**21**

**Harnessing Technology to Collect Behavioral Data**

***Jeffrey LaMondia***

*Auburn University, Auburn, USA*

***Konstadinos G. Goulias***

*University of California, Santa Barbara, USA*

***Jimmy Armoogum***

*IFSTTAR, Noisy le Grand, FRANCE*

**Abstract**

The workshop on harnessing technology to collect behavioral data began with the presentation of the resource paper: “Data Collection Technologies - Past, Present, and Future” by Sen and Bricka followed by a discussion of the paper’s ideas. This was followed by a brief presentation of an “Internet-Based Prompted Recall Activity Travel Survey Using GPS Data” by Mohammadian and a brainstorming session on the best methods for using new technologies, including policy interventions and experiments aided by technologies to measure behavioral data. Our discussion concluded with a research agenda.

1. **Introduction**

Researchers are currently faced with two travel behavior data collection trends: the decline in traditional survey response rates and the emergence of new forms of technology and their possible applications. These trends have led us to re-evaluate our methods for collecting behavioral data to determine which current or emerging methods provide the best response rates and which current or emerging methods provide the most (and appropriately) detailed information. Traditional data collection on travel behavior is currently facing substantial challenges, such as sampling bias (e.g., difficulties to have a comprehensive sampling frame), non-response (e.g., the selective dropout of respondents and the low response leading to small sample sizes), and so forth. However, new technologies may help up to mitigate errors due to these occurrences/phenomena. For example, satellite localisation and mobile phone may improve the accuracy of the temporal framework (i.e., collecting correct departure and arrival times, trip duration, and residency at specific locations) and of the geographical framework (i.e., more accurate location of activities, origins and destinations, trip distances along the network, routes selected) of each trip and of the entire activity-travel pattern in a day or even multiple days. In addition, computer assisted interview systems can allow for the detection of errors during interviews in real time with the possible provision of prompting and rectification questions. In addition to these improvements, the utilisation of new technologies may reduce respondent burden and the survey costs, which should have first-class impacts on data accuracy and quality. Still, researchers agree that there is much to learn about these methodologies to improve their quality and better understand their applications. As such, the workshop identified two main research objectives: (1) develop better means of interpolating small datasets to describe the entirety of populations and (2) develop simple and parsimonious methods for collecting behavioral data.

Emerging technologies offer many exciting possible opportunities to advance the way we collect behavioral data, but we must not ignore efforts to improve current technologies’ performance, efficiency, and effectiveness. Therefore, workshop participants agreed that the best approach to improve the quality and quantity of behavioral data is twofold: (1) to develop better techniques for combining complementary technologies, and (2) to carefully consider how respondents’ interpretations and perceptions of technologies impact their data collection. The aim is to reach desired populations, increase response rates, reduce burden and costs, and improve the quality of data collected. In order to accomplish this, we need to consider (and conduct further research on) five critical topics, discussed in the following sections: Current and Emerging Technologies, General Issues of Burdens/ Costs Associated with Using Technologies to Collect Behavioral Data, Appropriate technology for collecting specific data, Using Technology to Reach Specific Populations, and Opportunities for Combining Technologies.

1. **Summary of Current and Emerging Technologies**

In the discussion, we realized that during the past decades we experienced a dramatic increase in the number of current and emerging technologies that can be used to collect travel behavior data and that they go far beyond traditional paper/pencil surveys. Still, when evaluating each technology it is important to consider that:

* Each technology has its own set of benefits and challenges;
* Each technology collects specific types of information data;
* There is no ‘one-size-fits-all’ technology to collect behavioral data.

We mention here three aspects of these technologies.

**Computer assisted**

In the past two decades, computer assisted surveys have replaced PAPI (paper and pencil interview) surveys. The main benefit of computer assisted surveys is the real time and post processing of the data. Indeed, this stage can be done instantaneously when collecting the data (e.g., instant geocoding, consistency of the data by checking the speed). These computer assisted surveys take many different forms: Computer Aided Personal Interview (CAPI), Computer Aided Telephone Interview (CATI), Computer Aided Self Interview (CASI) and Computer Aided Web Interview (CAWI).

Drawbacks for CATI surveys are the accessibility of a sampling base of the full population (as the sources for a sampling frame are incomplete and with the diffusion of cell phone people are less equipped with land line phones that are the usual telecommunication employed in this type of interview, double counting if several devices, contacting a household with a land line phone or an individual with a mobile phone). The duration of the survey (as it should not exceed half an hour) and the number of persons (it is difficult to reach all members of the household with one phone call) to be interviewed are limited. The cost of a CAPI surveys is very high (much more than a CATI) but the response rates appear to be higher than any other type survey (this was not externally verified but reported at the workshop).

In addition, when surveys are made by interviewers, they may influence the response from interviewees (whatever the type of survey PAPI/CAPI or CATI). For the participants of the workshop, another issue was the cost of the survey which seems to be costly with these methods (except for CASI).In CASI surveys the participants may also introduce variability in responses due to personal interpretation of questions, high selective attrition rates, missing of follow up information, and no control over survey environment.

**Position tracking**

High resolution data loggers such as Global Positioning System (GPS),Global System for Mobile communications (GSM), Radio Data System (RDS) or Radio Frequency Identification (RFID), or WIFI especially inside buildings can be personal wearable and/or Vehicle-Installed. Each technology measures one’s location and movement patterns. The workshop participants also noted that RDS could be very useful for measuring freight long distance movement. New possibilities on position tracking may improve the resolution of the temporal framework parameters (e.g., departure and arrival times, duration of activities) and of the geographical framework (e.g., location of activities, origin and destination, trip distance, routes) of each trip/activity. The use of these devices considerably reduces respondents’ burden, and it is possible to extend the period under review (collection of more than one day of mobility) as well as give us an opportunity to design policy experiments and study behaviour in a before and after a policy actions situations. Moreover, it is a good way to collect mobility/activities of extremely mobile persons. Some of the drawbacks, however, include: Forgetting to wear the data logger, missing or misreported data, missing trip details, extreme post processing costs, possible learning burden, route matching and mode alternatives that challenge the analysts intuition.

The workshop participants also quickly mentioned a number of additional emerging technologies such as RFID in documents, Crowd sourcing (i.e. analyzing feedback that end users provide on reports and data) and Active data sourcing (i.e. group information sharing, such as Wikipedia). These sources of information provide additional data on the variety of communication people engage in and their ability to extend their reach beyond their physical capabilities in a day (called human extensibility).

**Internet and/or computer aided**

The attendees of the workshop confirmed a number of computer aided technologies that may help produce data to understand travel behavior. For example Computer Aided Transaction Logging (CATL) can collect the use of smart cards and their associated activities/travel, including origin/destination information, also expenditures (useful for the analysis of tourism) and once in place, it constantly generates data. Computer Aided Activity Logging (CAAL) may collect data about the search process and choice set development within websites providing insights of choice set development and decision-making processes, via unobtrusive data collection. It is, however, difficult to get access to this information (from blogs, twitter, Facebook, search for travel on internet) although the technology of analyzing tags is rapidly developing. A very important potential is the development of the Computer Aided Experimental Protocols (CAEP) that collect choice responses to virtual reality/simulator scenarios, stress levels, decision process patterns, stimuli and responses, and so forth. This is approach may capture the entire decision making processes at unprecedented level of detail and information.

Workshop participants were also especially concerned with how current and emerging technologies can be used to collect/enter activity data in real time and the relationship of this to some population groups for which this type of data collection may not be appropriate or even feasible.

1. **General Issues of Burdens/Costs Associated with Using Technologies to Collect Behavioral Data**

The topics of burden, costs, and nonresponsiveness were also addressed in the workshop. For example, who are the respondents and the nonrespondents of a “classical” mobility survey, and are they different from the respondents of mobility surveys using new technologies? The general impressions is that they seem to be different;, recent studies in the US and in France have shown that those households with low income and persons that are not very familiar with new technologies were less willing to participate in GPS surveys over traditional methods. Furthermore, we currently do not have consistent quantitative measures of burden. In fact, current measures are rather simplified (e.g., monetary cost by information gained or number of completed surveys). Workshop participants determined that there are two types of burden (and costs) associated with technologies that need to be considered and further explored:

* Burden on respondents
* Burden on researchers (intended as the data collector)

Experienced burden is based on what respondents and researchers perceive to be a realistic amount of effort required for a given situation. This varies greatly by population group and type of research study. Therefore, it would be helpful to develop measures in order to evaluate the appropriateness and efficiency of different technological methods.

A third aspect is the use of multiple technologies. Sometimes, a single technology might not be enough. As a result, we may need to optimize data collection by combining technologies. In this respect, respondent burden is critical because it has the potential to bias collected data. But within the population perceived burden is not consistent across all types of respondents. Learning curves and experience with technologies is different across population groups (for example: older populations not familiar with technology will face more burden than younger populations familiar with technology).

A fourth is the length of data collection that can impose added burden. Indeed, respondents have limited free time, many other commitments, and their value of time may be increasing. On-site or on-board surveys reduce respondent burden, unless relatively time-consuming. A day, a week, and longer surveys inherently increase respondent burden due to, increased stress, fatigue and loss of interest. But, incentives may mitigate/reduce respondents’ experienced burden.

The biases due to increased respondent burden can creep into the data collection in a number of ways. The respondents may simplify complex tasks to save time. This is especially true for longitudinal studies as respondents get tired on continually entering information. Moreover, the respondents may unknowingly misinterpret questions by fatigue.

We turn to the researcher burden which is conditional on expertise, skills, and available resources (both personnel and technological). While some advanced methods cost more up front, researchers can save on post processing because it requires less time and effort (for example: Advance phone applications). On the other hand, some simplified methods cost less up front, but require more post processing to get useable data (for example: GPS data collection).

Researchers are also concerned with formatting data collected from technology. The cleaning and entering collected data can lead to inconsistencies, errors (both human and computer), missing information, etc. But these concerns vary by type of technology used. Additionally, researchers are concerned with how effective technologies can be tied into modeling applications to make both more informative. For example can we use new technologies (and their associated data) to improve/ change the modeling process? In turn, can we use the modeling process to make the technologies more effective in collecting the data needed? All this is heavily dependent on the type of technology and the data it can provide and we address the issues in the following section.

1. **Appropriate Technology for Collecting Specific Data**

Behavioral research encompasses a broad range of topics; as a result, behavioral researchers are interested in using technology to collect an increasingly wide variety of data. Workshop participants were particularly interested in harnessing technology to collect detailed data on:

* Activity participation;
* Trip-making patterns;
* Origin/destinations;
* Mapping travel behavior;
* Attitudes towards different alternatives and scenarios;
* How respondents’ selected and compared alternatives;
* The entire search and decision-making process;
* Stress, thought processes, and stimuli associated with behavior.

Naturally, no single technology can collect data on all of these topics. Typically we think in terms of researchers needing to select those technologies that maximize the quality of data collected and minimize the burden on respondents and researchers. However, there is more to selecting a technology than these two criteria. In order to fully harness technology to collect behavioral data, we must consider many related issues:

* What specific type(s) of information are needed to address the specific question?
* How specific type(s) of people need to be targeted to address the specific question?
* How will these people interpret and perceive different technologies?
* How can technologies be adapted to minimize misinterpretations and error?
* How will the way questions are presented change across different technologies?
* What is the required timeframe (i.e. one day or long-term behavior)?
* What is the required scope (i.e. metropolitan planning versus specific group analyses)?
* Are these limitations with different technologies regarding stated preference or revealed preference questions?
* Will the way questions are presented or the type of technology itself affect behavior or responses?
* Does the technology allow follow-up questions on respondents’ thought processes or furtively collect information-seeking data during the survey?
* Does data needed to be collected in real-time, and which technologies support this?

All of these issues must be taken into account when selecting the most appropriate data collection technology, as they consider and resolve these issues differently. Additionally, we need to further study how technologies can be improved to better accommodate experimental design flexibility. GPS is currently one of the most commonly used technologies. Yet, the workshop participants recognized that there are still many issues with this technology that need to be addressed in order to harness its full potential including:

* GPS distances and times (RP) are quite different from reported distance and times (Perceptions, SP) and both provide useful information
* Extracting activity information from GPS requires a considerable amount of effort
* OD pairs are important and easy to determine using GPS, but we are still missing complete mode/route choice alternatives and characteristics
* Matching routes can be challenging

Web-based surveys are also common, and workshop participants identified issues with this technology as well and the most important are: a) the resulting sample may not be representative (actually we know it is not) of the population, and b) we have important differential attrition rates and at the same time we do not have address/phone number information for these participants for follow-ups or spatial analysis.

1. **Using Technology to Reach Specific Populations**

In addition to collecting detailed information, researchers are also concerned with being able to improve response rates, reach specific populations, and being able to extract samples to describe the full population. As such, we need to understand which technologies allow us to reach specific (and appropriate) populations. In the discussion so far, we have seen that specific population groups perceive and react to technologies differently. For example, older populations require more training to use GPS technology. Comprehension, familiarity, and ease-of-use are all important factors as to whether a specific population will embrace a technology. In fact, younger populations are interested in more advanced technology (i.e. smart phones, twitter, blogs) because they use this technology every day and it is part of their lifestyle and communication habits.

Not only do various groups perceive technologies differently, but they may actually more readily provide information if they enjoy using a specific technology. Given a choice, older populations may prefer paper/pencil surveys and younger populations may prefer web-based surveys. But all this may be changing rapidly as older populations engage with new technology (e.g., that fastest growing user group for Facebook are persons that are older than 65 years).

Once again, the participants of the workshop emphasized the difficulty of using a single technology to collect data on all population groups. This continually led to the question of how to choose which technology is most appropriate for each population group? Workshop participants concluded that age-based or wealth-based population groups may no longer be the most optimal method for disseminating technology. Instead, we need to consider clustering by lifestyles and to do this we face some challenges that are : (a) Lifestyles are comprised of socioeconomics AND common behaviors leading to a circular issues between surveys and explaining behavior; (b) forecasting and identifying lifestyles is a critical future research topic on which we know little; (c) new need to further distinguish between dwellings and households and; (d) individuals’ and households’ lifestyles will change over time as well, leading to lifestyle cycles. Different technologies fit different lifestyles better than different populations. Technology needs to fit the way of life of those you want to reach. Once researchers determine lifestyle clusters, they can use traditional probability samples of lifestyles to reach different groups, but we still need to consider how many lifestyle clusters are appropriate.

Ultimately, workshop participants recognized that much of the success of improving response rates lies in returning to simple solutions, like probability sampling techniques,. We also need to motivate respondents with interesting and tangible topics (although we need to be concerned about biases), and we need to explore the effect of incentives on response rates. In the USA, many surveys use incentives and continue to receive high(er) response rates. Individuals’ value of time is increasing, so we need to find appropriate incentives, which may not always be money. Moreover, we must keep in touch with interviewees, including many call-backs and visits.

1. **Opportunities for Combining Technologies**

Workshop participants concluded that each technology is successful at collecting a specific type of information and at reaching a specific population group. However, the caveat to this is that these technologies may not always coincide or be comprehensive to a research study. Therefore, we determined that it is important to identify better techniques for combining complementary technologies. The best approach to improve the quality and quantity of behavioral data was twofold: First, we need to develop better techniques for combining complementary technologies. Second, we need to carefully consider how respondents’ interpretations and perceptions of technologies impacts data collection.

One technology may not be enough to either reach desired populations or collect desired data. We have already seen examples of this in current research, for example, combining GPS records with GIS maps or with travel diaries, such as trip tracing software. All this offers additional insights, but also requires respondents to make “connections” among software modules and remain thorough. It is also important to consider different combination methods:

* Within individual, where each individual uses different methods to collect different data (for example GPS with travel diaries).
* Across groups, where each individual uses one method, which varies based on group membership (for example: older groups receive paper/pencil surveys and younger groups receive web-based surveys). Why not let them choose their survey instrument.

Workshop participants also considered the usefulness of letting respondents pick their technology from a few options. But this may lead to possible complications, including the need to harmonize the data from different type of data collection to respect the same methodologies. Furthermore, the issues associated with each technology, such as interpretations, perceptions, biases, only increase when we start combining them in a single experimental design. Specifically, researchers need to consider:

* How definitions, interpretations, and error vary across technologies
* Human and computer errors need to be considered
* Whether or not technologies are even compatible
* How to deal with data discrepancies when combining technologies

1. **Examples of Current Studies**

During the workshop we also looked for success stories with examples of combining technologies. In Santiago, Chile, researchers are combining data from GPS units installed in all transit system vehicles and traveler fare cards, which keep track of all transactions and travel patterns. In Cagliari, Italy, researchers are combining GPS location data from study participants’ phones with travel activity information through phone applications and maps. These two examples are illustrative of what we should expect to see as emerging trends in future studies and they demonstrate the potential for enhanced designs in data collection methods that make heavier use of technology. They also provide the context within which we developed a research agenda that concludes this report.

1. **Important Research Objectives**

The workshop participants identified a number of future research topics, falling into two categories that are about studies of technology methods directly and about research that uses these technologies to advance our insights into travel behavior. The studies of technology methods (and survey methodology-related) include: (a) Methods for harmonizing data to measure the same entity (construct validity, unit, etc.); (b)Methods to assess the quality of data; (c) Targeting ‘hard to reach’ groups, such as low income/ minority/disabled/ elderly; (d) Mapping of appropriate methods for different lifestyles; (e) Technologies to improve coverage schemes targeting behaviors of interest; (f) Methods to address confidentiality of data, methods to create anonymity and at the same time maintain utility of data and ways to deal with ‘privacy fundamentalists’ ; (g) Data combination opportunities and exploration of their potential, including qualitative analysis; (h) Education and training for practitioners; (i) Identification of opportunities emerging from each technology method; (j) Questions about the potential of methods of collecting information in affecting behavior when this is not desired; (k) Lessons learned from the use of technologies for surveys; (l) High resolution data means more data but computational issues and reduction into meaningful constructs requires further scrutiny.

Some of the travel behavior insight-related topics include: (a) Ways to use new technologies to predict changing behavior in real-time as situations change (in an experimental setting or in usual settings); (b) Identification of the appropriate and optimal techniques to use when we study travel behavior modification; (c) Development of paradigms and survey technologies and behavioral relationships that can be examined using them; (d) Identification of travel pattern complexity, changes in behavioral framework, and applications to describe decision-making processes; and (e) Studies of the potential to use the technology for increasing real-time data information diffusion and impacts on behavior.

**Acknowledgement**

The workshop chairs sincerely thank the resource paper authors and all of the workshop participants for their valuable insights and intellectually stimulating perspectives that greatly contributed to the success of the workshop and made this report possible.