Missing by Design: A Framework for Statistically Efficient Survey Design with Reduced Response-burden

hEART 2026 — Extended Abstract

Alorabi Reem[1], Prateek Bansal[2], and Makoto Chikaraishi*[1]

[1]Graduate School of Advanced Science and Engineering, Hiroshima University
[2]Department of Civil and Environmental Engineering, National University of Singapore

1. Introduction

Collecting behavioral data is an exercise in constrained judgment: ask too little and the model is blind; ask too much and the respondent walks away. Every question added to a travel survey sharpens the model to be estimated — and costs the respondent something in return: time, attention, and in the case of sensitive items such as income or employment status, a loss of trust. The consequences of asking too much are well documented. Respondent fatigue induces satisficing behavior that systematically degrades response quality (Krosnick, 1999), item nonresponse rises with survey length (Groves et al., 2009), and in travel and time-use surveys specifically, burden has been shown to affect both who participates and the care with which they respond (Stopher and Greaves, 2007).

The trade-off between statistical efficiency and respondent burden remains underexplored in the literature. Dillman et al. (2014) note that efforts to shorten surveys are typically driven by considerations of respondent experience rather than by a formal assessment of the statistical information required. Nevertheless, there is a substantial body of work in the literature on the "design of experiments" that focuses on maximizing inferential efficiency by optimizing combinations of attribute levels (Rose and Bliemer, 2009). However, these studies generally treat the number of tasks, and thus respondent burden, as predetermined, rather than as a design variable in its own right. Beyond experimental settings, the broader survey design literature offers limited guidance on the more fundamental question of "which questions should be asked to whom." This constitutes a challenging design problem that must be addressed jointly with the imputation of missing covariates during estimation, as missing responses on key variables often necessitate discarding entire observations, thereby undermining reliable parameter estimation and inference (Train, 2009; Pinjari and Bhat, 2011).Thus, there is a structural gap between the personalized survey questionnaire design and the model estimation, which we address in this study.

Rather than designing surveys for statistical representation, we design them for model estimation, explicitly aligning data collection with inferential objectives. We consider the survey design and model estimation in a unified framework. We treat partial observation not as an inconvenient

*Corresponding author: chikaraishim@hiroshima-u.ac.jp

property of data, but as a *controllable design variable*. The central question is not how to cope with missing data, but how to *design* it: which optional questions should be displayed to which respondents, under what rules, and in exchange for what inferential guarantees? This paper makes two main contributions.

The first contribution develops a framework for deciding **what to measure and from whom** under a constraint on respondent burden. In behavioral survey design, the selection of covariates has direct inferential consequences, yet lacks formal treatment. We formulate this as an optimization problem: given a behavioral model and a constraint on respondent burden, which questions should be asked to which respondents, and in what configuration to preserve inferential quality. The tradeoff between what is asked and what is learned is not incidental — it is the central objective of the framework.

The second contribution follows directly from the first. A design that withholds questions necessarily produces incomplete data — and incomplete data must be handled. Rather than delegating this responsibility to a separate imputation stage, our framework treats it as an integral part of the estimation problem. Missing covariate values are recovered jointly with the structural parameters of the behavioral model, so that the act of designing the missingness and the act of inferring under it are governed by the same probabilistic machinery. The framework does not choose what to leave unobserved and then hope for the best — it chooses what to leave unobserved precisely because it can recover it.

We instantiate the framework on daily time-use behavior modeled through a Multiple Discrete-Continuous Extreme Value (MDCEV) model, estimated on $N = 42{,}199$ respondents from the American Time Use Survey (ATUS). This application serves as a proof of concept: we develop the full framework and validate its inferential engine through systematic benchmark experiments, while the policy optimization loop — formulated here — is left for empirical evaluation as the immediate next step. The framework is general and applies to any behavioral model in which a Bayesian posterior can be computed over partially observed data.

2. Framework

The framework is built around a single organizing idea: the decision of what to collect from whom is not separate from the decision of how to estimate — it is the same decision, made jointly. To make this concrete, we describe the framework as a pipeline with three stages that operate in sequence and feed back into each other: a *design stage* that determines the observation pattern, an *inference stage* that estimates the behavioral model under that pattern, and an *optimization stage* that evaluates the result and improves the design. We describe the full pipeline first, then formalize each component.

2.1 The Full Pipeline

The pipeline begins before any respondent is surveyed. Covariates are partitioned into two classes: *mandatory* items collected from every respondent without exception, and *optional* items that are candidates for selective collection. Mandatory items are low burden and universally available — core demographics or basic identifiers. Optional items are costly, sensitive, or burden-intensive —

income, education, employment history, and similar socioeconomic attributes.

A *collection policy* then decides, for each respondent, which optional questions to ask — using only what the mandatory questions have already revealed about that respondent. When a respondent enters the survey, the mandatory questions are answered first; the policy then fires, generating a realized observation pattern R that specifies which optional questions are asked. Some optional variables are collected (Z_{obs}); the rest are intentionally left unobserved (Z_{mis}).

With the data in hand, the inference stage estimates the behavioral model. Crucially, Z_{mis} is not discarded or imputed separately — it is treated as a set of latent variables and sampled *jointly* with the structural model parameters θ during MCMC. The posterior over (θ, Z_{mis}) is therefore conditioned on everything that was observed, and imputation is a byproduct of estimation rather than a prerequisite for it.

The optimization stage evaluates how well this design performed. It scores the realized observation pattern against an objective that balances two quantities: how faithfully the design preserves behavioral inference, and how much burden it imposed on respondents. This score is returned to the *upper level* of a bilevel optimization — the design problem — which proposes an improved policy, while the *lower level* — the inference problem — solves the estimation for each candidate design. The loop repeats until the objective stabilizes. The output is a trained policy that can be applied to any future respondent: given their mandatory covariates, it specifies which optional questions to ask.

2.2 Covariate Partition and Observation Structure

Formally, let Y_i denote the behavioral outcome for respondent i and Z_i a vector of covariates, partitioned as:

$$Z_i = \big(Z_i^{\text{Mand}},\ Z_i^{\text{Opt}}\big), \tag{1}$$

where Z_i^{Mand} is always collected and Z_i^{Opt} is subject to the design. For each respondent i and optional variable j, the binary indicator $R_{ij} \in \{0, 1\}$ records whether variable j was collected. The full pattern $R = \{R_{ij}\}$ partitions Z^{Opt} into Z_{obs} and Z_{mis}. The behavioral model of interest is:

$$p(Y_i \mid Z_i^{\text{Mand}}, Z_i^{\text{Opt}}, \theta), \tag{2}$$

where θ is the structural parameter vector.

2.3 The Collection Policy

The collection policy is the mechanism through which the framework decides, for each respondent, which optional questions to display. Formally, it maps a respondent's mandatory covariates to a set of collection probabilities — one per optional question — that govern whether each question is administered:

$$\Pr\big(R_{ij} = 1 \mid Z_i^{\text{Mand}}\big) = f_j(Z_i^{\text{Mand}};\, \psi_j), \tag{3}$$

where $f_j(\cdot)$ is a function to be specified and ψ_j are its parameters. The choice of functional form depends on the nature of the optional variables being modeled: continuous optional variables suggest one specification, while ordinal or categorical optional variables may require a different treatment.

This choice is a modeling decision that we defer to the implementation stage; what matters structurally is that the policy is a function of Z_i^{Mand} alone — information already in hand at the time of survey administration.

The collection policy parameters $\psi = \{\psi_j\}$ are the design variables of the entire framework. Rather than searching over all binary observation matrices $R \in \{0,1\}^{N\times J}$ — a space of size 2^{NJ} that grows combinatorially with sample size and number of questions — the optimization operates over the continuous parameter space of ψ. This reduction in the design space is a necessary condition for the optimization to be well-posed. It does not, however, resolve the computational cost of the lower-level inference, which must be evaluated for each candidate policy. Scalable evaluation of the lower level — through posterior approximations or surrogate modeling — is an important open direction that we discuss in the conclusion.

2.4 Joint Inference and Imputation

When covariates are intentionally left unobserved, the likelihood $p(Y_i \mid Z_i, \theta)$ cannot be evaluated for respondents with incomplete Z_i. Classical approaches resolve this by treating imputation and estimation as sequential problems: missing values are deleted, substituted, or imputed in a preparatory stage, and the behavioral model is estimated on the completed data (Little and Rubin, 2002). This sequential treatment is fundamentally incompatible with a design optimization framework. When imputation precedes estimation, the uncertainty induced by designed missingness does not propagate into the parameter posteriors, and the inferential consequences of any given design decision cannot be measured consistently.

The Bayesian framework resolves this through the principle of data augmentation (Tanner and Wong, 1987): rather than treating Z_{mis} as an obstacle to be removed before estimation, it is introduced as a latent variable and recovered jointly with the structural parameters θ within a single inferential procedure. The joint posterior over all unknowns is:

$$p\big(\theta,\, \phi,\, Z_{\text{mis}} \mid Y,\, Z_{\text{obs}},\, R\big) \;\propto\; p(Y \mid Z_{\text{obs}}, Z_{\text{mis}}, \theta)\; p(Z_{\text{mis}} \mid Z_{\text{obs}}, \phi)\; p(\theta)\, p(\phi), \tag{4}$$

where ϕ parameterizes the conditional distribution of optional variables given mandatory ones. Posterior sampling proceeds by alternating between draws of (θ, ϕ) conditional on the current state of Z_{mis}, and draws of Z_{mis} conditional on current parameter estimates and observed data. Each draw of Z_{mis} constitutes a probabilistically coherent completion of the unobserved covariates, fully conditioned on everything the survey did collect. Imputation and estimation are not sequential operations — they are a single joint inference.

This has a direct structural consequence for the design framework. Because Z_{mis} and θ are recovered within the same posterior, the inferential cost of any design decision is endogenous to the estimation: a collection policy that withholds more optional variables yields a less constrained posterior over Z_{mis}, which propagates uncertainty into θ in a manner that is directly measurable and directly optimizable. Design decisions and their inferential consequences are therefore governed by the same probabilistic model — a coherence that no classical two-stage pipeline can provide.

2.5 The Bilevel Optimization

The optimization stage connects the design and inference layers througha bilevel problem. The upper level optimizes the policy parameters:

$$\max_{\psi} \quad \underbrace{\mathcal{I}(R_\psi)}_{\text{inferential quality}} \quad - \lambda \quad \underbrace{\mathcal{B}(R_\psi)}_{\text{respondent burden}} \quad , \tag{5}$$

subject to the lower-level posterior in equation (4). The two terms serve distinct roles: $\mathcal{I}$ measures how well the design preserves the behavioral estimands relative to the full-information benchmark, and $\mathcal{B}$ measures the expected survey burden across respondents. The weight λ is set by the analyst to reflect the acceptable cost of burden reduction against inferential fidelity.

For each candidate policy, the upper level generates an observation pattern R, the lower level solves the joint posterior in equation (4), and the objective is evaluated. A refined policy is then proposed and the loop repeats until convergence of theobjective. The result is a policy that specifies, for any new respondent, which optional questions should be administered given their mandatory covariates.

In the present paper, we present the full bilevel framework and validate its inferential engine — the lower-level posterior and its behavior under designed missingness — through systematic benchmark experiments. The execution of the upper-level optimization loop constitutes the immediate next step of this research agenda.

3. Proof-of-Concept Application

3.1 Model and Data

To demonstrate the proposed framework, we employ the MDCEV model (Bhat, 2005, 2008), which captures the joint discrete-continuous structure of time allocation: which activities receive positive time and how much. Each respondent allocates a fixed budget of 1,440 minutes across ten activity categories (personal care as outside good; leisure, work, household production, travel, eating and drinking, caring, shopping, civic/religious engagement, and education as inside goods). The model is estimated using Bayesian methods (Chikaraishi et al., 2010).

The dataset pools five years of ATUS (2020–2024), yielding $N = 42{,}199$ respondents after removing 23 cases with invalid time budgets. The covariate vector Z_i contains 24 variables: socioeconomic characteristics (age, gender, education, income, number of children, marital status, employment status, hours worked), household composition, geography (region, urban/rural), and calendar controls (day of week, season, year fixed effects for COVID and post-COVID period shifts). The full model has 196 parameters estimated across nine activity utility functions.

For the missing-by-design (MBD) experiments, five continuous socioeconomic variables are designated as optional covariates Z^{Opt}: *age, number of children, education level, income, and hours worked*. All remaining variables constitute Z^{Mand}.

3.2 Benchmark Estimation

The full-information model is estimated by maximum likelihood and achieves a log-likelihood of $-1{,}517{,}983$ ($\rho^2 = 0.028$), with 204 parameters estimated and 129 statistically significant.

The Bayesian version of the same full model is estimated using NUTS via PyMC. Figure 1 confirms that MLE estimates and Bayesian posterior means are in near-perfect agreement across all 196 parameters (ASC, β, and η types), with points lying tightly along the identity line. This is an important prerequisite: it establishes that the Bayesian posterior is a faithful representation of the full-information model, so that posterior-based diagnostics in the missing-data experiments are comparable to the MLE benchmark.

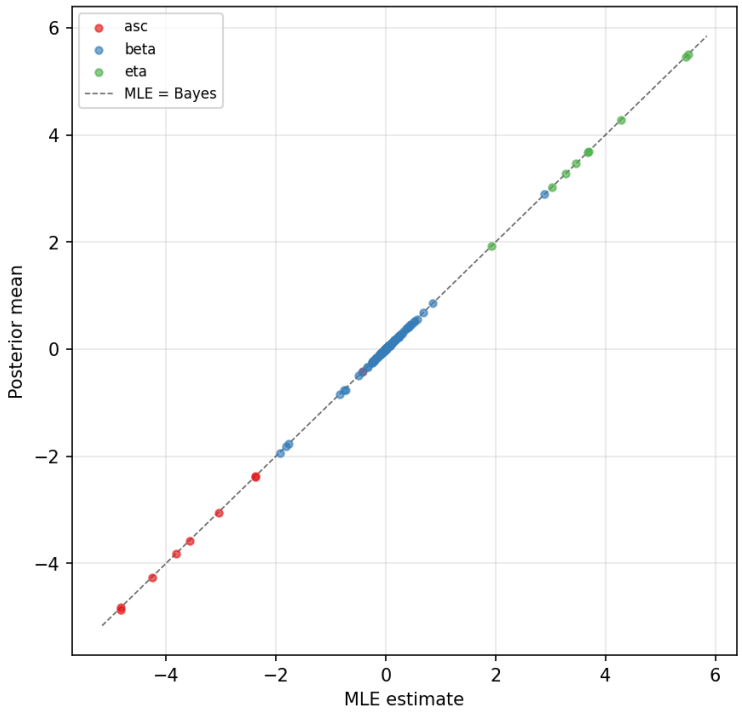


Figure 1: MLE point estimates versus Bayesian posterior means for all 196 parameters of the full-information MDCEV model. Near-perfect agreement along the identity line validates the Bayesian benchmark as the reference for all subsequent experiments.

4. Validating the Inferential Engine

Before optimizing the design, the lower-level inferential engine must be shown to behave reliably across a range of missingness rates. The following benchmark experiments use completely random (MCAR) missingness. MCAR is the natural baseline because it is the inferential analog of random assignment — it makes no assumptions about which respondents or covariates are more informative. Any policy-optimized design should be evaluated relative to what random assignment achieves at the same burden level.

Three experiments are reported in Table 1. All MCMC runs converge cleanly ($\hat{R}_{\max} \leq 1.01$, zero divergences in all cases).

Table 1: Inferential diagnostics across MCAR missingness experiments. All comparisons are relative to the full Bayesian benchmark. Sign agreement is the proportion of parameters whose posterior mean shares the correct sign. SD ratio is the ratio of posterior standard deviations (missing / full).

Experiment	Rate	Variables	Sign agree.	Mean SD ratio	Params lost sig.	$\hat{R}_{\max}$
Validation	2%	Age only	99.5%	1.002	0	1.003
Main	20%	All 5 SES	96.4%	1.037	4/196	1.008
Stress test	40%	All 5 SES	91.8%	1.100	—	1.009

Posterior standard deviations grow with missingness rate: from a mean SD ratio 1.037 at 20%, and 1.100 at 40%, with the progression nonlinear, confirming that estimation uncertainty accelerates as missingness increases. At 20% missingness across all five SES variables simultaneously, Figure 2 shows that parameter estimates remain tightly aligned with the full-information benchmark: posterior standard deviations widen by a mean ratio of 1.037 — 3.7% on average — and sign agreement holds at 96.4%, with just 4 of 196 parameters losing statistical significance.

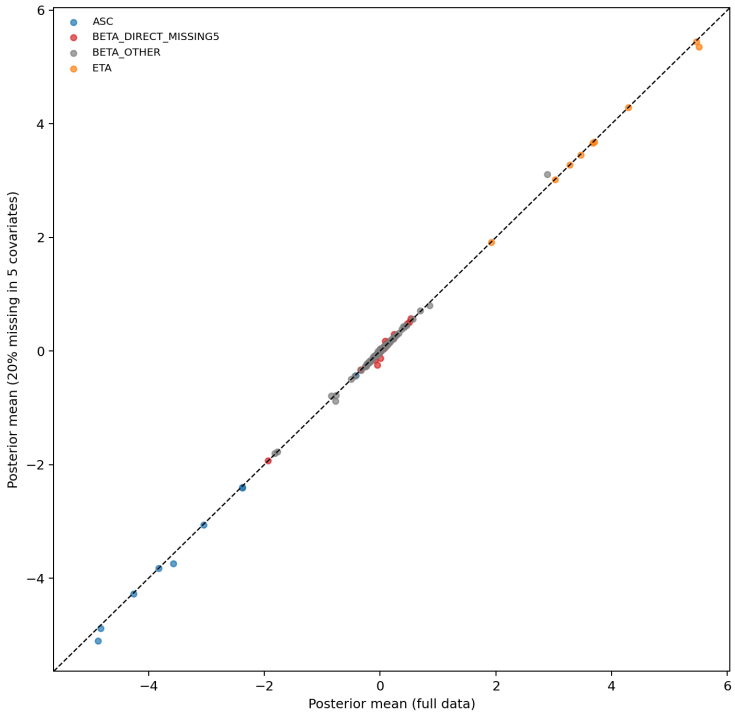


Figure 2: Parameter recovery at 20% MCAR missingness across all five optional SES variables. Each point is one of 196 MDCEV parameters, colored by type: activity-specific constants (ASC, blue), coefficients on the missing SES variables (BETA_DIRECT, red), coefficients on other covariates (BETA_OTHER, grey), and satiation parameters (ETA, orange). The tight alignment with the identity line confirms that designed missingness at this rate imposes only minimal inferential cost.

Even at 40% missingness — where posterior standard deviations widen to a mean ratio of 1.100, a 10% average increase —the Bayesian imputation engine continues to recover the behavioral parameters with meaningful fidelity — a result that speaks directly to the strength of the joint inference framework. When Z_{mis} is sampled from its conditional posterior rather than deleted or

substituted, the model extracts structural information from the correlation among observed variables and propagates it into the unobserved ones, attenuating what would otherwise be a severe inferential loss.

Two conclusions follow. First, the inferential engine is validated: Bayesian data augmentation handles designed missingness reliably across the full range of rates tested, providing the stable lower-level foundation the bilevel framework requires. Second, and more importantly, the MCAR results represent a lower bound on what the framework can achieve. Random missingness is the least informed possible design — it allocates survey burden without regard for which respondents or which questions carry the most inferential value. A policy-optimized design, by directing missingness toward covariates and respondents where the inferential cost is lowest, can be expected to preserve behavioral conclusions more effectively than MCAR at the same burden level. The benchmark experiments do not demonstrate the ceiling of the framework — they demonstrate its floor.

5. Discussion

The benchmark results establish a practical operating region for the framework: at 20% missingness posterior standard deviations widen by only 3.7% on average — a small increase in estimation uncertainty that leaves behavioral conclusions largely intact; at 40% that widening reaches 10% and accelerates nonlinearly. These differences in posterior standard deviations are precisely what the inferential quality term $\mathcal{I}$ in equation (5) is designed to capture and minimize: a design whose posterior standard deviations expand beyond what the burden saving justifies will be penalized by the upper-level objective. With the inferential cost of missingness now quantified, the question shifts from whether to reduce the survey, but how to allocate that reduction intelligently across respondents and questions — and the collection policy is the mechanism for that allocation. By conditioning question-display probabilities on mandatory covariates, the policy directs optional questions toward respondents for whom those questions carry the highest inferential value, and withholds them where the mandatory profile already constrains the likely value of the optional variable. This targeted allocation is precisely what MCAR cannot provide, which is why the benchmark results represent a lower bound rather than a ceiling.

The central computational challenge of the framework lies in the bilevel structure itself. For each candidate policy evaluated by the upper level, the lower level requires a full Bayesian posterior estimation — in the present application, NUTS over a 196-parameter MDCEV model with 42,000 respondents and a high-dimensional latent variable space for Z_{mis}. This is expensive, and repeating it across the many policy candidates that an optimization loop requires is the primary bottleneck standing between the framework as formulated and the framework as executable at scale. Resolving this will require either posterior approximations — such as variational inference — that reduce the cost of each lower-level evaluation, or surrogate modeling strategies that learn the shape of the objective function without evaluating the full posterior at every candidate. The upper-level optimization algorithm itself also remains an open choice: Bayesian optimization over the continuous policy parameter space is a natural candidate given the cost of each evaluation, but its behavior in this bilevel setting has not yet been characterized. These computational questions constitute the core technical agenda for the next stage of this research.

The framework is not specific to MDCEV or time-use data. The mandatory/optional partition,

the collection policy, and the bilevel objective are defined at the level of the design problem, independent of the behavioral model. The contribution is a framework for designing measurement — the MDCEV application is a proof of concept, not the boundary of the idea.

6. Conclusion

Survey incompleteness is conventionally treated as a defect to be repaired after data collection. This paper proposes a different framing: incompleteness is a design variable that can be optimized before collection, in a way that is coherent with the behavioral model it will serve. The missing-by-design framework developed here rests on two contributions. The first is a bilevel policy optimization that decides which optional questions to display to which respondents, governing that decision through a collection policy conditioned on always-observed mandatory covariates, with the explicit objective of trading respondent burden against inferential quality. The second is a Bayesian data augmentation approach in which missing covariate values are not imputed in a preparatory stage but are instead treated as latent variables and recovered jointly with the structural model parameters within a single posterior inference. Together, these contributions ensure that the act of designing missingness and the act of inferring under it are governed by the same probabilistic machinery — a coherence that no classical two-stage pipeline can provide.

Validated on a behaviorally rich MDCEV time-use application with over 42,000 respondents, the inferential engine demonstrates that 20% designed missingness across all five SES covariates reduces posterior precision by only 3.7% and preserves 96.4% of behavioral conclusions. Even at 40% missingness the Bayesian imputation engine continues to recover parameters with meaningful fidelity, demonstrating the robustness of the joint inference approach. Critically, these results represent the floor of the framework's performance, not its ceiling: MCAR is the least informed possible design, and a policy-optimized design operating at equal burden can be expected to do strictly better by directing the observation budget toward respondents and questions where inferential value is highest. Running that upper-level optimization and evaluating the resulting policy against the MCAR baseline constitutes the natural and immediate extension of this work.

Acknowledgements

The authors would like to acknowledge that this research was partially supported by JST FOREST (JPMJFR231G), and by the Council for Science, Technology and Innovation (CSTI) through the Cross-ministerial Strategic Innovation Promotion Program (SIP), 3rd period, "Smart Mobility Platform" (Grant Number: JPNP23023; funding agency: NEDO).

References

Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39:679–707.

Bhat, C. R. (2008). The multiple discrete-continuous extreme value (mdcev) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42:274–303.

Chikaraishi, M., Zhang, J., Fujiwara, A., and Axhausen, K. W. (2010). Exploring variation properties of time use behavior on the basis of, a multilevel multiple discrete-continuous extreme value model. *Transportation Research Record*, (2156):101–110.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana*, 17.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. Wiley, Hoboken, NJ, 2 edition.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50:537–567.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, 2 edition.

Pinjari, A. R. and Bhat, C. R. (2011). Activity-based travel demand analysis. In de Palma, A., Lindsey, R., Quinet, E., and Vickerman, R., editors, *A Handbook of Transport Economics*, pages 213–248. Edward Elgar Publishing, Cheltenham.

Rose, J. M. and Bliemer, M. C. (2009). Constructing efficient stated choice experimental designs. *Transport reviews*, 29(5):587–617.

Stopher, P. and Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, 41(5):367–381.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 2 edition.