The use of pooled RP-SP choice data to simultaneously identify alternative attributes and random coefficients on those attributes

Mehek Biswas [a]
Chandra R. Bhat [c]
Abdul Rawoof Pinjari [a,b,#]

[a] Department of Civil Engineering, Indian Institute of Science (IISc), Bengaluru 560012, India

[b] Centre for infrastructure, Sustainable Transportation and Urban Planning (CiSTUP), Indian Institute of Science, Bengaluru 560012, India

[c] Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX 78712, USA

[#] Corresponding author (abdul@iisc.ac.in)

ABSTRACT

Random utility maximization-based discrete choice models involve utility functions that are typically specified with explanatory variables representing alternative-specific attributes. It may be useful to specify some alternative-specific attributes as stochastic in situations when the analyst cannot accurately measure the attribute values considered by the decision maker. In addition, the parameters representing decision makers' response to the attributes may have to be specified as stochastic to recognize response heterogeneity in the population. Ignoring either of these two sources of stochasticity can lead to biased parameter estimates and distorted willingness-to-pay estimates. Further, in some situations the analyst may not even have access to measurements of important alternative-specific attributes to include them in the utility specification. In this study, we explore the feasibility of simultaneously inferring alternative attributes and the corresponding coefficients, as well as stochasticity in both – without the help of external measurement data on alternative attributes – using mixed logit models on pooled revealed preference (RP) and stated preference (SP) choice datasets. To do so, we first theoretically examine parameter identifiability for different specifications and distributional forms of alternative attributes and their coefficients. Next, we illustrate this through simulation experiments in a travel mode choice setting and demonstrate the conditions under which pooled RP-SP data can help disentangle stochastic alternative attributes from random coefficients. In addition, an empirical application is presented in the context of commute mode choice in Bengaluru, India, to demonstrate the importance of recognizing stochasticity in mode-specific in-vehicle travel times along with the random coefficient on in-vehicle travel times.

Keywords: pooled RP-SP data, travel mode choice, mixed logit, stochastic variables, random coefficients, parameter identification

1. INTRODUCTION

Pooling stated and revealed data sources has long been recognized by transportation researchers as a valuable approach for enriching behavioural information from travel choice models (Bradley and Daly, 1991; Hensher and Bradley, 1993; Morikawa, 1994; Hensher *et al.*, 1998; Bhat and Castelar, 2002; Cherchi and Ortuzar, 2006; Hensher, 2012; Helvestorn *et al.*, 2018, etc.). Combining these two data sources, also referred to as 'data enrichment' or 'data fusion', helps exploit the realism of revealed preference (RP) data along with the statistically appealing properties of stated preference (SP) data. SP data comes with its possible limitations related to the veracity of individual stated responses, which may be different in hypothetical settings than in real-life contexts. On the other hand, RP data may be associated with endogeneity bias when attributes that impact consumer choices are unobserved to the analyst and are correlated with the observed attributes. Moreover, SP data is able to reduce collinearity among attributes describing choice alternatives, which can potentially be a problem with RP data. The recognition of the strengths and drawbacks of each source has led to the understanding that pooling the two data sources can lead to an enhanced modelling of travel choice behaviours.

Consider, for example, a travel mode choice study that utilizes both RP and SP data. In this context, the attributes of choice alternatives presented to the respondent in SP choice scenarios are determined by (hence, known to) the analyst. For example, mode-specific travel times presented to a respondent in a stated mode choice experiment can be considered free of measurement errors and other sources of variability. In contrast, RP choice alternative attributes are often associated with several sources of variability. For example, there may be analyst's errors in measuring the travel times due to: (1) using level-of-service data between the centroids of aggregated spatial zones instead of precise, point-to-point measurements of the same (Bhatta and Larsen, 2011*;* Ortuzar and Ivelic, 1987; Train, 1978; Walker *et al.*, 2010), (2) errors in network coding and/or assumptions of travel speeds, (3) differences between travellers' perceptions and the analyst's measurements of travel times (Brey and Walker, 2011; Varotto *et al.*, 2017), and (4) day-to-day variability in travel times due to varying travel conditions on the network (Biswas *et al.*, 2024; Srinivasan *et al.*, 2014). In all these cases, it is very difficult for the analyst to accurately know and measure the specific travel time values considered by the travellers in making their travel choice decision.

To put the above discussion in a mathematical form, consider the following widely used additive random utility specification for an individual q for choosing mode i :

$$U_{qi} = \boldsymbol{\beta}_q' \mathbf{x}_{qi} + \varepsilon_{qi} \tag{1}$$

In the above expression for utility function U_{qi}, $\mathbf{x}_{qi}$ is a vector of alternative-specific attributes and decision-maker specific characteristics (including a constant); $\boldsymbol{\beta}_q$ is the corresponding vector of parameters that might vary across individuals; and ε_{qi} is a random error term that represents the difference between the utility perceived by the decision-maker and the utility specified by the analyst. The alternative attributes in $\mathbf{x}_{qi}$ are typically assumed to be known accurately to the analyst. However, this assumption may not hold in RP settings, because, as discussed earlier, the analyst might not accurately know the alternative-specific travel time values considered by the traveller.

In the discrete choice modelling literature, a few broad approaches have been used to incorporate variability due to measurement errors in the alternative attributes in $\mathbf{x}_{qi}$. One approach is the errors-in-variables (EIV) approach, which has been widely used in the context of regression models (Gleser, 1981; Carroll and Spiegelman, 1984). For example, Bhatta and Larsen (2011), Ortúzar and Ivelic (1987), and Diaz *et al.* (2015) specify error components in the utility functions to represent errors (or stochasticity) in alternative attributes such as travel time. In another study, Nirmale and Pinjari (2023) use the EIV approach to explore errors in choice environment variables that do not vary across alternatives (e.g., traffic environment variables such as relative speed and space gap in driver behaviour models). In another context, where data on exogenous attributes such as travel time are missing or unknown beyond certain interval bounds, the EIV method has been implemented through Rubin's multiple imputation approach (Steinmetz and Brownstone, 2005). The second approach to accommodate stochastic variables is the Integrated Choice and Latent Variable (ICLV) approach (Ben-Akiva *et al.*, 2002; Bhat and Dubey, 2014; Vij and Walker, 2016). This approach has been adopted to accommodate (a) analyst-based measurement errors for mode-specific travel times (Walker *et al.*, 2010), (b) errors in traveller reported travel times (Varotto *et al.*, 2017), and (c) missing data for demographic variables such as income (Sanko *et al.*, 2014). In the ICLV approach, the relevant explanatory variables are treated as latent (therefore,

stochastic) and are modeled using a separate latent variable equation. However, all the above studies specify the coefficients on the stochastic (or latent) variables as being deterministic. In this context, Diaz *et al.* (2015) and Nirmale and Pinjari (2023) highlight the issue of confounding of variability between the two sources – variability in $\boldsymbol{\beta}_q$ and stochasticity in $\mathbf{x}_{qi}$ – and the difficulty in identifying these with the help of choice data alone.

In another well-established and large stream of literature, the parameters in $\boldsymbol{\beta}_q$ corresponding to $\mathbf{x}_{qi}$ have been specified as random to capture unobserved taste heterogeneity among decision-makers (Bhat, 2001; Train, 2001, etc.). Such studies typically use either the mixed multinomial logit model (see, for example, Bhat, 2001; Bhat, 2003; Hensher and Greene, 2003; Hess and Polak, 2005; Mc Fadden and Train, 2000, and Brownstone *et al.*, 2000) or the mixed multinomial probit model (Daganzo, 1979; Keane, 1992; Bhat, 2011; Bhat and Sidharthan, 2012).

The streams of literature on addressing each of the two sources of variability discussed so far – variability in the alternative attributes in $\mathbf{x}_{qi}$ and the variability in the corresponding coefficients in $\boldsymbol{\beta}_q$ – have developed in independent directions, with little effort toward accounting for both sources of variability. This is because the EIV and ICLV approaches (for considering variable stochasticity) and mixed logit/probit models (for recognizing response stochasticity) do not allow the simultaneous identifiability of both stochasticity sources. In a recent study, however, Biswas *et al.* (2024) showed that the ICLV approach can be extended to allow the simultaneous identification of stochasticity in alternative-specific attributes (e.g., travel times) as well as random heterogeneity in response to the attributes in the context of traveller choices, as long as at least two different types of measurements (or data) are available in the RP setting – one for the alternative-specific attributes (e.g., travel time)[1] and one for traveller choices that depend on the attributes. However, measurements of the alternative-specific attributes under consideration may not be available in many situations. For example, if one is interested in accommodating mode-specific stochasticity in travel times for multiple modes of travel, the analyst may not always have access

[1] Typically, network skims (*i.e.*, travel times of shortest time paths between two locations) are used to generate travel times for mode choice models, assuming free-flow speeds in the network. Such data are not useful for identifying variability in travel times because they do not represent variability in travel conditions. One needs actual measurements of travel time that reflect travel conditions (and the variability therein) on the network.

to travel time measurements for all modes. Importantly, it is not easy to get measurements of some attributes influencing traveller choices, such as crowding levels in public transit vehicles, let alone variability in such attributes. For example, it is not uncommon that travellers do not respond to survey questions on their perceived crowding levels in transit systems, if they do not use public transit often. In the absence of measurements of crowding levels in transit systems, a large chunk of RP survey data samples may remain unutilized for estimating mode choice models with crowding level as an attribute entering the utility functions of transit modes. Similarly, it is not easy to get accurate measurements of mode-specific out-of-vehicle travel times (*OVTT*) or waiting times individuals face for estimating mode choice models. Therefore, it is common to make assumptions, albeit with errors that vary across individuals, about the values of *OVTT* or mode-specific waiting time values. The current study provides another approach to allow the simultaneous identifiability of both stochasticity sources by combining two sources of data and is especially valuable when external measurements are not available for certain attributes. In this context, SP data, which provides choices under specific "fixed" attribute values, can substitute for unavailable external measurements for the attribute. However, when SP data sources are used along with RP sources, particular issues arise in case of identification, which require theoretical examination of parameter identifiability.

In view of the above discussion, in the current study, we aim to explore the feasibility of using pooled RP-SP data as another approach to simultaneously infer the alternative attributes in $\mathbf{x}_{qi}$ and the corresponding coefficients ($\boldsymbol{\beta}_q$), as well as the stochasticity in both $\mathbf{x}_{qi}$ and $\boldsymbol{\beta}_q$ -- without the help of external measurement data for the alternative attributes. In this context, our hypothesis is that SP data, which typically include known and deterministic values of $\mathbf{x}_{qi}$ (since the alternative attribute values are carefully constructed and presented to the decision-maker)[2], allow the identification of $\boldsymbol{\beta}_q$ and corresponding heterogeneity. Once the response heterogeneity is identified from SP data, the alternative attributes in $\mathbf{x}_{qi}$ (and the variability therein) may be identified in the RP data collected from the 'field'. Essentially, SP data enable the estimation of

[2] It is assumed here that the decision-makers (*i.e.*, the survey respondents) do not distort the attribute values presented to them but directly utilize the presented values in their utility functions. There is a stream of literature on attribute non-attendance (Scarpa *et al.*, 2009; Hensher *et al.,* 2012), which highlights that some attributes presented in the SP settings may not even be considered by the decision-makers. While we do not delve into such issues in the current study, this might be an important confounding factor to be considered in future research.

random coefficients ($\boldsymbol{\beta}_q$) and, because the coefficients are treated as the same in both SP and RP settings, the RP data enable identification of alternative attributes in $\mathbf{x}_{qi}$. In this context, we formulate a mixed logit choice modelling framework for pooled RP-SP datasets with the alternative attributes in the RP data and the corresponding coefficients common to both RP and SP settings as unknown parameters to be estimated. Within the context of such mixed logit models for pooled RP-SP data, the following objectives are pursued:

- First, we conduct a theoretical investigation to ascertain whether the parameters describing the distributions of alternative attributes in the RP setting can be identified (and how many such parameters can be identified) along with the parameters describing the distribution of random coefficients on the attributes, using mixed logit models for pooled RP-SP data. This investigation is carried out in the context of mode choice models separately for the following two different types of alternative-specific (or mode-specific) attributes: (1) the attributes exhibit systematic variability across individuals in the data – due to variation that can be expressed as a function of an observed variable – as well as random variability, and (2) the attributes exhibit only random variability across individuals in the data. Further, in either case, the identification conditions are laid out for two widely used distributional assumptions on the alternative attributes and random coefficients on them – (1) normal distribution and (2) lognormal distribution.

  In most of the aforementioned cases, we establish that it is feasible to use pooled RP-SP data to infer the distributional parameters of a single RP attribute for all but one alternative in the choice set, along with the random coefficient on the attribute. Further, the identifiability of the parameters is better in situations where the alternative attribute of interest exhibits systematic variability across individuals in the data.

- Second, to augment the theoretical investigations, we conduct simulation experiments to examine the efficacy of our proposed approach (combining RP-SP datasets) in inferring the values of an alternative attribute for the choice alternatives in RP settings along with the random coefficient on that attribute and any other alternative attributes.
- Third, we present an empirical analysis of commute mode choice using pooled RP-SP data from Bengaluru, India, to demonstrate the feasibility of inferring in-vehicle travel times

(*IVTT*) in RP settings along with the corresponding coefficient. Furthermore, using this empirical analysis we highlight the importance of recognizing stochasticity in both mode-specific in-vehicle travel times and the corresponding coefficient.

The rest of the paper is structured as follows. Section 2 discusses the mixed logit modelling framework. Section 3 presents the theoretical conditions for identification of the model. Section 4 discusses the simulation experiments conducted for a mode choice setting. Section 5 presents the empirical results and findings in the context of travel mode choice in Bengaluru, India. Finally, Section 6 discusses the conclusions of this study and directions for future research.

2. MODEL FRAMEWORK AND ESTIMATION

In this section, we formulate a mixed logit modelling framework for travel mode choice on pooled RP-SP data. In this model, the mode-specific travel times for the RP choice occasions are treated as stochastic (since the trave time values considered by the travellers in the RP setting are unknown to the analyst) as well as the corresponding coefficient, and other coefficients may also be random. The SP choice alternative utility functions also have random coefficients but deterministic travel times and other explanatory variables.

2.1 Utility specifications for RP and SP settings

Define the utility associated with a mode choice alternative i for an individual q at choice occasion t, if it is an RP choice occasion, as:

$$U_{qit} = \beta_{0qi} + \beta_{TT,q} TT^{*}_{qit} + \boldsymbol{\varphi}' \mathbf{x}_{qit} + \varepsilon_{qit} \tag{2}$$

Alternatively, if t is an SP choice occasion, the utility associated with alternative i is defined as:

$$\bar{U}_{qit} = \bar{\beta}_{0qi} + \beta_{TT,q} \overline{TT}_{qit} + \boldsymbol{\varphi}' \bar{\mathbf{x}}_{qit} + \bar{\varepsilon}_{qit} \tag{3}$$

In the above equations, β_{0qi} and $\bar{\beta}_{0qi}$ are individual-level random effects (i.e., alternative-specific constants) representing individual q's preference for alternative i due to individual-level unobserved factors influencing the choices in RP and SP settings, respectively. The distribution of β_{0qi} across individuals is assumed to be represented by two moment parameters $\{\mu_{0i}, \sigma_{0i}\}$ and the

distribution of $\bar{\beta}_{0qi}$ is assumed to be represented by two moment parameters $\{\bar{\mu}_{0i}, \sigma_{0i}\}$; that is, $\beta_{0qi} = \mu_{0i} + \sigma_{0i} \times z_{qi}$ and $\bar{\beta}_{0qi} = \bar{\mu}_{0i} + \sigma_{0i} \times z_{qi}$. Note that the location parameters of the individual-level alternative-specific random effects are specified to be different between RP and SP choice occasions while the random components are kept same. Specifically, z_{qi} is a standard normal variate for alternative *i*, assumed to be the same across all RP and SP choice occasions of an individual *q* to generate covariance in utilities across different choice occasions of *q*.

The term TT^*_{qit} in Equation (2) is the mode-specific travel time considered by the individual q in an RP choice occasion t. As mentioned earlier, the specific TT^*_{qit} value considered by the individual is unknown to the analyst. Therefore, TT^*_{qit} is assumed to follow a distribution characterized by a vector of parameters $\boldsymbol{\chi}_i$. The distribution is assumed to be known to the analyst while the parameters of the distribution are to be estimated from pooled RP-SP data. [3]

Next, $\overline{TT}_{qit}$ in Equation (3) is the deterministic travel time value presented to the individual *q* for alternative i in an SP choice occasion. $\beta_{TT,q}$ is the individual-specific random coefficient on mode-specific travel times in both RP and SP choice utility functions. This random coefficient is assumed to follow a distribution characterized by a vector of parameters $\boldsymbol{\zeta}$. It is worth noting that $\beta_{TT,q}$ is assumed to be the same for both RP and SP choice occasions for any individual. This assumption is necessary for inferring travel times in the RP setting using the proposed approach. Of course, the same assumption need not be made for the coefficients on other alternative attributes. However, it is a common practice in pooled RP-SP data models to specify same coefficients on alternative attributes for RP and SP choice occasions. In fact, at least one parameter must be specified to be the same between RP and SP choice occasions for joint estimation while allowing scale differences between RP and SP settings.

[3] In many empirical contexts, the analyst would have access to some measurements of TT^*_{qit}, such as the individuals' reported travel times for their chosen modes. In this paper, we examine the extent to which the analyst can infer alternative attributes in RP settings using pooled RP-SP data in the absence of measurements in the RP setting. In future research, it would be useful to examine if the measurements available to the analyst can be utilized in the proposed RP-SP framework to improve the inference of alternative attributes.

Next, $\mathbf{x}_{qit}$ and $\overline{\mathbf{x}}_{qit}$ are vectors of other alternative-specific attributes (such as travel cost), and individual-specific attributes (such as employment status and gender) in the RP and SP utility functions respectively. Note that while the alternative-specific attributes in $\mathbf{x}_{qit}$ and $\overline{\mathbf{x}}_{qit}$ will likely be different between RP and SP choice occasions, the individual-specific attributes will, in general, be the same. The vector $\boldsymbol{\varphi}$ is the vector of coefficients on the variables in $\mathbf{x}_{qit}$ and $\overline{\mathbf{x}}_{qit}$.

To complete the utility specification, distributional assumptions are made for the stochastic components in the utility functions of both RP and SP settings. In this context, the additive kernel error terms ε_{qit} and $\overline{\varepsilon}_{qit}$ are assumed to be independent Gumbel distributed across all alternatives in both RP and SP choice occasions of all individuals. The scale of the error terms (ε_{qit}) is fixed to 1 for all alternatives in RP choice occasions for any q. To account for the scale difference between SP and RP choice settings (Ben-Akiva *et al.*, 1994), the scale of the error terms ($\overline{\varepsilon}_{qit}$) is specified as λ for all alternatives in SP choice occasions. λ is a parameter to be estimated.

Following the random utility maximization theory, an individual q at any choice occasion t is assumed to choose the alternative that provides the maximum utility. In this context, define an indicator variable y_{qit} which takes the value 1 if the individual q chose alternative i at choice occasion t (0 otherwise).

2.2 Mixed logit model for pooled RP-SP data

To write the model likelihood, we define some additional notation here. $\mathbf{J_q}$ denotes the set of available choice alternatives for individual q and $\left|\mathbf{J_q}\right| = J_q$. $\mathbf{J}$ denotes the full set of choice alternatives across all individuals and $\left|\mathbf{J}\right| = J$. R_q denotes the number of RP choice occasions corresponding to individual q. S_q denotes the number of SP choice occasions corresponding to individual q. Then, the total number of choice occasions for the individual is $T_q = R_q + S_q$. Stack the mode-specific stochastic travel time variables (TT^*_{qit}) for all modes available to the individual at an RP choice occasion t into a $J_q \times 1$ vector as: $\mathbf{TT}^*_{qt} = \left[TT_{q1t}{}^*, TT_{q2t}{}^*, ..., TT_{qJ_qt}{}^* \right]'$. Next, stack

$\mathbf{TT}_{qt}^{*}$ for all RP choice occasions of the individual into $\mathbf{TT}_{q,RP}^{*} = \{\mathbf{TT}_{q1}^{*},...,\mathbf{TT}_{qt}^{*},...,\mathbf{TT}_{qR_q}^{*}\}$. Further, stack the parameter vectors $\boldsymbol{\chi}_i$ of TT_{qit}^{*} for all choice alternatives into the vector $\boldsymbol{\chi} = [\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, ..., \boldsymbol{\chi}_i, ..., \boldsymbol{\chi}_J]$.

Next, consider the $2J \times 1$ vector of alternative-specific random effects (β_{0qi} and $\bar{\beta}_{0qi}$) for all alternatives in the RP and SP choice occasions. Recall that $\beta_{0qi} = \mu_{0i} + \sigma_{0i} \times z_{qi}$ and $\bar{\beta}_{0qi} = \bar{\mu}_{0i} + \sigma_{0i} \times z_{qi}$. Stack the standard normal variates (z_{qi}) in these random effects for all choice alternatives into a vector $\mathbf{z}_q = \{z_{q1}, z_{q2}, ..., z_{qi}, ..., z_{qJ}\}$. And stack the unique elements of parameter vectors $\{\mu_{0i}, \sigma_{0i}\}$ and $\{\bar{\mu}_{0i}, \sigma_{0i}\}$ for these random effects for all choice alternatives into a vector $\boldsymbol{\psi} = \{\{\mu_{01}, \bar{\mu}_{01}, \sigma_{01}\}, ..., \{\mu_{0i}, \bar{\mu}_{0i}, \sigma_{0i}\}, ..., \{\mu_{0J}, \bar{\mu}_{0J}, \sigma_{0J}\}\}$. For identification purposes, for each of the RP and SP settings, at least one location parameter of the individual-level alternative-specific random effects must be fixed (typically to zero). However, all J scale parameters corresponding to the random effects can be estimated to recognize covariance across utility functions of different choice occasions of the same individual.

Further, let $f(\cdot)$ denote the PDF of the random coefficient $\beta_{TT,q}$ on mode-specific travel times, $g(\cdot)$ denote the PDF of the stochastic travel times $\mathbf{TT}_{q,RP}^{*}$ in RP choice occasions, and $h(\cdot)$ denote the PDF of the standard normal variates ($\mathbf{z}_q$) used for random alternative-specific constants, respectively.

The full set of parameters to be estimated in the model for pooled RP-SP data are those in the vectors $\boldsymbol{\psi}$, $\boldsymbol{\chi}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\varphi}$, and the scalar λ. Among these, $\boldsymbol{\psi}$, $\boldsymbol{\chi}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\varphi}$ are relevant to RP data. And those in $\boldsymbol{\psi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\varphi}$, and λ are relevant to SP data. Stack the full set of parameters into a vector Θ, the subset relevant to RP data into Θ_{RP} and the subset relevant to SP data into Θ_{SP}. The conditional likelihood that individual q chooses alternative i at the t^{th} RP choice occasion, conditional on the stochastic components $\mathbf{z}_q$, $\beta_{TT,q}$, and $\mathbf{TT}_{qt}^{*}$ may be expressed as:

$$\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta_{RP},\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{qt}\right)=\frac{\exp\left(\mu_{0i}+\sigma_{0i}\times z_{qi}+\beta_{TT,q}TT^*_{qit}+\boldsymbol{\varphi}'\mathbf{x}_{qit}\right)}{\sum_{j=1}^{J_q}\exp\left(\mu_{0j}+\sigma_{0j}\times z_{qj}+\beta_{TT,q}TT^*_{qjt}+\boldsymbol{\varphi}'\mathbf{x}_{qjt}\right)} \tag{4}$$

Similarly, the conditional likelihood that individual q chooses alternative i at the t^{th} SP choice occasion, conditional on the stochastic components z_{qi} and $\beta_{TT,q}$, may be expressed as:

$$\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta_{SP},\mathbf{z}_q,\beta_{TT,q}\right)=\frac{\exp\left\{\lambda\left(\bar{\mu}_{0i}+\sigma_{0i}\times z_{qi}+\beta_{TT,q}\overline{TT}_{qit}+\boldsymbol{\varphi}'\bar{\mathbf{x}}_{qit}\right)\right\}}{\sum_{j=1}^{J_q}\exp\left\{\lambda\left(\bar{\mu}_{0j}+\sigma_{0j}\times z_{qj}+\beta_{TT,q}\overline{TT}_{qjt}+\boldsymbol{\varphi}'\bar{\mathbf{x}}_{qjt}\right)\right\}} \tag{5}$$

The conditional likelihood function individual q chooses alternative i on any choice occasion, regardless of whether it is an RP occasion or an SP occasion, can be written in a compact form as:

$$\begin{aligned}\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta,\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{qt}\right)=&\left[\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta_{RP},\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{qt}\right)\right]^{Y_{qt,RP}}\\&\times\left[\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta_{SP},\mathbf{z}_q,\beta_{TT,q}\right)\right]^{(1-Y_{qt,RP})}\end{aligned} \tag{6}$$

In the above equation, $Y_{qt,RP}=1$ for an RP choice occasion and $Y_{qt,RP}=0$ for an SP choice occasion $(Y_{qt,SP}=1-Y_{qt,RP})$. Using the above expression, the conditional likelihood function for the observed choices on all choice occasions of an individual q can be written as below:

$$\mathcal{L}_q\left(\Theta,\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{q,RP}\right)=\prod_{t=1}^{T_q}\left[\prod_{i=1}^{J_q}\mathcal{L}_{qit}\left(y_{qit}=1\,|\,\Theta,\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{qt}\right)\right]^{y_{qit}} \tag{7}$$

Next, the unconditional likelihood function for the observed choices across all choice occasions of an individual q can be formulated as below:

$$\mathcal{L}_q\left(\Theta\right)=\int_{\mathbf{TT}^*_{q,RP}}\int_{\beta_{TT,q}}\int_{\mathbf{z}_q}\mathcal{L}_q\left(\Theta,\mathbf{z}_q,\beta_{TT,q},\mathbf{TT}^*_{q,RP}\right)h(\mathbf{z}_q)\,f(\beta_{TT,q}\,|\,\boldsymbol{\zeta})\,g(\mathbf{TT}^*_{q,RP}\,|\,\boldsymbol{\chi})\,d(\mathbf{z}_q)\,d(\beta_{TT,q})\,d(\mathbf{TT}^*_{q,RP}) \tag{8}$$

[4]

[4] The likelihood expression in Equation (8) is a single level integral since the unobserved heterogeneity in the attributes (e.g., travel time) and in their parameters is considered at the individual level, as opposed to the choice occasion level. This formulation is suitable for datasets with a single RP observation per individual (as is the case with the empirical data used in the current study), under the assumption that the unobserved heterogeneity does not vary across the different SP choice occasions of an individual. However, if there are multiple RP choice occasions per individual, then the unobserved heterogeneity in the attributes is likely to be at the choice occasion level. Then, the integration in

The open-form integral in the above likelihood function may be simulated as below:

$$SL_q = \frac{1}{R}\sum_{r=1}^{R}\left[\mathcal{L}_q\left(\Theta, \mathbf{z}_q^r, \beta_{TT,q}^r, \mathbf{TT}_{q,RP}^{*r}\right)\right] \tag{9}$$

where, SL_q is the simulated likelihood function for the q^{th} individual's choices on all of their choice occasions; $\beta_{TT,q}^r$ denotes the r^{th} draw $(r = 1, 2, ..., R)$ from $f(\beta_{TT,q} \mid \boldsymbol{\zeta})$, $\mathbf{TT}_{q,RP}^{*r}$ denotes the r^{th} set of draws $(r = 1, 2, ..., R)$ from $g(\mathbf{TT}_{q,RP}^{*} \mid \boldsymbol{\chi})$; $\mathbf{z}_q^r$ denotes the r^{th} set of draws $(r = 1, 2, ..., R)$ from $h(\mathbf{z_q})$; and R denotes the total number of such draws used for simulation. Finally, the simulated log-likelihood function for all individuals in the data can be written as:

$$SLL = \sum_{q} \ln\left(SL_q\right) \tag{10}$$

In this paper, the above simulated log-likelihood function was computed using Halton draws (Train, 2000; Bhat, 2003) using a code written in the GAUSS mathematical programming platform.

3. PARAMETER IDENTIFICATION: THEORETICAL ANALYSIS

This section presents a theoretical analysis of the identifiability of parameters in the RP-SP model with a stochastic alternative-specific attribute in the RP setting and a generic (common to all utility functions) random coefficient on the alternative-specific attribute.[5] In this context, it should be noted that the SP data helps inform or identify the parameters of the random coefficient on the alternative attribute of interest, since the corresponding variables in the SP setting are assumed to

---

the likelihood expression would involve a two-level integral – one level for the choice occasion specific unobserved effects and another level for individual specific unobserved effects. Such a likelihood expression is provided below:

$$\mathcal{L}_q(\Theta) = \int_{\mathbf{z}_q}\int_{\beta_{TT,q}}\prod_{t=1}^{T_q}\left[\int_{\mathbf{TT}_{q,RP}^{*}}\mathcal{L}_q\left(\Theta, \mathbf{z}_q, \beta_{TT,q}, \mathbf{TT}_{q,RP}^{*}\right) g(\mathbf{TT}_{q,RP}^{*} \mid \boldsymbol{\chi})\, d(\mathbf{TT}_{q,RP}^{*})\right] f(\beta_{TT,q} \mid \boldsymbol{\zeta})\, h(\mathbf{z}_q)\, d(\beta_{TT,q})\, d(\mathbf{z}_q)$$

The reader is referred to Bhat and Castellar (2002) for such likelihood expressions involving two-level integrals.

[5] Note for a given alternative-specific attribute, there can be as many different variables as the number of choice alternatives. For example, in the context of travel time in mode choice models, there can be as many travel time variables as the number of mode choice alternatives – with each mode-specific travel time variable entering the corresponding mode's utility function. And we consider a generic (i.e., common) random coefficient for all the variables of a given alternative-specific attribute.

be free of measurement errors. Given the random coefficient parameters can be identified using the SP data alone, one can examine how much can be inferred about the alternative-specific attributes using the RP data. It is as if the random coefficient estimated from the SP data serves as an explanatory variable for estimating the parameters of the corresponding stochastic alternative-specific attributes using RP data. Therefore, the identification analysis can be limited to the RP data setting alone, assuming that the parameters of the random coefficient on the stochastic variable under consideration are already known (from SP data)[6]. For the stochastic alternative-specific attribute in the RP setting, we consider two different possibilities, as discussed next.

The first case is when *the stochastic alternative-specific attribute exhibits both systematic and random variation across individuals in the data*. That is, the corresponding stochastic variables vary across individuals due to an observed variable in addition to random variation that cannot be attributed to an observed variable. For example, one can express in-vehicle travel time of a mode i for an individual q, $IVTT_{qi}$ as a function of travel distance, an observable variable that varies across individuals in the data. That is, $IVTT_{qi} = \theta_{qi} \times d_q$, where d_q is the observed travel distance and θ_{qi} is the randomly varying mode-specific inverse speed (Note: Inverse speed of a mode is the time it takes to traverse unit distance by that mode).

In the second case, *the stochastic alternative-specific attribute exhibits only random variation (i.e., it does not exhibit systematic variation) across individuals in the data*. For example, waiting time or out-of-vehicle travel time ($OVTT$) might exhibit only random variation without any systematic variation that can be attributed to an observed variable.

In either case, the following two fundamental principles of identification in RUM-based discrete choice models apply: (1) The location of at least one of the random utility functions should be fixed (i.e., only differences in utility matter) and (2) the scale of the random utility functions of the model cannot be identified. Also, in either case, one should examine the structure of the covariance matrix of the differenced random utility components (i.e., after taking the difference of the utility functions with respect to one, base alternative utility function). This should be done to eliminate any further linear dependencies in the covariance matrix of utility differences to avoid

[6] While we separate the SP and RP data settings for the purpose of this discussion, the estimation of parameters in the SP and RP settings takes place jointly on the pooled RP-SP data.

unidentified models. The discussion of the identification conditions is presented next for each of the two cases.

3.1 Stochastic alternative-specific attribute exhibits both systematic and random variation across individuals in the data

In the presence of systematic variation of stochastic alternative-attribute across individuals in the data, the corresponding terms in the covariance matrix of utility differences also exhibit systematic variation across individuals. This systematic variation obviates the need for examining the structure of the matrix for eliminating linear dependencies. Only the two fundamental principles of identification mentioned earlier suffice for this case. Next, the identification conditions for this setting are discussed for two distributional assumptions for the stochastic alternative-specific attribute and its coefficient – normal and lognormal.

3.1.1 Normal distributed stochastic alternative-specific attribute with normal distributed coefficient

Considering a total of J choice alternatives, each with a normal distributed alternative-specific stochastic attribute and a generic, normal distributed random coefficient. In this situation, up to $J-1$ alternative-specific location parameters and $J-1$ alternative-specific scale parameters are identifiable for the alternative-specific stochastic variables. That is, one location parameter and one scale parameter per choice alternative are identifiable for all but one of the J alternatives. This is based on the fundamental principle that only differences in utility matter. The reader is referred to the corresponding discussion in Appendix A for additional details. Among the J alternative-specific scale parameters, the one with the lowest variance (scale) should be normalized so that the resulting covariance matrix of utilities is positive semidefinite (Walker, 2001).

3.1.2 Lognormal distributed stochastic alternative-specific attribute with lognormal distributed coefficient

In this case, similar to the case of normal distributed stochastic variable and coefficient, up to $J-1$ alternative-specific location parameters are identifiable for the alternative-specific stochastic variables. Further, $J-1$ alternative-specific scale parameters are identifiable in this setting. In some situations, however, it may be possible to identify all J alternative-specific scale parameters. This is because, as discussed in Appendix A, the difference of two lognormal random variables

does not yield a distribution with a known analytic form. Since the resulting distribution might need more than two parameters to describe it fully, it may be possible to estimate one more than $J-1$ alternative-specific scale parameters. This is demonstrated later (Section 4) using simulation experiments. However, it is safe to fix at least one of the J scale parameters to be sure of a theoretically identified model.

3.2 Stochastic alternative-specific attribute exhibits only random variation across individuals in the data

In this case, first we examine the situation when the coefficient on the alternative-specific stochastic attribute is fixed and same for all individuals in the data. That is, consider the following utility function for an RP setting, where the mode-specific travel time (TT_{qi}^{*}) exhibits only random variation without systematic variation across individuals, and the coefficient (β_{TT}) of TT_{qi}^{*} is same for all individuals:

$$U_{qi} = \beta_{0i} + \beta_{TT} TT_{qi}^{*} + \varepsilon_{qi} \tag{11}$$

With the above utility specification, even though the β_{TT} value can be estimated from relevant SP data, it is not possible to identify any parameters of the distributions describing TT_{qi}^{*} in RP data. To understand this, one can view β_{TT} estimate from the SP data as an explanatory variable for estimating TT_{qi}^{*} (or their parameters) using RP data. Just as an explanatory variable without any variation in the data does not help estimate its coefficients, when β_{TT} does not vary across individuals, one cannot estimate any parameter describing TT_{qi}^{*}. Therefore, it is important that the coefficient on the stochastic alternative-specific attribute be specified as random, when there is no systematic variation across individuals in the alternative-specific attribute, to be able to infer the distribution parameters of the alternative-specific attribute.[7]

In the rest of this section, we consider only the situations when the coefficient on the alternative-specific attribute is random (i.e., it varies across individuals). In this case, since the

[7] However, the coefficient on the alternative-specific attributes need not vary across individuals when the alternative attributes (e.g., TT_{qi}^{*}) themselves demonstrate some systematic variation across individuals (as in Section 3.1).

variables for the alternative-specific attribute do not exhibit any systematic variation across individuals, one must examine the structure of the covariance matrix of utility differences to determine the number of parameters one can estimate. Specifically, one should determine the maximum number of unique parameters that give rise to the covariance matrix of utility differences. To do so, one should find the rank of the Jacobian of the vector of unique elements in covariance matrix of utility differences (the Jacobian is with respect to the parameters to be estimated). This is called the *rank condition* (Bunch, 1992; Walker 2001). The rank of such a Jacobian matrix depends on the following: (a) The number of choice alternatives (J), (b) the number of utility functions in which the alternative-specific stochastic attribute enters (L), and (3) the distributional assumptions made on the alternative-specific stochastic attribute and the corresponding random coefficient. Appendix B derives and discusses the identification conditions for the following two distributional assumptions for different values of J and L: (1) normal distributed alternative-specific attribute and random coefficient, and (2) lognormal distributed alternative-specific attribute and random coefficient.

Table 1 presents a summary of the identification conditions discussed for the various cases discussed in this section.

Table 1. Summary of identification conditions when the alternative-specific stochastic attribute exhibits only random variation across individuals (i.e., no systematic variation)

Distributional assumptions	**No. of choice alternatives (*L*) with a stochastic alternative attribute for different choice set sizes (J)**	**Rank of Jacobian matrix**	**No. of identifiable parameters in the covariance matrix of utility functions**
Normal distributed alternative-specific stochastic attribute with normal distributed random coefficient	$L = J\ ;\ J \le 5, J \ne 1$	$\frac{J(J-1)}{2}$	$\frac{J(J-1)}{2}-1$
	$L = J\ ;\ J > 5$	Difficult to derive a generic expression for the rank of the Jacobian matrix. Need to derive the rank on a case-to-case basis.	
	$L < J\ ;\ J \le 5, J \ne 1$ (The alternative-specific attribute enters the utility functions of other, $J-L$ alternatives in a deterministic form)	$\frac{J(J-1)}{2}$	$\frac{J(J-1)}{2}-1$
	$L < J\ ;\ J > 1$ (The alternative-specific attribute does not enter the utility functions of other, $J-L$ alternatives)	$\frac{L(L+1)}{2}$	$\frac{L(L+1)}{2}-1$
Lognormal distributed alternative-	$L = J\ \forall J \le 5, J \ne 1$	$\frac{J(J-1)}{2}$	$\frac{J(J-1)}{2}-1$

specific stochastic attribute with lognormal distributed random coefficient	$L = J\ (J > 5)$	Difficult to derive a generic expression for the rank of the Jacobian matrix. Need to derive the rank on a case-to-case basis.	
	$L < J\ ;\ J > 1$ (The alternative-specific attribute enters the utilities of other, $J - L$ alternatives in a deterministic form. OR the alternative-specific attribute does not enter the utilities of $J - L$ alternatives)	$\frac{L(L+1)}{2}$	$\frac{L(L+1)}{2} - 1$

3.3 Discussion

The theoretical analysis in Sections 3.1 and 3.2 provides insights into the identifiability of parameters of alternative-specific stochastic attributes in RP data, given that the random coefficient on the attribute is identified using SP data. Based on this analysis, some virtues and downsides of using of using pooled RP-SP data for inferring stochastic alternative-specific attributes, while recognizing random coefficients on such attributes, are discussed here.

For settings where the alternative-specific stochastic attributes involve both random and systematic variation, the identification conditions discussed in Section 3.1 (and Appendix A) are encouraging for mode choice models when the analyst can infer attributes such as the distance-dependent *IVTT* for up to (at least) $J - 1$ modes if the attributes are normal (lognormal) distributed. The analyst can typically measure the *IVTT* of at least one travel mode (e.g., non-motorized modes) without error and infer the distance-dependent *IVTT* of other modes using pooled RP-SP data. Next, for inferring alternative-specific stochastic attributes such as *OVTT* that are free of systematic variation, the identification conditions derived in Section 3.2 (and Appendix B) can be applied. However, there are some limitations to this method. First, when the stochastic attribute is free of systematic variation, pooling RP-SP data allows for the inference of only a single alternative-specific attribute per choice alternative (not for multiple attributes), and with limits on the number of parameters one can estimate for such an attribute across different choice alternatives. Second, if multiple alternative-specific stochastic attributes need to be inferred for each choice alternative, each variable requires deterministic variation based on a different observable variable. It is not possible to use a single observable variable (such as d_q) to infer multiple alternative-specific attributes. Such model specifications can potentially lead to parameter (un)identifiability issues.

Despite the above-mentioned limitations, the ability to identify both alternative-specific attributes and their coefficients – while recognizing stochasticity in both attributes and the coefficients – is a step forward in the choice modelling literature.

4. SIMULATION EXPERIMENTS

To augment the theoretical investigations, we conduct simulation experiments using a mixed logit model of commute mode choice for a pooled RP-SP data setting based in Bengaluru, India. Using these experiments, we examine the efficacy of our proposed approach (combining RP-SP datasets) in inferring alternative-specific attributes for the choice alternatives in RP settings along with the randomness in the coefficient on such attributes. The experiments also aid in verifying the parameter identifiability conditions laid out in Section 3.1 for the case when the alternative-attributes to be inferred exhibit both systematic and random variation across individuals in the data and are also associated with a random coefficient. Additionally, the experiments help in demonstrating the effects of ignoring variability in alternative-specific attributes when such variability is present.

4.1 Simulation design

We simulated synthetic data considering three different designs for commute mode choice setting. The three designs are variants of each other. Of these, the first and second designs, which involve normal distributed alternative-specific attributes and random coefficients, are discussed in this section. The third design, which involves lognormal distributed attributes and random coefficients, is discussed in Appendix C. For each of the three designs we simulated 200 RP-SP datasets, each comprising 5000 individuals. For each individual in each dataset, four RP choice occasions and one SP choice occasion are generated.

4.1.1 Design-I: Normal distributed stochastic alternative-specific attribute in $J-1$ alternatives with normal distributed random coefficient

In this design, RP and SP data are simulated for a commute mode choice context with five travel mode alternatives – bus, personal car, personal two-wheeler (TW), metro and walk – in Bengaluru, India. The following equations denote the utility functions for the RP alternatives:

$$U_{qi} = \beta_{0i,RP} + \sigma_i z_{qi} + \beta_{TT,q} TT^*_{qi} + \beta_c TC_{qi} + \varepsilon_{qi}\ ;\ \forall i \in \{bus, car, TW, metro\}$$
$$U_{qi} = \beta_{0i,RP} + \sigma_i z_{qi} + \beta_{TTw} TT_{qi} + \varepsilon_{qi}\ ;\ i \in \{walk\} \quad (12)$$

In these utility functions, $\beta_{0i,RP}\ \forall i \in \{car, TW, metro, walk\}$ is the location parameter of the alternative specific constant for a given mode *i*. $\sigma_i z_{qi}\ \forall i \in \{bus, car, TW, metro, walk\}$ is a normal distributed term representing individual-specific random effects for mode *i*, where σ_i is the standard deviation parameter and z_{qi} is a standard normal variate that is same across all RP and SP utility functions of the alternative for the individual. Note that the location parameter of the alternative specific constant for the bus mode is fixed to be zero for identification purposes.

Next, the RP setting travel times for motorized modes ($TT^*_{qi}\ \forall i \in \{bus, car, TW, metro\}$) are expressed as a product of the inverse speed θ_{qi} (time required to traverse unit distance) by the mode and the travel distance, d_q, as below[8]:

$$TT^*_{qi} = \theta_{qi} \times d_q \quad (13)$$

The mode specific inverse speeds θ_{qi} (min/km) are considered random (normal distributed) to allow variability in travel times (for all modes except walk). This specification (Equation (13)) enables the alternative attribute TT^*_{qit} to vary across individuals due to an observed variable (d_q), thus introducing systematic variation, in addition to random variation that cannot be attributed to d_q. The location and scale parameters of θ_{qi} for each mode are to be estimated for the RP setting using the proposed model framework. True values for the location parameters of inverse speeds are assumed based on the average speed of the corresponding mode in Bengaluru and reasonable values for the scale parameters of inverse speeds are assumed based on the order of travel time variability among the modes in the city.

For the walk mode, a fixed walk speed of 4 kmph (i.e., inverse speed of 15 min/km) is assumed to generate walk travel times (TT_{qi}). The inverse speed for walk mode is assumed to be

[8] In our simulation experiments, we assumed that the travel distance between a given origin and destination as the same for all modes. This assumption can be easily relaxed to allow the travel distance to be mode-specific, which we allow in the empirical analysis in Section 5. Also, it is possible to incorporate route choice along with mode choice in a joint modelling framework, albeit that would increase the choice dimensions being modeled.

known to the analyst, while the location parameters and scale parameters of θ_{qi} for all other modes are to be estimated. Data on d_q for each individual is generated from a truncated normal distribution with a minimum and maximum of 1 km and 25 km, respectively, and an average travel distance of 7 km (to represent commute travel distances in Bengaluru).

Next, travel costs for each mode alternative i (except for walk mode, which does not involve monetary costs), which are assumed to be non-stochastic, are denoted by TC_{qi}. Data for TC_{qbus} are generated using distance-based bus fare charts in the city, while that for TC_{qmetro} are constructed based on the fare chart for the metro mode. Next, data for TC_{qcar} and TC_{qTW} are generated based on vehicle maintenance costs, fuel price in Bengaluru, and average mileage of a hatchback car model.

A generic random coefficient $\beta_{TT,q}$, which is assumed to follow $N(-1, 0.5^2)$, is specified on the mode-specific travel times for all motorized modes. A separate deterministic coefficient β_{TTw} (true value assumed to be -0.4) is taken as the coefficient for walk travel time. Considering an average value-of-in-vehicle-time of 133 (in INR/hour), the value of the coefficient on travel cost (β_c) is assumed to be -0.45. All coefficients are assumed to be individual-specific (that is, they do not vary across the SP and RP occasions for an individual). Finally, the kernel error terms in the RP setting, $\varepsilon_{qi}\ \forall i \in \{bus, car, TW, metro, walk\}$, are assumed to be independent Gumbel distributed with scale parameter 1.

The following equations denote the utility functions for the SP alternatives:

$$\begin{aligned} \bar{U}_{qi} &= \bar{\beta}_{0i,SP} + \sigma_i z_{qi} + \beta_{TT,q}\overline{TT}_{qi} + \beta_c \overline{TC}_{qi} + \bar{\varepsilon}_{qi}\ ;\ \forall i \in \{bus, car, TW, metro\} \\ \bar{U}_{qi} &= \bar{\beta}_{0i,SP} + \sigma_i z_{qi} + \beta_{TTw}\overline{TT}_{qi} + \bar{\varepsilon}_{qi}\ ;\ i \in \{walk\} \end{aligned} \tag{14}$$

In these SP utility functions, $\bar{\beta}_{0i,SP}\ \forall i \in \{car, TW, metro, walk\}$ is the location parameter of the alternative specific constant for a given mode i. The $\sigma_i z_{qi}\ \forall i \in \{bus, car, TW, metro, walk\}$ terms, and the travel time and cost coefficients ($\beta_{TT,q}$, β_{TTw}, β_c) are the same as those defined for the RP setting Equations (12). However, Unlike the randomly distributed travel times (TT_{qi}^{*}) in the RP

setting, the travel times ($\overline{TT}_{qi}$) in the SP setting are assumed to be non-stochastic, albeit they are generated in the same way as the RP travel times. $\overline{TC}_{qi}$ are the non-stochastic travel costs for the SP setting generated in the same way as that for the RP setting. The kernel error terms $\bar{\varepsilon}_{qi} \forall i \in \{bus, car, TW, metro, walk\}$ are assumed to be independent Gumbel distributed with scale parameter 0.7 (i.e., the ratio of scales of the SP and RP kernel error terms is 0.7).

Let J denote the total number of choice alternatives available to an individual in the data and L denote the number of alternatives with stochastic travel time. In this simulation design, $J = 5$ and $L = 4$, because 4 of the 5 mode alternatives are associated with stochastic travel times.

4.1.2 Design-II: Normal distributed stochastic alternative-specific attribute in J alternatives with normal distributed random coefficient

Design-II is similar to Design-I in all respects except that the walk mode is not included in the choice set. Here, $J = 4$ and $L = 4$. In this design, since the walk mode is not in the choice set, it is assumed that the location parameter of the metro mode inverse speed is assumed to be known to the analyst (that is the analyst does not have to estimate this parameter). This assumption is necessary to meet the basic identification requirement that only utility differences matter. However, it is assumed that the analyst has to estimate the scale parameters of travel time distributions for all four modes in the choice set. This design is implemented to demonstrate that such a model with normal distributed travel times cannot be identified (more on this later).

4.2 Evaluation and discussion

4.2.1 Evaluation of models estimated on simulated data from Design-I

We estimated the following two models on all 200 simulated RP-SP datasets from Design-I: Model-I and Model-II. Model-I is a mixed logit model with the same structure as the true data generation process (DGP) used to simulate data from Design-I. That is, $L = J - 1$ alternatives (motorized alternatives) involve normal distributed travel time (normal distributed inverse speeds, to be precise) and a normal distributed random coefficient on travel time. The location and scale parameters of all these distributions need to be estimated. However, it is assumed that the analyst knows the travel time of the J^{th} alternative (walk mode) and estimates a deterministic coefficient on its travel time. Model-II simplifies Model-I by ignoring the variability in travel times of all the

motorized modes (i.e., the scale parameter of the normal distributions is assumed to be zero and only the location parameter is estimated for all the motorized modes).

The parameter recovery for each of the models is examined using the following metrics:

(1) *Absolute Percentage Bias* (APB): For a given parameter in the model, APB is the absolute value of the difference between the true parameter value and the mean of the parameter estimates across the 200 simulated datasets – expressed as a percentage of the true parameter value.

(2) *Asymptotic Standard Error* (ASE): ASE for a given parameter is the average (across the 200 simulated datasets) of the standard errors of the parameter's estimated values.

(3) *Finite Sample Standard Error* (FSSE): FSSE for a given parameter is the standard deviation of the parameter's estimated values across the 200 datasets.

Table 2 reports a summary of the above metrics separately for the two models, along with the true parameter values used in the DGP for simulating the 200 datasets. As can be observed from the results for Model-I, the model was able to recover the assumed true parameters accurately and precisely. This includes the scale and location parameters of the normal distributed inverse speeds for all the $J-1$ motorized modes and the scale and location parameters of the corresponding random coefficient. The identification of these parameters was possible because of the presence of at least one mode (walk mode) in the RP choice set with travel times known to the analyst, and the SP data helping in the identification of the random coefficient. In addition, all J scale parameters corresponding to the alternative-specific random effects were identified due to the panel nature of the RP-SP data. These results corroborate the theoretical discussion in Section 3.1.1 for a pooled RP-SP data setting with an alternative-specific attribute that has both systematic and random variation (and the random variation is normal distributed). That is, if the alternative-specific attribute (in the RP setting) and the random coefficient on it are normal distributed, the analyst can combine SP and RP datasets to estimate the location and scale parameters of the alternative-specific attribute for up to $J-1$ alternatives along with the paramters of the corresponding random coefficient.

Table 2. Simulation results for mixed logit models on pooled RP-SP data (normal travel time and normal random coefficient on travel time)

Variable description	True value	Model-I				Model-II				$H_0 : \hat{\beta}_{Model-I} = \hat{\beta}_{Model-II}$
		Mean estimate	APB	ASE	FSSE	Mean estimate	APB	ASE	FSSE	
Location parameters for alternative-specific random effects (bus mode is base)										
Car	1.80	1.623	9.82	0.1622	0.0996	0.805	55.28	0.1002	0.0436	4.29
Two-wheeler	0.30	0.333	10.89	0.1275	0.0665	0.169	43.59	0.0986	0.0497	1.02
Metro	0.50	0.424	15.19	0.0985	0.0555	0.431	13.73	0.0865	0.0242	0.05
Walk	1.50	1.341	10.57	0.1854	0.1379	0.629	58.05	0.1497	0.0225	2.99
Scale parameters for alternative-specific random effects (normal distributed)										
Bus	1.00	0.869	13.10	0.0092	0.0107	0.528	47.15	0.0681	0.0692	4.96
Car	1.85	1.552	16.09	0.1292	0.0763	1.393	24.68	0.0729	0.0701	1.07
Two-wheeler	1.55	1.328	14.29	0.1765	0.0768	1.001	35.42	0.0735	0.0543	1.71
Metro	1.35	1.115	17.42	0.1390	0.0802	0.412	69.48	0.1625	0.0215	3.29
Walk	1.10	1.040	5.44	0.1428	0.1101	0.677	38.49	0.0732	0.1291	2.26
Parameters of mode-specific inverse speeds (for travel times) in the RP setting										
RP bus inverse speed – Location parameter	1.85	1.793	3.06	0.0790	0.0537	1.645	11.06	0.0244	0.0283	1.79
RP bus inverse speed – Scale parameter	0.40	0.351	12.13	0.0897	0.0328	0.000	NA	NA	NA	NA
RP car inverse speed – Location parameter	1.25	1.177	5.81	0.0681	0.0534	1.080	13.61	0.0274	0.0437	1.32
RP car inverse speed – Scale parameter	0.20	0.170	15.06	0.0662	0.0238	0.000	NA	NA	NA	NA
RP TW inverse speed – Location parameter	1.10	1.066	3.12	0.0697	0.0523	0.921	16.29	0.0286	0.0123	1.92
RP TW inverse speed – Scale parameter	0.30	0.276	7.87	0.0582	0.0199	0.000	NA	NA	NA	NA
RP metro inverse speed – Location parameter	1.50	1.450	3.33	0.0730	0.0523	1.428	4.79	0.0241	0.0196	0.29
RP metro inverse speed – Scale parameter	0.15	0.165	10.16	0.0633	0.0220	0.000	NA	NA	NA	NA
Coefficients on level of service variables										
Travel time for motorized modes – Location parameter	-1.00	-0.856	14.36	0.0554	0.0339	-0.602	39.83	0.0281	0.0386	4.09
Travel time for motorized modes – Scale parameter	0.15	0.128	14.71	0.0110	0.0066	0.099	33.83	0.0095	0.0132	2.00
Walk travel time	-0.40	-0.343	14.31	0.0231	0.0136	-0.250	37.61	0.0097	0.0133	3.71
Travel cost	-0.45	-0.386	14.18	0.0238	0.0149	-0.274	39.07	0.0095	0.0116	4.37
SP scale parameter (RP scale is assumed to be 1)	0.70	0.663	5.22	0.0505	0.0317	0.547	21.87	0.0250	0.0299	2.06
Average value across all parameter estimates		NA	10.73	0.0864	0.0511	NA	33.55	0.0595	0.0386	NA

Next, observe from the results for Model-II in Table 2 that the location and scale parameters of the random coefficient on travel time display a statistically significant bias toward zero when compared to the corresponding estimates from Model-I. In addition, the coefficients for other parameters in the utility function also display a similar trend of bias toward zero. This is evident from the t-statistics reported in the last column of the table for the null hypothesis that the parameter estimates from Model-II are the same as those from Model-I. Also, this finding is consistent with the findings in a recent paper by Biswas *et al.* (2024) pertaining to bias due to ignoring stochasticity in alternative attributes whose stochasticity is additive in nature (e.g., normally distributed). As discussed in that study, the additive stochasticity in alternative attributes, when ignored, gets lumped into the kernel error terms and increases the variance of the kernel error terms. Since the model's kernel error scale is not identified, the remaining parameter estimates of the model, which are confounded by the scale of the model, get biased toward zero.

Finally, although not reported in the tables, in terms of model fit, Model-II was found to be inferior to Model-I across all the simulated datasets. The inferior fit to data and the large bias in parameter estimates of Model-II highlight the importance of recognizing stochasticity in alternative attributes.

4.2.2 Evaluation of models estimated on simulated data from Design-II

We estimated mixed logit models (Model III) on simulated data from Design-II with the same structure as the true DGP used in that design. That is, $L = J$ alternatives involved normal distributed travel time (normal distributed inverse speeds, to be precise) and a normal distributed coefficient on travel time. For identification purposes, the location parameter of metro model inverse speed was assumed to be known (i.e., the location parameters of inverse speeds for only $J-1$ mode-specific inverse speeds were estimated). However, the scale parameters of all J distributions were estimated. Such a model ran into parameter (un)identifiability issues. Only after fixing at least one mode-specific inverse speed scale as a known parameter, the model was identified, and we could retrieve the other parameters with similar levels of accuracy and precision as that of Model-I in Table 1. These results again corroborate the findings in Section 3.1.1 that the location and scale parameters of up to only $J-1$ normal distributed mode-specific inverse speeds can be estimated along with the parameters of the corresponding random coefficient.

In addition, we carried out a third set of simulation experiments for lognormal distributed stochastic alternative-specific attribute in J alternatives with lognormal distributed random coefficient. The corresponding simulation design and evaluation of the results are presented in Appendix C of the paper.

Note that all the simulation experiment results reported in this paper are based on estimations carried out using 400 Halton draws to simulate the stochastic terms in the likelihood expressions. We explored increasing the number of Halton draws to 600 for a few simulated datasets and did not find substantial differences in the parameter estimates from those obtained using 400 draws. Also, the parameters of the proposed model were recovered well using 400 draws.

5. EMPIRICAL ANALYSIS

In this section, we present an empirical analysis of commute mode choice using pooled RP-SP data from Bengaluru, India, to demonstrate the feasibility of inferring in-vehicle travel times (*IVTT*) in RP settings along with the corresponding coefficient. Furthermore, using this empirical analysis we highlight the importance of recognizing stochasticity in both mode-specific in-vehicle travel times and the corresponding coefficient.

5.1 Empirical Data

The empirical data used for this analysis was obtained from a survey conducted from February to April 2022 to understand the travel behaviour of the residents of the Bruhat Bengaluru Mahanagara Palike (BBMP) area of Bengaluru. For this analysis, we used data of only those who reported commuting as the most frequent purpose of their travel. Such commuters were asked about their most frequently used travel mode for their home-to-work commute, along with information on the travel origin (home) and destination (work) locations for their commute, travel distance, and travel times and costs for their most frequently used travel modes.

In addition to the RP questions, the survey included an SP section for mode choice, which comprised four different SP choice scenarios. The first three of the four SP scenarios corresponded to a non-pandemic situation, where it was stated that the risk of the pandemic was minimal, and the entire population was vaccinated. The fourth scenario described travel situations in a pandemic

setting. In the current empirical analysis, the three non-pandemic SP scenarios were used for every respondent in the data sample.

For administering both the RP and SP components of the survey in a single meeting with the respondent, a repository of SP scenarios was pre-generated based on b-efficient designs using the Ngene software (ChoiceMetrics, 2012). Different sets of scenarios were pre-generated for each of the following travel distance bands: 0-5 km, 5-10 km, 10-15 km, 15-20 km, 20-25 km and 25-30 km. It was assumed that the maximum travel distance for an individual within the survey region would be within 30 km. The distance bands were used as the basis to compute mode-specific attributes such as *IVTT*, *OVTT*, and travel cost for the SP scenarios. Soon after a respondent completed the RP section of the survey, appropriate SP scenarios, drawn on-the-fly from the pre-generated repository, were presented to the respondent based on their reported travel distances and mode availability in the RP section. The RP distance-band based SP scenario generation was done to reduce the risk of endogeneity due to pivoting off directly from RP attributes (such as *IVTT*, *OVTT*, and cost) for generating SP attributes (Guavera and Hess, 2019), while also presenting hypothetical mode choice settings that were not too different from the respondents' commute settings.

The final estimation sample comprised data from a total of 914 respondents, with one RP choice occasion and three SP choice occasions per respondent. A total of ten modal alternatives were considered in the RP section of the survey – (1) own car (or just "car" from here on), (2) own two-wheeler (or just "two-wheeler" from here on), (3) auto-rickshaw, (4) bus, (5) metro, (6) walk, (7) bicycle, (8) ride-hailing (Ola or Uber) cars, (9) ride-hailing (Ola or Uber) two-wheelers, and (10) other modes. However, less than ten respondents indicated they would choose either of the two ride-hailing modes, bicycle, and other modes. This may be because ride-hailing is not commonly used, and bicycles are rarely used for commuting purposes in Bengaluru. Therefore, only the first six modes were considered for the current analysis. The RP mode shares in the sample are as follows: 6.4% for car, 51.0% for two-wheeler, 3.3% for auto-rickshaw, 22.4% for bus, 14.7% for metro, and 2.2% for walk. Additional details of the empirical data, including the sociodemographic characteristics of the respondents, the rules used to determine availability of different modes for each respondent, and the mode-specific level-of-service variables (travel times and costs) used in one of the models estimated in the study are presented in Appendix D.

5.2 Empirical results and findings

We estimated the following three different empirical models in this study:

(1) Model A, which is the proposed model that combines SP and RP data to estimate the distribution of mode-specific *IVTT* (mode-specific inverse speeds, to be precise) in RP settings as well as random coefficient on *IVTT* in both RP and SP settings. In this model, the inverse speed of the metro mode was fixed to 1.67 minutes per km, considering the average speed of metro in Bengaluru as 36 kmph, and the walk inverse speed for the walk mode was assumed to be 15 minutes per km. For both these modes, the variability in inverse speeds was assumed to be zero. This was done for two reasons – first, to assist in the estimation (identifiability) of the model, and the other, because of negligible variability in metro in-vehicle travel times and that of walking times. For the other four modes – car, bus, two-wheeler, auto rickshaw – we estimated the distribution of mode-specific inverse speeds using this model, assuming that the inverse speeds are power lognormal distributed.[9] The coefficient on *IVTT* is assumed to follow a lognormal distribution, whose parameters are estimated using empirical data.

(2) Model B, which is a simpler version (and a special case) of Model A, ignores the variability in mode-specific inverse speeds, and only estimates a single value of mode-specific inverse speed (as opposed to a distribution) for each of the car, bur, two-wheeler, and auto rickshaw modes. Similar to Model A, this model assumes a fixed inverse speed of 1.67 minutes per km for the metro mode and 12 minutes per km for the walk mode. Further, the coefficient on *IVTT* is assumed to follow a lognormal distribution, whose parameters are estimated using empirical data.

(3) Model C, which does not involve any estimation of mode-specific inverse speeds, uses exogenously obtained travel times for RP data instead of estimating the mode-specific

[9] The power lognormal distribution is a three-parameter distribution – with location, scale, and power parameter as its three parameters. A special case of this distribution, when the power parameter is 1, is the familiar lognormal distribution. When the power parameter is greater than 1, the power lognormal distribution has an advantage over the lognormal distribution in that it has a thinner right tail (i.e., low PDF values for large values) than that of the lognormal distribution. This property facilitates an easier estimation than models that use the lognormal distribution (Bhat and Lavieri, 2018). The reader is referred to Bhat and Lavieri (2018), who originally introduced the use of this distribution in discrete choice models, for a discussion on the properties and advantages of this distribution when compared to the lognormal distribution. For the estimation of the proposed model in this study, we fixed the power parameter terms to a specific value in the range from 1 to 3 and estimated the other parameters of the model. After exploring different values of the power parameter, we found that the power parameter value of 2.4 yielded the best fitting model.

travel time (or inverse speed) distributions. Such mode-specific travel time data were obtained from Google Application Programming Interfaces (APIs), a commonly used source of travel time data in the recent past. In this model, similar to the other two models, the coefficient on *IVTT* is assumed to follow a lognormal distribution, whose parameters are estimated using empirical data.

In all the above three models, exogenously obtained data on *OVTT* and travel costs are used as discussed in Appendix D (i.e., *OVTT* and travel costs are assumed to be exogenously known to the analyst).[10] In all the three models, the *OVTT* variable enters the utility functions in the form of *OVTT/distance* with lognormal distributed coefficient. Further, in all the three models, the auto-rickshaw mode of transportation is considered the base alternative for the introduction of the effects of alternative-invariant exogenous variables. Note that we considered correlation between the coefficients of *IVTT* and *OVTT/distance* variables, but did not recover any significant correlation. There may be some correlation among the travel times (e.g., *IVTT*) of different modes in Model A. For example, for a given origin-destination (OD) pair, it is likely that travel times of some modes might covary because of OD and route-specific unobserved effects. However, it is difficult to identify such correlations because the generic random coefficient on travel time across the different modes would pick up the correlations.

All the three models were estimated using 400 Halton draws to simulate the random terms in the likelihood expressions. Increasing the number of draws to 800 did not yield significantly different parameter estimates from those of models estimated using 400 draws.

The estimation results for each of these models are presented in Table 4 and Table 5. Table 4 presents the goodness-of-fit measures and implied values of time. As can be observed from the goodness-of-fit measures in this table, the proposed model provides better fit than the other two models (in terms of AIC, BIC, as well as Rho-squared values). In addition, the mean monetary values of *IVTT* and *OVTT* savings (at average travel distance of 9.85 km) from the proposed model are 199.28 INR/hour and 470.75 INR/hour, respectively, whereas those for Model B and Model C

[10] It may be argued that the *OVTT* and travel cost values considered by the travellers need not be known accurately by the analyst, and therefore the analyst should infer these values, too. Doing so, along with inferring the *IVTT* is an avenue for future research.

appear to be too high for an Indian city context.[11] Based on these comparisons, Model A can be adjudged the best of the three models estimated in the study.

Table 5 presents the parameter estimates of all the three models. Since Model A provided better fit and more reasonable values of monetary values of *IVTT* and *OVTT* savings than the other two models, we focus the discussion on this model. The first and second broad row panels of Table 5 provide the estimates of the mode-specific constants for RP and SP datasets respectively. These estimates do not have any substantive interpretation but adjust to best fit the sample shares after accommodating the effects of other explanatory variables. The next set of rows report the scale parameter estimates corresponding to the mode-specific random effects that persist across the different choice occasions of an individual, which could be estimated due to the presence of multiple choice occasions per individual in the RP-SP data. These random effects capture correlations among the mode-specific utility functions of different choice occasions of an individual.

Table 4. Empirical data fit and money values of time for pooled RP-SP models of commute mode choice

	Model A (proposed model)	**Model B (RP travel time variability ignored)**	**Model C (with exogenous RP travel times)**
Goodness-of-fit measures			
Log likelihood at convergence	-1,495.19	-1,525.31	-1,562.58
Log likelihood for constants-only model	-2,655.23	-2,655.23	-2,655.23
Mc Fadden's Rho-squared	0.44	0.43	0.41
Akaike Information Criterion (AIC)	3,102.37	3,154.62	3,221.16
Bayesian Information Criterion (BIC)	3,372.17	3,405.15	3,452.42
Money value of time measures			
Mean value of *IVTT* (INR /hour)	199.28	209.78	314.91
Mean value of *OVTT* at avg. travel distance (INR /hour)	470.75	935.63	1241.38

[11] It is not easy to compare these values against the money values of time reported in the literature for Indian cities, as the other studies do not consider variability in travel times along with the random coefficients. Besides, most studies in the Indian context use the basic multinomial logit (MNL) model for mode choice analysis. Therefore, we estimated an MNL model and obtained the resulting money values of *IVTT* and *OVTT* as 83 INR/hr and 171 INR/hr, which are in line with the money values of *IVTT* and *OVTT* reported in other studies on Indian cities (Athira *et al.*, 2016).

Next, observe from the fourth and fifth broad row panels in Table 5 that the proposed model (Model A) enables the simultaneous estimation of the distributions of mode-specific *IVTT* (or, mode-specific inverse speeds) in RP settings while allowing the coefficient on *IVTT* to be randomly distributed. The estimates obtained for the mode-specific inverse speed distributions are reasonable. For example, it takes an average of 2.23 minutes by car to travel 1 km while it takes 3.65 minutes by bus to travel the same distance.

The fifth broad row panel in Table 5 provides the parameter estimates on the level-of-service variables – *IVTT*, *OVTT*/distance, and travel cost. As discussed earlier, all three models allow the coefficients on *IVTT* and *OVTT*/distance to be randomly (lognormal) distributed. Next, note from the parameter estimates of Model B and Model C in Table 5 that the expected value and standard deviation of the random coefficient on *IVTT* display a bias toward zero when compared to the corresponding estimates from Model A. In addition, the coefficients for other parameters in the utility function also display a similar trend of bias toward zero. This finding is consistent with the findings from our simulation experiments in Section 4 and also with those in a recent paper by Biswas *et al.* (2024). [12]

Furthermore, socio-demographic variables such as gender, income and age are included in the model specification (see the sixth, seventh, and eighth broad row panels of Table 5). The results for the gender variable suggest that commuting men are more likely to prefer car, two-wheeler and metro modes relative to commuting women whereas there is no difference in preferences between men and women for the walking mode. This may be because men are likely to make longer trips compared to women (Saigal *et al.*, 2021). Among the motorized modes, car and two-wheelers appear to be the most preferred by men over other modes, perhaps because men tend to have a greater access (than women) to vehicles in Indian households (Shirgaokar *et al.*, 2018; Jain and

[12] Each of the three models reported in Table 5 constrain the coefficients on the LOS attributes to be same between the RP and SP settings. As discussed in Section 2.1, this constraint is necessary for only the coefficient of *IVTT* (since *IVTT* in the RP setting is inferred from the pooled data model). That is, there is no need to impose such a constraint on the coefficients of other LOS attributes – *OVTT*/distance, walk travel time, and travel cost – in the model. Therefore, we estimated additional empirical specifications (for the proposed model) that allowed the coefficients on LOS attributes other than *IVTT* to be different between the RP and SP settings. However, doing so did not yield significant improvement in model fit when compared to the corresponding specification reported in Table 5. Since the data fit did not improve when we allowed different coefficients for SP and RP settings, we retained the specification that constrained the SP and RP coefficients to be same for all LOS attributes (while allowing the scales of the kernel error terms and means of the alternative-specific constants to be different between the SP and RP settings).

Tiwari, 2019). Next, from the parameter estimates for the income variables, note that low-income commuters are less likely than high-income commuters to prefer car and two-wheeler modes. Also, the middle-income commuters' preference for the metro mode is higher than that of high-income commuters. In addition, middle income commuters demonstrate a greater preference for two-wheelers than higher income commuters, perhaps because the latter segment is more likely to prefer cars over two-wheelers. Next, individuals aged between 26 to 45 years are less likely to prefer bus, metro, and two-wheelers as they are more likely to prefer car when compared to younger individuals (in the age group 19 to 25 years). This is expected since younger individuals would be more likely to use public transit modes and two-wheelers when compared to older individuals who would place a higher value on comfort and convenience as well be more likely to have greater access to personal cars. Finally, individuals of age above 45 years are the least likely to prefer the bus mode among all individuals, perhaps because they prefer modes that offer more comfortable travel than buses.

The last reported parameter estimate for all the three models is the scale of the utility functions in the SP setting (Note: the scale for the RP setting is fixed to be 1). The estimated parameter value is smaller than 1, which implies that the unobserved factors influencing choices in the SP setting exhibit lower variability than those in the RP setting.

Table 5. Empirical parameter estimates for pooled RP-SP models of commute mode choice[13]

Variable description	Model A (proposed model)		Model B (RP travel time variability ignored)		Model C (with exogenous RP travel times)	
	Par. est.	t-stat.	Par. est.	t-stat.	Par. est.	t-stat.
Location parameters for normal distributed alternative-specific random effects (Auto-rickshaw mode is the base)						
Metro – RP	11.535	10.38	11.064	10.16	8.664	4.41
Walk – RP	6.296	1.93	6.211	1.37	4.979	1.65
Car – RP	4.915	7.03	4.883	7.44	4.275	2.47
Two-wheeler – RP	7.098	9.72	7.068	10.32	7.066	3.96
Bus – RP	7.910	8.22	7.637	8.27	7.348	3.90
Metro – SP	8.433	11.95	8.255	12.34	8.153	3.51
Walk – SP	9.097	2.39	9.056	1.57	8.510	2.27
Car – SP	7.205	10.67	7.117	10.77	7.104	3.12
Two-wheeler – SP	4.305	9.85	4.255	10.57	3.410	1.60
Bus – SP	9.093	13.35	9.011	12.82	8.721	3.69
Scale parameters for normal distributed alternative-specific random effects						
Metro	4.315	5.18	4.250	3.49	2.521	4.41
Walk	1.062	2.39	1.043	1.87	1.017	2.95
Car	3.540	2.55	3.354	2.58	3.340	2.50
Two-wheeler	1.037	2.83	0.841	1.30	0.760	2.01
Auto-rickshaw	4.664	6.61	4.329	2.72	4.073	3.07
Bus	5.815	10.01	5.806	10.36	3.880	5.36
Mean and standard deviations of mode-specific inverse speeds (for IVTT) in RP setting (min/km) [14]						
Car – Expected value	2.227	53.27	0.333	1.06	NA	NA
Car – Standard deviation	0.372	2.52	NA	NA	NA	NA
Two-wheeler (TW) – Expected value	1.894	10.82	0.826	3.40	NA	NA
Two-wheeler (TW) – Standard deviation	0.525	1.59	NA	NA	NA	NA
Auto-rickshaw – Expected value	3.256	18.76	1.048	5.39	NA	NA
Auto-rickshaw – Standard deviation	1.295	4.60	NA	NA	NA	NA
Bus – Expected value	3.652	27.20	1.117	5.21	NA	NA
Bus – Standard deviation	1.709	6.63	NA	NA	NA	NA

[13] '*' in a specific cell indicates that the corresponding parameter estimate was not statistically significant and hence dropped from the specification. 'NA' in a specific cell indicates 'Not Applicable'.

[14] For Model A, inverse speeds are considered power lognormal distributed. Location and scale parameters were estimated while fixing the power parameter values. For ease of interpretation, the values reported here are the resulting expected values and standard deviations of the power-lognormal distribution; not the location and scale parameter estimates.

Table 5 (Contd.). Empirical parameter estimates for pooled RP-SP models of commute mode choice

Variable description	Model A (proposed model)		Model B (RP travel time variability ignored)		Model C (with exogenous RP travel times)	
	Par. est.	t-stat.	Par. est.	t-stat.	Par. est.	t-stat.
Coefficients on level-of-service variables (lognormal distributed coefficients on IVTT and OVTT/distance)						
IVTT (min) for motorized modes – Expected value	-0.109	-4.01	-0.058	-6.38	-0.058	-1.52
IVTT (min) for motorized modes – Standard deviation	0.120	3.49	0.118	5.62	0.112	1.03
Walk travel time (min)	-0.179	-1.60	-0.134	-0.78	-0.074	-0.90
OVTT/distance (min/km) for motorized modes – Expected value	-0.259	-7.83	-0.257	-4.46	-0.228	-6.15
OVTT/distance (min/km) for motorized modes – Standard deviation	0.239	5.61	0.216	3.13	0.156	2.70
Travel cost (INR)	-0.033	-4.18	-0.017	-2.12	-0.011	-2.24
Gender (Female is the base, Auto-rickshaw is base mode)						
Metro	2.099	2.24	1.771	1.30	1.687	1.54
Car	4.883	4.75	2.690	1.99	1.818	1.56
Two-wheeler	4.750	4.50	2.446	1.89	1.584	1.34
Income dummy variables (High income is base category, Auto-rickshaw is base mode)						
Medium income - Metro	4.536	3.53	2.823	2.11	2.328	2.21
Medium income - Walk	-3.436	-1.90	-3.394	-1.44	-1.681	-1.12
Medium income – Car	1.050	1.08	*	*	*	*
Medium income - Two-wheeler	1.831	1.78	1.776	1.49	*	*
Low income – Walk	-2.605	-1.39	-2.577	-1.24	-2.188	-1.34
Low income – Car	-2.240	-2.14	-1.938	-1.22	-1.461	-1.55
Low income – TW	-2.383	-2.24	-2.156	-1.31	-2.132	-2.04
Age (Age 19-25 years is the base category; Auto-rickshaw is base mode)						
26-45 years – Bus	-3.109	-2.53	-3.094	-2.08	-2.085	-2.02
26-45 years – Metro	-2.068	-1.84	-1.930	-1.47	-1.362	-1.42
26-45 years – Walk	*	*	*	*	-1.700	-1.00
26-45 years – Car	3.530	3.26	3.335	2.40	2.559	2.63
26-45 years – TW	-5.858	-4.08	-5.804	-3.12	-3.257	-2.93
>45 years – Bus	-1.317	-1.20	-1.287	-1.10	-1.283	-1.23
>45 years – TW	-1.600	-1.58	-1.555	-1.26	-1.096	-1.13
SP scale parameter (RP scale is assumed to be 1; t-stat is against 1)	0.356	10.53	0.198	11.91	0.150	19.22

Table 6 reports the mode-specific speeds for each of the three empirical models estimated in the study. For Model A, these values were obtained by first simulating the power lognormal distributions of the inverse speeds (whose distributional parameters were estimated using the proposed model), then taking an inverse of the simulated values to simulate the speeds, and then taking an average and standard deviation of the simulated speeds. For Model B, the estimated mode-specific fixed inverse speeds were inverted to obtain the speeds. For Model C, the speeds were computed from the exogenously obtained travel times and distances from Google APIs. Note from the modal speed values inferred from Model A that the two-wheeler average speed is the highest among all the modes, while the bus mode has the lowest average speed. Further, the estimated average speeds are in the ballpark range of the speeds typically observed in Bengaluru (Infrastructure Development Corporation (Karnataka) Ltd., 2020). In addition, the standard deviations and coefficients-of-variation in the inferred mode-specific speeds offer insight into the variation in the mode-specific speeds profiles perceived by travellers. Such insights cannot be obtained from Model B, which ignores the variability in mode-specific inverse speeds. On the other hand, recall that Model C utilizes exogenous data on travel times. It is apparent from the last row of the Model C column that the exogenous data is based on speeds that are relatively higher than those inferred from Model A. For example, the data used for Model C suggests an average of 22 kmph for bus travel speeds. However, other estimates in Bengaluru (Infrastructure Development Corporation (Karnataka) Ltd., 2020) suggest lower speeds that are nearer to the speeds inferred from Model A.

Table 6. Mode-specific speeds

	Model A (proposed model)	**Model B (RP travel time variability ignored)**	**Model C (with exogenous RP travel times)**
Car	Mean = 27.7 kmph; SD = 4.8	Mean = 43.0 kmph	Mean = 25.6 kmph
Two-wheeler	Mean = 34.2 kmph; SD = 9.9	Mean = 26.3 kmph	Mean = 35.0 kmph
Auto-rickshaw	Mean = 21.5 kmph; SD = 8.9	Mean = 21.0 kmph	Mean = 22.0 kmph
Bus	Mean = 20.3 kmph; SD = 10.2	Mean = 19.6 kmph	Mean = 22.7 kmph

In summary, the empirical results are consistent with the findings from the simulation experiments and demonstrate that combining RP and SP datasets enables the simultaneous estimation of alternative attributes (and the stochasticity therein) in RP settings as well as a random coefficient on such attributes. Additionally, ignoring stochasticity in alternative attributes, when present, can cause bias in parameter estimates of the random coefficient on these attributes as well as other parameter estimates in the model. Furthermore, the mode-specific speeds inferred from the proposed model (Model A) are closer to the speeds typically observed in Bengaluru than those inferred from Model B or those obtained exogenously (from Google APIs).

6. CONCLUSION

In the current study, we explore the feasibility of using pooled revealed preference and stated preference (RP-SP) data models to simultaneously infer alternative attributes and the corresponding coefficients as well as the stochasticity in both alternative attributes and their coefficients. To do so, we formulate a mixed logit choice modelling framework for pooled RP-SP datasets with the alternative attributes in the RP data and the corresponding coefficients common to both RP and SP settings as unknown parameters to be estimated. Using such a mixed logit model framework for a mode choice setting, we conduct theoretical investigations to examine whether the parameters describing the distributions of alternative attributes in the RP setting can be identified (and how many such parameters can be identified) along with the parameters describing the distribution of random coefficients on the attributes. We carried out the investigations separately for the following two different types of alternative-specific (mode-specific) attributes: (1) the attributes exhibit systematic variability across individuals in the data – due to variation that can be expressed as a function of an observed variable – as well as random variability, and (2) the attributes exhibit only random variability across individuals in the data. Further, for each case, we lay out the identification conditions for two widely used distributional assumptions on the alternative attributes and random coefficients on them – (1) normal distribution and (2) lognormal distribution.

Our investigations revealed that for settings where the alternative-specific stochastic attributes involve both random and systematic variation, the analyst can infer the distribution of the attributes (e.g., distance-dependent *IVTT*) for up to (at least) $J-1$ choice alternatives if the

attributes are normal (lognormal) distributed. To apply such a framework to model mode choice while inferring the *IVTT* of different travel modes, the analyst can measure the *IVTT* of at least one mode (e.g., non-motorized modes) without error and infer the distance-dependent *IVTT* of other modes using pooled RP-SP data. Next, we derived separate identification conditions for inferring alternative-specific stochastic attributes such as *OVTT* that do not exhibit systematic variation. Further, we discuss the limitations of this method pertaining to the identification of stochastic attributes free of systematic variation and if multiple alternative-specific stochastic attributes need to be inferred for each choice alternative.

To augment the theoretical investigations, we conduct simulation experiments to examine the efficacy of combining RP-SP datasets in inferring the values of an alternative attribute for the choice alternatives in RP settings along with the random coefficient on that attribute and any other alternative attributes. Our findings from these experiments corroborated those from the theoretical investigations carried out for parameter identification. Finally, we present an empirical analysis of commute mode choice using pooled RP-SP data from Bengaluru to demonstrate the feasibility of inferring mode-specific *IVTT* in RP settings along with the corresponding random coefficient.

In summary, this study formulates a methodology to infer stochastic alternative attributes and identify the randomness in their coefficients through pooling SP and RP datasets. By doing so, the study addresses the problem of confounding between these two sources of stochasticity, which has not been addressed well in the literature. Nevertheless, this study is not without limitations. First, the study assumes that all attributes presented in SP choice occasions are considered by the survey respondents. This assumption ignores the issue of attribute non-attendance, where some attributes may not be considered by all respondents. Second, the empirical application in this study demonstrates the application of the proposed method for inferring mode specific *IVTT* values when estimating mode choice models. It would be useful to expand the empirical application to infer *OVTT* values and crowding levels in transit modes. These issues comprise important avenues for future research.

ACKNOWLEDGEMENTS

The authors are thankful to Malik Najeebul Feroz for his assistance in data analysis, model estimation, and for being part of several discussions throughout the course of working on the paper.

Karthik Srinivasan, Ganesh Ambi Ramakrishnan, and Deepa L provided valuable inputs on the mode choice survey questionnaire and design. Sandhya Sourirajan assisted with coding the survey for computer (tablet)-assisted administration, while several interns helped conduct the survey. The authors acknowledge the support provided by the Ministry of Education (MoE) of the Government of India through its Scheme for Promotion of Academic and Research Collaboration (SPARC) program in facilitating this research collaboration across various academic institutions, as well as the Ministry of Electronics and Information Technology, Government of India for providing funding which was used in the data collection for this research. In addition, the second author acknowledges support from the Satish Dhawan-IoE-Visiting Chair Professor position at the Indian Institute of Science for this collaboration. Anonymous reviewers provided helpful comments on an earlier version of the manuscript.

Appendix A: Parameter identification when alternative-specific stochastic attribute exhibits both systematic and random variation across individuals in the data

Consider the following utility expressions for an RP alternative i for individual q:

$$U_{qi} = \beta_{0i} + \beta_{TT,q} TT^*_{qi} + \varepsilon_{qi} \quad \text{(A1)}$$

In this equation, TT^*_{qi} is the stochastic alternative-specific travel time variable with a random coefficient $\beta_{TT,q}$. The parameters describing the distribution of $\beta_{TT,q}$ are estimated with the help of SP data. ε_{qi} is a Gumbel distributed error term specified as IID across all RP alternatives and across all individuals. Denote its variance as $\rho^2 g$, where ρ is the scale of the kernel error terms and $g = \frac{\pi^2}{6}$. To rewrite the above utility expression, express TT^*_{qi} as a function of an observed variable travel distance d_q and randomly varying inverse speed θ_{qi} (i.e., $TT^*_{qi} = d_q \theta_{qi}$), and denote the random coefficient $\beta_{TT,q}$ as y_q. The utility expression in Eq. (A1) can now be written as:

$$U_{qi} = \beta_{0i} + y_q d_q \theta_{qi} + \varepsilon_{qi} \quad \text{(A2)}$$

Using the above notation, consider a choice setting with three alternatives, with the following utility expressions:

$$
\begin{aligned}
U_{q1} &= \beta_{01} + y_q d_q \theta_{q1} + \varepsilon_{q1} \\
U_{q2} &= \beta_{02} + y_q d_q \theta_{q2} + \varepsilon_{q2} \\
U_{q3} &= \qquad y_q d_q \theta_{q3} + \varepsilon_{q3}
\end{aligned}
\tag{A3}
$$

For this setting, we discuss parameter identifiability for two different sets of distributional assumptions – (1) normal distribution and (2) lognormal distribution – on θ_{qi} and y_q.

<u>*Stochastic alternative-specific attribute and its coefficient follow normal distribution*</u>

Consider that $\theta_{qi} \sim N\left(\mu_{\theta i}, \sigma_{\theta i}^2\right)$ and $y_q \sim N\left(\mu_y, \sigma_y^2\right)$, where μ_y and σ_y are estimable (therefore, known) using SP data. In this case, the utility expressions can be rewritten to separate the location parameters ($\mu_{\theta i}$) from the scale parameters ($\sigma_{\theta i}$), as:

$$
\begin{aligned}
U_{q1} &= \beta_{01} + y_q d_q (\mu_{\theta 1} + \sigma_{\theta 1} z_{q1}) + \varepsilon_{q1} \\
U_{q2} &= \beta_{02} + y_q d_q (\mu_{\theta 2} + \sigma_{\theta 2} z_{q2}) + \varepsilon_{q2} \\
U_{q3} &= \qquad y_q d_q (\mu_{\theta 3} + \sigma_{\theta 3} z_{q3}) + \varepsilon_{q3}
\end{aligned}
\tag{A4}
$$

Next, writing the utility equations in differenced form, with respect to U_{q3}, we have:

$$
\begin{aligned}
U_{q1} - U_{q3} &= \beta_{01} + y_q d_q (\mu_{\theta 1} - \mu_{\theta 3}) + y_q d_q (\sigma_{\theta 1} z_{q1} - \sigma_{\theta 3} z_{q3}) + (\varepsilon_{q1} - \varepsilon_{q3}) \\
U_{q2} - U_{q3} &= \beta_{02} + y_q d_q (\mu_{\theta 2} - \mu_{\theta 3}) + y_q d_q (\sigma_{\theta 2} z_{q2} - \sigma_{\theta 3} z_{q3}) + (\varepsilon_{q2} - \varepsilon_{q3})
\end{aligned}
\tag{A5}
$$

The above utility differences can be rewritten, by expanding y_q as $\mu_y + \sigma_y z_{qy}$, as below:

$$
\begin{aligned}
U_{q1} - U_{q3} &= \beta_{01} + \mu_y d_q (\mu_{\theta 1} - \mu_{\theta 3}) + \sigma_y z_{qy} d_q (\mu_{\theta 1} - \mu_{\theta 3}) + y_q d_q (\sigma_{\theta 1} z_{q1} - \sigma_{\theta 3} z_{q3}) + (\varepsilon_{q1} - \varepsilon_{q3}) \\
U_{q2} - U_{q3} &= \beta_{02} + \mu_y d_q (\mu_{\theta 2} - \mu_{\theta 3}) + \sigma_y z_{qy} d_q (\mu_{\theta 2} - \mu_{\theta 3}) + y_q d_q (\sigma_{\theta 2} z_{q2} - \sigma_{\theta 3} z_{q3}) + (\varepsilon_{q2} - \varepsilon_{q3})
\end{aligned}
\tag{A6}
$$

The deterministic components of the above utility differences are $\beta_{01} + \mu_y d_q (\mu_{\theta 1} - \mu_{\theta 3})$ and $\beta_{02} + \mu_y d_q (\mu_{\theta 2} - \mu_{\theta 3})$. It can be observed from these deterministic components that only two of the three location parameters – $\mu_{\theta 1}, \mu_{\theta 2}$, and $\mu_{\theta 3}$ – can be estimated. In a general case with J choice alternatives, one can estimate up to $J-1$ location parameters.

Now, let us turn to the random components of the utility differences, with the differences taken with respect to the third alternative utility function. The corresponding covariance matrix of random utility differences is:

$$\Omega_{\Delta 3}=\begin{bmatrix} \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 3}^2 d_q^2 \\ +\mu_{\theta 1}^2\sigma_y^2 d_q^2+\sigma_{\theta 1}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 1}^2 d_q^2 \\ -2\mu_{\theta 3}\mu_{\theta 1}\sigma_y^2 d_q^2+2\rho^2 g \end{Bmatrix} & \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2 \\ +\mu_y^2\sigma_{\theta 3}^2 d_q^2+\mu_{\theta 1}\mu_{\theta 2}\sigma_y^2 d_q^2 \\ -\mu_{\theta 1}\mu_{\theta 3}\sigma_y^2 d_q^2-\mu_{\theta 3}\mu_{\theta 2}\sigma_y^2 d_q^2+\rho^2 g \end{Bmatrix} \\ \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2 \\ +\mu_y^2\sigma_{\theta 3}^2 d_q^2+\mu_{\theta 1}\mu_{\theta 2}\sigma_y^2 d_q^2 \\ -\mu_{\theta 1}\mu_{\theta 3}\sigma_y^2 d_q^2-\mu_{\theta 3}\mu_{\theta 2}\sigma_y^2 d_q^2+\rho^2 g \end{Bmatrix} & \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 3}^2 d_q^2 \\ +\mu_{\theta 2}^2\sigma_y^2 d_q^2+\sigma_{\theta 2}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 2}^2 d_q^2 \\ -2\mu_{\theta 3}\mu_{\theta 1}\sigma_y^2 d_q^2+2\rho^2 g \end{Bmatrix} \end{bmatrix} \quad \text{(A7)}$$

This covariance matrix can be separated into two components – one that varies across individuals based on the observed variable d_q and another that does not vary across individuals – as below:

$$\Omega_{\Delta 3}=\begin{bmatrix} \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 3}^2 d_q^2 \\ +\mu_{\theta 1}^2\sigma_y^2 d_q^2+\sigma_{\theta 1}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 1}^2 d_q^2 \\ -2\mu_{\theta 3}\mu_{\theta 1}\sigma_y^2 d_q^2 \end{Bmatrix} & \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2 \\ +\mu_y^2\sigma_{\theta 3}^2 d_q^2+\mu_{\theta 1}\mu_{\theta 2}\sigma_y^2 d_q^2 \\ -\mu_{\theta 1}\mu_{\theta 3}\sigma_y^2 d_q^2-\mu_{\theta 3}\mu_{\theta 2}\sigma_y^2 d_q^2 \end{Bmatrix} \\ \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2 \\ +\mu_y^2\sigma_{\theta 3}^2 d_q^2+\mu_{\theta 1}\mu_{\theta 2}\sigma_y^2 d_q^2 \\ -\mu_{\theta 1}\mu_{\theta 3}\sigma_y^2 d_q^2-\mu_{\theta 3}\mu_{\theta 2}\sigma_y^2 d_q^2 \end{Bmatrix} & \begin{Bmatrix} \mu_{\theta 3}^2\sigma_y^2 d_q^2+\sigma_{\theta 3}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 3}^2 d_q^2 \\ +\mu_{\theta 2}^2\sigma_y^2 d_q^2+\sigma_{\theta 2}^2\sigma_y^2 d_q^2+\mu_y^2\sigma_{\theta 2}^2 d_q^2 \\ -2\mu_{\theta 3}\mu_{\theta 1}\sigma_y^2 d_q^2 \end{Bmatrix} \end{bmatrix}+\begin{bmatrix} 2\rho^2 g & \rho^2 g \\ \rho^2 g & 2\rho^2 g \end{bmatrix}$$

(A8)

The part of the covariance matrix that exhibits variation across individuals (due to d_q) has all the unknown parameters related to the distributions of the alternative-specific stochastic TT_{qi}^* variables. Since this part varies across individuals, there is no need to examine the structure of the covariance matrix for any linear dependencies in this part. In such cases, the basic principle that only differences in utility matter (and the overall scale of the model is not identified) is invoked to determine the identifiability of parameters. To do so, one needs to examine the structure of the random utility differences and assess how many scale parameters can be estimated. In this context, the random components of the utility differences in Equation (A5) that include the scale parameters are $y_q d_q(\sigma_{\theta 1} z_{q1}-\sigma_{\theta 3} z_{q3})$ and $y_q d_q(\sigma_{\theta 2} z_{q2}-\sigma_{\theta 3} z_{q3})$. Since the difference of two normal distributed random variables $(\sigma_{\theta 1} z_{q1}-\sigma_{\theta 3} z_{q3})$ is another normal distributed random variable, only one scale parameter for each the differences $(\sigma_{\theta 1} z_{q1}-\sigma_{\theta 3} z_{q3})$ and $(\sigma_{\theta 2} z_{q2}-\sigma_{\theta 3} z_{q3})$ can be

estimated. Alternatively, only two of the three scale parameters – $\sigma_{\theta 1}, \sigma_{\theta 2}$, and $\sigma_{\theta 3}$ – can be estimated and by normalizing one scale parameter. In a general case with J choice alternatives, one can estimate up to $J-1$ scale parameters for the alternative-specific stochastic variables.

In summary, up to $J-1$ alternative-specific location parameters and up to $J-1$ alternative-specific scale parameters are estimable when a normal distributed alternative-specific attribute enters each of the J utility functions with some systematic variation across individuals. On the other hand, if the alternative-specific attribute (with systematic variation) enters only L of the J utility functions ($L < J$) then up to L alternative-specific location parameters and up to L alternative-specific scale parameters are estimable.

Stochastic alternative-specific attribute and its coefficient follow lognormal distribution

Consider the following utility structure, where $\exp(\mu_{\theta i} + \sigma_{\theta i} z_{qi})$ is the alternative-specific inverse speed, assumed to be lognormal distributed (i.e., $\theta_{qi} = \exp(\mu_{\theta i} + \sigma_{\theta i} z_{qi})$), and $\exp(y_q)$ is the lognormal distributed random coefficient on travel time, whose parameters are known from the SP data. All other terms in the utility structure are as defined earlier in this appendix.

$$\begin{aligned} U_{q1} &= \beta_{01} + \exp(y_q) d_q \exp(\mu_{\theta 1} + \sigma_{\theta 1} z_{q1}) + \varepsilon_{q1} \\ U_{q2} &= \beta_{02} + \exp(y_q) d_q \exp(\mu_{\theta 2} + \sigma_{\theta 2} z_{q2}) + \varepsilon_{q2} \\ U_{q3} &= \exp(y_q) d_q \exp(\mu_{\theta 3} + \sigma_{\theta 3} z_{q3}) + \varepsilon_{q3} \end{aligned} \tag{A9}$$

Even this case, as in the case of normal distributed attributes and coefficient, the systematic variation across individuals due to the observed variable d_q obviates the need for checking the rank of the covariance matrix of utility differences. Therefore, one can follow the basic principle that only differences in utilities matter to assess how many parameters of the alternative-specific stochastic variables can be estimated. To do so, let us setup the utility differences with respect to the third alternative as below:

$$\begin{aligned} U_{q1} - U_{q3} &= \beta_{01} + \exp(y_q) d_q \left\{ \exp(\mu_{\theta 1} + \sigma_{\theta 1} z_{q1}) - \exp(\mu_{\theta 3} + \sigma_{\theta 3} z_{q3}) \right\} + (\varepsilon_{q1} - \varepsilon_{q3}) \\ U_{q2} - U_{q3} &= \beta_{02} + \exp(y_q) d_q \left\{ \exp(\mu_{\theta 2} + \sigma_{\theta 2} z_{q2}) - \exp(\mu_{\theta 3} + \sigma_{\theta 3} z_{q3}) \right\} + (\varepsilon_{q2} - \varepsilon_{q3}) \end{aligned} \tag{A10}$$

Note from the above set of utility differences that the parameters to be estimated ($\mu_{\theta i}$, $\sigma_{\theta i}$ $\forall i$) are entangled in the differences of lognormal distributions:

$\left\{\exp(\mu_{\theta i}+\sigma_{\theta i}z_{qi})-\exp(\mu_{\theta 3}+\sigma_{\theta 3}z_{q3})\right\}; i\in\{1,2\}$. In a general case with J alternatives there would be $J-1$ such differences. While the difference of two lognormal distributions is not a known distribution, one can estimate $J-1$ location parameters and $J-1$ scale parameters, given each of the above differences results in another (unknown) distribution that needs at least two parameters to describe it. However, since the difference of two log-normal distributions is an (unknown) distribution that might need more than two parameters to fully describe it, one might be able to estimate all J scale parameters in a few settings. Nevertheless, the analyst can safely assume that estimating $J-1$ location parameters and $J-1$ scale parameters would not lead to theoretical identification issues.

Multiple stochastic alternative-specific attributes enter a choice alternative's utility function

The above discussion – for both normal and lognormal distributional assumptions – was only for situations when each choice alternative's utility function has a single alternative-specific attribute that is stochastic, albeit the number of such stochastic variables across all choice alternatives can be as many as J. However, there may be situations where the RP utility function of each choice alternative has multiple stochastic alternative-specific attributes, as in the utility function below:

$$U_{qi}=\beta_{0i}+\sum_{k}y_{kq}d_{kq}\theta_{kqi}+\varepsilon_{qi} \tag{A11}$$

In this equation, the utility function of alternative i has k stochastic alternative-specific attributes ($d_{kq}\theta_{kqi}$), each with its random coefficient (y_{kq}) that is identifiable from SP data. And each of these attributes is associated with a separate observed variable (d_{kq}) that introduces systematic variation across individuals and a random variable θ_{kqi} that exhibits random variation across individuals. The parameters of $\theta_{kqi}\forall k$ ought to be estimated using RP data. In such situations, since each alternative-specific stochastic attribute is associated with a separate observed variable (d_{kq}) that introduces systematic variation across individuals, one can estimate up to $J-1$ alternative-specific location parameters and up to $J-1$ alternative-specific scale parameters of θ_{kqi} for each of the k stochastic attributes (if the attribute appears in all J utility functions).

In other situations, the systematic variation across individuals for the different alternative-specific stochastic attributes is only due to a single observed variable (say d_q), as given below:

$$U_{qi} = \beta_{0i} + \sum_{k} y_{kq} d_q \theta_{kqi} + \varepsilon_{qi} \quad \text{(A12)}$$

In such situations, one cannot estimate more than a single location parameter and a single scale parameter for the distributions of alternative-specific attributes of a given alternative, and only up to $J-1$ alternative-specific location parameters and up to $J-1$ alternative-specific scale parameters can be estimated for all stochastic attributes of all alternatives.

The above discussion holds for both the distributional assumptions discussed earlier – normal and lognormal – for stochastic alternative attributes and their coefficients.

Appendix B: Parameter identification when alternative-specific stochastic attribute does not exhibit systematic variation across individuals in the data

As in Appendix A, consider the following utility expressions for an RP alternative i for individual q, with mode-specific stochastic travel time TT_{qi}^{*} and its random coefficient $\beta_{TT,q}$:

$$U_{qi} = \beta_{0i} + \beta_{TT,q} TT_{qi}^{*} + \varepsilon_{qi} \quad \text{(B1)}$$

However, unlike in Appendix A, the stochastic travel time variable (TT_{qi}^{*}) is not associated with any systematic variation across individuals. It exhibits only random variation across individuals. All other terms in the above equation are as defined for Equation (A1) of Appendix A.

Denoting the random coefficient $\beta_{TT,q}$ as y_q, the utility expression in Equation (B1) can be rewritten as below:

$$U_{qi} = \beta_{0i} + y_q TT_{qi}^{*} + \varepsilon_{qi} \quad \text{(B2)}$$

For this setting, we discuss parameter identifiability for two different sets of distributional assumptions – (1) normal distribution and (2) lognormal distribution – on TT_{qi}^{*} and y_q.

Stochastic alternative-specific attribute and its coefficient follow normal distribution

Consider that $TT_{qi}^{*} \sim N\left(\mu_i, \sigma_i^2\right)$ or $TT_{qi}^{*} = \mu_i + \sigma_i z_{qi}$, where z_{qi} is a standard normal variate. And let $y_q \sim N\left(\mu_y, \sigma_y^2\right)$ or $y_q = \mu_y + \sigma_y z_{qy}$, where z_{qy} is a standard normal variate. The parameters

$\{\mu_y, \sigma_y\}$ describing y_{qi} are estimable (therefore, known) using SP data, whereas the parameters $\{\mu_i, \sigma_i\}$ describing TT_{qi}^* need to be estimated using RP data. The resulting utility functions can be written as below:

$$U_{qi} = \beta_{0i} + (\mu_y + \sigma_y z_{qy})(\mu_i) + (\mu_y z_{qi} + \sigma_y z_{qy} z_{qi})(\sigma_i) + \varepsilon_{qi} \tag{B3}$$

In the above utility structure, the deterministic component is $\beta_{0i} + \mu_i \mu_y$. From such a deterministic component, it is not possible to identify μ_i because μ_y does not exhibit variation across individuals (unlike the situation in Appendix A where the observed attribute d_q which varies across individuals is part of the deterministic component). Therefore, μ_i, as with σ_i, should be considered as an unknown parameter in the variance-covariance matrix of utility differences.

The variance-covariance elements of the random components of the above utility functions are as follows (assuming that $TT_{qi}^* \forall i \in \mathbf{J}$, y_q, and $\varepsilon_{qi} \forall i \in \mathbf{J}$ are all independent of each other):

$$\begin{aligned} Var(U_{qi}) &= Var(y_q TT_{qi}^*) + Var(\varepsilon_{qi}) \\ &= s_{ii} + \rho^2 g, \text{where } s_{ii} = \mu_i^2 \sigma_y^2 + \sigma_i^2 \sigma_y^2 + \mu_y^2 \sigma_i^2. \end{aligned} \tag{B4}$$

$$\begin{aligned} Cov(U_{qi}, U_{qj}) &= E(y_q TT_{qi}^* \, y_q TT_{qj}^*) - E(y_q TT_{qi}^*) E(y_q TT_{qj}^*) \\ &= s_{ij} \text{ where } s_{ij} = \sigma_y^2 \mu_i \mu_j \end{aligned} \tag{B5}$$

Consider a setting with three choice alternatives (i.e., $J = 3$). In such a setting, there would be three normal distributed alternative-specific stochastic variables (TT_{qi}^*) – one in each utility function. The resulting covariance matrix of the utility expressions in Equation (B2) is:

$$\Omega = \begin{bmatrix} \left\{ \begin{matrix} \mu_1^2\sigma_y^2 + \sigma_1^2\sigma_y^2 \\ +\mu_y^2\sigma_1^2 + \rho^2 g \end{matrix} \right\} & \sigma_y^2\mu_1\mu_2 & \sigma_y^2\mu_1\mu_3 \\ \sigma_y^2\mu_2\mu_1 & \left\{ \begin{matrix} \mu_2^2\sigma_y^2 + \sigma_2^2\sigma_y^2 \\ +\mu_y^2\sigma_2^2 + \rho^2 g \end{matrix} \right\} & \sigma_y^2\mu_1\mu_3 \\ \sigma_y^2\mu_3\mu_1 & \sigma_y^2\mu_3\mu_2 & \left\{ \begin{matrix} \mu_3^2\sigma_y^2 + \sigma_3^2\sigma_y^2 \\ +\mu_y^2\sigma_3^2 + \rho^2 g \end{matrix} \right\} \end{bmatrix} \tag{B6}$$

Now, define a linear operator M_j that transforms the J utilities into $(J-1)$ utility differences taken with respect to the j^{th} alternative. M_j is an identity matrix of size $(J-1)\times(J-1)$ where the j^{th} column is substituted by a column of -1s. Let us consider the first alternative as the base alternative. The corresponding linear operator matrix can be defined as below:

$$M_1 = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \tag{B7}$$

The covariance matrix $\Omega_{\Delta 1}$ of utility differences (with respect to the first alternative) is given by $M_1 \Omega M_1'$. Such a covariance matrix of utility differences for the utility form in Equation (B2) is:

$$\Omega_{\Delta 1} = \begin{bmatrix} s_{11} + s_{22} - 2s_{12} + 2\rho^2 g & s_{11} - s_{12} - s_{13} + s_{23} + \rho^2 g \\ s_{11} - s_{12} - s_{13} + s_{23} + \rho^2 g & s_{11} + s_{33} - 2s_{13} + 2\rho^2 g \end{bmatrix} \tag{B8}$$

To determine the maximum number of unique parameters that result in the above covariance matrix of utility differences, one should find the rank of the Jacobian of the vector of unique elements in above matrix with respect to the parameters in the matrix. This is called the *rank condition* (Bunch, 1991; Walker 2001). The vector of unique elements of $\Omega_{\Delta 1}$ is given below:

$$\begin{aligned} vecu(\Omega_{\Delta 1}) &= \begin{bmatrix} s_{11} + s_{22} - 2s_{12} + 2\rho^2 g \\ s_{11} + s_{33} - 2s_{13} + 2\rho^2 g \\ s_{11} - s_{12} - s_{13} + s_{23} + \rho^2 g \end{bmatrix} \\ &= \begin{bmatrix} \mu_1^2\sigma_y^2 + \sigma_1^2\sigma_y^2 + \mu_y^2\sigma_1^2 + \mu_2^2\sigma_y^2 + \sigma_2^2\sigma_y^2 + \mu_y^2\sigma_2^2 - 2\sigma_y^2\mu_1\mu_2 + 2\rho^2 g \\ \mu_1^2\sigma_y^2 + \sigma_1^2\sigma_y^2 + \mu_y^2\sigma_1^2 + \mu_3^2\sigma_y^2 + \sigma_3^2\sigma_y^2 + \mu_y^2\sigma_3^2 - 2\sigma_y^2\mu_1\mu_3 + 2\rho^2 g \\ \mu_1^2\sigma_y^2 + \sigma_1^2\sigma_y^2 + \mu_y^2\sigma_1^2 - \sigma_y^2\mu_1\mu_2 - \sigma_y^2\mu_1\mu_3 + \sigma_y^2\mu_2\mu_3 + \rho^2 g \end{bmatrix} \end{aligned} \tag{B9}$$

The elements in the above vector (and those in the covariance matrix $\Omega_{\Delta 1}$ of utility differences), comprise seven unknown parameters – three μ_i, three σ_i^2, and one ρ^2. The Jacobian of the above vector with respect to these seven parameters is given below:

$$Jacobian\left(vecu(\Omega_{\Delta 1})\right) = \begin{bmatrix} \begin{Bmatrix} 2\mu_1\sigma_y^2 \\ -2\sigma_y^2\mu_2 \end{Bmatrix} & \begin{Bmatrix} 2\mu_2\sigma_y^2 \\ -2\sigma_y^2\mu_1 \end{Bmatrix} & 0 & \sigma_y^2+\mu_y^2 & \sigma_y^2+\mu_y^2 & 0 & 2g \\ \begin{Bmatrix} 2\mu_1\sigma_y^2 \\ -2\sigma_y^2\mu_3 \end{Bmatrix} & 0 & \begin{Bmatrix} 2\mu_3\sigma_y^2 \\ -2\sigma_y^2\mu_1 \end{Bmatrix} & \sigma_y^2+\mu_y^2 & 0 & \sigma_y^2+\mu_y^2 & 2g \\ \begin{Bmatrix} 2\mu_1\sigma_y^2 \\ -\sigma_y^2\mu_2 \\ -\sigma_y^2\mu_3 \end{Bmatrix} & \begin{Bmatrix} -\sigma_y^2\mu_1 \\ +\sigma_y^2\mu_3 \end{Bmatrix} & \begin{Bmatrix} -\sigma_y^2\mu_1 \\ +\sigma_y^2\mu_2 \end{Bmatrix} & \sigma_y^2+\mu_y^2 & 0 & 0 & g \end{bmatrix} \quad \text{(B10)}$$

The rank of the above Jacobian matrix is: $Rank\left[Jacobian\left(vecu(\Omega_{\Delta 1})\right)\right]=3$, indicating that a total of three unknown parameters can be estimated. Since the scale (ρ^2) of the overall model must be normalized, a total of two unknown parameters can be estimated out of the remaining six μ_i and σ_i^2 parameters describing the alternative-specific stochastic variables. This exercise can be repeated for situations with different choice set sizes (J). Our explorations suggest that one can derive the following generic expression for only situations when $J \le 5$, for which the number of unknown parameters in the Jacobian is greater than the number unique elements in $vecu(\Omega_{\Delta 1})$ (i.e., the number of columns of the Jacobian is greater than its number of rows):

$$Rank\left[Jacobian\left(vecu(\Omega_{\Delta 1})\right)\right]=\frac{J(J-1)}{2} \;\; \forall J \le 5, J > 1$$

Therefore, after accounting for one normalization for the overall scale of the model, the number of estimable parameters of normal distributed alternative-specific stochastic variables is $\frac{J(J-1)}{2}-1$ when $J \le 5$. Note that, when $J=2$, the rank of the Jacobian matrix is 1, and thus,

no parameters describing alternative-specific stochastic variables can be identified in a binary choice setting.

For situations when $J > 5$, the number of unknown parameters in the Jacobian becomes less than the number of unique elements in $vecu(\Omega_{\Delta 1})$ (i.e., the number of rows in the Jacobian is greater than the number of columns). The structure of such Jacobian matrix is such that it is difficult to derive a generic expression for its rank. One must derive its rank on a case-to-case basis separately for each value of $J > 5$.

Note that the above discourse is for a setting where the utility functions of all choice alternatives involve a stochastic alternative-specific attribute (TT^*_{qi}) with a random coefficient. However, in some situations, only a subset of the utility functions might include the stochastic attribute while the other utility functions include only a deterministic component of the corresponding attribute (with the same random coefficient across all alternatives). For example, in a mode choice model mode-specific access times might be stochastic for only transit and shared mobility alternatives and deterministic for personal vehicle alternatives. To consider such situations, let L ($L < J$) be the number of utility functions in which the alterative-specific attribute enters in a stochastic form along with the random coefficient. In such situations, for values of $J \leq 5$, the rank of the Jacobian matrix can be derived as $\frac{J(J-1)}{2}$ as long as the alternative-specific attributes enter the other $J - L$ utility functions in a deterministic manner (with the same random coefficient $\beta_{TT,q}$). For $J > 5$, the rank of the Jacobian matrix has to be derived on a case-to-case basis.

Next, consider a situation where $L < J$ number of utility functions have the alterative-specific attribute in a stochastic form along with the random coefficient and the remaining $J - L$ utility functions do not include the corresponding attribute even in the deterministic form. For example, the waiting time variable is not relevant (as the waiting time is zero) for the walk mode and personal vehicle modes. Similarly, transfer time variable may be zero for non-transit modes, but a stochastic variable for transit modes. In such situations, the rank of the Jacobian matrix for determining the number of estimable parameters of the L stochastic alternative attributes can be derived as $\frac{L(L+1)}{2}$ for any J.

Stochastic alternative-specific attribute and its coefficient follow lognormal distribution

Consider that $TT_{qi}^{*} \sim LN\left(\mu_i, \sigma_i^2\right)$ or $TT_{qi}^{*} = \exp\left(\mu_i + \sigma_i z_{qi}\right)$, where z_{qi} is a standard normal variate, and $y_q \sim LN\left(\mu_y, \sigma_y^2\right)$. The parameters $\left\{\mu_y, \sigma_y\right\}$ describing y_{qi} are estimable (therefore, known) using SP data and the parameters $\left\{\mu_i, \sigma_i\right\}$ describing TT_{qi}^{*} need to be estimated using RP data. For these distributional assumptions, the utility functions can be written as:

$$\begin{aligned} U_{qi} &= \beta_{0i} + y_q TT_{qi}^{*} + \varepsilon_{qi} \\ &= \beta_{0i} + \exp(\mu_y + \sigma_y z_{qy}) \exp(\mu_i + \sigma_i z_{qi}) + \varepsilon_{qi} \end{aligned} \tag{B11}$$

For the resulting utility functions, μ_i cannot be separated out from the random component of the utility. Therefore, μ_i, as with σ_i, should be considered as an unknown parameter in the variance-covariance matrix of utility differences.

For the above utility structure, the variance-covariance elements of the random components are as follows (assuming that $TT_{qi}^{*}\ \forall i \in \mathbf{J}$, y_q, and $\varepsilon_{qi} \forall i \in \mathbf{J}$ are all independent of each other):

$$\begin{aligned} Var(U_{qi}) &= e^{2\mu_i + \sigma_i^2}(e^{\sigma_i^2} - 1) e^{2\mu_y + \sigma_y^2}(e^{\sigma_y^2} - 1) + \rho^2 g \\ &= s_{ii} + \rho^2 g, \text{ where } s_{ii} = c e^{2\mu_i + \sigma_i^2}(e^{\sigma_i^2} - 1) \text{ and } c = e^{2\mu_y + \sigma_y^2}(e^{\sigma_y^2} - 1). \end{aligned} \tag{B12}$$

$$\begin{aligned} Cov(U_{qi}, U_{qj}) &= E(y_q TT_{qi}^{*} y_q TT_{qj}^{*}) - E(y_q TT_{qi}^{*}) E(y_q TT_{qj}^{*}) \\ &= E(TT_{qi}^{*}) E(TT_{qj}^{*}) [E(y_q^2) - E^2(y_q)] \\ &= s_{ij}, \text{where } s_{ij} = c e^{\mu_i + \frac{\sigma_i^2}{2}} e^{\mu_j + \frac{\sigma_j^2}{2}} \text{ and } c = e^{2\mu_y + \sigma_y^2}(e^{\sigma_y^2} - 1). \end{aligned} \tag{B13}$$

The structure of the covariance matrix $\Omega_{\Delta 1}$ of utility differences is similar to the one written in the case with the normal distribution assumption for the stochastic variable and its coefficient. Specifically, the vector of unique elements in the covariance matrix of utility differences for a three-alternative choice set is:

$$vecu(\Omega_{\Delta 1}) = \begin{bmatrix} s_{11} + s_{22} - 2s_{12} + 2\rho^2 g \\ s_{11} + s_{33} - 2s_{13} + 2\rho^2 g \\ s_{11} - s_{12} - s_{13} + s_{23} + \rho^2 g \end{bmatrix} \tag{B14}$$

except that the terms s_{ii} and s_{ij} in this matrix are defined differently from those in the matrix of Equation (B9) for normal distributional assumption. In this case, we have seven unknown parameters – three μ_i, three σ_i^2, and one ρ^2. To make the notation compact, let $m_i = e^{2\mu_i}\ \forall i \in \mathbf{J}$ and let $\text{var}_i = e^{\sigma_i^2}(e^{\sigma_i^2} - 1) \forall i \in \mathbf{J}$. Substituting these for the elements in $vecu(\Omega_{\Delta 1})$, we have:

$$vecu(\Omega_{\Delta 1}) = \begin{bmatrix} cm_1^2\ \text{var}_1^2(\text{var}_1^2 - 1) + cm_2^2\ \text{var}_2^2(\text{var}_2^2 - 1) - 2cm_1m_2\ \text{var}_1\ \text{var}_2 + 2\rho^2 g \\ cm_1^2\ \text{var}_1^2(\text{var}_1^2 - 1) + cm_3^2\ \text{var}_3^2(\text{var}_3^2 - 1) - 2cm_1m_3\ \text{var}_1\ \text{var}_3 + 2\rho^2 g \\ cm_1^2\ \text{var}_1^2(\text{var}_1^2 - 1) - cm_1m_2\ \text{var}_1\ \text{var}_2 - cm_1m_3\ \text{var}_1\ \text{var}_3 + cm_2m_3\ \text{var}_2\ \text{var}_3 + \rho^2 g \end{bmatrix} \tag{B15}$$

Further, let $s_i = \text{var}_i^2(\text{var}_i^2 - 1) \forall i \in \mathbf{J}$. The Jacobian of the above vector with respect to the seven unknown parameters is given below:

$$Jacobian(vecu(\Omega_{\Delta 1})) = \begin{bmatrix} \begin{pmatrix} 2cm_1s_1 \\ -2cm_2\sigma_1\sigma_2 \end{pmatrix} & \begin{pmatrix} 2cm_2s_2 \\ -2cm_1\sigma_1\sigma_2 \end{pmatrix} & 0 & \begin{pmatrix} 2cm_1^2\sigma_1(\sigma_1^2 - 1) \\ +2cm_1^2\sigma_1^3 \\ -2cm_1m_2\sigma_2 \end{pmatrix} & \begin{pmatrix} 2cm_2^2\sigma_2(\sigma_2^2 - 1) \\ +2cm_2^2\sigma_2^3 \\ -2cm_1m_2\sigma_1 \end{pmatrix} & 0 & 2g \\ \begin{pmatrix} 2cm_1s_1 \\ -2cm_3\sigma_1\sigma_2 \end{pmatrix} & 0 & \begin{pmatrix} 2cm_3s_3 \\ -2cm_1\sigma_1\sigma_3 \end{pmatrix} & \begin{pmatrix} 2cm_1^2\sigma_1(\sigma_1^2 - 1) \\ +2cm_1^2\sigma_1^3 \\ -2cm_1m_3\sigma_3 \end{pmatrix} & 0 & \begin{pmatrix} 2cm_3^2\sigma_3(\sigma_3^2 - 1) \\ +2cm_3^2\sigma_3^3 \\ -2cm_1m_3\sigma_1 \end{pmatrix} & 2g \\ \begin{pmatrix} 2cm_1s_1 \\ -cm_2\sigma_1\sigma_2 \\ -cm_3\sigma_1\sigma_3 \end{pmatrix} & \begin{pmatrix} -cm_1\sigma_1\sigma_2 \\ +cm_3\sigma_2\sigma_3 \end{pmatrix} & \begin{pmatrix} -cm_1\sigma_1\sigma_2 \\ +cm_2\sigma_2\sigma_3 \end{pmatrix} & \begin{pmatrix} 2cm_1^2\sigma_1(\sigma_1^2 - 1) \\ +2cm_1^2\sigma_1^3 \\ -cm_1m_2\sigma_2 \\ -cm_1m_3\sigma_3 \end{pmatrix} & \begin{pmatrix} -cm_1m_2\sigma_1 \\ +cm_2m_3\sigma_3 \end{pmatrix} & \begin{pmatrix} -cm_1m_3\sigma_1 \\ +cm_2m_3\sigma_2 \end{pmatrix} & g \end{bmatrix} \tag{B16}$$

The rank of the above Jacobian is: $Rank\left[Jacobian\left(vecu(\Omega_{\Delta 1})\right)\right] = 3.$ Thus, after normalizing the scale of the overall model, a total of two unknown parameters can be estimated out of the remaining six μ_i and σ_i^2 parameters describing the alternative-specific stochastic variables.

Repeating the above exercise for situations with different choice set sizes (J), as in case of the normal distributional assumption, one can derive the following expression for the rank of the Jacobian matrix for only situations when $J \le 5$: $Rank\left[Jacobian\left(vecu(\Omega_{\Delta 1})\right)\right] = \frac{J(J-1)}{2}$. For

situations when $J > 5$, the rank of the Jacobian matrix has to be derived on a case-to-case basis, for reasons discussed earlier in the discussion for the normal distributional assumption.

Now, let us consider situations when only L ($L < J$) of the utility functions includes alternative-specific stochastic attributes along with the random coefficient. In such situations, regardless of whether the remaining $J - L$ utility functions include the corresponding attribute (in a deterministic form), one can derive the following generic expression for the rank of the Jacobian matrix: $Rank\left[Jacobian\left(vecu(\Omega_{\Delta 1})\right)\right] = \frac{L(L+1)}{2}$. That is, after normalizing the scale of the model, one can estimate $\frac{L(L+1)}{2} - 1$ number of parameters describing the L alternative-specific stochastic attributes.

Multiple stochastic alternative-specific attributes enter a choice alternative's utility function

The above discussion – for both normal and lognormal distributional assumptions – was only for situations when each choice alternative's utility function has a single alternative-specific stochastic attribute that does not exhibit stochastic variation, albeit the number of such stochastic variables across all choice alternatives can be as many as J. However, there may be situations where the RP utility function of each choice alternative has multiple stochastic alternative-specific attributes without any systematic variation across individuals. Such a model would be unidentified, because the variance-covariance matrix of utility differences does not allow the estimation of parameters for multiple stochastic attributes in each alternative.

Appendix C: Additional Simulation Experiments Considering Lognormal Distributed Alternative-specific Attributes and Random Coefficients

Design-III: Lognormal distributed stochastic alternative-specific attribute in J alternatives with lognormal distributed random coefficient

Design-III involves four modes in the choice set – bus, car, TW, metro – with stochastic travel times for all four modes and a generic random coefficient on travel time (i.e., $J = 4$ and $L = 4$). However, unlike in Design-I or Design-II, data on mode-specific inverse speeds (θ_{qi}) are generated using lognormal distributions for travel times. Also, the random coefficient on travel time ($\beta_{TT,q}$)

is generated from a lognormal distribution (negative of the lognormal distribution draws are used for the coefficient on travel time). The parameters of the normal distributions used to generate these lognormal distributions are the same as the parameters of the normal distributions used in Design I. Another difference between this design and Design-I is that the location parameter of the metro mode inverse speed is assumed to be known to the analyst (that is the analyst does not have to estimate this parameter). This assumption is necessary to meet the basic identification requirement that only utility differences matter. All other assumptions and parameters in this design are the same as those in Design-I in Section 4.

Evaluation of models estimated on simulated data from Design-III

We estimated the following two models on all simulated datasets from Design-III: Model-IV and Model-V. Model-IV is a mixed logit model with the same structure as the true DGP used to simulate data from Design-III. That is, $L = J$ alternatives involved lognormal distributed travel time (lognormal distributed inverse speeds, to be precise) and a lognormal distributed coefficient on travel time (negative of the lognormal distribution draws are used for the coefficient on travel time). The location parameter of inverse speed for any one of the modes was assumed to be known to the analyst (since only differences in utility matter) while the scale parameters of all these distributions were estimated. Specifically, we assumed that the location parameter of metro inverse speed was known to the analyst. Model-V simplifies Model-IV by ignoring the variability in travel times (or inverse speeds) of all the modes (i.e., scale parameters of the parent normal distributions of the inverse speeds were assumed to be zero and only the location parameters were estimated).

Table C1 reports a summary of these estimation results. As can be observed from the results for Model-IV, we are able to recover the assumed true parameters accurately and precisely. This includes the location parameters for inverse speed distributions of $J-1$ alternatives, scale parameters of inverse speed distributions of all J alternatives, and the parameters of the corresponding random coefficient. As can be noted, we could estimate the scale parameters of inverse speed distributions of all J alternatives without running into parameter (un)identifiability issues. This is unlike in the case of normal distributed inverse speeds for which we could estimate only up to $J-1$ alternatives' scale parameters. The likely reason behind the identification of all J scale parameters is that the difference of two lognormal random variables does not yield a distribution with a known analytic form. Since the resulting distribution might need more than two

parameters to describe it fully (unlike in the case of the difference of two normal distributions), it may be possible to estimate one more than $J-1$ alternative-specific scale parameters. Additionally, all J scale parameters corresponding to the alternative-specific random effects are identified due to the panel nature of the data. However, this result might not hold in general, that is, for all possible sets of true parameter values. Therefore, as discussed in Section 3.1.2, it is safer that at least one of the alternative-specific inverse speed scale parameters is known *a priori* to the analyst (or is assumed to be zero) and estimate the scale parameters for up to $J-1$ alternatives to be more certain of an identified model.

Next, the results for Model-V in Table C1 indicate that the location and scale parameters of the random coefficient on travel time display a statistically significant bias toward zero when compared to the corresponding estimates from Model-IV. This finding concurs with the discussion and findings in Biswas et al. (2024) on bias in parameter estimates when stochasticity in alternative attributes with multiplicative stochasticity (e.g., lognormal distributions) is ignored. They discuss that in such cases, either one of the location and scale parameters or both of these may be biased toward zero. It is also possible that one of the two parameter estimates gets biased toward zero while the other gets biased away from zero. In the current simulation experiments, we observe that both the location and scale parameters of the coefficient on travel time are biased toward zero. In addition, the coefficients for the alternative-specific constants, travel cost, and the SP scale parameter also display a similar bias. The location parameters of the inverse speeds of car and two-wheeler (which are also the degenerate values for the corresponding inverse speeds, given their scale parameters are set to be zero) display a bias away from zero while that for bus displays a bias toward zero. Finally, although not reported in the tables, in terms of model fit, Model-V was found to be inferior to Model-IV across all simulated datasets. These results, and the high bias in parameter estimates of Model-V, highlight the importance of recognizing stochasticity in alternative attributes.

Table C1. Simulation results for mixed logit models on pooled RP-SP data (lognormal travel time and lognormal random coefficient on travel time)

Variable description	True value	Model-IV				Model-V				$H_0 : \hat{\beta}_{Model-IV} = \hat{\beta}_{Model-V}$
		Mean estimate	APB	ASE	FSSE	Mean estimate	APB	ASE	FSSE	
Location parameters for alternative-specific random effects (bus mode is base)										
Car	1.80	1.988	10.46	0.1874	0.0642	0.729	59.51	0.0788	0.3097	6.20
Two-wheeler	0.30	0.315	5.05	0.1176	0.0745	0.219	26.92	0.0634	0.3175	0.72
Metro	0.50	0.499	0.14	0.1442	0.1238	0.245	50.98	0.0569	0.3368	1.64
Scale parameters for alternative-specific random effects (normal distributed)										
Bus	1.00	0.906	9.36	0.1296	0.0664	0.196	80.39	0.0614	0.4197	4.95
Car	1.85	1.891	2.21	0.1548	0.1086	0.641	65.36	0.0405	0.0443	7.81
Two-wheeler	1.55	1.395	9.98	0.1078	0.0973	0.501	67.65	0.0442	0.0419	7.67
Metro	1.35	1.193	11.62	0.1233	0.1969	0.096	92.89	0.0416	0.3854	8.43
Parameters of mode-specific inverse speeds (for travel times) in the RP setting										
RP bus inverse speed – Location parameter	1.85	1.687	8.79	0.0508	0.1044	1.558	15.79	0.0040	0.0088	2.54
RP bus inverse speed – Scale parameter	0.40	0.363	9.20	0.0347	0.0442	0.000	NA	NA	NA	NA
RP car inverse speed – Location parameter	1.25	1.112	11.07	0.0925	0.1945	1.284	2.68	0.0045	0.0055	1.86
RP car inverse speed – Scale parameter	0.20	0.203	1.75	0.0502	0.0850	0.000	NA	NA	NA	NA
RP TW inverse speed – Location parameter	1.10	0.969	11.90	0.0502	0.0209	1.275	15.93	0.0059	0.0078	6.06
RP TW inverse speed – Scale parameter	0.30	0.278	7.25	NA	0.0522	0.000	NA	NA	NA	NA
RP metro inverse speed – Location parameter	1.50 (*Fixed*)	NA	NA	NA	NA	NA	NA	NA	NA	NA
RP metro inverse speed – Scale parameter	0.15	0.157	4.49	0.0457	0.0240	0.000	NA	NA	NA	NA
Coefficients on level of service variables										
Travel time – Location parameter	-1.00	-1.072	7.19	0.0759	0.0456	-0.327	67.27	0.0158	0.0086	9.60
Travel time – Scale parameter	0.15	0.146	2.93	0.0121	0.0160	0.018	87.72	0.0028	0.0375	10.28
Travel cost	-0.45	-0.416	7.64	0.0304	0.0168	-0.147	67.36	0.0069	0.0029	8.63
SP scale parameter (RP scale is assumed to be 1)	0.70	0.759	8.42	0.0611	0.0360	0.258	63.13	0.0158	0.0107	7.93
Average value across all parameter estimates		NA	7.19	0.0843	0.0762	NA	54.54	0.0316	0.1384	NA

Appendix D: Empirical data description

Here, we provide details on the RP data used in the empirical analysis of this study.

Socio-demographics

In addition to commute mode choice, the survey collected information on demographic characteristics such as age, gender, education level, and household income. Table D1 presents the relevant descriptive statistics. The sample has a considerably higher share of men (71.6%) than women (28.4%), which is reflective of the labor force participation trends in Bengaluru (Census of India, 2011). A similar trend has been documented for the entire country (National Statistical Survey Office, 2021). In terms of age distribution, there is a good representation of individuals from different age categories, with about 19.8% from the 19-25 years category, 36.9% from the 26-35 years category, 22.7% from 36-45 years, 20.6% from 45 – 60 years, and about 3% above 60 years. In terms of education level, around 36.5% of the sample reported an education level up to high school. This is not unexpected for a sample of employed individuals since approximately 50% of the employed population in India records an education of high school and lower (National Statistical Survey Office, 2012). In terms of income level, 53.2% of the sample belonged to income category of less than Rs. 40,000 per month, 19.3% belonged to the middle- and upper middle-income category (between Rs. 40,000 and Rs. 200,000 per month; see People Research on India's Consumer Economy, 2023). Interestingly, one fourth of the sample did not reveal income information. In terms of vehicle ownership, more than 82% of the individuals come from households with a two-wheeler while about 32% come from households that own a car. The average commute distance in the sample is 9.85 km, with the following sample shares in different commute distance bands: 4% from less than 2 km, 27% from 2-5 km, 33% from 5-10 km, 20% from 10-15 km, 9% from 15-20 km, 5% from 20-25 km, and 1% from 25-30 km. The average commute distance is representative of that travelled by a typical individual in urban areas in the Indian context for commute trips (Nayka and Sridhar, 2019).

Mode availability

The assumptions used to decide the mode availability for the individuals in the sample are discussed here. Walk was considered a feasible alternative only when the commute distance was less than 5 km. The highest walk distance encountered in this data was 3.9 km. However, walk commute of up to an hour is feasible in the Indian context (Tiwari *et al.*, 2016). Therefore, assuming that a distance of 5 km would require an hour to cover while walking at an average speed

of 5 km per hour, we considered the availability threshold for walk as 5 km. As a result, the walk mode was an available/feasible alternative for 257 individuals in our sample of 914 individuals. Next, public transit modes (i.e., bus and metro) were considered available (as the primary mode of travel), if the total first and last mile access distance between an individual's home and workplace was less than 5 km. Additionally, if the *OVTT* of metro mode was found to be unreasonably high (for example, more than an hour), while the *IVTT* for the same was found to be low, such trip options were considered unavailable for travellers. Based on these considerations, bus and metro were considered available for 781 and 239 individuals in the sample, respectively. Personal modes (cars and two-wheelers) were considered feasible for an individual if the individual's household owned these modes (based on the response of the individual in the RP section of the survey). Auto-rickshaw was considered available for all the individuals in the sample, owing to the easy availability of the mode in Bengaluru.

Table D1 Descriptive statistics of estimation Sample (N= 914)

Mode shares in the sample	
Auto- rickshaw	3.3%
Metro	14.7%
Bus	22.4%
Walk	2.2%
Two-wheeler	51.0%
Car	6.4%
Exogenous variables	
Individual specific attributes	
Gender	
Male	71.6%
Female	28.4%
Age	
Age 19 - 25 years	19.8%
Age 26 - 35 years	36.9%
Age 36 - 45 years	22.7%
Age 45 - 60 years	20.6%
Age greater than 60 years	3.0%
Education	
Less than 12th grade	36.5%
Diploma	13.6%
Undergraduate degree	35.5%
Graduate and above	14.4%
Employment status	
Employed in government sector	8.1%
Employed in private sector	54.5%
Self-employed/Business	37.4%
Household characteristics	
Monthly income	
Less than ₹40,000	53.2%
Between ₹40,000 and ₹100,000	17.4%
Between ₹100,000 and ₹200,000	1.9%

More than ₹200,000	2.4%
Don't know	25.1%
Household two-wheeler count	
Zero two-wheeler	17.4%
One two-wheeler	51.5%
Two or more two-wheelers	31.1%
Household car ownership	
Zero car	68.4%
One car	27.0%
Two or more cars	4.6%

Data on level-of-service (LoS) variables

Secondary data sources and reasonable assumptions were utilized to compute the level-of-service (LoS) attributes – *IVTT*, *OVTT*, and travel costs for the different modes considered in the analysis. Specifically, for a given trip in the data, for each of the motorized modes of travel other than metro mode, relevant Google APIs (Application Programming Interfaces) were used to extract the *IVTT* and *OVTT* values between the travel origin and destination locations. Metro mode travel time was calculated assuming a speed of 36 kmph. Walk mode travel time was calculated based on an assumed walking speed of 5 kmph.

Travel costs for the transit modes were computed from the fare charts published by the respective transit agencies for bus and metro models based on the distance of travel between the origin and destination transit stops. For the auto-rikshaw mode, the current government fare structure was used, albeit with an inflation factor (the inflation factor was computed based on the ratio of the traveler-reported costs and government stipulated fares for trips that involved the use of autorickshaw mode). Travel costs for the personal car and two-wheeler modes were computed based on the trip distance, prevalent fuel price, and assumed mileages for cars and two-wheelers (the average mileage of a hatchback car was assumed to be 15 km per litre in Bengaluru and that of a two-wheeler was assumed to be 40 km per litre and the prevalent fuel price was 90 Rupees per litre).

The average travel time (and the sample standard deviation for time) and the average travel cost (and the sample standard deviation for cost) reported for each of the modes are computed by taking the average (and the sample standard deviation) across all the individuals for whom the respective modes are considered available. The average travel times for public transit modes are the highest – 42.54 minutes for the bus mode and 42 minutes for the metro mode, while that of car and two-wheeler (private modes) are 24 minutes and 20 minutes respectively. Next, the average travel cost for the bus mode is the least among the motorized

modes of transportation, followed by two-wheelers and metro. As expected, the average travel cost by private cars is at a higher end with a value of INR 65. The auto-rickshaw mode is reported to have the highest average travel cost with a value of just over INR 125.

Table D2 Details of travel related attributes of modes

Travel attributes	Auto-rickshaw	Metro	Bus	Walk	Two-wheeler	Private car
Travel time						
Mean	26.9	42.0	42.5	36.5	20.4	24.1
Standard deviation	20.7	22.5	21.9	11.5	15.7	18.6
Travel cost						
Mean	125.8	30.0	18.5	NA	21.4	64.9
Standard deviation	87.9	11.2	7.0	NA	15.2	46.9

REFERENCES

Athira, I. C., Muneera, C. P., Krishnamurthy, K., & Anjaneyulu, M. V. L. R. (2016). Estimation of value of travel time for work trips. *Transportation Research Procedia*, 17, 116-123.

Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., & Rao, V. (1994). Combining Revealed and Stated Preferences Data. *Marketing Letters*, 5(4), 335–349.

Bhat, C. R. (2000). Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. *Transportation Science*. 34(2), 228-238.

Bhat, C. R., & Castelar, S. (2002). A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco Bay area. *Transportation Research Part B: Methodological*. 36(7), 593-616.

Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37(9), 837-855.

Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7), 923-939.

Bhat, C. R., & Sidharthan, R. (2012). A new approach to specify and estimate non-normally mixed multinomial probit models. *Transportation Research Part B: Methodological*, 46(7), 817-833.

Bhat, C. R., & Dubey, S. K. (2014). A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B: Methodological*, 67, 68-85.

Bhat, C. R., & Lavieri, P. S. (2018). A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions. *Theory and Decision*, 84(2), 239-275.

Bhatta, B. P., & Larsen, O. I. (2011). Errors in variables in multinomial choice modeling: A simulation study applied to a multinomial logit model of travel mode choice. *Transport policy*, 18(2), 326-335.

Biswas, M., Bhat, C. R., Ghosh, S., & Pinjari, A. R. (2024). Choice models with stochastic variables and random coefficients. *Journal of Choice Modelling*, 51, 100488.

Bradley, M. A., & Daly, A. J. (1991). "Estimation of Logit Choice Models Using Mixed Stated Preference and Revealed Preference Information." Paper presented to the6th International Conference on Travel Behavior, Quebec, May 22–24, 1991.

Brey, R., & Walker, J. L. (2011). Estimating time of day demand with errors in reported preferred times: an application to airline travel. *Procedia-Social and Behavioral Sciences*, 17, 150-168.

Brownstone, D., Bunch, D. S., & Train, K., (2000). Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transportation Research Part B: Methodological*, 34(5), 315-338.

Bunch, D. S. (1991). Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, 25(1), 1-12.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*. 71(1), 19-25.

Census of India (2011). https://censusindia.gov.in/census.website/

Cherchi, E., Ortúzar, J. de D., (2008). Predicting best with mixed logit models: understanding some confounding effects. In: Inweldi, P.O. (Ed.), *Transportation Research Trends*. Nova Science Publishers, Inc., New York, pp. 215–235.

ChoiceMetrics (2012). Ngene 1.1.1 User Manual & Reference Guide, Australia.

Daganzo, C. (1979). *Multinomial probit: the theory and its application to demand forecasting*. Elsevier.

Díaz, F., Cantillo, V., Arellana, J., & de Dios Ortúzar, J. (2015). Accounting for stochastic variables in discrete choice models. *Transportation Research Part B: Methodological*. 78, 222-237.

Gleser, L. J. (1981). Estimation in a multivariate" errors in variables" regression model: large sample results. *The Annals of Statistics*, 24-44.

Guevara, C. A., & Hess, S. (2019). A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. *Transportation Research Part B: Methodological*, 123, 224-239.

Hensher, D. A., & Bradley, M. (1993). Using stated response choice data to enrich revealed preference discrete choice models. *Marketing Letters*, 4(2), 139-151.'

Hensher, D., Louviere, J., & Swait, J. (1998). Combining sources of preference data. *Journal of Econometrics*, 89(1-2), 197-221.

Hensher, D. A., & Greene, W. H. (2003). The mixed logit model: the state of practice. *Transportation*, 30, 133-176.

Hensher, D. A. (2012). Accounting for scale heterogeneity within and between pooled data sources. *Transportation Research Part A: Policy and Practice*, 46(3), 480-486.

Hensher, D. A., Rose, J. M., & Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2), 235-245.

Helveston, J. P., Feit, E. M., & Michalek, J. J. (2018). Pooling stated and revealed preference data in the presence of RP endogeneity. *Transportation Research Part B: Methodological*, 109, 70-89.

Hess, S., & Polak, J. W. (2005). Mixed logit modelling of airport choice in multi-airport regions. *Journal of Air Transport Management*, 11(2), 59-68.

Infrastructure Development Corporation (Karnataka) Limited (2020). Comprehensive Mobility Plan for Bengaluru. Bangalore Metro Rail Corporation Limited, Directorate of Urban Land Transport, Urban Development Dept., GoK.

Jain, D., & Tiwari, G. (2019). Explaining travel behaviour with limited socio-economic data: Case study of Vishakhapatnam, India. *Travel Behaviour and Society*, 15, 44-53.

Keane, M. P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2), 193-200.

McFadden, D., & Train, K. (2003). Mixed MNL models for discrete response. Journal of Applied *Econometrics*, 15(5), 447-470.

Morikawa, T. (1994). Correcting state dependence and serial correlation in the RP/SP combined estimation method. *Transportation*, 21(2), 153-165.

National Statistical Survey Office. (2012). Employment Unemployment Survey. Ministry of Statistical Programme and Implementation, GoI.

National Statistical Survey Office. (2021). Periodic Labor Force Survey, Ministry of Statistical Programme and Implementation, GoI.

Nayka, S., & Sridhar, K. S. (2019). Determinants of intra urban mobility: A study of Bengaluru. Working Papers No. 437, Institute for Social and Economic Change, Bangalore.

Nirmale, S. K., & Pinjari, A. R. (2023). Discrete choice models with multiplicative stochasticity in choice environment variables: Application to accommodating perception errors in driver behaviour models. *Transportation Research Part B: Methodological*, 170, 169-193.

Ortúzar J de D., & Ivelic A. M. Effects of using more accurately measured level of service variables in the specification and stability of mode choice models. Proceeding 15th PTRC Summer Annual Meeting, 1987, P290, 117–130. PTRC, London.

People Research on India's Consumer Economy (2023). The Rise of India's Middle Class: A Force to Reckon With. https://www.price360.in/expertview/the-rise-of-indias-middle-class-a-force-to-reckon-with/

Saigal, T., Vaish, A.K. & Rao, N.V.M. (2021). Gender and class distinction in travel behavior: evidence from India. *Ecofeminism and Climate Change* , 2(1), 42-48.

Sanko, N., Hess, S., Dumont, J., & Daly, A. (2014). Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of Choice Modelling*. 12, 47-57.

Scarpa, R., Gilbride, T. J., Campbell, D., & Hensher, D. A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2), 151-174.

Shirgaokar, M. (2014). Employment centers and travel behavior: Exploring the work commute of Mumbai's rapidly motorizing middle class. *Journal of Transport Geography*, 41, 249-258.

Srinivasan, K.K., Prakash, A.A., & Seshadri, R. (2014). Finding most reliable paths on networks with correlated and shifted lognormal travel times. *Transportation Research Part B: Methodological*, 66, 110-128.

Steimetz, S. S., & Brownstone, D. (2005). Estimating commuters' "value of time" with noisy data: a multiple imputation approach. *Transportation Research Part B: Methodological*, *39*(10), 865-889.

Tiwari, G., Jain, D., & Rao, K. R. (2016). Impact of public transport and non-motorized transport infrastructure on travel mode shares, energy, emissions and safety: Case of Indian cities. *Transportation Research Part D: Transport and Environment*, 44, 277-291.

Train, K. (1978). A validation test of a disaggregate mode choice model. *Transportation Research*, 12(3), 167-174.

Train, K. (2000). Halton sequences for mixed logit. Working paper. Department of Economics, Institute for Business and Economic Research, UC Berkeley.

Train, K. (2001). A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit. *University of California, Berkeley*, 1-13.

Varotto, S. F., Glerum, A., Stathopoulos, A., Bierlaire, M., & Longo, G. (2017). Mitigating the impact of errors in travel time reporting on mode choice modelling. *Journal of Transport Geography*, 62, 236-246.

Vij, A., & Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192-217.

Walker, J. L. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables* (Doctoral dissertation, Massachusetts Institute of Technology).

Walker, J., Li, J., Srinivasan, S., & Bolduc, D. (2010). Travel demand models in the developing world: Correcting for measurement errors. *Transportation Letters*, 2(4), 231-243.