Obaid, R. N., Hobson, K. A., Schröter, D. C., Dodds, E. L., Vo, A. T. & Ramlow, M. (2020). Material incentives and other potential factors associated with response rates to internet surveys of American Evaluation Association members: Findings from a randomized experiment. *American Journal of Evaluation, 41*, 277-296.

Couper, M. P., Kapteyn, A., Schonlau, M. & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research, 36*, 131-148.

de Haas, M. C., Hoogendoorn, R. G., Scheepers, C. E. & Hoogendoorn-Lanser, S. (2018). Travel mode choice modeling from cross-sectional survey and panel data: The inclusion of initial nonresponse. *Transportation Research Procedia, 32*, 268-278.

Drucker, J. & Khattak, A. J. (2000). Propensity to work from home: Modeling results from the 1995 Nationwide Personal Transportation Survey. *Transportation Research Record. 1706*(1), 108-117.

Edwards, P., Roberts, I., Clarke, M., Diguiseppi, C., Pratap, S., Wentz, R. & Kwan, I. (2002). Increasing response rates to postal questionnaires: Systematic review. *British Medical Journal, 324*, 1183-1185.

Hardigan, P. C., Succar, C. T. & Fleisher, J. M. (2012). An analysis of response rate and economic costs between mail and web-based surveys among practicing dentists: A randomized trial. *Journal of Community Health, 37*, 383-394.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. Chapters in *Annals of Economic and Social Measurement, 5*(4), 475-492. National Bureau of Economic Research, Inc.

Heckman, J., Tobias, J. L., & Vytlacil, E. (2001). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal, 68*(2), 211-223.

Kim, S. H., Mokhtarian, P. & Circella, G. (2019). *The Impact of Emerging Technologies and Trends on Travel Demand in Georgia: Final Report.* Georgia Department of Transportation. Available from the authors and at http://g92018.eos-intl.net/G92018/OPAC/Index.aspx.

Lavrakas, P. J. (2008). Panel Survey. Chapters in *Encyclopedia of Survey Research Methods.* Thousand Oaks, California: Sage Publications, Inc.

Lin, Y. H., Yang, C. M., Hurng, B. S., Liu, I. W., Wu, S. I. & Chiou, S. T. (2011). Practical strategies to improve the response rate for a household interview survey. *Taiwan Journal of Public Health, 30*, 290-299.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I. & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research, 50*, 79-104.

McFadden, D. (2000). Disaggregate behavioral travel demand's RUM side: A 30-year retrospective. International Association for Travel Behavior (IATB) Conference, Gold Coast, Queensland, Australia, July 2-7, 2000.

Miller, C. A., Guidry, J. P. D., Dahman, B., Thomson, M. D. (2020). A tale of two diverse Qualtrics samples: Information for online survey researchers. *Cancer Epidemiol Biomarkers & Prevention, 29(4),* 731–735.

Mokhtarian, P. (2016). Discrete choice models' ρ^2: A reintroduction to an old friend. *Journal of Choice Modelling, 21*, 60-65.

National Research Council. (2013). *Nonresponse in Social Science Surveys : A Research Agenda*. Washington, DC: The National Academies Press.

Neufeld, A. J. & Mokhtarian, P. L. (2012). *A Survey of Multitasking by Northern California Commuters: Description of the Data Collection Process.* UC Davis: Institute of Transportation Studies. Retrieved from https://escholarship.org/uc/item/9f49x4h8

Parady, G., Ory, D., & Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling, 38*, 100257.

Parsons, N. L. & Manierre, M. J. (2014). Investigating the relationship among prepaid token incentives, response rates, and nonresponse bias in a web survey. *Field Methods, 26*, 191-204.

Ruiz, T. & Habib, K. N. (2016). Scheduling decision styles on leisure and social activities. *Transportation Research Part A, 88*, 304-317.

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (2021). Supplementing transportation data sources with targeted marketing data: Applications, integration, and validation. *Transportation Research Part A, 149*, 150-169.

Shaw, F. A., Wang, X., Mokhtarian, P. & Watkins, K. (2022). Using machine learning to enrich transportation surveys through variable transfer: with a sample application for psychometric variables. Paper in preparation. Available from the authors.

Shih, T. H. & Xitao, F. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods, 20*, 249-271.

Smith, V. K., Larson, K. L. & York, A. (2020). Using quality signaling to enhance survey response rates. *Applied Economics Letters, 27*, 951-954.

Stavropoulou, C. (2011). Non-adherence to medication and doctor-patient relationship: Evidence from a European survey. *Patient Education and Counseling, 83*, 7-13.

Sun, H., Wang, H., & Wan, Z. (2019). Model and analysis of labor supply for ride-sharing platforms in the presence of sample self-selection and endogeneity. *Transportation Research Part B, 125*, 76-93.

Tobias, E., Ralf, M. & Christian, B. (2013). On the impact of response patterns on survey estimates from access panels. *Survey Research Methods, 7*, 91-101.

van de Ven, W.P.M.M., & van Praag, B.M.S. (1981). The demand for deductibles in private

health insurance: A probit model with sample selection. *Journal of Econometrics, 17*, 229-252.

van Herick, D., & Mokhtarian, P. L. (2020). How much does the method matter? An empirical comparison of ways to quantify the influence of residential self-selection. *Travel Behaviour and Society, 18*, 68-82.

Wang, X (2021). Respondent Recruitment to Consecutive Travel Surveys: Exploring Sample Representativeness and Travel Behavior Model Quality Using Sample Selection Models. Master's thesis, Georgia Institute of Technology.

Wang, X., Shaw, F. A., Mokhtarian, P., Circella, G. & Watkins, K. (2022). Combining disparate surveys across time to study satisfaction with life: The effects of study context, sampling method, and transport attributes. *Transportation,* https://doi.org/10.1007/s11116-021-10252-x.

Wittwer, R. & Hubrich, S. (2015). Nonresponse in household surveys: A survey of nonrespondents from the repeated cross-sectional study "Mobility in Cities – SrV" in Germany. *Transportation Research Procedia, 11*, 66-84.

Wolf, H. K., Kuulasmaa, K., Tolonen, H., Sans, S., Molarius, A. & Eastwood, B. J. (2005). Effect of sampling frames on response rates in the WHO MONICA risk factor surveys. *European Journal of Epidemiology, 20*, 293-299.

Young, B., Bedford, L., Das Nair, R., Gallant, S., Littleford, R., Robertson, J. F. R., Schembri, S., Sullivan, F. M., Vedhara, K., Kendrick, D. & ECLS study team (2020). Unconditional and conditional monetary incentives to increase response to mailed questionnaires: A randomized controlled study within a trial (SWAT). *Journal of Evaluation in Clinical Practice, 26*, 893-902.

APPENDIX A. Marginal distribution of selected variables (random selection)

As discussed in Section 6.1, the household representative filter results in biases for individual-level variables. We would expect a more representative follow-up survey sample if the NHTS were to ask for every household member's willingness to participate in a follow-up survey. We simulate such a scenario by randomly selecting one adult from each household as the household representative and predicting their response to the follow-up survey. Table 9 presents the marginal distributions for randomly selected NHTS respondents (column 3a), the corresponding follow-up survey prediction (column 7a), and the effect size between the prediction and the population distribution (column 9a). Compared to the household representatives prediction (column 9), the new effect sizes calculated from the randomly selected NHTS respondents are generally reduced, especially for the largest effect sizes (e.g., age, education).

Table 9 Marginal distribution of selected individual-level variables (HH reps and random selection)

Column number	1	2	3	3a	7	7a	9	9a
Dataset	*ACS GA*[1]	*NHTS GA*[2]	*NHTS HH reps*[2]	*NHTS random*[2]	*Follow-up survey final predicted (HH reps)*[2‡]	*Follow-up survey final predicted (random)*[2‡]	*Effect size (HH reps)*[3]	*Effect size (random)*[4]
Age								
18-24	0.13	0.13	0.043	0.097	0.025	0.087	0.43**	0.26*
25-34	0.18	0.17	0.16	0.18	0.11	0.13		
35-44	0.18	0.19	0.20	0.20	0.17	0.17		
45-54	0.18	0.17	0.20	0.16	0.20	0.16		
55-64	0.16	0.17	0.20	0.18	0.23	0.19		
65+	0.17	0.17	0.20	0.19	0.26	0.25		
Gender								
Male	0.48	0.48	0.41	0.45	0.44	0.45	0.08	0.06
Female	0.52	0.52	0.59	0.55	0.56	0.55		
Education								
Less than a high school graduate	0.062	0.070	0.051	0.072	0.038	0.058	0.61***	0.44**
High school graduate or GED	0.36	0.25	0.20	0.23	0.17	0.22		
Some college or associates degree	0.30	0.30	0.31	0.30	0.29	0.29		
Bachelor's degree	0.17	0.21	0.24	0.22	0.24	0.23		
Graduate degree or professional degree	0.10	0.17	0.21	0.18	0.26	0.21		
Worker	0.59	0.62	0.63	0.62	0.56	0.55	0.06	0.08
Hispanic	0.078	0.083	0.075	0.078	0.058	0.062	0.08	0.06
Asian/Pacific Islander	0.046	0.039	0.034	0.034	0.017	0.028	0.14*	0.09
Black	0.31	0.31	0.32	0.33	0.26	0.23	0.12*	0.17*
Native American	0.0090	0.0036	0.0037	0.0029	0.0028	0.0031	0.07	0.06
White	0.62	0.63	0.62	0.61	0.70	0.72	0.17*	0.19*

Commute mode								
Private vehicle	0.94	0.93	0.92	0.92	0.93	0.93	0.16*	0.17*
Taxi	0.0030	0.0077	0.0050	0.011	0.0059	0.0068		
Public transit	0.022	0.032	0.041	0.035	0.032	0.028		
Walk	0.016	0.018	0.022	0.019	0.018	0.016		
Bike	0.0025	0.0065	0.0090	0.0083	0.0085	0.0058		
Other	0.013	0.0074	0.0050	0.0073	0.0077	0.010		
Commute time								
0-10 min	0.21	0.21	0.21	0.21	0.21	0.21	0.17*	0.17*
10-20 min	0.30	0.26	0.27	0.26	0.27	0.25		
20-30 min	0.21	0.20	0.19	0.19	0.17	0.20		
30-60 min	0.23	0.27	0.27	0.27	0.28	0.27		
60-90 min	0.033	0.048	0.047	0.042	0.050	0.041		
90+ min	0.015	0.026	0.017	0.024	0.020	0.027		

Notes: For each variable, the sum of category shares might not equal 1 due to rounding errors. Column numbers in Table 9 match the counterparts in Table 7.

[1] 2018 ACS individual weights are applied.

[2] NHTS individual weights, based on 2015 ACS individual weights, are applied.

[3] Comparison between the population distribution and follow-up survey predicted distribution (HH reps, columns 1 and 7a).

[4] Comparison between the population distribution and follow-up survey predicted distribution (random, columns 1 and 7b)

* Small effect size ($w = 0.10$). ** Medium effect size ($w = 0.30$). *** Large effect size ($w = 0.50$).

‡ Calculated with $P(y_i^S = 1, y_i^O = 1)$.

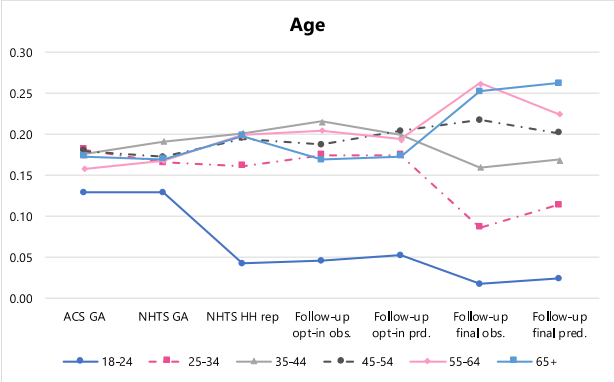
APPENDIX B. Changing trajectories of marginal distributions

To further illustrate the changing trajectories of the marginal distributions from the population to the predicted follow-up survey respondents, we select two individual-level variables (i.e., age, gender) and two household-level variables (i.e., household size, household income) and visualize them in Figure 3 (for each figure, read lines from left to right).

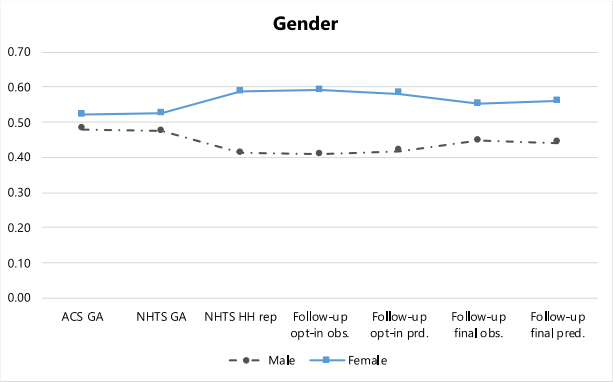
Regarding the two individual-level variables, we see large differences between the NHTS Georgia population and NHTS household representatives. Specifically, household representatives underrepresent younger groups (i.e., 18-24 and 25-34) and males, meaning that middle-aged/older people (45+) and females are more likely to answer the household-related questions in the retrieval survey. In the observed opt-in follow-up survey sample, we see slightly increased shares of young and middle-aged people, which indicates that the self-selection bias partially offsets the HH representative bias. However, the non-response bias results in an even worse underrepresentation of younger people and overrepresentation of older people in the observed final follow-up survey. The marginal distribution of gender is relatively stable after the household representative filter (except for the small increase of males in the sample), which indicates small self-selection biases, non-response biases, and prediction errors.

The two household-level sociodemographic variables, namely, household size and household income, have fluctuating trajectories. Regarding household size, we see similar marginal distributions of the population (ACS) and the NHTS Georgia sample/household rep sample. The main distribution divergence occurs between the NHTS Georgia/household rep sample and the observed opt-in follow-up survey respondents. As we have discussed in Section 5.1, larger households are less willing to participate in a follow-up survey due to the heavy burden of survey completion that accompanies more family members. After the opt-in process, the proportion of households with three or more members keep shrinking, while two-member households take the largest share in the final follow-up survey sample due to non-response biases and prediction errors.

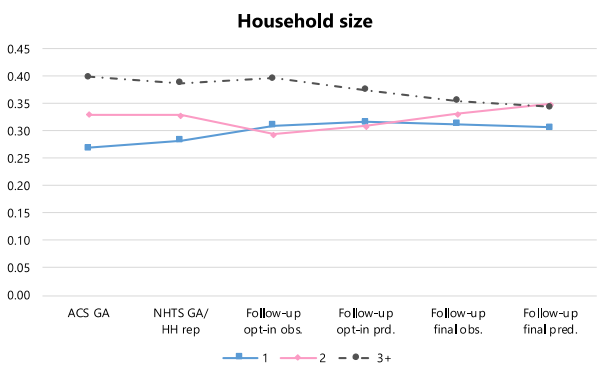
Regarding household income, we see that the NHTS Georgia/household rep sample overrepresents the lower income group (less than $24,999) and underrepresents some middle/high-income groups ($50,000 to $ 99,999, $150,000 or more). The household income distributions of the observed opt-in follow-up sample diverge from the household income distribution of the NHTS Georgia/household rep sample, which indicates self-selection biases. Interestingly, the traits of observed final follow-up survey respondents partially *correct* some of the divergences, i.e., the marginal distribution of the *final* follow-up survey respondents is close to the population marginal distribution. In other words, the non-response biases partially offset the self-selection bias.



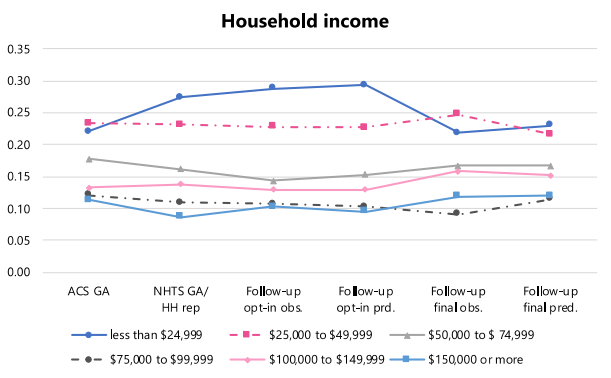
(a)



(b)



(c)



(d)

Figure 3 Changing trajectories of the marginal distributions