Exhibit D – Research Project Requirement Template

Data Collection, Weighting, and Modeling Techniques to Estimate Unbiased Population Parameters

Recipient/Grant (Contract) Number: The University of Texas at Austin/Grant # 69A3552344815 and 69A3552348320

Center Name: National Center for Understanding Future Travel Behavior and Demand (TBD)

Research Priority: Improving Mobility of People and Goods

Principal Investigator(s): Chandra R. Bhat

Project Partners: N/A

Research Project Funding: $150,000 ($75,000 Federal + $75,000 matching funds)

Project Start and End Date: 6/1/2025 – 5/31/2026

Project Description: Empirical research studies across multiple fields employ data from large surveys for their analysis. In doing so, studies must address such sampling-related issues as non-response, missing data, unequal sampling, and other survey biases. The voluntary nature of most surveys means that, in many empirical applications, data are not randomly selected from the population. Instead, researchers only observe the responses of those who choose to respond to the survey, potentially resulting in sample selection biases. A variety of modeling approaches have been proposed to accommodate such selection biases. Specifically, sampling weights have long been considered essential when undertaking descriptive statistical analysis (such as determining population averages) on data with unequal sampling probabilities. However, for causal effects modeling, if individuals have nearly equal sampling selection probabilities given their values of exogenous variables, then both weighted and unweighted estimators are consistent, and the lower variance of the unweighted estimator is preferred. But, when the probability of selection differs significantly among individuals due to a selection mechanism that is endogenous (that is, the probability of selection is not completely explainable based on exogenous variables), using sampling weights (representing the inverse probability of sample selection) can yield consistent estimates of population parameters, while unweighted estimators are generally inconsistent.

A critical issue, however, is that the true probability of selection is generally unknown in cases of nonresponse. In such cases, weights are not based on the true probability of selection. Instead, they are estimated using post-data collection comparisons with population statistics to match the proportion of respondents in each demographic group with their population proportions in an external independent control. The basic idea is that, by employing such weighting, one essentially gets back to the case of an equal probability sample, which takes care of any selection bias. However, unobserved factors may also play a significant role in response decisions (and thus, sampling probabilities), and such unobserved factors may also be correlated with the main outcome of interest. Such situations cannot be addressed through post-data collection weights, which rely on the assumption that selection is based solely on observed characteristics. Relatedly, while descriptive statistics are often considered separately from model-based approaches, since weights are needed even when sampling is based only on exogenous variables using standard formulas, these same statistics can be calculated using model-based approaches. While either approach yields an unbiased result when sampling is based only on exogenous variables, the traditional weight-based approach cannot accommodate unobserved self-selection effects while the model-based approach can accommodate such selection on unobserved variables.

Therefore, in this study, we consider the ways that appropriate sampling strategies and modeling techniques can be used to improve estimation results when the collection of a representative sample is unnecessary or impractical. Through theoretical and simulation-backed support, we underscore the importance of adopting appropriate sampling and estimation methods when sampling is based (a) only on observed exogenous variables, or (b) also on unobserved variables. In the context of exogenous sampling, we demonstrate that range variation in exogenous variables needs to be the key in survey designs (not necessarily population representativeness) to estimate individual-level causal relationships. Further, we demonstrate that weighting approaches are unable to accommodate endogenous selection, when sampling is based on unobserved variables. Instead, we propose a joint modeling approach that can recover the true population parameters when the joint distribution of exogenous variables in the population is known. We also demonstrate that this method can improve upon existing methods that do not account for endogenous selection even when only the population marginal distribution of exogenous variables is known. This analysis should be of interest to all empirical researchers working in the area of survey research and associated data modeling.

US DOT Priorities: This project supports USDOT's *transformation* priority by advancing the methodological foundations of survey-based research, a cornerstone of transportation data and decision-making. By rigorously evaluating the use of sampling weights, modeling techniques, and joint estimation approaches to address selection bias, the project enhances the reliability of insights derived from large-scale survey data. These advances will enable more accurate estimation of population parameters and causal effects, providing researchers and policymakers with robust evidence to guide transformative investments, planning strategies, and policy design in a rapidly evolving transportation landscape.

Outputs: Expected research outputs include a journal publication and additional conference presentations. These outputs will highlight specific practices that researchers and policymakers can use during data collection processes and data analysis to ensure that samples are collected efficiently and analysis methods that minimize bias in estimation are used. These results and implications will be developed into a series of tangible recommendations for practical surveying and modeling applications that can be readily disseminated and applied by researchers and policymakers.

Outcomes/Impacts: Expected research outcomes include increased understanding and awareness of the implications of sampling biases on modeling results, which impact a variety of transportation outcomes. By exploring the specific process through which these biases occur and identifying new modeling strategies that can be applied to alleviate the impacts of sampling biases, researchers, transportation officials, and policymakers will be better equipped to align policies, regulations, and legislation to the needs of the general public as revealed through these transportation surveys. The ability to accurately measure and predict a wide range of transportation behaviors will be improved through the processes developed.

The impacts of improving sampling and modeling procedures to reduce sampling biases are widespread. Using these techniques, transportation demand modeling and demand management policies can be improved to better predict behavioral responses to infrastructure changes across the entire network and entire population, efforts aimed at promoting the adoption of transportation technologies can be better aligned with the needs of all potential users, and research and development efforts can be better generalized beyond the collected sample. These impacts are particularly important in an era when survey response rates are declining and recruiting respondents has become particularly challenging for specific subregions or population subgroups, exacerbating challenges associated with data representativeness from these samples.

Final Research Report: A URL link to the final report will be provided upon completion of the project.