2018-2019 Travel Behavior Inventory

Dataset User Guide

This document provides explanations for select variables in the datasets delivered for the Travel Behavior Inventory study. While a comprehensive list of survey data variables is provided in the survey data codebook, this document expands on those definitions where additional explanation is valuable. In many cases, referring to the original survey questionnaires and methodology memo also provides helpful context for the exact data collection procedures.

ITEM 1: Data Privacy

This guide accompanies datasets that contain sensitive and confidential data, including personally identifiable information (PII). PII includes data such as name, contact info, home location, and other personal identifiers (see Appendix at end of this document). Combinations of additional variables may also constitute PII. ***Please follow your organization's privacy policy, this study's privacy policy, and generally accepted best practices such as state laws like CCPA for data privacy and security. An annotated copy of the privacy policy under which participant data was collected for this project has been provided as well.***

RSG has removed any PII used solely for survey administration from the final datasets. These removed variables include participant names, email addresses, and phone numbers. Additionally, participant passwords have been replaced with household and person IDs (hh_id/person_id) to ensure that no information in the final dataset can be identified using the passwords the participant used.

ITEM 2: Study and Dataset Overview

The primary means of data collection for the TBI was RSG's travel survey smartphone app, rMove^{TM}. Households in which all adults (age 18+) own smartphones were asked to download rMove and use it for seven consecutive days. Households in which all adults (age 18+) do not own smartphones were asked to complete a one-day travel diary either online or over the telephone using RSG's rSurvey™ platform; call center operators utilized the same online survey instrument to complete the travel diary with participants over the phone. Data from all three data collection methods (online, call center, and smartphone app) were integrated into a single dataset during post-processing.

This dataset includes data collected for the Travel Behavior Inventory from October 2018 through September 2019. The study included two parts:

1. **Part one**, also called the "**recruit survey**," collected information about household composition, demographics, and typical travel behavior.

2. **Part two**, also called the "**travel diary**," required participants to record their travel during an assigned travel period. RSG's travel survey app, rMove™, collected data for 88% of study trips, and RSG's online travel diary, rSurvey™, collected the remaining trip data.

"Complete" participants met the following conditions:

1. The household completed either a recruitment/demographic survey online, in rMove, or through the call center.
2. All household members completed all travel diary surveys on at least one day throughout their travel period.
   - Note**:** Some households who participated solely via rMove did not provide complete demographic details (e.g., educational attainment) for all household members. These households are retained in the dataset at this time.

Participation Groups

Respondents could participate in one of three ways:

- **Group 1:** The household completed Part 1 online or through the call center. All adults (age 18+) then completed Part 2 (the travel diary) using rMove. Child trips were reported by proxy using rMove. The travel period was seven days.

- **Group 2:** The household completed Part 1 online or through the call center. All household members then completed Part 2 online or through the call center. Child trips were reported by proxy in most cases. The travel period was one day.

- **Group 3:** The household completed Part 1 using rMove. All adults (age 18+) then completed Part 2 using rMove. Child trips were reported by proxy using rMove. The travel period was seven days.
   - **Note:** Part 1 was significantly shorter for Group 3 than for other groups and contained only basic demographic information along with any information required for study administration. All other demographic information was collected throughout the travel period.

Distinct survey data tables:

The final project dataset includes six data tables. These tables include all user-input survey variables, passively collected location data, certain survey metadata (e.g., survey completion timestamps), and variables derived to support data analysis (e.g. number of trips during the travel period.

DATA TABLE	RECORD COUNT
Household	7,837 households
Person	16,152 persons completing one or more days
Vehicle	13,431 vehicles
Day	84,582 days, of which 76,415 are complete travel days
Trip	351,177 trips, of which 315,272 are on complete travel days
Location	6,830,105 trip location/GPS points (latitude/longitude/timestamp)

Time and location standards:

All **timestamps** reflect the local time zone for the study region (Central Standard time), regardless of where the trip took place geographically (e.g., if a trip took place in another time zone, the timestamps for that trip are still in Central Standard time).

ITEM 3: Joining the data tables

These data tables can be joined into a single database as needed. The datasets can be joined to the others as follows:

TABLE NAME	VARIABLE(S) TO JOIN TO OTHER SURVEY DATA TABLES
Household	hh_id
Person	hh_id, person_id
Vehicle	hh_id
Day	hh_id, person_id, day_num
Trip	hh_id, person_id, day_num, trip_id
Location	trip_id
Note: Further details about these primary keys are in the table-specific sections of this guide below.	

ITEM 4: Missing Values and Gaps in the Survey Data

A survey dataset cell may be missing data for one of six reasons:

1. **A value or response was not required under the circumstances.**

Example: Participants who traveled by bus were not asked if they were the driver or passenger on the trip.

Coded as: 995 for categorical variables, blank/NA for continuous variables

2. **A respondent indicated that the question was not applicable and skipped that question.**

Example: Some participants didn’t share if they’re typical mode to school changed seasonally because they take online classes only.

Coded as: 996 (often labeled as “Not applicable”)

3. **A respondent indicated that they didn’t know the answer and skipped that question.**

Example: Some participants who made a vehicle trip and paid to park the vehicle may not remember the amount they paid.

Coded as: 998 (Don’t know)

4. **A respondent indicated that they preferred not to answer a question and skipped that question.**

Example: Some participants chose not to provide their household income.

Coded as: 999 (Prefer not to answer)

5. **A respondent did not answer part of a survey that was required.**

Example: A participant did not complete a trip survey and their travel mode is missing.

Coded as: -9998 (Non-response) for categorical variables, blank/NA for continuous variables.

6. **A technological or logical error occurred.**

Example: A logic or data collection error occurred, resulting in missing responses for otherwise required data/questions.

Coded as: -9999 (Error).

Other notes about missing survey data:

- For a survey to be complete, all required data from the survey must be present.
- Continuous variables (e.g., trip distance, trip duration) are not coded with missing value codes and are instead left empty when missing to avoid interfering with statistical calculations. In some cases, a variable was added to flag for missing data (e.g., missing_work_location).
- Due to the large size of the location table, missing values were left exactly as they were collected. Speed, heading, and accuracy can all potentially contain missing values that are either stored as “-1”, NA, or 0. Analysis on those fields should filter to where the values are greater than zero.

ITEM 5: Outliers

Continuous variables (e.g., trip distance, trip duration, parking cost) in the dataset may contain outliers. Data users should be aware of these outliers when calculating summary statistics (e.g., mean) for these variables.

ITEM 6: Household-Level Data Table Description

The household-level dataset has one row per household.

Unique identifier (Primary Key): hh_id

Example: 181035337.

Travel dates (first_travel_date, last_travel_date)

These dates indicate the first and last days of a participant's travel period.

Home location variables (home_lat, home_lon, sample_home_lat, sample_home_lon)

Reported home locations are either reported by the participant in the Part 1 recruit survey (participation groups 1 and 2) or derived from trip ends where the participant reports "Home" as the trip purpose (participation group 3). This derivation process for participation group 3 is described in more detail below as other factors like the dwell time at the destination are also taken into account. The home location coordinates used to mail invitations (sample_home_lon, sample_home_lat) are also provided as a secondary reference for home location. Participants with reported home locations outside of the study area have been removed from the dataset.

Home location geography variables (home_bg, home_county, home_state, sample_home_bg, sample_home_county, sample_home_state)

Derived home locations are joined to state, county, and block group geographies using a spatial join. The state, county, and block group shapefiles were retrieved from the U.S. Census Bureau 2018 data files.

Household income variables (income_detailed, income_broad)

Household income was first asked via a detailed question with ten income categories plus a "prefer not to answer" option ("income_detailed"). Participants who selected "prefer not to answer" were asked a follow-up question using five broader income categories plus a "prefer not to answer" option. The responses to both questions are combined in the variable "income_broad."

Number of workers (num_workers)

Workers are defined as participants who responded that they are: 1) employed full-time (paid), 2) employed part-time (paid), or 3) self-employed. This variable provides a count of workers in the household.

Number of students (num_students)

Students are defined here as adults who are in school either 1) full-time or 2) part-time and any child (under age 18). This variable provides a count of students in the household.

ITEM 7: Person-Level Data Table Description

The person-level dataset contains one row per person. This dataset includes rows for all household members, including adults and children.

Unique identifier (Primary Key): person_id (derived using hh_id + person_num)

Example: 18103533701 = hh_id (181035337) + person_num (01).

School and work location variables (school_lat, school_lon, work_lat, work_lon)

Participants in participation groups 1 and 2 reported their school and work locations as part of the Part 1 recruit survey. School and work locations are derived from trips where the participant reports “school” and “work,” respectively, for participation group 3.

Missing derived school and work locations (missing_school_location, missing_work_location)

When a student (full-time or part-time) is missing a school location, missing_school_location is flagged as 1. When a worker (employed full-time, part-time, self-employed, or volunteer) is missing a work location, missing_work_location is flagged as 1. These situations occur when the person does not take any trips to work or school during their travel period.

School and work location geography variables (school_county, school_bg, school_state, work_county, work_bg, work_state)

Work and school locations are joined to state, county, and block group geographies using a spatial join. The county and block group shapefiles were retrieved from the U.S. Census Bureau.

Shared mobility use (e.g., uses_tnc, tnc_freq)

All adults (age 18 and older) were asked several questions about their typical transportation habits. The questions asked whether participants use shared mobility travel modes (e.g., carshare, bikeshare, smartphone-app ride service) and how often they use them, as well as how often they use transit.

Number of trips reported by person during the travel period (num_trips)

The number of trips (num_trips) is the count of total trip records associated with each person’s ID (person_id) for all travel days.

Number of complete days for persons during the travel period (num_person_days)

The total number of days where a person met the completion criteria for the travel day.

ITEM 8: Vehicle-Level Data Table Description

The vehicle-level dataset contains one row per reported vehicle within each completed household.

Unique identifier (Primary Key): hh_id + vehicle_num

Unique identifiers for each household vehicle can be created by combining hh_id with the vehicle number. The vehicle number (vehicle_num) is the sequence number of the vehicle in the household.

Vehicle year, make, and model (year, make, model)

Households provided the year, make, and model of their vehicle in the recruitment survey based on a database of vehicle models from 1980-2020. Vehicles with a model year of earlier than 1980 are given the value of 1980. Participants could also write in the year, make, and model of their vehicle if they did not see it in the database.

ITEM 9: Day-Level Data Table Description

The day-level dataset has one row per participant per travel day in the study. For participants in groups 1 and 3 who had a 7-day travel period, there are seven records per participant in this table. For participants in group 2 who had a 1-day travel period, there is one record per participants in this table.

Unique identifier (Primary Key): person_id + day_num

Unique identifiers for each person-day can be created by combining person_id with the day number. The day number (day_num) is the sequence number of the travel date.

Number of total trips (num_trips_day)

The number of total trips collected on the travel day.

Reason didn't travel on travel date (no_travel_delivery, no_travel_home_school, no_travel_house_work, no_travel_kids_break, no_travel_no_transport, no_travel_no_work, no_travel_other, no_travel_sick, no_travel_telework, no_travel_weather)

If rMove did not collect any trips (or collected only trips that the user deleted) (participation groups 1 and 3) or the participant did not report any trips (participation group 2), the participant was asked why they did not travel. Note that for participation groups 1 and 3 who used rMove this question was asked based on the status of that person's travel record at the end of the day, and thus may not be consistent with a user's number of trips for a given day if users added trips to their travel record after completing the daily survey or if trips were retained/removed in the trip cleaning process. In these cases, RSG inserted values of -9998 or 995 if the questions were either unanswered or not required.

Day is complete (day_complete)

A flag to indicate which person travel days are complete, according to the completion criteria described above in Item 2.

Number of trips reported by person on travel date (num_trips)

The number of trips (num_trips) is the count of total trip records associated with each person's ID (person_id) on each travel date.

ITEM 10: Trip-Level Data Table Description

The trip-level dataset contains one row per recorded person-trip

Unique identifier (Primary Key): trip_id (derived using hh_id + person_num + trip_num)

Example: 18103533701001 = hh_id (181035337) + person_num (01) + trip_num (001).

Trip number (trip_num)

Trip numbers are sorted according to trip departure time within a person-level record. Trip numbers are unique to each person, with trip records beginning at trip_num = 001.

Origin and destination trip purpose (o_purpose, d_purpose, o_purpose_category, d_purpose_category, o_purpose_imputed, d_purpose_imputed, o_purpose_category_imputed, d_purpose_category_imputed)

Respondents report the purpose of the trip destination in each trip survey. The origin purpose is derived from the destination purpose of the previous trip, except for the first trip in in the travel period or where an rMove trip falls after a trip with item non-response. When the trip purpose at the destination was not reported, that trip's destination purpose and the origin purpose of the subsequent trip are coded -9998.

When purpose was not asked because an analyst split a trip during data cleaning (creating a new destination along a trip), purpose values are derived where possible based on proximity (within 150 meters) to estimated home, work, or school locations. If the location is not proximate to home, work, or school locations, the purpose is set to "other."

The purpose category variables (o_purpose_category, d_purpose_category) contain aggregated purpose values based on the type of purpose at the origin/destination of each trip. Dataset users are welcome to perform their own recoding of the purpose categories as well.

Imputed purposes and purpose categories are also included in this dataset. These imputed purposes indicate cases where a purpose reported by the user is assumed to be inaccurate based on information about that person's other trips (primarily to home, work, and school locations). More information about the process to identify and correct mis-reported purposes and impute new purposes is described in the Technical Report.

Departure and arrival time (depart_time, arrive_time, depart_time_imputed)

Departure and arrival time indicate when a trip began and ended. For groups 1 and 3, rMove collected these times passively except for where users added their own trips. For group 2, these times were reported. Timestamps are in Central Standard time (local to the study area).

rMove users can travel a significant distance before rMove recognizes that they are making a trip, and this can yield invalid or extreme values for trip duration and speed. The field depart_time_imputed provides updated values that can be analyzed where the original speed or duration appear invalid. This is described in the Technical Report.

Trip duration in minutes (duration)

Travel time is derived using the difference between the start and end timestamp of the trip.

Trip distance in miles (distance)

For trips that rMove collects, the trip distance is derived from the sum of all distances between points included in the trip. Trips added by the user in rMove typically only have two points, and thus the distance may not represent the full path distance in this case.

For trips entered online, the trip distance is derived based upon the driving distance calculated by the Bing Maps API using the "shortest" (or fastest) path between the origin and destination. If a driving distance could not be calculated between two locations (typical examples include water- or air-based travel or travel on military bases), this field is left empty.

Implied speed in miles per hour (speed_mph_imputed)

Speed is equal to the derived trip distance over the trip duration (based on depart_time_imputed) in units of miles per hour.

Travel mode (mode_1, mode_2, mode_3, mode_type, mode_type_detailed, mode_type_p_1, mode_type_p_2, mode_type_p_3, mode_type_p_4, mode_type_predicted, predicted_mode_probability)

Respondents could select more than one mode in rMove. If the respondent only selected one mode for the trip, only mode_1 is populated. The other mode columns (mode_2/3) are populated if respondents selected more than one travel mode. The order in which respondents select modes is not recorded, so mode_1 does not necessarily contain the first mode the respondent selected; thus, all modes should be considered.

Online diary participants were only allowed to choose one mode.

Mode_type and mode_type_detailed synthesizes mode_1 to mode_3 down to a single, easier-to-use variable for analytical purposes (so that data users can avoid always referencing all modes on a multi-modal trips). Lower values of mode_type are prioritized over higher mode_type values in the derivation. For example, rail trips, with mode_type 1, are prioritized over walk trips, with mode_type 12. When transit trips were unlinked using the Google API during cleaning, the non-transit legs of the trip were recoded using Google's suggested mode (most frequently "walk" or "car"). The mode_type synthesis prioritized this recoded mode over the original reported mode.

Table 1 below shows the full crosswalk of which detailed modes correspond to which mode_types. Mode type is imputed in cases where the reported mode is unlikely given other reported information, or the mode was not reported in the trip survey. More information about the process is described in the Technical Report.

TABLE 1: MODE HIERARCHY FOR DETERMINING MODE TYPE

Detailed Mode	Detailed Mode Value	Detailed Mode Type Value	Mode Type Value	Mode Type
Northstar	172	1	1	Rail
Light rail (e.g., Blue Line, Green Line)	39	2	1	
Other rail	42	3	1	
School bus	24	4	2	School Bus
Bus Rapid Transit (BRT) (e.g., A Line, C Line, Red Line)	61	5	3	Public Bus
Express bus	55	6	3	
Local bus	23	7	3	
Dial-A-Ride (e.g., Transit Link)	27	8	3	
Metro Mobility	171	9	3	
Employer shuttle/bus	62	10	4	Other Bus
University shuttle/bus	38	11	4	

Detailed Mode	Detailed Mode Value	Detailed Mode Type Value	Mode Type Value	Mode Type
Other private shuttle/bus (e.g., a hotel's, an airport's)	26	12	4	
Vanpool	21	13	4	
Other bus	28	14	4	
Intercity rail (e.g., Amtrak)	41	15	5	Long distance passenger mode
Intercity bus (e.g., Greyhound, Bolt Bus)	25	16	5	
Airplane/helicopter	31	17	5	
Lyft Line, Uberpool, or other shared ride	37	18	6	Smartphone Ridehailing Service
Uber, Lyft, or other smartphone-app ride service	49	19	6	
Regular taxi	36	20	7	For-Hire Vehicle
Other hired car service (e.g., black car, limo)	60	21	7	
Household vehicle 1	6	22	8	Household Vehicle
Household vehicle 2	7	23	8	
Household vehicle 3	8	24	8	
Household vehicle 4	9	25	8	
Household vehicle 5	10	26	8	
Household vehicle 6	11	27	8	
Household vehicle 7	12	28	8	
Household vehicle 8	13	29	8	
Other vehicle in household	16	30	8	
Other motorcycle	47	31	9	Other Vehicle
Car from work	33	32	9	
Friend/relative/colleague's car	34	33	9	
Rental car	17	34	9	
Carpool match (e.g., Waze Carpool)	76	35	9	
Carshare service (e.g., HOURCAR, Car2Go, Zipcar, Maven)	18	36	9	
Peer-to-peer car rental (e.g., Turo, Getaround)	59	37	9	
Other vehicle	22	38	9	
Bicycle owned by my household	2	39	10	Micromobility
Borrowed bicycle (e.g., from friend)	3	40	10	
Bike-share (regular bicycle)	69	41	10	
Bike-share (electric bicycle)	70	42	10	
Other rented bicycle	4	43	10	
Personal scooter or moped (not shared)	77	44	10	
Scooter share: electric push scooter	71	45	10	

Detailed Mode	Detailed Mode Value	Detailed Mode Type Value	Mode Type Value	Mode Type
Scooter share: non-electric push scooter	72	46	10	
Moped share (e.g., Scoot)	73	47	10	
Segway	74	48	10	
Other scooter or moped	75	49	10	
Skateboard/rollerblade	43	50	10	
Boat/ferry/water taxi	32	51	11	Other
Golf cart	44	52	11	
ATV or snowmobile	45	53	11	
Medical transportation service (non-emergency)	63	54	11	
Other	5	55	11	
Walk, jog, or roll using a wheelchair	1	56	12	Walk

Member 1 through N was on trip (hh_member_1-N, where N is the maximum household size)

Participants had the option to report whether other household members traveled with them on a trip. When other household members are reported on a trip, the corresponding hh_member variable is coded as 1.

Invalid trip purpose flag (invalid_purpose_flag)

A trip with a purpose of going to school is flagged if the respondent is not a student, and a trip with a purpose of going to work is flagged if the respondent is not a worker.

Synthetic trip flag (synthetic_trip)

School trips are under-reported for children in the dataset whose trips are reported by an adult in the household using rMove. Synthetic school trips are added to the trip table to help correct under-reporting of child school trips. This is described in the Technical Report.

ITEM 11: Location-Level Data Table Description

The location-level dataset has one row per location point collected along a trip. Only trips collected in rMove have locations in this table, so participation group 2 trips do not have locations in the location table. This dataset uses metric units (to be consistent with the units used in the raw location data from the smartphones) and the missing value codes as sent by each device are retained.

Unique location identifiers

Locations are unique by the collected_time timestamp in conjunction with a trip_id.

Time collected (collected_time)

The timestamp at which the point was collected by rMove.

Location accuracy in meters (accuracy)

For points collected on an Android device: Accuracy is the measurement in meters of the radius in which there is 68% confidence that the point lies. In other words, if you draw a circle centered at this location's latitude and longitude, and with a radius equal to the accuracy, then there is a 68% probability that the true location is inside the circle.

For points collected on an iOS device: Accuracy is the measurement in meters of the radius in which the location is likely to be. Apple currently provides no documentation on the level of confidence of this radius.

Heading in degrees (heading)

Heading is the direction in degrees (due north being 0 or 360) collected by rMove. This is the direction in which the smartphone was traveling when the location point was collected. When heading could not be detected, this variable is empty/missing.

ITEM 12: Derived and recoded variables in this dataset

This dataset includes a combination of variables that were actively collected via survey questions, passively collected via rMove or other metadata, implicitly assigned (e.g., administrative variables such as ID numbers), and derived or recoded (calculated from some combination of other variables). Key derived or recoded variables in this dataset are summarized below.

Household-level Derived Variables

- Number of total trips during travel period
- Home location (based on trip end purposes)
- Home geographies (block group, county, state)
- Aggregate income (based on the initial and follow-up income questions)
- Number of adults
- Number of children
- Number of workers
- Number of students

Person-level Derived Variables

- Number of person trips during travel period
- Number of complete days
- Work/school locations (based on trip end purposes)
- Work/school geographies (block group, county, state)
- School mode
- Work mode

Day-level Variables

- Number of trips per day
- Day completion status

Trip-level Variables:

- Trip speed
- Trip path distance (based on the GPS location data)
- Trip origin and destination geographies (block group, county, state)
- Origin purpose (typically based upon the prior trip’s destination purpose)
- Destination purpose where not asked (analyst split trips)
- Mode type and purpose categories
- Imputed trip purpose
- Access and egress flags
- Trip quality flag

APPENDIX A. SENSITIVE AND PERSONALLY IDENTIFIABLE INFORMATION (PII)

TABLE 2: SENSITIVE AND PERSONALLY IDENTIFIABLE INFORMATION (PII) IN THE DATASET

Variable Name(s)	Data Level	Confidential Data Type
home_lat, home_lon, sample_home_lat, sample_home_lon	Household	PII
income_detailed, income_broad	Household	Sensitive
ethnicity_afam, ethnicity_aiak, ethnicity_asian, ethnicity_hapi, ethnicity_hisp, ethnicity_mideast, ethnicity_other, ethnicity_other_specify, ethnicity_white	Person	Sensitive
school_lat, school_lon	Person	PII
work_lat, work_lon	Person	PII
o_purpose_other, d_purpose_other	Trip	Text entry: May contain sensitive and PII
o_lat, o_lon	Trip	PII
d_lat, d_lon	Trip	PII
lat, lon	Location	PII