Background, Design, and Implementation of the Phase 1 WholeTraveler Survey

C. Anna Spurlock, Lawrence Berkeley National Laboratory
(caspurlock@lbl.gov)
August 14th, 2020

1. Design of the WholeTraveler Survey Instrument

The WholeTraveler Survey was conducted in two contiguous phases. Phase 1 involved an online-only survey while Phase 2 (which started immediately after the completion of the Phase 1 survey for any respondent that opted in to Phase 2) involved additional GPS data collection.

The WholeTraveler team developed the data collection instrument leveraging expertise across a spectrum of disciplines. Members of the team represent a wide range of disciplines and experience, including behavioral economics, anthropology, geographic information systems, data science, sociology, psychology, engineering, and robotics.

1.1. The Phase 1 Instrument

The WholeTraveler team coordinated with the survey implementation contractor, Resource Systems Group (RSG), to code and test the Phase 1 online survey instrument. The survey instrument is available at https://livewire.energy.gov/ds/wholetraveler/. Questions developed for Phase 1 include the following categories:

- Location of primary destination
- Commute mode and habits
- Mode access and travel characteristics and preferences
- Shopping travel and e-commerce behavior and preferences
- Awareness of, exposure to, use/adoption of, and interest in emerging technologies and services such as:
    - Automated vehicles
    - Mobility as a service options (single-rider and carpool)
    - Alternative fuel vehicles (plug-in electric and hybrid)
    - Analogue technologies (Solar PV, smartphones)
    - Other services (Amazon Prime account)
- Vehicle ownership, including make and model of the most frequently driven vehicle
- Use of various modes
- Personality and risk characteristics
- Standard demographic information (age, gender, race/ethnicity, educational attainment, household income, language spoken at home, employment status, household size)
- Life history calendar

As many of the subjects of focus in the WholeTraveler effort are technologies or practices that have only emerged in recent years, a WholeTraveler effort sought out prior similar research projects for comparison. The research team conducted a thorough investigation of existing datasets and similar efforts and found a paucity of recent, relevant inquiry at a commensurate scale. Based on our review of data collection activities at the time the survey was designed there were two main data sources we found that touched on some of the similar topics as WholeTraveler, such as TNCs/ride-hailing. Below we summarize key ways in which WholeTraveler differs from those survey efforts.

- **National Household Travel Survey (NHTS**): While the 2017 release of the NHTS has some added questions on emerging modes (e.g., TNCs), the NHTS does not track the same respondents over time and so the long-term lifecycle patterns or changes within respondent over time cannot be readily assessed. In addition, the NHTS lacks data on some of the psychological, preference, and characteristics elicited in the WholeTraveler survey.
- **California Millennial Panel Study (UC Davis survey**): This survey has some detailed data similar to elements collected in WholeTraveler, and some complementary data not collected by WholeTraveler. This effort is undertaking the longer-term investment of tracking individuals over time to create a panel. This survey effort covers more of California than the WholeTraveler survey, however, the data collected at the time the WholeTraveler survey was done consisted of approximately 2400 valid responses for all of California. This is in contrast to over 900 for the WholeTraveler sample for just the Bay Area. This means WholeTraveler has more concentrated coverage. In addition, this survey included less detail on some of the technologies and services explored in WholeTraveler, including e-commerce behavior.[1]

At the time the WholeTraveler data was collected it was the only known research effort that integrated a range of emerging transportation technologies and practices from the perspective of individual respondents (i.e., eliciting attitudes about not only EVs or only AVs, as some other efforts had done, but about EVs, AVs, e-commerce, ride-hailing and car-sharing).

One of the elements included in the Phase 1 survey that does is not captured in any of the larger scale data colletions that address emerging technologies is the approach used to collecting life history data, capturing travel options and behaviors experienced over the course of respondents' lives, from early adulthood onward. This approach allows for identifying how major life events – changing residence location, employment, marriage, a growing family – affects travel behavior. The approach used for this was the Life History Calendar.

The team conducted a pilot test of survey implementation. Feedback and comments were elicited from researchers across SMART Mobility, and from experts in transportation behavior research. Following extensive testing and review, adjustments were made to improve the survey, including user experience and streamlining of the data collection process.

The WholeTraveler Transportation Behavior Study, including all data collection methodologies, data transfer and storage protocols, and data analysis plans, were reviewed by the Lawrence Berkeley National Laboratory

[1] More detailed information here: https://ncst.ucdavis.edu/wp-content/uploads/2015/09/NCST_Report_Millennials_Part_II_2017_March_31_FINAL.pdf

(LBNL) Human Subjects Committee (HSC), which is LBNLs Institutional Review Board (IRB). LBNL holds Office of Human Research Protections Federalwide Assurance number FWA 00006253. The project protocol was approved by the HSC on August 30th, 2017 (approval number 366H001-29AU18).

1.2. Sampling, Recruitment and Implementation Methodology

The sampling methodology used was a random addressed-based sample of active addresses in the core 9 counties of the San Francisco Bay Area (Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa, Clara, Solano, Sonoma). RSG worked with Marketing Systems Group, which licenses the bi-monthly updated US Postal Service's Computerized Delivery Sequence (CDS) database, to purchase the random sample of households to recruit.

An invitation letter was sent to the address-based random sample inviting them to take part in the Phase 1 survey, which could be completed online using a laptop or desktop computer only (the survey could not practically be optimized for smartphones or tablets due to the large table structure of the Life History Calendar and a few of the other questions). Compensation in the form of a $10 Amazon Gift Card was provided to all those who completed all of the required questions in the Phase 1 survey.

The initial target was 900 responses to Phase 1, and given the intial recruitment startegy (a single recuitment letter, in an envelope, with a $10 incentive), the response rate was anticipated to be about 3%. Given these initial expectations, the random address-based sample purchased was 30,000 addresses.

However, once data collection started, it became clear that the response rate was lower than anticipated (closer to 1.5%). For this reason, a postcard reminder was sent to those within the first 30,000 sample that had not yet completed the survey, and a second 30,000 random sample was obtained and recruited using a slightly improved invitation letter, followed by the same reminder postcard.

1.3. Data Collection Outcomes

Data collection for the Phase 1 survey took place between March 2018 and June 2018. The final number of fully complete responses was 997, with an additional 48 that completed the entire survey other than the life history calendar. This resulted in a total of 1,045 responses. These responses were split almost exactly 50/50 between the first and second 30,000 address-based samples. The median response time was about 27 minutes.

The final set of respondents tended to be better educated than the San Francisco Bay Area general population. Income levels tended to be commensurate with the U.S. American Community Survey (ACS) for the region, though Solano County respondents tended to have slightly lower incomes on average than indicated by the ACS census data, while San Francisco County respondents tended to have higher incomes on average than the population based on the census. Alameda and San Francisco Counties tended to have higher population representation rates relative to the full region average, while Solano and Napa tended to have relatively less coverage in the sample realtive to the full region average. Figure 1 provides a visualization of the geographic distribution of respondents to the survey.

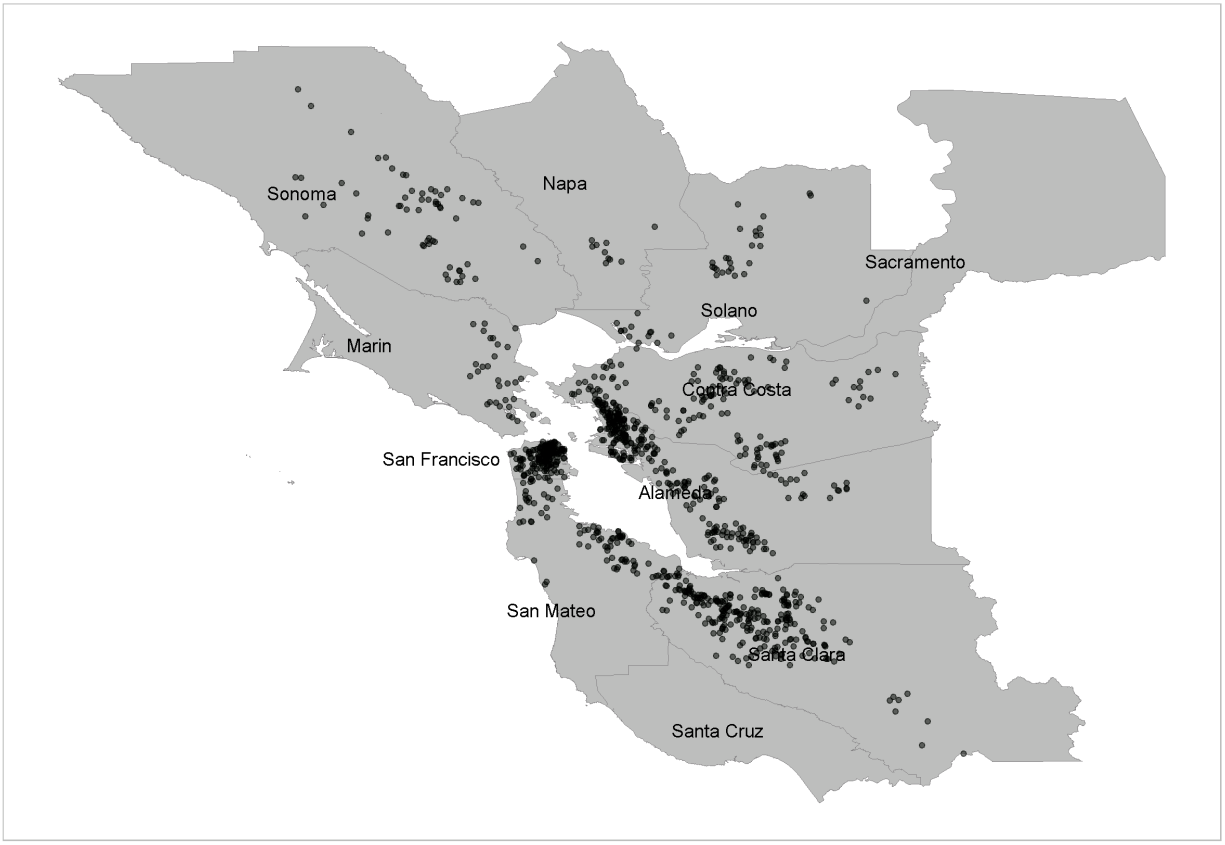

Figure 1 Geographic distribution of residential locations in sample

2. Ancillary Data Collection

Ancillary data were collected and merged with the survey data form publicly available sources to augment the information from the survey itself. In this section more information is provided on several of these data sources.

1.4. U.S. Census Bureau

We match the locations of reported residence and primary destination with census block shapefiles and retrieve census-block-level information of total population using the 2012-2016 ACS 5-year estimate data using the Census Data API. Current land area information for each block is then retrieved from TIGER/Line Shapefiles. Dividing the total population by current land area, we calculate population density for each census block. These data are available to be merged with the WholeTraveler data. They are provided at https://livewire.energy.gov/ds/wholetraveler/, in the Phase 1 data packet (WT_phase1_ancillary_locational.csv).

1.5. Walk Score

Walk Score is a number between 0 and 100 that measures the walkability of any address, constructed by a private company. Here is the explanation of their methodology:

> Walk Score measures the walkability of any address using a patented system. For each address, Walk Score analyzes hundreds of walking routes to nearby amenities. Points are awarded based on the distance to amenities in each category. Amenities within a 5-minute walk (0.25 miles) are given maximum points. A decay function is used to give points to more distant amenities, with no points given after a 30-minute walk.
>
> Walk Score also measures pedestrian friendliness by analyzing population density and road metrics such as block length and intersection density. Data sources include Google, Education.com, Open Street Map, the U.S. Census, Localeze, and places added by the Walk Score user community. (https://www.walkscore.com/methodology.shtml).

A summary of the qualitative description of the Walk Score applying to ranges of the Score is provided in Table 1. The data we use were manually collected directly from their partner site, Redfin.com, an online real estate brokerage site. The table below contains description of various ranges of walk score. These data are available to be merged with the WholeTraveler data. They are provided at https://livewire.energy.gov/ds/wholetraveler/, in the Phase 1 data packet (WT_phase1_ancillary_locational.csv).

Table 1 Description of walk score ranges

90–100	Walker's Paradise	Daily errands do not require a car
70–89	Very Walkable	Most errands can be accomplished on foot
50–69	Somewhat Walkable	Some errands can be accomplished on foot
25–49	Car-Dependent	Most errands require a car
0–24	Car-Dependent	Almost all errands require a car

1.6. Google Map Platform API

The Distance Matrix API is a service of Google Map Platform that provides travel distance and time for a set of origins and destinations. The API returns information based on the recommended route between start and end points, as calculated by the Google Maps API.

For the calculation of distances, one may specify the transportation mode to use. The following travel modes are supported:

- *Driving* (default) indicates distance calculation using the road network.
- *Walking* requests distance calculation for walking via pedestrian paths & sidewalks (where available).
- *Bicycling* requests distance calculation for bicycling via bicycle paths & preferred streets (where available).
- *Transit* requests distance calculation via public transit routes (where available).

If the mode is set to transit or driving, one can optionally specify either a *departure_time* or an *arrival_time*. The responses of queries may contain the total fare (that is, the total ticket costs) on the route. This property is only returned for transit requests and only for transit providers where fare information is available. Traffic information is used when the travel mode parameter is driving, and the request includes a valid *departure_time* parameter. The *departure_time* can be set to the current time or some time in the future. It cannot be in the past.

We apply Distance Matrix API on four sets of locations – residence, primary destination, the Bay Area Rapid Transit (BART) station closest to the residence, the BART station closest to the primary destination. The latter two are selected based on Great Circle (WGS84 ellipsoid) distance between residence/primary destination and BART stations. API queries are run for all four modes, two *departure_time* parameters ("2018-07-18 08:00:00 PST8PDT" and "2018-07-18 17:00:00 PST8PDT"), three sets of segments and two directions ("leave" and "return").

The first set of segments is residence->start-point BART station (start-point BART station-> residence for "return"). The second set of segments is end-point BART station->primary destination (primary destination ->end-point Bart station for "return"). The third set of segments is residence->primary destination (primary destination -> residence for "return").

These data are available to be merged with the WholeTraveler data. They are provided at https://livewire.energy.gov/ds/wholetraveler/, in the Phase 1 data packet (WT_phase1_ancillary_locational.csv).

1.7. Google Maps API Platform – More detailed Public Transit Trip Data

In a second round of ancillary data collection using the Google Maps API, more data were collected in greater detail on public transit trips between the respondent's home and primary destination (with trips in both directions and at multiple times of day). These data capture the detailed steps, including transit service, transit line, transit stop, trip duration, step of trip duration, number of steps in the trip, etc. Some information is redacted from these data in order to ensure anonymity of the respondents. Particularly for bus modes in the trip steps, bus stops are provided for stops along the route, but the bus stop name of the stop closes to the home or primary destination location is redacted.

These data are available to be merged with the WholeTraveler data. They are provided at https://livewire.energy.gov/ds/wholetraveler/, in the Phase 1 data packet (WT_phase1_ancillary_GoogleAPI_public_transit_commute.csv).

1.8. EPA Fuel Economy

FuelEconomy.gov is an Internet resource that helps consumers make informed fuel economy choices when purchasing a vehicle. It is maintained by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy with data provided by the U.S. Environmental Protection Agency (EPA).

Information of EPA miles per gallon (MPG) estimates for vehicles from 1984 to the present are available, as well as real-world MPG shared by other users. We match the year, make, model, fuel type of the vehicle that is driven most frequently by the respondent's household to the FuelEconomy.gov database and retrieve MPG metrics. Note that some vehicles have more than one match due to incomplete information.

These data are available to be merged with the WholeTraveler data. They are provided at https://livewire.energy.gov/ds/wholetraveler/, in the Phase 1 data packet (WT_phase1_ancillary_fuelecon.csv).

3. Data pre-processing

In analyses using data from the Phase 1 survey, the following data validation and cleaning determinations have been made. In order to avoid basing analysis on responses from people who are likely to have clicked quickly through the survey without reading and responding to the questions, the determination was made to drop respondents that completed the entire survey (including the Life History Calendar section) in less than 12 minutes. Figure 2 below shows a histogram of the survey response times with an indication of where this cut-off is. This results in 18 observations being dropped. In addition, one respondent reported a birth year of 1900, which results in an age at the time the survey was completed of 118. This individual was dropped in some cases as well. These cleaning steps resulted in 98% of the original 1,045 respondent dataset.

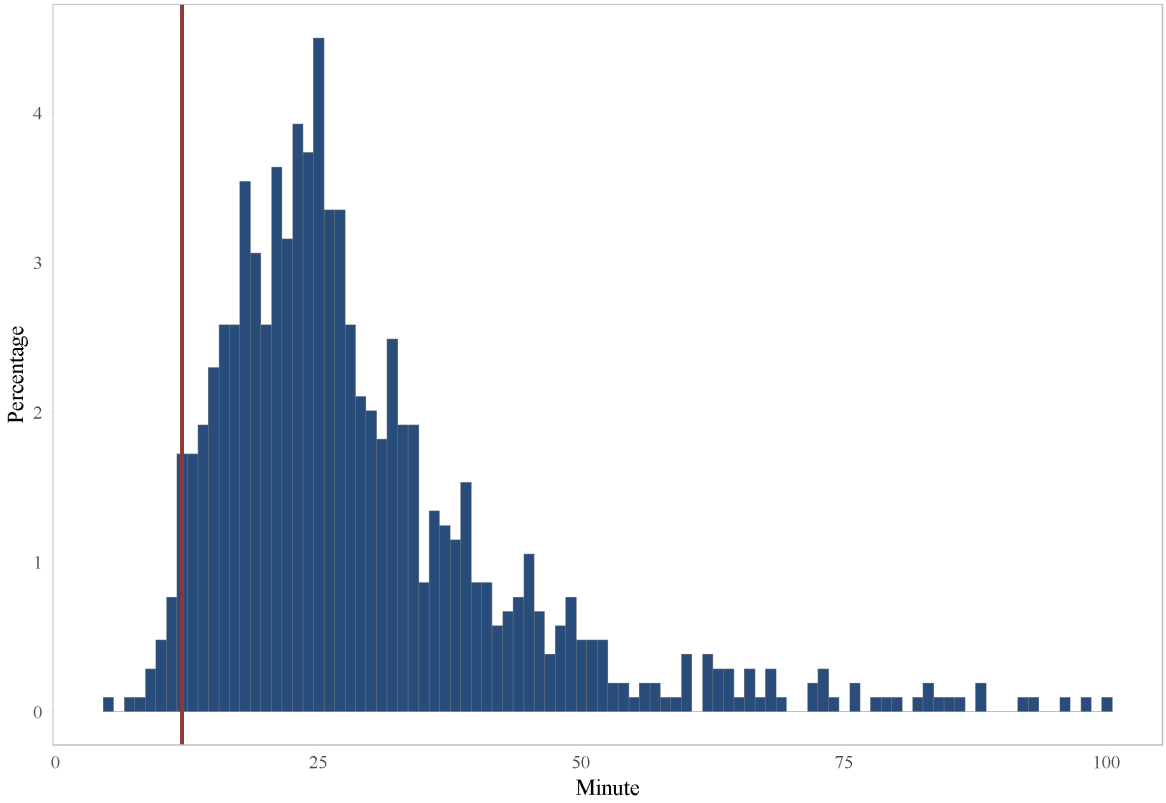

Figure 2 Histogram of survey completion times

Note: Observations with values greater than 100 minutes are not shown; Survey responses completed the full survey in

less than 12 minutes are those to the left of the vertical red line.

4. Summary Statistics and Data Visulatization

In this section we provide some selected summary tables and data visualizations to provide a sense of some of the characteristics of the sample. In all cases the tables and figures presented were generated after dropping the 19 respondents described in the Section 3.

Table 2 and Table 3 provide summary statistics for the cleaned subsample (1,026 respondents) for socio-demographic information.

Table 2 Summary Statistics of Socio-Demographic Information (1 of 2)

	Count	Percent of total cleaned sample (1026)	Percent of those that responded to question (excluding "Prefer not to answer or N/A")
Gender			
Male	502	48.9%	50.7%
Female	487	47.5%	49.2%
Other	1	0.1%	0.1%
Prefer not to answer or N/A	36	3.5%	
Household Income (before taxes)			
0 <= HH income < 10,000	14	1.4%	1.6%
10,000 <= HH income < 15,000	11	1.1%	1.3%
15,000 <= HH income < 25,000	26	2.5%	3.0%
25,000 <= HH income < 35,000	26	2.5%	3.0%
35,000 <= HH income < 50,000	60	5.8%	6.9%
50,000 <= HH income < 75,000	95	9.3%	10.9%
75,000 <= HH income < 100,000	109	10.6%	12.5%
100,000 <= HH income < 150,000	189	18.4%	21.6%
150,000 <= HH income < 200,000	129	12.6%	14.7%
200,000 <= HH income < 300,000	132	12.9%	15.1%
300,000 <= HH income < 400,000	52	5.1%	5.9%
HH income >= 400,000	32	3.1%	3.7%
Prefer not to answer or N/A	151	14.7%	
Ethnicity			
White	622	60.6%	59.6%
Hispanic, Latino, or Spanish origin	71	6.9%	6.8%
Black or African American	27	2.6%	2.6%
Asian	269	26.2%	25.8%

Middle Eastern or North African	12	1.2%	1.2%
American Indian or Alaska Native	9	0.9%	0.9%
Native Hawaiian or Other Pacific Islander	14	1.4%	1.3%
Some other race or origin	19	1.9%	1.8%
Prefer not to answer or N/A	69	6.7%	

Table 3 Summary Statistics of Socio-Demographic Information (2 of 2)

	Count	Percent of total cleaned sample (1026)	Percent of those that responded to question (excluding "Prefer not to answer or N/A")
Highest Education Level Completed			
No Diploma	4	0.4%	0.4%
HS degree	15	1.5%	1.5%
Some College	81	7.9%	8.1%
Associate's Degree	50	4.9%	5.0%
Bachelor's Degree	388	37.8%	38.9%
Master's Degree	315	30.7%	31.6%
Professional Degree	57	5.6%	5.7%
Doctoral Degree	88	8.6%	8.8%
Prefer not to answer or N/A	28	2.7%	
Employment Status			
Employed for wages	701	68.3%	64.3%
Self-employed	117	11.4%	10.7%
Out of work and looking for work	35	3.4%	3.2%
Out of work but not currently looking	11	1.1%	1.0%
A homemaker	35	3.4%	3.2%
A student	52	5.1%	4.8%
Military	3	0.3%	0.3%
Retired	124	12.1%	11.4%
Unable to work	13	1.3%	1.2%
Prefer not to answer or N/A	31	3.0%	
Any Children < 8yrs in the Household			
Yes	158	15.4%	15.4%
No	868	84.6%	84.6%
Prefer not to answer or N/A	0	0.0%	
Number of household members with drivers licenses			
0	7	0.7%	0.9%
1	78	7.6%	9.7%
2	582	56.7%	72.6%
3	100	9.7%	12.5%

4 or more	35	3.4%	4.4%
Prefer not to answer or N/A	224	21.8%	
Languages Spoken in the Home			
English only	713	69.5%	73.8%
At least one language other than English	253	24.7%	26.2%
Prefer not to answer or N/A	60	5.8%	

Figure 3 Shows the average number of vehicles owned by respondent households in each county in the Bay Area.

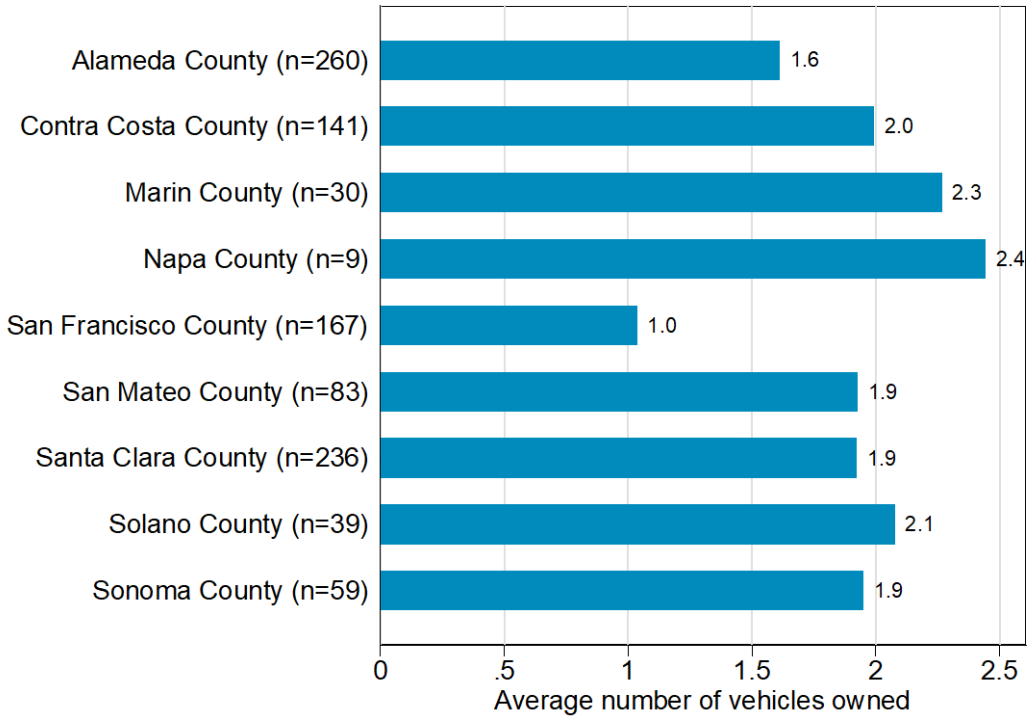

Figure 3 Average number of vehicles owned by household by residential county

In the survey respondents were asked to report a single primary destination (where they travel to most frequently). Travel to this destination is referred to as their "commute" but it should be noted that this destination may not be an employment-related location. To summarize this, Figure 4 shows the number of days per week respondents reported commuting to their primary destination broken out by the type of location they report is their primary destination.

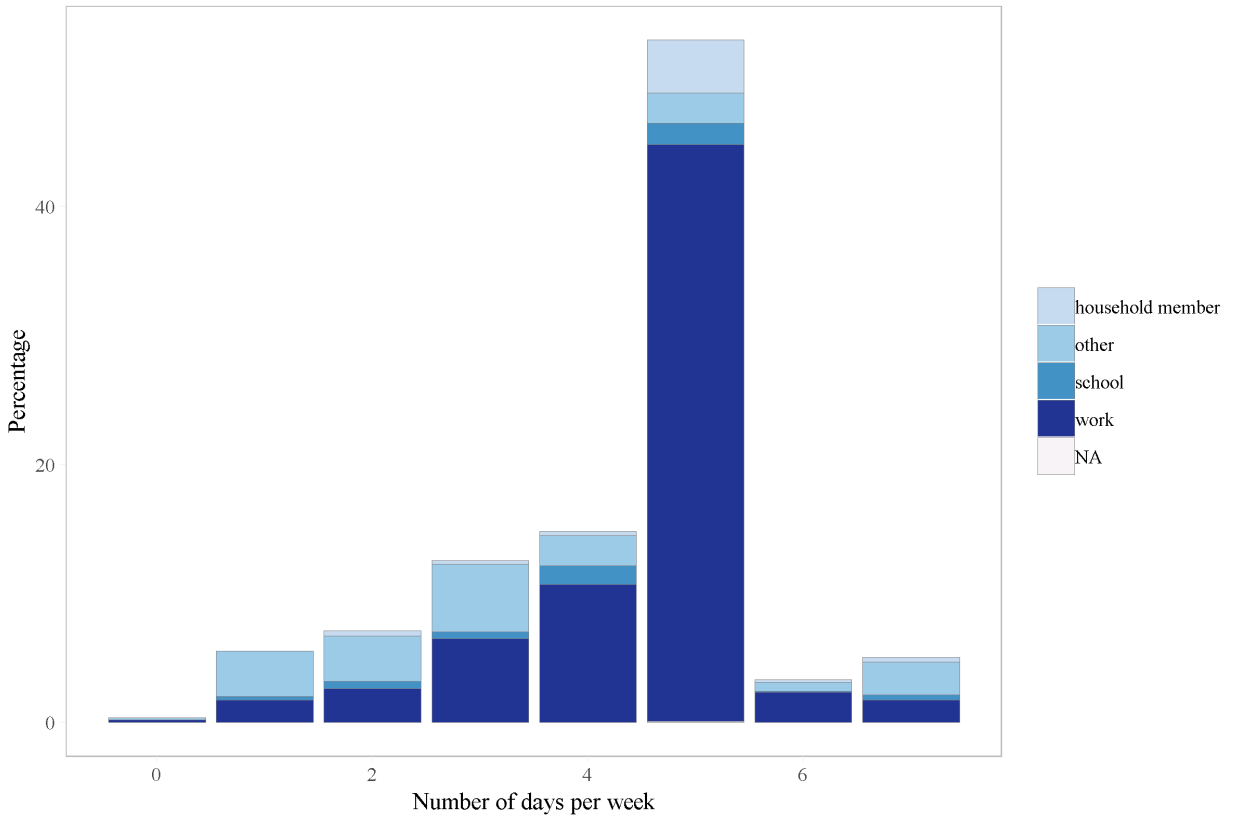

Figure 4 Number of days per week commutes to reported primary destination by primary destination type
Note: 17 respondents chose more than one primary destination type.

Figure 5 shows all the transportation modes asked about in the survey broken out by how recently the respondent indicated they last took that mode for their commute to their primary destination.

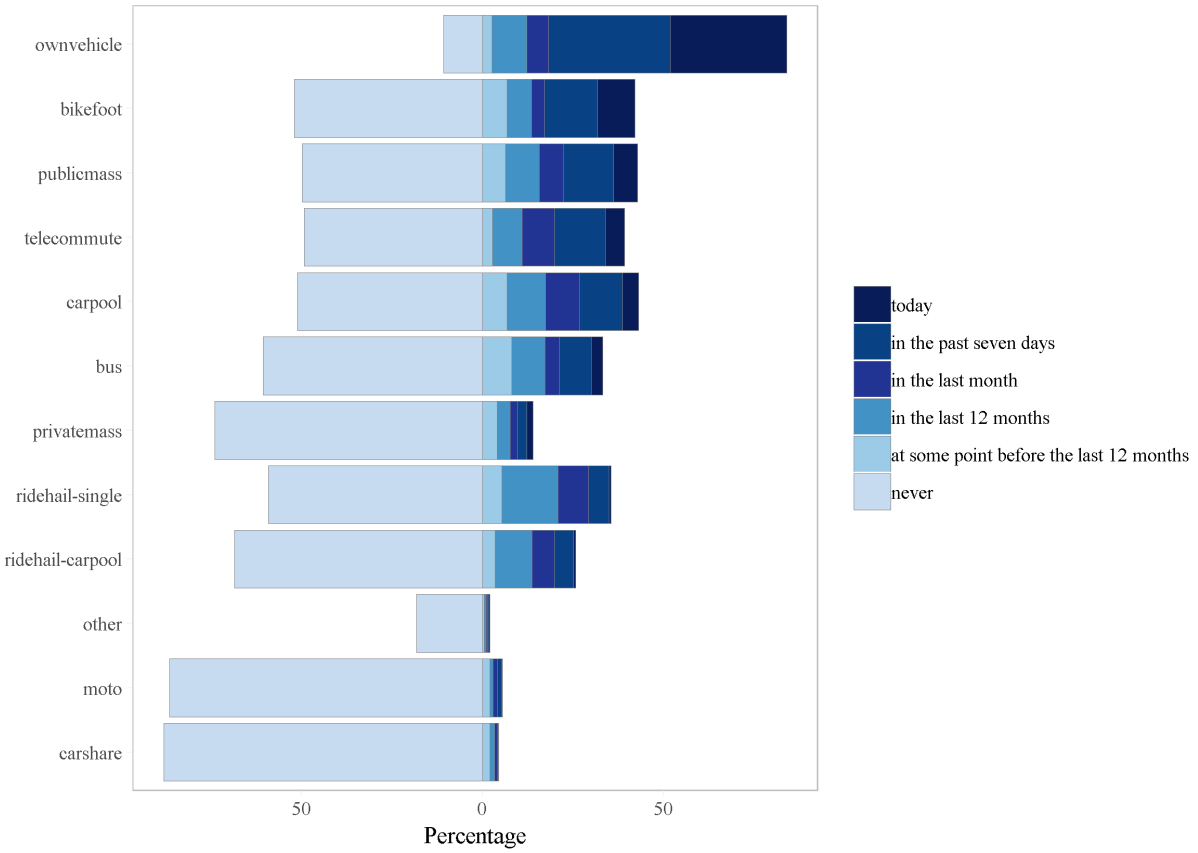

Figure 5 Transportation modes by recency of last use for commute to reported primary destination
Note: Those that responded with N/A are not shown.

5. Publications using the WholeTraveler Transportation Behavior Study Data

Spurlock, C. Anna, James Sears, Gabrielle Wong-Parodi, Victor Walker, Ling Jin, Margaret Taylor, Andrew Duvall, Anand Gopal, and Annika Todd. "Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the San Francisco Bay Area." *Transportation Research Part D: Transport and Environment* 71 (2019): 283-301. https://doi.org/10.1016/j.trd.2019.01.014 (publicly available here: https://eta.lbl.gov/publications/describing-users-understanding)

Spurlock, C. Anna, Annika Todd-Blick, Gabrielle Wong-Parodi, and Victor Walker. "Children, Income, and the Impact of Home Delivery on Household Shopping Trips." *Transportation Research Record* (2020): 0361198120935113. https://doi.org/10.1177/0361198120935113 (published via open access option)