Sampling of Alternatives in MEV Models: Evidence From Migration Aspiration Models with Large Choice Sets

Evangelos Paschalidis [*] Andreas B. Vortisch [†]
Michel Beine [†] Michel Bierlaire [*]

December 9, 2025

Report TRANSP-OR 251204
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*]École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, {evangelos.paschalidis, michel.bierlaire}@epfl.ch

[†]University of Luxembourg, Department of Economics and Management, Luxembourg {andreas.vortisch@iab.de, michel.beine@uni.lu}

Abstract

Certain case studies require the estimation of discrete choice models with a very large number of alternatives and attributes, which in turn increases the computational cost. This challenge becomes even more pronounced in the family of Multivariate Extreme Value (MEV) models such as the cross-nested logit (CNL) model due to the additional complexity in the model specification. In this study, we investigate the use of sampling of alternatives to make the estimation of MEV models more tractable. Using migration aspiration data, we assess how different sampling protocols and sample sizes approximate the parameters of the full choice set model. Our results show that importance sampling strategies informed by observed choice frequencies outperform random sampling, providing more stable estimates even with small subsets of alternatives. These findings highlight that carefully designed sampling can substantially reduce estimation time without compromising accuracy, thereby making MEV models a practical tool for studying migration destination choice.

Keywords: Sampling of alternatives, stratified importance sampling, cross-nested logit model, migration

1 Introduction

Various choice modelling applications require the estimation of models with a large number of alternatives and attributes; residential location, workplace, route choice, or activity locations are some typical examples (Lee and Waddell, 2010; Ma et al., 2017; Lai and Bierlaire, 2015). For these types of case studies, the computational cost can be very high, resulting in long estimation times; hence, model estimation using the full choice set is impractical. McFadden (1978) demonstrated that it is possible to obtain consistent parameter estimates of a logit model using sampling of alternatives, i.e., a subset of the full choice set. Estimation under sampling involves the use of an adjusted version of the likelihood function with the addition of a correction term to the utility function. The initial implementation of McFadden (1978) has been further advanced to account for more complex discrete choice models. Two of the most notable contributions are Guevara and Ben-Akiva (2013a) and Guevara and Ben-Akiva (2013b), who extended the approach for logit mixtures and generalised multivariate extreme value (MEV) models, respectively. With respect to the latter, Bierlaire et al. (2008) initially introduced the sampling of alternatives framework for MEV models (which was later implemented by Guevara and Ben-Akiva (2013b)). Daly et al. (2014) suggested a simplification of the Guevara and Ben-Akiva (2013b) method for easier practical implementation in the case of the nested logit model. Finally, Lee and Waddell (2010) proposed an approach by correcting for the sample proportion within each nest.

Regarding implementation, the first step of the sampling of alternatives requires the design of the sampling protocol. At this stage, the researcher decides (a) which procedure to follow for the sampling of the alternatives (whether some alternatives should be given a higher probability to be sampled than others) and (b) the number of alternatives to sample from the full choice set. Sampling of alternatives inevitably leads to some bias in the parameter estimates (Daly et al., 2014); however, the design of an effective sampling protocol could limit such deviations even when fewer alternatives are sampled, which, in turn, can reduce the estimation time. In the literature, several approaches have been proposed with respect to the development of the sampling protocols. The simplest procedure is pure random sampling; each alternative has an equal probability of being included in the choice set. The main limitation of random sampling is that the sampled alternatives can be irrelevant to the chosen; therefore, the model assigns a higher probability to the chosen alternative, compared to the other sampled alternatives, diminishing the overall explanatory power and failing to capture meaningful trade–offs (Tsoleridis et al., 2022). The limited number of close substitutes for the chosen alternative, in small choice sets, means that a random sampling protocol generally requires larger choice sets to reach the same level of accuracy as other sampling procedures, such as importance sampling (Nerella and Bhat, 2004). Importance sampling is a commonly used approach to overcome the limitations of random sampling. The modeller can assign a higher probability to certain alternatives to be sampled. One particular implementation is the stratified importance sampling which involves the stratification of the alternatives, typically based on certain attributes. Once the strata are defined, the next step requires a decision on the number of alternatives to be sampled from each stratum. For example, drawing many alternatives from a stratum with a relatively small number of alternatives significantly increases their probability of being observed in the sampled choice set.

The concept of importance sampling is particularly relevant in the context of location choice models, where the number of potential alternatives may be prohibitive for full choice set model estimation. In these applications (for example, residential location choice), the sampling of alternatives is typically implemented with distance-based criteria. For instance, Farooq and Miller (2012) suggested sampling the majority of alternatives within a given radius from the origin. Leite Mariante et al. (2018) also constrained the choice sets based on detour factors, while studies such as Li et al. (2005) or Tsoleridis et al. (2022) defined strata based on some distance–based metrics to then implement stratified importance sampling. For MEV models, Guevara and Ben-Akiva (2013b) explicitly used stratified importance sampling to ensure that each of the nests in the model specification was adequately represented. Frejinger et al. (2009) suggested an importance sampling protocol for a logit model with a path size formulation where every link was drawn from a probability distribution given a weight associated with it. Lai and Bierlaire (2015) later extended the framework for the cross–nested logit (CNL) model, building on Guevara and Ben-Akiva (2013b).

Except for the typical case studies on location or destination choices, such as residential location or workplace, migration aspiration choices are another example for which sampling of alternatives could be considered. The topic of migration aspirations involves a large choice set (close to 200 sovereign countries) and a long list of attributes. The question of the number of people who migrate, their individual characteristics, and their destination have been extensively investigated in the literature using discrete choice models (Bekaert et al., 2021; Docquier et al., 2020; Bertoli and Ruyssen, 2018; Gubert and Senne, 2016; Lovo, 2014 to name a few), and in particular the logit model. The logit model specification assumes independence from irrelevant alternatives (IIA), which implies that cross-elasticities due to the change of an attribute in one destination are identical for all alternatives. However, individuals are expected to substitute their choice in favour of certain locations. This behaviour could be due to factors such as language, religion, visa restrictions, and others. Certain studies have implemented more elaborate model specifications to address the deviations from the IIA. The most popular approach to capture substitution patterns between countries is the nested logit model (Bertoli and Moraga, 2013; Ortega and Peri, 2013; Buggle et al., 2020; Monras, 2018; Langella and Manning, 2021). More recently, Beine et al. (2025), implemented a cross-nested logit (CNL) model to investigate the migration aspirations of Indian individuals while some similar specifications were also adopted later in other studies such as Beine et al. (2024) and Baud et al. (2024). However, the benefits of the CNL model in the context of migration aspirations come at the cost of a high computational cost and longer estimation time.

Drawing on the CNL model specifications used in previous research on migration aspirations, we examine a similar case study to assess how sampling of alternatives affects parameter estimates, model fit, and estimation time. Despite the extensive use of sampling of alternatives, especially in logit and logit mixture models (McConnell and Tseng, 1999; Nerella and Bhat, 2004; Lemp and Kockelman, 2012; Leite Mariante et al., 2018; Becker et al., 2025; Dekker et al., 2025, to name a few), MEV models have received less attention. Apart from the original work by Guevara and Ben-Akiva (2013b), only a few examples can be found, such as the study of Zhong et al. (2025) who implemented sampling of alternatives in the framework of a multilevel nested logit model for the joint decision of workplace, residential choice, and type of house or the work of Lai and Bierlaire (2015) who used sampling of alternatives in the context of route choice. However, to the best of our knowledge, no studies have examined in detail how different sampling protocols affect CNL models using empirical data. In this paper, we aim to address this gap in the literature. Our main research objectives are as follows:

- Assess the influence of sampling protocols on model accuracy: We expect that assigning higher sampling probabilities to alternatives more relevant to the observed choices improves model accuracy for a given sample size. In the context of migration aspirations, this can be implemented by oversampling destinations that are chosen more frequently in the data.

- Examine the effect of sample size on model accuracy and estimation time: For a given sampling procedure, we anticipate that larger sample sizes improve model accuracy, at the expense, however, of longer estimation times. It is therefore important to evaluate the extent of this trade–off and identify the sample size that offers balance between accuracy and computational cost.

- Investigate how the complex substitution patterns of the CNL model are preserved: The scale parameters of the nests may be affected, given that some nests may be over– or underrepresented.

The remainder of the paper is organised as follows. Section 2 describes the general methodology of the study. Section 3 outlines the experimental design. Sections 4–7 present the main analysis and results. The paper concludes with the lessons learned and recommendations (Section 8), and a brief discussion (Section 9).

2 Methodology

2.1 Formulation of the CNL model

Under the assumption that migration aspiration choices are based on utility maximisation, we can define the utility of a destination in the form of a linear function as:

$$U_{in} = V_{in} + \varepsilon_{in}, \tag{1}$$

where V_{in} represents the deterministic part of the utility related to an individual *n* for a potential location *i*, while ε is a disturbance term. In order to capture more complex patterns among the disturbance terms (i.e. correlations among the unobserved characteristics of destinations), we adopt the generating function G of the CNL model specification – which is a Multivariate Extreme Value model (MEV) that stems from the random utility approach. For an individual n, the probability of choosing i (among J) destinations is expressed as:

$$P_n(i|\mathcal{C}_n) = P_n(U_{in} \geq U_{jn} \forall j \in \mathcal{C}_n) = \frac{e^{V_{in} + \ln G_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn} + \ln G_{jn}}} \tag{2}$$

where G_{in} is a generating function that is specific to each member of the MEV family. For the CNL model, we define $\ln G_{in}$ as:

$$\ln G_{in} = \ln \left(\sum_{m=1}^{M} \left(\mu \alpha_{im} e^{V_{in}(\mu_m - 1)} \left(\sum_{j \in \mathcal{C}_n} \alpha_{jm} e^{\mu_m V_{jn}} \right)^{\frac{\mu - \mu_m}{\mu_m}} \right) \right) \tag{3}$$

where the choice set $\mathcal{C}_n$ is divided into M overlapping subsets of destinations ($m = 1, \ldots, M$), $\alpha_{jm} \geq 0$, $\frac{\mu}{\mu_m} \leq 1$ ($\mu > 0$ and $\mu_m > 0$) and $\forall j$, $\exists$m such that $\alpha_{jm} \geq 0$. In Eq. 3, the parameters μ_ms capture the similarity between the alternatives within nest m. The α_{jm} is a participation parameter that denotes the extent to which a destination *j* is part of the nest m. In the CNL model, μ_m and α_{jm} jointly capture the correlation between the destinations. For identification, one of the scale parameters must be normalised; it is common practice to set μ = 1.

2.2 Sampling of alternatives

The sampling of alternatives is a technique used to approximate the likelihood function without enumerating the entire choice set. For each individual n, we denote by $\mathcal{C}_n$ their full choice set, which contains J_n alternatives. We can assume, without loss of generality, the same full choice set $\mathcal{C}$ for all individuals. The choice probability for the alternative i is then

$$P_n(i \mid \mathcal{C}; \theta) \tag{4}$$

where θ is a vector of parameters to be estimated. The maximum likelihood estimation procedure for $\mathcal{C}$ amounts to solving the following optimization problem:

$$\max_{\theta} \sum_{n=1}^{N} \ln P_n(i \mid \mathcal{C}; \theta) \tag{5}$$

Sampling of alternatives assigns to the individual n a subset of alternatives denoted as $\mathcal{D}_n$ with $\tilde{J}_n$ alternatives ($\tilde{J}_n \leq J_n$). We also denote as $\pi_n(\mathcal{D}_n \mid i)$ the probability of generating the subset $\mathcal{D}_n$, given that the choice is i. Following the aforementioned approach, it is only possible to maximise the conditional likelihood function rather than the true likelihood

$$\max_{\theta} \sum_{n=1}^{N} \ln P_n(i \mid \mathcal{D}_n, \mathcal{C}; \theta). \tag{6}$$

The joint probability of constructing a subset $\mathcal{D}_n$ and the choice probability is derived as

$$\pi_n(i \mid \mathcal{D}_n; \theta) = \pi_n(\mathcal{D}_n \mid i)\; P_n(i \mid \mathcal{C}; \theta) \tag{7}$$

Following the Bayes' theorem, we get

$$\pi_n(i \mid \mathcal{D}_n; \theta) = \frac{\pi_n(\mathcal{D}_n \mid i)\; P_n(i \mid \mathcal{C}; \theta)}{\sum_{j \in \mathcal{D}_n} \pi_n(\mathcal{D}_n \mid j)\; P_n(j \mid \mathcal{C}; \theta)} \tag{8}$$

The logit model probability is defined as

$$P_n(i \mid \mathcal{C}; \theta) = \frac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}} e^{\mu V_{jn}}} \tag{9}$$

If we insert Eq. 9 into Eq. 8, the denominator of Eq. 9 cancels out and we obtain

$$\pi_n(i \mid \mathcal{D}_n; \theta) = \frac{e^{V_{in} + \ln \pi_n(\mathcal{D}_n \mid i)}}{\sum_{j \in \mathcal{D}_n} e^{V_{jn} + \ln \pi_n(\mathcal{D}_n \mid j)}} \tag{10}$$

The probability in Eq. 10 no longer depends on the full choice set $\mathcal{C}$. The logit model can now be estimated on the sampled alternatives by including the correction term $\ln \pi_n(\mathcal{D}_n|j)$ in the utility functions. The correction term can also be considered as a penalty added to the utility, since $\pi_n(\mathcal{D}_n|i)$ is constrained between 0 and 1 and hence its logarithm will always have a negative value (a lower probability of sampling a choice set results in a larger penalty in the utility function). The interested reader is referred to Ben-Akiva and Lerman (1985) for further details.

The derivation for MEV models is similar; by substituting Eq. 2 into Eq. 8, we obtain

$$\pi_n(i \mid \mathcal{D}_n; \theta) = \frac{e^{V_{in} + \ln G_{in} + \ln \pi_n(\mathcal{D}_n|i)}}{\sum_{j \in D_n} e^{V_j n + \ln G_{jn} + \ln \pi_n(\mathcal{D}_n|j)}} \tag{11}$$

However, Eq. 3 and Eq. 11 suggest that the choice probability still depends on the entire choice set, since G_{jn} is expressed as a function of the full choice set $\mathcal{C}_n$. Guevara and Ben-Akiva (2013a) proposed an approximation of the G_{jn} by including an additional weight in its specification as

$$\sum_{j \in C_m} \alpha_{jm}^{\frac{\mu_m}{\mu}} \exp(\mu_m V_j) \approx \sum_{j \in C_m \cap D_{G_n}} w_j \alpha_{jm}^{\frac{\mu_m}{\mu}} \exp(\mu_m V_j) \tag{12}$$

where:

$$w_j = \frac{1}{\pi(j)}$$

is the probability to sample the alternative j in the sample D_{G_n} and C_m is the nest m. Equations 11 and 12 indicate that the implementation of sampling of alternatives in MEV models requires a repetition of the sampling procedure; the subset D_n is used to approximate the denominator of the choice model, while the subset D_{G_n} is used to approximate each nest. There are two important differences when implementing the sampling of these two subsets. First, it is not necessary to partition the full choice set in D_{G_n}. Alternatives that are alone in a nest do not contribute to the calculation of the MEV terms, and can therefore be excluded. Second, the chosen alternative does not play any role in the sampling procedure of D_{G_n}, for example, it may not belong to any of the nests. Hence, unlike in the sampling procedure of D_n, it is not required to be sampled. The exact implementation of the sampling procedure is discussed in more detail in Section 2.3.

2.3 The sampling algorithm

Random sampling and importance sampling are two of the most common sampling approaches. In the latter, the modeller splits the alternatives into strata using a set of deterministic rules, typically derived from the characteristics of the problem at hand. The sample size within each stratum is then decided in order to give a higher probability for some alternatives to be sampled, over others. It should be noted that each alternative can be assigned to one stratum only. The process of sampling of alternatives can be summarised as follows (the interested reader is referred to Bierlaire and Paschalidis (2023) for more details):

1. Partition the full choice set into K strata of size $R_k : J = \sum_{k=1}^{K} R_k$.

2. Define a number r_k which represents the number of alternatives to be sampled from each stratum, $D_n : \sum_{k=1}^{K} r_k$.

3. Denote $k(i)$ the stratum containing the chosen alternative i.

4. Randomly draw $r_{k_i} - 1$ alternatives among the non chosen ones in stratum $k(i)$ and add i to obtain $D_n(i)$.

5. Randomly draw r_k alternatives in each stratum k, $k \neq k(i)$ to obtain D_n.

6. Compute the correction factor $\ln \pi(D|i) = \ln R_{k(i)} - \ln r_{k(i)}$ ($\frac{R_k(i)}{r_k(i)} \propto \pi(D_n|i)$).

This process is initially used to generate the sample and the correction terms used in the utility functions, as in Eq. 10. For the CNL (or any MEV model), the process is repeated to generate D_{G_n}. The steps of this second sampling procedure are similar; however, with some key differences as follows:

- It is not necessary to partition the full choice set; alternatives that are alone in a nest do not contribute to the generating function and can be excluded. In this case, the full choice set is the set of alternatives that belong to at least one nest with more than one alternative.

- The chosen alternative does not play any role in the sampling procedure. As a consequence, the modeller does not have to include the choice in the generated samples.

- Every nest must be represented. A common issue that may arise, especially in the case of pure random sampling at lower sample sizes, is that no alternatives are sampled for one or more nests. In such cases, it is not possible to estimate the scale parameter of these nests.

- The correction term is calculated as $w_{jn} = \frac{1}{\Pr(j)} = \frac{R_k(j)}{r_k(j)}$.

The above-mentioned process has been implemented in Biogeme (Bierlaire, 2003; Bierlaire and Paschalidis, 2023).

3 Data, experimental design and preliminary analysis

3.1 Data

In the current study, we used data from the Gallup World Poll survey to obtain migration aspiration choices. The data are representative of about 95% of the world's population over the age of 16 years, at the national level, and cover several topics such as religion, politics, economy, and others. Migration aspirations are collected with the question: "Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?" Destination preferences are only inquired conditional on an existing aspiration to emigrate with the question: "To which country would you like to move?" For our analysis, we considered data between the years 2009 and 2017. Moreover, the Gallup World Poll survey data were further augmented with destination–specific attributes closely following the work of Beine et al. (2025). The main additional destination attributes are measures for the diaspora in a country, the distance from the respondents' origin country, GDP per capita, and others.

We extracted data from five different origin countries, namely Austria, Denmark, Mexico, New Zealand, and Afghanistan. Our objective was to consider datasets that exhibit different patterns in the choice of destination. In particular, the Austria and Denmark samples did not show a very strong preference for particular destinations. By contrast, in the samples of Mexico and New Zealand, there was a clear preference for one specific destination, the United States and Australia, respectively. In these cases, aspirations may be shaped by factors such as historical migration ties and established migrant networks (Mexico–United States) or geographical isolation (New Zealand–Australia). Finally, no dominant options were observed in the Afghanistan sample; however, respondents preferred several countries with different characteristics not observed in the other samples, including several non–OECD destinations. The latter is also not part of any of the nests in the CNL model specification adopted in this study (see Section 3.3 for more details).

Table 1 outlines some basic characteristics per origin country with respect to the number of aspirant emigrants, the proportion they represent in the total sample, and the number of different countries individuals mentioned as potential destinations. In particular, the column related to the percentage of aspirants (aspirants %) indicates the proportion of individuals aspiring to migrate in each of the samples. For example, in the Austria sample, 7.76% of the individuals (equally 670 out of the total of 8632) aspired to move abroad. The column "Distinct destinations" presents the total number of different countries that were reported as a potential destination in each of the five samples.

Table 1: Sample size characteristics

Country	**Aspirants (N)**	**Aspirants (%)**	**Distinct destinations**
Austria	670	7.762	76
Denmark	916	10.79	74
Mexico	1626	17.74	64
New Zealand	695	9.944	42
Afghanistan	2212	25.61	40

Table 2 presents the number of times a country with some specific attributes was chosen. These attributes are OECD membership, Schengen membership, English as an official language, and their counterparts. These are the same attributes used for the nest specification of the CNL models in existing research (Beine et al., 2024; Baud et al., 2024). Overall, the data suggest a pronounced preference of OECD countries over non-OECD countries. Regarding the Mexico sample, there is also a preference towards non-Schengen countries. However, as we see in Table 3, this outcome arises due to the significant preference of this sample for choosing the United States as a potential destination.

Table 2: Sample choices

Country	**English**	**non-English**	**OECD**	**non-OECD**	**Schengen**	**non-Schengen**
Austria	284	386	566	104	307	363
Denmark	339	577	816	100	481	435
Mexico	969	657	1408	218	414	1212
New Zealand	577	118	646	49	80	615
Afghanistan	783	1429	1365	847	419	1793

Table 3 summarises the 10 most popular destination choices in each sample. We observed that specific destinations such as the United States, Spain, Germany, France, and the United Kingdom were among the most preferred destinations, regardless of the origin country. In addition, some origin–specific patterns occurred in the data. For example, the Mexico sample had a strong preference for choosing the United States (41.14%) followed by Canada (13.16%), while each of the remaining destinations was chosen by less than 10% of the sample. Similarly, most of the New Zealand sample chose Australia (54.96%) followed by the United Kingdom (10.36%). However, such cases where certain destinations dominated were not found across all samples. For example, in the Austria sample, although the United States are the most preferred destination, they are chosen by less than 12%. Moreover, in the Afghanistan sample, we observed the choice of some destinations (mainly Asian or Muslim–majority countries) that did not appear in any of the other samples. This finding further indicates the heterogeneity in the pattern of destination choice among different origin countries.

Table 3: Destination percentages across different samples

(a) Austria

Destination	%
United States	11.8%
Australia	10.6%
Spain	8.81%
Switzerland	7.01%
Germany	6.87%
Canada	5.22%
United Kingdom	4.18%
Italy	4.18%
France	3.88%
New Zealand	3.58%

(b) Denmark

Destination	%
United States	17.03%
Spain	11.1%
Sweden	9.28%
United Kingdom	7.53%
Germany	6.99%
Norway	6.44%
Italy	5.46%
France	4.69%
Australia	4.37%
Canada	3.06%

(c) Mexico

Destination	%
United States	41.1%
Canada	13.2%
Spain	7.01%
Germany	6.64%
France	4.86%
Italy	2.46%
Brazil	2.28%
China	1.85%
Japan	1.72%
United Kingdom	1.72%

(d) New Zealand

Destination	%
Australia	55.0%
United Kingdom	10.4%
United States	8.20%
Canada	5.76%
France	2.88%
Italy	2.01%
Spain	2.01%
Germany	1.15%
Netherlands	1.01%
Samoa	0.72%

(e) Afghanistan

Destination	%
United States	14.2%
Germany	13.3%
Iran	11.8%
Turkey	9.99%
Saudi Arabia	9.63%
Canada	7.05%
United Kingdom	5.42%
Australia	5.20%
Pakistan	4.52%
United Arab Emirates	3.21%

3.2 Overview of the experimental design

This section outlines the main elements of the experimental design of the study (Table 4). To thoroughly investigate the impact of sampling of alternatives, we considered five different samples, related to migration aspirations. Moreover, we implemented both random sampling and importance sampling considering different sample sizes for sampling $\mathcal{D}_n$ and $\mathcal{D}_{G_n}$. Finally, a series of different metrics was used to evaluate the results.

Table 4: Overview of experimental design

Factors	**Levels**
Data	Origin countries: Austria, Denmark, Mexico, New Zealand, Afghanistan
Full choice set models	–
Sampling protocols	Random sampling, importance sampling (four realisations)
Sample sizes $\mathcal{D}_n$	20, 40, 60
Sample sizes $\mathcal{D}_{G_n}$	20, 40, 60, all
Performance indicators	Computation time, precision of parameter estimates, goodness–of–fit

The dimensions of the experimental design can be briefly summarised as:

- Data: We considered five different origin countries, namely, Austria, Denmark, Mexico, New Zealand and Afghanistan. Migration aspiration choices were obtained from the Gallup World Poll survey. The details of the data have already been presented in Section 3.1.

- Full choice set models: We estimated the full choice set model for each of the five data sets and used the results as benchmark. The results are presented in Section 3.3

- Sampling protocols: We implemented random sampling and four different variations of importance sampling. The sampling protocols are described in Section 3.4.

- Sample size: For each protocol, we tested three different sample sizes for the choice model subsets $\mathcal{D}_n$, and four different sample sizes for the nests' subsets $\mathcal{D}_{G_n}$. This design resulted in 12 different combinations of sample sizes per protocol and per sample. More details on the sampling size and its implications are presented in Section 3.4.

- Evaluation: We evaluated the models at different levels, such as computation time, precision of parameter estimates, and goodness–of–fit. The exact implementation is discussed in Section 3.5.

The attribute levels of the factors in Table 4 were considered in all possible combinations per sample; for each sample and each set of attribute levels, we performed 100 model estimations.

3.3 Full choice set models: specification and results

Before implementing the sampling of alternatives, we estimated the CNL models using the full choice set (199 alternatives). For simplicity, we adopted an existing model specification (Beine et al., 2024; Baud et al., 2024) for all five samples, since the aim was not to estimate the most representative model for each origin country, but rather to explore the impact of sampling of alternatives. The CNL model consists of three nests

- The English–speaking nest: English is an official language of the destination.

- The OECD nest: The destination is a member of the Organization for Economic Co-operation and Development.

- The Schengen nest: The destination belongs to the Schengen area.

The participation parameters α_{jm}s were not estimated, but instead we assigned their values equally between nests. For example, the destination Greece has $\alpha_{\text{OECD}} = \alpha_{\text{Schengen}} = 0.5$ and $\alpha_{\text{EnglishSpeaking}} = 0$. In the same way, for destination Brazil it is $\alpha_{\text{OECD}} = \alpha_{\text{Schengen}} = \alpha_{\text{EnglishSpeaking}} = 0$. In practice, this means that certain destinations were not assigned to any of the nests.

We assumed a linear utility specification. The explanatory variables were the distance to the destination, the (ln of the) population at the destination, the (ln of the) gross domestic product (GDP) at the destination (in USD), the diaspora, and the constants representing the nests. Some of the independent variables were interacted with the educational level (elementary, secondary and college) of the respondents. It should be mentioned that we differentiated our model specification, compared to the previous studies, by removing the observations of individuals not aspiring to migrate abroad (the so-called 'stayers'). The latter represented the vast majority of all the samples and were excluded in order to reduce the estimation time. All models (both full choice and sampled alternatives) were estimated with Biogeme 3.2.15a01 (Bierlaire, 2003).

The results of the full choice set model for Austria as the origin country are presented in Table 5. The results of the rest of the full choice set models can be found in the Appendix - Section A.1. The interpretation of the models is beyond the scope of the current study and is not explicitly presented. The interested reader is referred to studies that used analogous model specifications for a comprehensive analysis and interpretation of the parameter estimates (Baud et al., 2024; Beine et al., 2025; Beine et al., 2024).

Table 5: Parameter estimates Austria - Full choice set model

Parameter	Value	Rob str err.	Rob t-stat.	Robust p-value
$\beta_{\text{GDP\_eduCollege}}$	0.582	0.167	3.48	0.001
$\beta_{\text{GDP\_eduElementary}}$	0.93	0.379	2.45	0.014
$\beta_{\text{GDP\_eduSecondary}}$	0.645	0.0925	6.97	0.000
$\beta_{\text{englishSpeakingCountries}}$	0.301	0.429	0.701	0.483
$\beta_{\text{logdiaspora\_eduCollege}}$	0.289	0.0368	7.85	0.000
$\beta_{\text{logdiaspora\_eduElementary}}$	0.247	0.066	3.74	0.000
$\beta_{\text{logdiaspora\_eduSecondary}}$	0.229	0.026	8.83	0.000
β_{logdist}	0.547	0.0721	7.59	0.000
β_{logpopul}	0.154	0.0292	5.29	0.000
$\beta_{\text{oecdCountries}}$	1.17	0.36	3.25	0.001
$\beta_{\text{schengenCountries}}$	1.51	0.239	6.31	0.000
μ_{English}	1	0.102	9.77	0.000
μ_{OECD}	1.5	0.207	7.24	0.000
μ_{Schengen}	1.42	0.135	10.5	0.000
LL	-2443.88			
AIC	4915.754			
BIC	4978.856			
N	670			

3.4 Sampling protocols

In the present study, we performed both random sampling and importance sampling. With respect to the latter, we mainly considered OECD membership as the main attribute for the generation of the strata, given the stronger preference towards OECD countries (as shown in Table 3). Hence, the sampling procedure to generate each choice set D_n for the correction of utilities, was carried out by oversampling the OECD countries, to assign to these alternatives a higher probability of being selected. Regarding the sampling procedure of the D_{G_n}, we generated two main strata, namely, *English–speaking* and *OECD*. All sampling protocols were tested for 20, 40, and 60 alternatives in the utility function in the utility function (henceforth U) and 20, 40, 60, and all alternatives in the generating function of the CNL model (henceforth G). Each model was estimated 100 times for every sampling protocol and sample size.

3.4.1 Random sampling (RS)

In random sampling, each alternative has the same probability of being included in the choice set. However, one of its disadvantages is the difficulty in recovering the scale parameters of nests that are underrepresented (or, in extreme cases, not represented at all) due to the nature of the sampling procedure.

3.4.2 Importance sampling 1 protocol: OECD membership (IS1)

The first sampling protocol (IS1) is based solely on OECD membership. Following the observed choices (Table 2), the OECD countries are potentially more relevant and competitive to the choice. Hence, we assigned a higher probability to these countries of being sampled. To better illustrate this protocol, in Table 6, we present the probability that a country is sampled, conditional on being an OECD member or not, for Austria being the origin country. Similarly, we can compute the probabilities for any country of origin.

Table 6: Probabilities to be included in the choice set – Origin: Austria

Group (Total)	**Variable**	**20**	**40**	**60**
OECD (37)	Sampled alternatives	10	20	30
	Probability in $\mathcal{D}$	0.270	0.541	0.811
non-OECD (162)	Sampled alternatives	10	20	30
	Probability in $\mathcal{D}$	0.062	0.123	0.185

Given the specification of the CNL model, the sampling for the correction of G was based on the criteria presented in Table 7.

Table 7: Probabilities to be included in the G set – Origin: Austria

Group (Total)	**Variable**	**20**	**40**	**60**
OECD, Schengen, non-English (34)	Sampled alternatives	10	20	30
	Probability in $\mathcal{D}_G$	0.294	0.588	0.882
English speaking (57)	Sampled alternatives	10	20	30
	Probability in $\mathcal{D}_G$	0.175	0.351	0.526

Note: Because the origin country is excluded from the choice set (stayers are not modelled), the number of available alternatives to sample for $\mathcal{D}_G$ is 91 (out of the 92 possible) for Austria, Denmark, Mexico, and New Zealand. For Afghanistan, the total is 92 since it does not belong to any of the nests.

3.4.3 Importance sampling 2 protocol: OECD membership & 2 most chosen destinations (IS2)

This importance sampling protocol (IS2) further builds on the hypothesis that the sampling of more relevant, to the choice, alternatives results in a better approximation of the true model with a smaller choice set. In practice, IS2 is almost identical to IS1; however, we allowed the two most popular destinations of the sample (recall that these are presented in Table 3) to always be included in the choice set. Regarding the sampling for the nests, we followed the stratification presented in Table 7 but deterministically sampled the two most chosen alternatives in each sample.

3.4.4 Importance sampling3 protocol: OECD membership & countries selected by 5% of the sample (IS3)

The third importance sampling protocol (IS3) follows a very similar philosophy. In this case, we always included all alternatives that were chosen by at least 5% of the respondents. Based on the data, this results in 4-6 countries always included in $\mathcal{D}$. For the nests, we followed an approach similar to IS2; however, in this case we deterministically included destinations that were part of a nest and chosen by at least 5% of the respondents.

3.4.5 Importance sampling 4 protocol: Sampling based on choice "popularity"

Finally, we implemented a procedure solely based on the hypothesis that sampling should be informed by the frequency of the observed choices. Hence, we sampled the alternatives according to their proportion of being chosen in each of the samples. In order to implement this procedure, we segregated the countries into four groups:

- Group 1: Countries always in the sampled choice set (these are the same as in the third importance sampling protocol)
- Group 2: Countries chosen below 5% but above 1% of the time
- Group 3: Countries chosen below 1% of the time
- Group 4: Non-chosen countries

In this protocol, the number of alternatives to be sampled within each group, except for Group 1, is less straightforward. For smaller sample sizes, we used heuristics such as $r_{\text{Group2}} \approx r_{\text{Group3}} + r_{\text{Group4}}$ and $r_{\text{Group3}} = r_{\text{Group4}}$, where r is the number of alternatives sampled. For larger sample sizes, we sampled all alternatives in Group 2 and still followed the $r_{\text{Group3}} = r_{\text{Group4}}$ heuristic. Regarding the nest sampling, we adopted the same approach as in IS3.

3.4.6 Sampling using incoherent procedures

The sampling procedures discussed thus far rely on the principle that frequently chosen alternatives should have a greater probability of being sampled. We attempted to challenge this idea by proposing a protocol that does not follow this pattern. In particular, we implemented and tested the following protocols:

- Incoherent procedure 1 (IC1): In this procedure, we followed the split of Protocol 2, however, we oversampled the non–OECD countries with a 80%–20% split approximately. For the nests, we sampled approximately 2/3 of the total alternatives from the English nest.
- Incoherent procedure 2 (IC2): This procedure is similar to IC1; however, we deterministically included the five most popular destinations that were chosen by less than 1% of the sample. For the nests, we implemented the same rule but for the five most popular destinations that were chosen by less than 1% of the sample and belonged to at least one nest.

The objective of the incoherent procedures was to show that importance sampling works effectively only when alternatives more relevant to the observed choices are sampled.

3.5 Model evaluation

3.5.1 Estimation time

The computational time was recorded as CPU time in seconds. All models were estimated using the SCITAS High Performance computer of the Swiss Federal Technology Institute of Lausanne (EPFL). We estimated all models using 55 CPU cores and 384GB RAM.

3.5.2 Accuracy of parameter estimates

In order to assess the quality of the parameter estimates, we used a series of metrics that can be found in existing studies that implemented the sampling of alternatives technique (Bierlaire et al., 2008; Guevara and Ben-Akiva, 2013b; Lai and Bierlaire, 2015). These are

- the estimate β^* of the full choice set model
- the empirical mean $\widehat{\beta}_k$ over the 100 estimations
- the empirical standard deviation $\widehat{\sigma_k}$ over the 100 estimations
- t−test: the ratio $\frac{\beta^* - \widehat{\beta}_k}{\widehat{\sigma_k}}$
- Mean absolute error (MAE): $\frac{\sum_{i=1}^{K} |\beta_k - \beta^*|}{K}$

3.5.3 Aggregate–level metrics

In addition to the indicators presented in Section 3.5.2, we conducted some additional analyses to assess the quality of the models at the aggregate level. Specifically:

- We examined the difference in the log–likelihood ($\mathcal{LL}$) of the full choice set model between the parameter estimates β^* of the full choice set model, and the estimated parameters, $\widehat{\beta}_k$.
- We conducted out–of–sample validation; for each protocol p and sample size s (including the full choice set), we performed 100 estimations using 80% of the observations. The log–likelihood value $\mathcal{LL}_{p,s}$ was then computed on the remaining 20% of the data, using the parameter estimates $\widehat{\beta}_{p,s,80\%}$ and considering all available alternatives. This metric was used to identify which sampling protocol achieved the highest $\mathcal{LL}_{p,s}$ value.

- We analysed the impact of sampling on the predicted shares, defined as the proportion of times each destination country is chosen according to the model.

4 The impact on estimation time

In this section, we explore the average estimation time for various sample sizes in the utility function (U) and the generating function of the CNL model (G), and how this compares to the estimation time of the full choice set model. Table 8 presents the results with respect to the Austria sample. The cells coloured in grey highlight the occurrences where the estimation time via sampling was on average higher compared to the full choice set model, which was typically the case for large samples of alternatives. We also performed the same type of analysis for Mexico and Afghanistan in Section A.2 of the Appendix (Tables A.5 and A.6, respectively).

The first observation regards the average estimation time of random sampling, which is typically less than the importance sampling protocols. However, as we present in the following sections, this time gain is not translated into accurate parameter estimates. Between importance sampling protocols, the estimation time increased when some of the alternatives were deterministically included in the choice set. However, these sampling protocols were also those that performed better in terms of approximating the parameters of the full choice set.

Regarding sample size, we initially observed a counterintuitive result; in some cases, estimation with sampled choice sets took longer than with the full choice set. The reason lies in the bookkeeping required when using sampled alternatives. In a standard full choice set estimation, each attribute column is tied to a fixed alternative. By contrast, with sampled sets, each column corresponds to the alternative drawn for that observation. This means that values from a single attribute column may come from alternatives belonging to different nests. To resolve this, an additional identifier must be stored to link each draw back to its original alternative–a step not needed in conventional estimation. This extra bookkeeping increases computational cost.

Turning to the results, when sampling 20 alternatives in U, estimation was generally faster than with the full choice set, except when all alternatives were sampled simultaneously in G. With 40 alternatives in U, the pattern was less consistent; estimation time did not always improve relative to the full model. In contrast, when up to 40 alternatives were sampled in G, the estimation was on average faster than with the full choice set. For the Austria sample, the gains were only marginal, while for the other two samples the average estimation time was reduced by 15-40%, depending on the protocol. Finally, when 60 alternatives were sampled in U, time savings appeared only when 20 alternatives were sampled in G. Some improvements were also observed for the Mexico sample and, under certain protocols, for the Afghanistan sample with 40 alternatives in G.

Another interesting observation regards the value of the estimation time for the same number of total alternatives but distributed differently across U and G. One such example is $\mathsf{U}20-\mathsf{G}60$ (20 alternatives sampled in U and 60 alternatives sampled in G), $\mathsf{U}40-\mathsf{G}40$, and $\mathsf{U}60-\mathsf{G}20$, where the total number of alternatives sampled is 80 in all cases. Our results suggested that the estimation time is higher in a balanced approach (i.e., $\mathsf{U}40-\mathsf{G}40$) while the fastest estimation time was observed in the $\mathsf{U}60-\mathsf{G}20$ case. In conclusion, the most significant reduction in estimation time was observed when 20 alternatives were sampled in U.

Table 8: Estimation times - Austria sample

Full choice set time (s)	Sample size	RS		IS1		IS2		IS3		IS4	
		mean	std	mean	std	mean	std	mean	std	mean	std
32.45	U20–G20	7.356	1.348	8.418	0.7457	8.768	1.287	9.000	0.8495	10.01	1.201
	U20–G40	15.03	1.216	16.29	1.371	16.71	1.120	16.83	1.547	19.31	2.041
	U20–G60	21.93	1.878	24.81	2.110	25.47	2.668	25.54	1.828	29.32	3.405
	U20–G91	34.54	3.090	38.84	3.185	40.58	3.919	40.91	2.409	45.78	5.275
	U40–G20	13.91	2.218	15.74	1.132	15.95	1.571	16.08	1.053	16.74	1.196
	U40–G40	27.51	1.981	30.20	2.507	30.60	2.266	30.45	1.989	32.66	2.668
	U40–G60	41.57	3.275	46.88	3.880	47.57	3.674	45.97	2.992	51.10	5.125
	U40–G91	68.43	5.802	75.34	6.078	76.35	6.104	74.59	4.894	80.74	7.409
	U60–G20	20.90	2.515	23.45	1.778	23.08	1.621	23.30	1.408	23.53	1.555
	U60–G40	40.47	2.696	44.10	3.384	44.48	3.015	44.34	2.795	45.57	3.718
	U60–G60	61.90	5.060	68.89	7.378	70.01	6.466	68.27	4.378	69.24	6.634
	U60–G91	99.82	7.012	110.2	8.556	111.0	9.409	108.7	6.514	111.6	10.03

Note: The cells coloured in grey indicate the cases where estimation time with sampling of alternatives was longer than the estimation time of the full choice set model.

5 The impact of sample size on the parameter estimates

The present section investigates the impact of the sample size on the accuracy of the parameter estimates. For each parameter, we calculated the MAE and $t-$test metrics (as described in Section 3.5). To maintain conciseness, we focus solely on the most interesting and insightful findings for each of the sampling protocols. Moreover, it should be highlighted that comparisons between protocols are outside the scope of this section; a thorough investigation of this aspect is presented in Section 6.

5.1 Random sampling (RS) protocol

Table 9 shows the MAE values of the β_{logpopul} parameter of the Austria sample. The patterns in this table are representative of the MAE behaviour for many parameters across all samples. In particular, we observed that increasing the number of parameters in U led to a general decrease in MAE. Moreover, for a fixed number of alternatives in U (for example, 20 alternatives), MAE tended to decrease as the number of alternatives in G increased. Regarding the latter, the most considerable reduction typically occurred between 20 and 40 alternatives, while the rate of improvement decreased when more alternatives were sampled in G. However, it should be noted that this trend did not apply in all cases. An example is the parameter β_{logpopul} of the New Zealand sample (Table A.7), where including all alternatives in G did not improve the MAE value. Another interesting case concerns the behaviour of the MAE with respect to the Schengen nest scale parameter of the Afghanistan model (Table 10); MAE improved considerably when all alternatives were sampled in G. It is worth mentioning that the value of this parameter is considerably higher, compared to the other parameters of the model (Table A.1). Although, from an interpretation point of view, the value of the scale parameter may be problematic, this finding may highlight the potential limitation of random sampling in capturing certain scale parameters related to nests.

Table 9: Austria sample – MAE of β_{logpopul} parameter

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	**MAE**	**Estimation time longer than full choice set**
20	20	0.0523	No
20	40	0.0268	No
20	60	0.0253	No
20	91	0.0245	Yes
40	20	0.0342	No
40	40	0.0134	No
40	60	0.0155	Yes
40	91	0.0130	Yes
60	20	0.0223	No
60	40	0.0111	Yes
60	60	0.0111	Yes
60	91	0.0090	Yes

Table 10: Afghanistan sample – MAE of μ_{Schengen} parameter

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	**MAE**	**Estimation time longer than full choice set**
20	20	10.83	No
20	40	10.21	No
20	60	10.05	No
20	92	2.230	Yes
40	20	10.84	No
40	40	10.17	No
40	60	10.01	No
40	92	1.974	Yes
60	20	10.79	No
60	40	10.17	No
60	60	10.01	No
60	92	1.682	Yes

Similarly to the MAE, the t–test analysis indicated that a "the more, the better" approach may hold for some of the parameters; however, it was not always the case. Although in certain occasions the MAE and t–test showed consistent patterns ($t_{\text{englishSpeakingCountries}}$ in Table 11), there were also situations where reducing the number of alternatives in $\mathbb{U}$ and $\mathbb{G}$ led to smaller values of t–test (hence not significant values). This outcome is counterintuitive and contradicts some of the findings of the MAE analysis. As a reference, in Table A.8, it can be seen that the MAE values for $\beta_{\text{logdiaspora\_eduSecondary}}$ follow different patterns, compared to the t−test ($t_{\text{logdiaspora\_eduSecondary}}$ in Table 11). The source of this inconsistency could be related to the way the t–test is computed (by dividing the difference between the true value and the mean of the estimated values by the standard deviation of the estimated values); typically, for smaller sample sizes, the standard deviation of the estimated parameters is higher, which also results in lower t–test values.

Table 11: Denmark sample – t–test of $\beta_{\text{logdiaspora\_eduSecondary}}$ parameter and t–test of $\beta_{\text{englishSpeakingCountries}}$ parameter

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	$t_{\text{logdiaspora\_eduSecondary}}$	$t_{\text{englishSpeakingCountries}}$	**Estimation time longer than full choice set**
20	20	0.5370	-2.579	No
20	40	1.517	-1.375	No
20	60	2.016	-1.304	No
20	91	2.399	-1.081	No
40	20	0.4573	-2.129	No
40	40	0.6926	-1.210	No
40	60	0.9950	-0.8817	No
40	91	1.716	-0.7642	Yes
60	20	0.3959	-1.795	No
60	40	-0.0856	-1.076	No
60	60	0.6369	-0.6584	Yes
60	91	0.9503	-0.4039	Yes

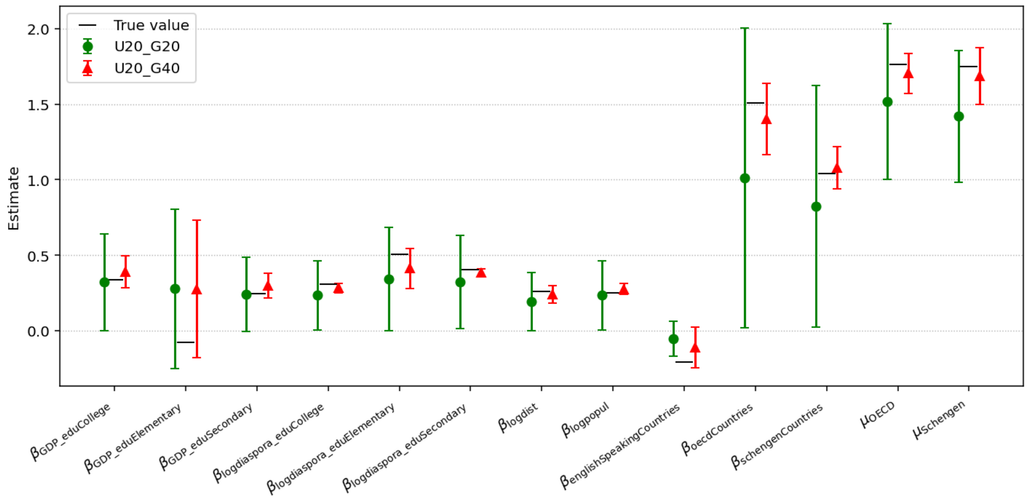

Figure 1: Denmark sample – Comparison of means and std. deviation of parameters across different sample sizes

We further illustrate the inconsistency between MAE and t–test in Figure 1 and Table 12. Figure 1 depicts the mean and standard deviation of the parameter estimates of the Denmark sample model (assuming normality for the empirical distribution), for the sample sizes U20−G20 and U20−G40. Although the two sample sizes approximate better different parameters (based on the mean values), it is obvious that sampling 40 alternatives in G, resulted in a smaller standard deviation of the parameters, compared to sampling 20 alternatives in G. In this example, we notice for the parameters related to the diaspora at the destination that even though their means are closer to the true value when 40 alternatives are sampled in G, the t−test is higher, compared to 20 alternatives sampled in G. Although only one case ($\beta_{\text{logdiaspora\_eduCollege}}$) can be considered marginally significant, this outcome still suggests that the t–test on its own may be a misleading evaluation metric. We argue that in the context of sampling of alternatives, focusing on parameters individually may pose a challenge in identifying the most accurate model; to that end, aggregate metrics should also be considered to obtain more robust conclusions.

Table 12: Denmark sample – Comparison of t–test across different sample sizes

Parameter	U20–G20	U20–G40
$\beta_{\text{logdiaspora\_eduCollege}}$	0.646	1.740
$\beta_{\text{logdiaspora\_eduElementary}}$	0.944	1.387
$\beta_{\text{logdiaspora\_eduSecondary}}$	0.537	1.517

Regarding the scale parameter μ_{Schengen} of the Afghanistan sample, the t−test had a different pattern, compared to MAE (Table 13). In particular, for a fixed number of alternatives sampled in U, the lowest t−test values were observed for 20 and 92 (all) alternatives sampled in G, while higher values were observed for 40 and 60 alternatives sampled in G. This outcome might again indicate that the t−test itself could be unreliable, when the number of alternatives in G is low. It should be noted that regardless of sample size, t−test $\geq |1.96|$ for μ_{Schengen}, which implies that random sampling may not be suitable to accurately capture the parameter estimate of the full choice set model.

Table 13: Afghanistan sample – t–test of μ_{Schengen} parameter

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	**t**	**Estimation time longer than full choice set**
20	20	37.74	No
20	40	86.11	No
20	60	80.17	No
20	92	3.978	Yes
40	20	37.71	No
40	40	94.49	No
40	60	87.69	No
40	92	4.301	Yes
60	20	39.60	No
60	40	124.9	No
60	60	96.02	No
60	92	3.052	Yes

5.2 Importance sampling 1 (IS1) protocol

In the importance sampling 1 (IS1) protocol, MAE values tended to decrease when more alternatives were sampled in $\mathbb{U}$ (Table A.9), which is consistent with the findings presented for RS. Moreover, for a fixed number of alternatives in $\mathbb{U}$, MAE tended to decrease further for more alternatives in $\mathbb{G}$. Deviations were observed for certain parameters such as the $\beta_{\text{logdiaspora\_eduSecondary}}$ of the Denmark sample (Table 14). These exceptions were most often observed when 20 alternatives are sampled in $\mathbb{U}$, suggesting that the cause might be the insufficient sample size.

Table 14: Denmark sample – MAE of $\beta_{\text{logdiaspora\_eduSecondary}}$ parameter

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	**MAE**	**Estimation time longer than full choice set**
20	20	0.0081	No
20	40	0.0061	No
20	60	0.0088	No
20	91	0.0072	No
40	20	0.0079	No
40	40	0.0041	No
40	60	0.0043	Yes
40	91	0.0033	Yes
60	20	0.0093	No
60	40	0.0036	No
60	60	0.0023	Yes
60	91	0.0022	Yes

The t–test resulted, as in RS, in mixed outcomes. For instance, in the Mexico sample, the t−test values behaved in a manner similar to MAE; that is, the values improved for larger samples, for example the $\beta_{GDPCollege}$ parameter (Table A.10). However, there were also cases of inconsistent behaviour. For example, we observed mixed patterns in the $\beta_{logpopul}$ parameter of the New Zealand sample (Table 15), whereas the $\mu_{Schengen}$ parameter of the same sample showed patterns similar to the $\mu_{Schengen}$ of the Afghanistan sample in the RS protocol (Table 13). The rationale behind the latter finding is again related to the calculation formula of the t−test.

Table 15: New Zealand sample – t–test of the $\beta_{logpopul}$ parameter and t–test of the $\mu_{Schengen}$ parameter

Alternatives in U	**Alternatives in** G	$t_{logpopul}$	$t_{\mu_{Schengen}}$	**Estimation time longer than full choice set**
20	20	-1.358	0.4092	No
20	40	-1.163	-0.6532	No
20	60	-1.177	-1.058	No
20	91	-2.009	0.2128	No
40	20	-0.2818	0.5630	No
40	40	0.2004	-0.9868	No
40	60	0.0149	-1.502	Yes
40	91	-1.163	0.1208	Yes
60	20	0.5810	0.6415	No
60	40	1.451	-1.178	Yes
60	60	1.556	-1.927	Yes
60	91	-0.6274	0.2032	Yes

5.3 Importance sampling 2 (IS2) protocol

The results in IS2 largely resembled those in IS1. With respect to MAE, the values generally decreased with sample size (Table A.11) but some exceptions still remained, mostly when only 20 alternatives were sampled in U (Table 16). In addition, no major changes were observed regarding the Afghanistan scale parameter $\mu_{Schengen}$ (Table A.12).

Table 16: Denmark sample – MAE of $\beta_{\text{logdiaspora\_eduSecondary}}$ parameter

Alternatives in $\mathcal{U}$	**Alternatives in** $\mathcal{G}$	**MAE**	**Estimation time longer than full choice set**
20	20	0.0064	No
20	40	0.0074	No
20	60	0.0073	No
20	91	0.0087	No
40	20	0.0071	No
40	40	0.0038	No
40	60	0.0038	Yes
40	91	0.0033	Yes
60	20	0.0069	No
60	40	0.0028	Yes
60	60	0.0025	Yes
60	91	0.0020	Yes

Similarly, the t−test tended to decrease with increasing sample size. Some typical examples are the μ_{OECD} parameter of the Mexico sample and the $\mu_{Schengen}$ parameter of the New Zealand sample in Tables A.13 and A.14 respectively. Some of the parameters that deviated from this pattern were the $\beta_{\text{population}}$ in the New Zealand sample (Table 17) and the μ_{Schengen} in the Afghanistan sample (Table A.15).

Table 17: New Zealand sample – t–test of the β_{logpopul} parameter

Alternatives in $\mathcal{U}$	**Alternatives in** $\mathcal{G}$	t	**Estimation time longer than full choice set**
20	20	-1.723	No
20	40	-1.491	No
20	60	-1.592	No
20	91	-1.613	Yes
40	20	-1.277	No
40	40	-0.6135	No
40	60	-0.9444	Yes
40	91	-1.015	Yes
60	20	-0.3229	No
60	40	-0.3716	Yes
60	60	-0.2552	Yes
60	91	-0.5755	Yes

5.4 Importance sampling 3 (IS3) protocol

The IS3 protocol produced outcomes broadly consistent with the previous importance sampling strategies, again showing decreasing errors with larger sample sizes (Table A.16), although the same deviations from this pattern remained, for example, the parameter $\beta_{\text{logdiaspora\_eduSecondary}}$ of the Denmark sample (Table A.17). Such inconsistencies tended to be observed more often when only 20 alternatives were sampled in U), suggesting that too few alternatives may not suffice to capture the full choice set parameters. Moreover, the discrepancies between the MAE and t−test remained. An example is illustrated in Table 18 regarding the parameter $\beta_{\text{GDP\_eduCollege}}$ of the Mexico sample.

Table 18: Mexico sample – MAE and t–test of the $\beta_{\text{GDP\_college}}$ parameter

Alternatives in U	**Alternatives in** G	t	**MAE**	**Estimation time longer than full choice set**
20	20	0.0296	0.0177	No
20	40	0.3120	0.0164	No
20	60	0.3913	0.0173	No
20	91	0.4288	0.0189	No
40	20	-0.5228	0.0116	No
40	40	-0.0305	0.0101	No
40	60	0.3111	0.0087	No
40	91	0.3842	0.0097	Yes
60	20	-0.8693	0.0110	No
60	40	-0.1591	0.0065	No
60	60	0.0373	0.0070	Yes
60	91	0.4333	0.0066	Yes

5.5 Importance sampling 4 (IS4) protocol

The results from IS4 provided very similar insights to IS3 and, therefore, we do not explicitly present any tables. Overall, the metrics improved with the sample size of alternatives, but once again we observed several exceptions to this pattern.

6 The impact of the sampling protocol on the parameter estimates

In the current section, we investigate the impact of the sampling protocols on the parameter estimates by comparing the MAE and the t–test of certain parameters that we also examined in Section 5. To maintain succinctness, we only present the most representative results and focus on the sample size $40\mathrm{U}-40\mathrm{G}$, as the results in Sections 4 and 5 suggested that this is typically one of the most stable and time-efficient sample sizes in our specific case study.

In general, we observed that RS resulted in the highest MAE values, compared to importance sampling protocols (Figure 2), while IS3 and IS4 performed the best in most cases. This finding is even more profound in certain cases, such as in the Mexico sample (Figure 3). Furthermore, in the Afghanistan sample, the deterministic addition of alternatives in the choice set considerably reduced the MAE value (Figure A1).

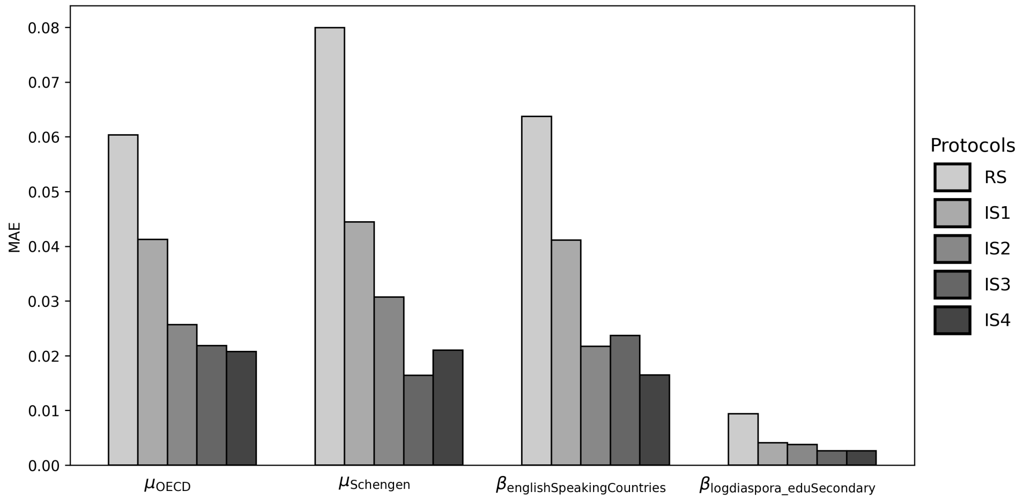

Figure 2: Denmark sample - MAE for U40-G40 sample size

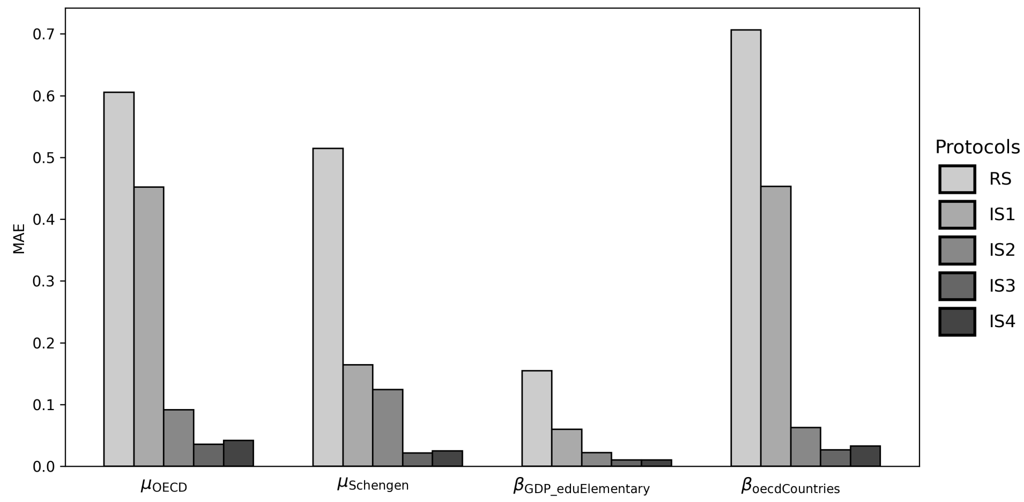

Figure 3: Mexico sample - MAE for U40-G40 sample size

With respect to the t–test, RS again resulted in the highest values, while IS3 generally resulted in the lowest values, followed by IS4. However, we also found instances where these two protocols performed worse than the others for certain parameters (Figures A2–A4). In certain cases, such as the Mexico sample, it was possible to obtain $|t-\mathrm{test}|<1.96$ only for IS3 and IS4 while in the Afghanistan sample this was achieved only with IS4. These findings may indicate that importance sampling protocols that favour the most chosen alternatives can be more effective compared to random sampling or more naïve importance sampling protocols. The t–test results need careful interpretation because, as noted in Section 5, they can be unstable across different sample sizes.

7 Evaluation with aggregate–level metrics

The analysis in Sections 5 and 6 provided useful insights and patterns regarding the impact of sample size and protocol on parameters individually. Nevertheless, the findings remain somewhat inconclusive, as some inconsistencies were observed between different tests. In addition, the focus was limited to specific parameters, which does not allow for robust conclusions about the overall performance of the models. In this section, we assess the impact of the sampling protocols at the aggregate level. Specifically, we analyse the goodness–of–fit and the impact on the predicted destination demand. The analysis refers to sample size U40−G40 unless otherwise stated.

7.1 The impact of the sampling protocol on of the log-likelihood

The idea behind the log–likelihood ($\mathcal{LL}$) analysis is that if the $\mathcal{LL}$ value obtained using parameters from a model with sampled alternatives is close to the $\mathcal{LL}$ of the full choice set model, then those parameters provide a good approximation of the full model's estimates. It should be mentioned that, in addition to the five protocols that we examined previously, we also considered the two incoherent protocols presented in Section 3.4.6 for three of the samples. As a reminder, in the incoherent protocols we deliberately over–sampled alternatives that are less relevant to the observed choices, rather than favouring those more frequently chosen. As shown in Table 19, the $\mathcal{LL}$ values exhibited different patterns between samples and protocols. For example, the sampling protocol did not make any substantial difference in the Austria and Denmark samples. However, in the Mexico and Afghanistan samples, we noted that RS and IS1 performed worse, compared to the other importance sampling protocols. This finding is consistent with the analysis presented in Section 6; protocols IS3 and IS4 notably improved MAE for certain parameters in these two samples. It is worth mentioning that the incoherent protocols performed similarly or worse than RS in the Mexico and Afghanistan samples. This finding suggests that although importance sampling can be a very powerful approach, it should be performed properly; otherwise, the quality of the results may be worse than random sampling.

Table 19: Comparison of log-likelihood values across protocols and samples

	Austria	**Denmark**	**Mexico**	**New Zealand**	**Afghanistan**
Full choice set	-2443.880	-3021.658	-4382.408	-1434.214	-6413.896
RS	-2444.212	-3022.204	-4410.459	-1437.103	-6452.602
IS1	-2443.932	-3021.751	-4387.233	-1435.560	-6443.742
IS2	-2443.905	-3021.692	-4382.906	-1434.307	-6419.206
IS3	-2443.886	-3021.673	-4382.455	-1434.268	-6419.099
IS4	-2443.889	-3021.675	-4382.452	-1434.293	-6413.928
IC1	-2444.200	No models estimated	-4410.650	No models estimated	-6455.013
IC2	-2444.260	No models estimated	-4412.110	No models estimated	-6454.372

Table 20 presents the $\mathcal{LL}$ scores for the sample size U20–G20. Overall, we found that the sample size did not considerably affect the $\mathcal{LL}$ for the more sophisticated importance sampling protocols, whereas higher deviations were observed in the RS and IS1 protocols. The main exception was the Afghanistan sample, where only IS4 was close to the original $\mathcal{LL}$ value. Similar findings were noted for the sample sizes U20–G20 and U60–G20 in Tables (Tables A.19 and A.20, respectively). In conclusion, we observed that even though metrics such as MAE tended to be higher for U20–G20, U20–G40 and U60–G20, compared to U40–G40, this did not have pronounced implications on the $\mathcal{LL}$ values, especially when more sophisticated sampling protocols were implemented. However, it may be best to avoid random sampling, as the sample size can affect the accuracy of the models.

Table 20: Comparison of log-likelihood values across protocols and samples (U20–G20)

	Austria	**Denmark**	**Mexico**	**New Zealand**	**Afghanistan**
Full choice set	-2443.880	-3021.658	-4382.408	-1434.214	-6413.896
RS	-2480.088	-3086.763	-4530.415	-1443.946	-6747.995
IS1	-2443.932	-3022.514	-4407.462	-1437.483	-6476.679
IS2	-2443.905	-3022.036	-4388.016	-1435.475	-6431.101
IS3	-2443.949	-3021.890	-4382.781	-1434.846	-6430.649
IS4	-2444.025	-3021.961	-4382.963	-1435.029	-6414.258
IC1	-2445.456	No models estimated	-4471.642	No models estimated	-6492.240
IC2	-2446.068	No models estimated	-4493.866	No models estimated	-6500.815

In Section 5.1, we discussed the discrepancies between MAE and the t–test, and the difficulty in identifying the most accurate model, with respect to the sample size, based on parameters individually. In Tables 20 and A.19, we observed that the $\mathcal{LL}$ values related to the Denmark sample for the sample sizes $\mathtt{U20{-}G20}$ and $\mathtt{U20{-}G40}$ are -3086.763 and -3023.360, respectively. This outcome suggests that despite the difference in performance in approximating the parameters individually, the sample size $\mathtt{U20{-}G40}$ approximated the $\mathcal{LL}$ value of the full choice set model (-3021.658) considerably better.

7.2 Out–of–sample validation

For three of the samples, namely, Austria, Mexico and Afghanistan, we performed an additional out–of–sample validation. In particular, we considered 80% of the observations to estimate the model. Then we used the parameter estimates on the remaining 20% of the observations and computed the $\mathcal{LL}$ for each of the samples, considering the full choice set (Table 21). The results of the Austria sample suggested that the sampling protocol did not have any impact on $\mathcal{LL}$, as the differences were negligible. With respect to the Afghanistan and Mexico samples, we found that RS, IS1, and the incoherent samples performed worse than the other importance sampling protocols. As a general conclusion, the out–of–sample validation results further confirmed our findings in Section 7.1.

Table 21: Estimation–Validation comparison of log-likelihood values across protocols and samples (U40–G40)

	Austria	**Mexico**	**Afghanistan**
Full choice set	-465.627	-832.204	-1316.904
RS	-465.640	-840.957	-1326.131
IS1	-465.596	-836.998	-1322.926
IS2	-465.640	-834.970	-1319.612
IS3	-465.601	-834.597	-1320.905
IS4	-465.741	-834.607	-1320.136
IC1	-465.619	-841.249	-1326.811
IC2	-465.615	-841.583	-1326.353

Table 22 presents the out–of–sample validation results for the U20—G20 sample size. The main differences in performance (compared to the U40—G40 sample size) were mainly observed for random sampling, while for the rest of the sampling protocols, the differences were less obvious. We also observed the same trends for sample sizes U20—G40 and U60—G20 (Tables A.21 and A.22 respectively).

Table 22: Estimation–Validation comparison of log-likelihood values across protocols and samples (U20–G20)

	Austria	**Mexico**	**Afghanistan**
RS	-468.404	-866.932	-1348.312
IS1	-465.617	-841.220	-1329.702
IS2	-465.658	-837.195	-1321.750
IS3	-465.688	-834.888	-1323.127
IS4	-466.099	-835.074	-1319.984
IC1	-465.871	-854.668	-1334.663
IC2	-465.976	-858.141	-1336.640

7.3 The impact on shares

Table 23 shows the results of the analysis on the predicted shares, for each of the sampling protocols. For the interest of space, we only present the results of the 10 most popular destinations with respect to the Austria sample, as these were observed in the data. Overall, we did not find major differences between the sampling protocols, which is consistent with the analysis performed previously with respect to this sample.

Table 23: Comparison of shares across protocols – Austria sample (U40–G40)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	13.97%	13.91%	14.19%	14.04%	13.96%	13.91%
Australia	9.53%	9.49%	9.54%	9.54%	9.55%	9.49%
Germany	8.23%	8.20%	8.18%	8.14%	8.21%	8.20%
Canada	5.57%	5.56%	5.60%	5.58%	5.55%	5.56%
Switzerland	5.04%	5.05%	4.97%	4.99%	5.03%	5.05%
Spain	4.72%	4.71%	4.66%	4.68%	4.73%	4.71%
France	4.29%	4.28%	4.26%	4.26%	4.28%	4.28%
Italy	3.53%	3.54%	3.51%	3.51%	3.53%	3.54%
New Zealand	3.14%	3.14%	3.11%	3.13%	3.14%	3.14%
United Kingdom	2.20%	2.20%	2.22%	2.20%	2.18%	2.20%

Table 24 reports the results of the shares for the Mexico sample. In this case, RS emerged as the least efficient protocol, resulting in the largest deviation from the full choice set model for the United States, which was by far the most popular destination in the sample. However, as the protocols became more sophisticated, the deviations from the original shares gradually decreased. The results suggest that the presence of dominating choices and their impact on sampling should be further investigated, especially under random sampling.

Table 24: Comparison of shares across protocols – Mexico sample (U40–G40)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	45.76%	52.50%	47.23%	46.30%	45.93%	45.88%
Germany	7.29%	5.73%	6.50%	6.89%	7.17%	7.18%
Canada	5.62%	5.25%	5.59%	5.68%	5.65%	5.63%
France	4.88%	3.93%	4.45%	4.68%	4.82%	4.82%
Spain	3.96%	3.12%	3.58%	3.76%	3.94%	3.91%
Italy	3.14%	2.70%	2.98%	3.06%	3.10%	3.11%
United Kingdom	2.68%	2.40%	2.60%	2.68%	2.68%	2.67%
Japan	1.28%	1.17%	1.18%	1.27%	1.26%	1.28%
Brazil	0.74%	0.87%	0.81%	0.75%	0.74%	0.75%
China	0.37%	0.46%	0.43%	0.38%	0.37%	0.38%

Finally, Table 25 presents the shares for the Afghanistan sample. In this case, larger deviations were noted compared to the full choice set model for certain destinations (e.g., USA and Germany), particularly in RS and IS1. Both these protocols did not always capture the order of preferences. For example, Germany is less popular than Iran and Pakistan in RS and IS; however, this result is not true, as shown in the full choice set model and the rest of the sampling protocols (Table 3). By contrast, for destinations such as Iran or Australia, the results remained consistent across all protocols. These discrepancies were likely due to the sampling procedures: apart from the destinations deterministically included in IS3 and IS4, the remaining countries were stratified based on OECD membership. This approach did not reflect the choice patterns of this specific sample, as over-sampling OECD countries was not representative when individuals relied on other criteria for their decisions. IS4 did not suffer from this limitation, as all destinations were selected according to their observed choice frequency. This outcome was consistent with the results shown in Figures A1 and A3, where IS4 performed better in all parameters.

Table 25: Comparison of shares across protocols – Afghanistan sample

Choice	**Full choice set**	**RS**	**IS1**	**IS2**	**IS3**	**IS4**
USA	18.02%	20.53%	18.64%	17.50%	17.29%	17.95%
Germany	12.43%	8.43%	9.29%	12.38%	12.15%	12.41%
Iran	10.45%	10.77%	10.89%	10.44%	10.86%	10.40%
Türkiye	8.66%	8.08%	8.69%	8.68%	8.70%	8.65%
Pakistan	8.32%	9.56%	10.24%	9.84%	9.95%	8.32%
Saudi Arabia	8.18%	8.23%	7.91%	7.55%	7.92%	8.16%
United Kingdom	6.31%	6.75%	6.56%	6.45%	6.33%	6.34%
Canada	4.52%	4.47%	4.36%	4.35%	4.31%	4.54%
Australia	3.32%	3.19%	3.14%	3.17%	3.14%	3.34%
United Arab Emirates	1.62%	1.54%	1.53%	1.52%	1.57%	1.63%

The impact on shares was also investigated for other sample sizes (U20–G20 and U60–G20). The results are presented in the Appendix section A.6 (Tables A.23–A.28). The RS protocol for sample size U20–G20 did not accurately approximate the full choice set model shares on certain occasions, however, the remaining of the protocols resulted in shares similar to the sample size U40–G40.

8 Findings

This section summarises the main results of the analysis and highlights the key lessons learned. It also provides a set of practical recommendations based on how different sampling procedures and sample sizes affected the estimation of the CNL models.

8.1 Summary of findings and lessons learned

The models were evaluated at different levels, in particular, estimation time, the accuracy of the parameters, and at the aggregate level.

- Impact of sample size on the accuracy of parameters: For several of the parameters, increasing the sample size reduced the value of MAE suggesting that increasing the sample size could be beneficial for the accuracy. However, this trend did not hold for all parameters and several exceptions were observed. Moreover, we found that, especially in lower sample sizes, MAE and t–ratio led to conflicting results. Given the number of parameters estimated, the individual check of parameters poses challenges in the reliable evaluation of a model. For instance, it is not clear how the decrease of MAE affects the performance of a model in practice or how much a low MAE value is "low enough".

- Impact of the sampling procedure on the accuracy of the parameters: The sampling procedure had a considerable impact on the accuracy of the parameters. We found that sampling alternatives more relevant to the chosen alternatives, i.e. deterministically including certain alternatives in the choice set, improved the accuracy of parameters with respect to the MAE metric. Unlike the sample size, the improvement in parameters with respect to the sampling procedure was clearer.

- Impact of the sampling procedure on the model fit: The choice of the sampling protocol had a considerable impact on how the log–likelihood of the full choice set model specification is approximated. This finding was especially profound in the case of data sets with specific choice patterns. For example, in the Mexico sample, the choices were skewed towards certain destinations, and in the Afghanistan sample, the nesting structure might not have been appropriate, given that we observed a different pattern in preferences, compared to the other data sets. In these cases, a good implementation of importance sampling was crucial for the performance of the models, especially for smaller sample sizes. This finding was further supported by the results related to the incoherent sampling procedures. The latter performed similarly (or worse) to random sampling, demonstrating the importance of a well–designed sampling protocol procedure.

- Impact on shares: In most of the samples, the predicted shares of the sampled models were close to those of the full choice set, even for smaller sample sizes. However, in the Mexico and Afghanistan samples, larger differences appeared. For Mexico, this was due to the strong dominance of one destination, while for Afghanistan, it likely reflected a nesting structure that did not fully match the observed preferences. The design of the sampling protocol was crucial to accurately reflect the shares of the full choice set model; random sampling was the least efficient approach.

- Model evaluation: Evaluating models at the aggregate level, such as log–likelihood or predicted shares, provided more reliable insight into overall performance. Aggregate measures were less sensitive to parameter–specific variations and offered a clearer view of how well the sampled models approximated the full choice set.

- Impact on estimation time: A reasonable expectation would be the reduction of the average estimation time when sampling of alternatives is implemented; however, this was not the case in our results. As we discussed in Section 4, This is due to the additional bookkeeping needed when sampled alternatives are used. Unlike the full choice set, where each attribute column is fixed to one alternative, sampling requires tracking which alternative each draw belongs to, adding extra computational overhead. Moreover, one should keep in mind that this bookkeeping process applies to both sampled data sets, which further increases the estimation time.

8.1.1 Recommendations

Based on the findings and lessons learned from this exercise, there is a series of recommendations for researchers and practitioners for the implementation of the sampling of alternatives.

- Despite its simplicity, random sampling should be in general avoided. Although random sampling performed well in some of the data sets, such as the Austria sample, it struggled to capture certain parameters in datasets where more particular behaviour was observed, such as the Mexico and Afghanistan samples.

- Except for the representativeness of alternatives, from a behavioural point of view, random sampling may also cause estimation–related issues regarding the nests. Given the pure stochastic nature of the approach, there is always the possibility that alternatives of a specific nest are not sufficiently or not at all sampled. Importance sampling is a very effective way to reduce the magnitude of this problem.

- The sampling procedure implemented in an importance sampling protocol matters. The results of the incoherent sampling procedures suggested that a poorly designed importance sampling procedure may perform even worse than random sampling. The researcher should take into account the patterns in the data during the design of the sampling procedure.

- The deterministic sampling of certain popular alternatives can improve the performance of the models; based on our results, this practice, if possible, is strongly advised.
- For certain protocols, smaller sample sizes (for example U20–G20 in our case) resulted in shares comparable to higher sample sizes. However, in some case studies, the accuracy of certain individual parameters is crucial. Our analysis showed that the sample size can affect the parameter values; it is better to avoid very small sample sizes, even though it might be tempting sometimes.
- Beyond a certain sample size, the estimation time increases for marginal benefits in the performance of the models. It is preferred to implement a well–designed importance sampling protocol with fewer alternatives than random sampling with more alternatives. However, given the previous recommendation, there is an optimal point with respect to how small the sample size should be.
- If the model evaluation compared to the full choice set model is of interest, this should be performed considering more than one indicator. Our individual parameter analysis revealed cases where conflicting results were obtained, depending on the evaluation metric used. The use of measures at the aggregate level may be an overall more reliable approach to evaluate the larger picture of a model.

9 Conclusion

This paper presented an empirical investigation of sampling of alternatives in CNL models, focusing on a case study of migration aspirations. Starting from a hypothesis that sampling alternatives more relevant to the observed choices would enhance the overall performance of the models, compared to more naïve protocols, we implemented a series of sampling protocols, varying both in the sampling procedure and the sample size, to examine the following research objectives:

- Assess the influence of sampling procedures on model accuracy: Our results confirmed that importance sampling has the potential to outperform random sampling. Moreover, we can obtain different results for different importance sampling procedures.

- Examine the effect of sample size on model accuracy and estimation time: Sample size was mainly identified as an issue in random sampling. However, in certain cases, the combination of sample size and inefficient importance sampling procedure was also problematic. Sample size had a considerable effect on estimation time; sampling many alternatives actually resulted in longer estimation times, compared to the full choice set model.

- Investigate how the complex substitution patterns of the CNL model are preserved: Our analysis showed that well–designed importance sampling protocols largely preserve the scale parameters of the nests. In certain samples, random sampling failed to capture the aggregate shares; this is another implication of a poorly designed sampling protocol.

Despite the promising and interesting results, there are a number of limitations and additional considerations that should be noted. First, the nature of migration data means that some destinations dominate the choice sets for certain origin countries. This was especially clear in the Mexico sample, but present to some extent in all cases. Such skewed distributions made it possible to design frequency-based sampling strategies in this study, but this may not be the case in other applications where the number of alternatives is larger and patterns in choices less clear, such as in residential location or workplace choice problems. However, this fact opens up the possibility for additional research towards this direction, for example in the context of residential location choice models. Such an extension could also allow one to explore protocols that combine frequency-based rules with distance-based sampling to ensure that both attractive and nearby alternatives are represented. Second, all sampling procedures were based on the assumption that the full choice set can be enumerated. However, this does not apply in certain applications, such as route choice, where different approaches such as the variational autoencoder proposed by Yao and Bekhor (2022) may be considered. Third, our experiments suggested that although at the aggregate level sampling of alternatives can give good results (for example, with respect to the $\mathcal{LL}$ of a model), different parameters may respond differently to sampling using different metrics of evaluation; we highlighted the inconsistencies between MAE and t–test in several cases. This is particularly important for applications where the accuracy of individual parameters is crucial for the analysis, for example the value–of–time. These cases may require larger or more carefully structured choice sets to be accurately recovered. Fourth, it should be mentioned that the representativeness of our results is limited to the OECD–based stratification approach that we implemented. Despite the adequate performance of these protocols, additional research is required to explore alternative stratification criteria. This aspect can be further explored with the sample size of each stratum. For example, in our case, we followed a balanced approach in sampling from the strata; however, alternative sample sizes can be considered. Finally, our heuristics for designing the choice frequency–based importance sampling (IS4) were arbitrary, and no sensitivity analysis was performed to test the impact of the cutoff criteria for the generation of the strata on the results. Hence, additional efforts may be required in this direction, to develop a systematic methodology.

In conclusion, the study provided some evidence that importance sampling protocols tailored to observed data can help approximate full choice set models in migration aspiration research. However, additional work is needed to assess their robustness in different contexts and model specifications.

Acknowledgements

This research was funded in whole or in part, by the Luxembourg National Research Fund (FNR), grant reference MIGDCM, and the Swiss National Science Foundation (project number: 200021E_205280).

References

Baud, C., Paschalidis, E., Vortisch, A. B., Beine, M. and Bierlaire, M. (2024). Migration patterns of Europeans after brexit, *Proceedings of the 24th Swiss Transportation Research Conference*.
URL: *www.strc.ch*

Becker, O., Börger, T. and Meyerhoff, J. (2025). Sampling of alternatives in spatial decision contexts with logit and logit mixture models: Simulation and application to freshwater recreation in Germany, *Resource and Energy Economics* p. 101535.

Beine, M., Bierlaire, M. and Docquier, F. (2025). New York, Abu Dhabi, London or Stay at Home? Using a Cross-Nested Logit Model to Identify Complex Substitution Patterns in Migration, *Journal of Economic Geography* **25**: 685â–712.

Beine, M., Bierlaire, M., Paschalidis, E., Varotto, S. and Vortisch, A. B. (2024). The Impact of a Possible Trump Reelection on Mexican Immigration Pressures in Alternative Countries, *Available at SSRN* .

Bekaert, E., Ruyssen, I. and Salomone, S. (2021). Domestic and international migration intentions in response to environmental stress: A global cross-country analysis, *Journal of Demographic Economics* **87**: 383–436.

Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*, Vol. 9.

Bertoli, S. and Moraga, J. F.-H. (2013). Multilateral resistance to migration, *Journal of development economics* **102**: 79–100.

Bertoli, S. and Ruyssen, I. (2018). Networks and migrants' intended destination, *Journal of Economic Geography* **18**: 705–728.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the Swiss Transport Research Conference*, Ascona, Switzerland.

Bierlaire, M., Bolduc, D. and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples, *Transportation Research Part B: Methodological* **42**: 381–394.

Bierlaire, M. and Paschalidis, E. (2023). Estimating MEV models with samples of alternatives, *Technical Report TRANSP-OR 231225*, Lausanne, Switzerland.

Buggle, J. C., Mayer, T., Sakalli, S. and Thoenig, M. (2020). The Refugee's dilemma: evidence from Jewish migration out of Nazi Germany, *The Quarterly Journal of Economics* **138**: 1273â–1345.

Daly, A., Hess, S. and Dekker, T. (2014). Practical solutions for sampling alternatives in large-scale models, *Transportation research record* **2429**: 148–156.

Dekker, T., Bansal, P. and Huo, J. (2025). Revisiting McFadden's correction factor for sampling of alternatives in multinomial logit and mixed multinomial logit models, *Transportation Research Part B: Methodological* **192**: 103129.

Docquier, F., Tansel, A. and Turati, R. (2020). Do emigrants self-select along cultural traits? Evidence from the MENA countries, *International Migration Review* **54**: 388–422.

Farooq, B. and Miller, E. J. (2012). Towards integrated land use and transportation: A dynamic disequilibrium based microsimulation framework for built space markets, *Transportation Research Part A: Policy and Practice* **46**: 1030–1053.

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological* **43**: 984–994.

Gubert, F. and Senne, J.-N. (2016). Is the European Union attractive for potential migrants?: An investigation of migration intentions across the world, *OECD Social, Employment and Migration Working Papers* (No. 188).

Guevara, C. A. and Ben-Akiva, M. E. (2013a). Sampling of alternatives in logit mixture models, *Transportation Research Part B: Methodological* **58**: 185–198.

Guevara, C. A. and Ben-Akiva, M. E. (2013b). Sampling of alternatives in multivariate extreme value (MEV) models, *Transportation Research Part B: Methodological* **48**: 31–52.

Lai, X. and Bierlaire, M. (2015). Specification of the cross-nested logit model with sampling of alternatives for route choice models, *Transportation Research Part B: Methodological* **80**: 220–234.

Langella, M. and Manning, A. (2021). Income and the desire to migrate, *CEP Discussion Papers* (No. 1794).

Lee, B. H. Y. and Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives, *Transportation* **37**: 587–601.

Leite Mariante, G., Ma, T.-Y. and Van Acker, V. (2018). Modeling discretionary activity location choice using detour factors and sampling of alternatives for mixed logit models, *Journal of Transport Geography* **72**: 151–165.

Lemp, J. D. and Kockelman, K. M. (2012). Strategic sampling for large choice sets in estimation and application, *Transportation Research Part A: Policy and Practice* **46**: 602–613.

Li, M.-T., Chow, L.-F., Zhao, F. and Li, S.-C. (2005). Geographically stratified importance sampling for the calibration of aggregated destination choice models for trip distribution, *Transportation research record* **1935**: 85–92.

Lovo, S. (2014). Potential migration and subjective well-being in Europe, *IZA Journal of Migration* **3**: 1–18.

Ma, T.-Y., Mariante, G. and Van Acker, V. (2017). Location choice modeling based on mixed logit model and sampling of alternatives.

McConnell, K. E. and Tseng, W.-C. (1999). Some preliminary evidence on sampling of alternatives with the random parameters logit, *Marine Resource Economics* **14**: 317–332.

McFadden, D. (1978). Modeling the Choice of Residential Location, *In: Karlquist, Lundqvist, Snickers, Weibull (Eds.) Spatial interaction theory and planning models* pp. 75–96.

Monras, J. (2018). Economic shocks and internal migration.

Nerella, S. and Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models, *Transportation Research Record* **1894**: 11–19.

Ortega, F. and Peri, G. (2013). The effect of income and immigration policies on international migration, *Migration Studies* **1**: 47–74.

Tsoleridis, P., Choudhury, C. F. and Hess, S. (2022). Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities, *Journal of Choice Modelling* **42**: 100336.

Yao, R. and Bekhor, S. (2022). A variational autoencoder approach for choice set generation and implicit perception of alternatives in choice modeling, *Transportation Research Part B: Methodological* **158**: 273–294.

Zhong, Y., Lu, J. and Gong, Z. (2025). Population migration and housing market dynamics in california agglomerations, *International Journal of Sociology and Social Policy* pp. 1–18.

A Appendix

A.1 Parameter estimates - full choice set models

Table A.1: Parameter estimates Afghanistan - Full choice set model

Parameter	Value	Rob str err.	Rob t-stat.	Robust p-value
$\beta_{\text{GDP\_eduCollege}}$	1.09	0.168	6.46	0.000
$\beta_{\text{GDP\_eduElementary}}$	0.671	0.0795	8.44	0.000
$\beta_{\text{GDP\_eduSecondary}}$	1.05	0.0894	11.8	0.000
$\beta_{\text{englishSpeakingCountries}}$	-1.09	0.171	-6.36	0.000
$\beta_{\text{logdiaspora\_eduCollege}}$	0.305	0.0282	10.8	0.000
$\beta_{\text{logdiaspora\_eduElementary}}$	0.361	0.0149	24.2	0.000
$\beta_{\text{logdiaspora\_eduSecondary}}$	0.315	0.0158	19.9	0.000
β_{logdist}	-0.203	0.0565	-3.59	0.000
β_{logpopul}	0.431	0.0352	12.2	0.000
$\beta_{\text{oecdCountries}}$	1.85	0.087	21.2	0.000
$\beta_{\text{schengenCountries}}$	-1.01	0.0972	-10.4	0.000
μ_{English}	1	0.0867	11.5	0.000
μ_{OECD}	1.37	0.0531	25.9	0.000
μ_{Schengen}	12.2	0.624	19.6	0.000
LL	-6413.9			
AIC	12855.79			
BIC	12935.62			
N	2212			

Table A.2: Parameter estimates Denmark - Full choice set model

Parameter	Value	Rob str err.	Rob t-stat.	Robust p-value
$\beta_{\text{GDP\_eduCollege}}$	0.34	0.182	1.88	0.061
$\beta_{\text{GDP\_eduElementary}}$	-0.0737	0.511	-0.144	0.885
$\beta_{\text{GDP\_eduSecondary}}$	0.247	0.106	2.33	0.020
$\beta_{\text{englishSpeakingCountries}}$	-0.204	0.253	-0.806	0.420
$\beta_{\text{logdiaspora\_eduCollege}}$	0.311	0.0409	7.61	0.000
$\beta_{\text{logdiaspora\_eduElementary}}$	0.507	0.182	2.79	0.005
$\beta_{\text{logdiaspora\_eduSecondary}}$	0.406	0.0312	13	0.000
β_{logdist}	0.26	0.0425	6.1	0.000
β_{logpopul}	0.25	0.0282	8.86	0.000
$\beta_{\text{oecdCountries}}$	1.51	0.229	6.58	0.000
$\beta_{\text{schengenCountries}}$	1.04	0.193	5.41	0.000
μ_{English}	1	0.0688	14.5	0.000
μ_{OECD}	1.76	0.194	9.09	0.000
μ_{Schengen}	1.75	0.284	6.14	0.000
LL	-3021.66			
AIC	6071.317			
BIC	6138.797			
N	916			

Table A.3: Parameter estimates Mexico - Full choice set model

Parameter	Value	Rob str err.	Rob t-stat.	Robust p-value
$\beta_{\text{GDP\_eduCollege}}$	0.737	0.119	6.19	0.000
$\beta_{\text{GDP\_eduElementary}}$	0.468	0.132	3.54	0.000
$\beta_{\text{GDP\_eduSecondary}}$	0.659	0.0987	6.68	0.000
$\beta_{\text{englishSpeakingCountries}}$	-0.514	0.337	-1.52	0.127
$\beta_{\text{logdiaspora\_eduCollege}}$	0.0623	0.0225	2.76	0.006
$\beta_{\text{logdiaspora\_eduElementary}}$	0.153	0.026	5.91	0.000
$\beta_{\text{logdiaspora\_eduSecondary}}$	0.105	0.0217	4.84	0.000
β_{logdist}	-0.695	0.0916	-7.59	0.000
β_{logpopul}	0.377	0.0516	7.32	0.000
$\beta_{\text{oecdCountries}}$	2.3	0.206	11.2	0.000
$\beta_{\text{schengenCountries}}$	0.671	0.16	4.2	0.000
μ_{English}	1	0.081	12.4	0.000
μ_{OECD}	2.25	0.257	8.73	0.000
μ_{Schengen}	2.63	0.473	5.56	0.000
LL	-4382.41			
AIC	8792.816			
BIC	8868.33			
N	1626			

Table A.4: Parameter estimates New Zealand - Full choice set model

Parameter	Value	Rob str err.	Rob t-stat.	Robust p-value
$\beta_{GDP\_eduCollege}$	0.3	0.176	1.71	0.088
$\beta_{GDP\_eduElementary}$	-0.0831	0.197	-0.422	0.673
$\beta_{GDP\_eduSecondary}$	0.0229	0.119	0.192	0.848
$\beta_{englishSpeakingCountries}$	0.464	0.507	0.916	0.36
$\beta_{logdiaspora\_eduCollege}$	0.226	0.0506	4.46	8.15E-06
$\beta_{logdiaspora\_eduElementary}$	0.35	0.0583	6	2.02E-09
$\beta_{logdiaspora\_eduSecondary}$	0.313	0.051	6.15	7.67E-10
$\beta_{logdist}$	-0.462	0.11	-4.21	2.57E-05
$\beta_{logpopul}$	0.22	0.0451	4.89	9.98E-07
$\beta_{oecdCountries}$	2.52	0.44	5.73	9.95E-09
$\beta_{schengenCountries}$	0.888	0.297	2.99	0.00279
$\mu_{English}$	1.06	0.108	9.78	0
μ_{OECD}	2.28	0.665	3.42	0.000627
$\mu_{Schengen}$	2.28	0.376	6.06	1.34E-09
LL	-1434.21			
AIC	2896.428			
BIC	2960.042			
N	695			

A.2 Estimation times

Table A.5: Estimation times - Mexico sample

Full choice set time (s)	Sample size	RS		IS1		IS2		IS3		IS4	
		mean	std	mean	std	mean	std	mean	std	mean	std
116.89	U20–G20	9.825	2.687	13.04	1.980	17.66	2.249	18.15	2.676	16.41	3.615
	U20–G40	22.23	2.338	27.34	2.964	37.31	4.476	36.06	3.980	32.81	8.130
	U20–G60	35.34	3.451	43.58	5.462	59.26	6.778	58.91	5.997	51.98	10.85
	U20–G91	57.03	5.431	71.23	9.621	94.30	10.30	92.15	8.139	78.68	15.34
	U40–G20	18.17	4.321	26.32	3.006	31.24	3.291	32.15	3.504	33.63	8.624
	U40–G40	41.26	4.371	59.01	6.986	68.91	8.150	68.82	13.99	70.32	22.12
	U40–G60	66.39	6.612	95.21	13.42	111.0	15.51	109.0	10.92	114.7	67.57
	U40–G91	109.2	12.38	154.4	23.20	178.1	19.00	177.4	20.01	177.4	36.60
	U60–G20	28.11	6.849	36.76	3.788	45.55	4.600	46.66	3.698	47.69	11.56
	U60–G40	62.73	6.441	80.27	8.392	100.2	11.46	98.68	10.61	97.11	19.17
	U60–G60	99.42	9.569	139.6	18.35	167.0	20.25	161.3	19.95	154.4	25.47
	U60–G91	167.4	18.02	233.5	29.78	264.1	31.36	256.6	29.55	245.4	46.33

Table A.6: Estimation times - Afghanistan sample

Full choice set time (s)	Sample size	RS		IS1		IS2		IS3		IS4	
		mean	std	mean	std	mean	std	mean	std	mean	std
238.71	U20–G20	10.02	3.188	14.74	3.282	47.31	24.44	51.01	20.60	62.26	29.66
	U20–G40	25.38	3.556	30.48	6.003	103.6	48.55	97.56	38.52	121.3	54.36
	U20–G60	37.52	3.706	53.00	16.99	158.0	63.66	160.4	70.23	178.4	80.02
	U20–G92	262.7	110.7	239.2	91.69	249.6	100.4	258.9	99.32	312.6	148.0
	U40–G20	19.63	6.395	26.64	7.100	78.49	36.64	84.97	31.24	97.30	39.16
	U40–G40	44.93	4.343	56.87	9.451	171.5	93.75	175.9	78.97	201.0	74.03
	U40–G60	73.21	7.757	90.02	14.77	251.3	91.51	288.8	132.2	329.1	114.7
	U40–G92	495.3	231.3	415.9	199.9	451.2	182.3	456.3	202.5	528.9	199.1
	U60–G20	30.88	9.351	40.55	5.655	123.7	55.83	135.8	61.20	131.8	58.00
	U60–G40	67.34	7.476	82.50	11.81	253.7	92.67	277.9	126.1	279.5	120.8
	U60–G60	118.4	15.69	129.6	16.62	481.8	216.7	427.9	187.7	445.9	202.3
	U60–G92	675.3	285.6	690.9	277.5	718.9	332.8	685.7	305.5	720.7	322.1

A.3 Tables - parameter estimates

Table A.7: New Zealand sample – MAE of $\beta_{\mathtt{logpopul}}$ parameter (RS)

Alternatives in $\mathbb{U}$	**Alternatives in $\mathbb{G}$**	**MAE**	**Estimation time longer than full choice set**
20	20	0.0987	No
20	40	0.0866	No
20	60	0.0947	No
20	91	0.1009	No
40	20	0.0725	No
40	40	0.0393	No
40	60	0.0447	Yes
40	91	0.0560	Yes
60	20	0.0430	No
60	40	0.0211	No
60	60	0.0211	Yes
60	91	0.0315	Yes

Table A.8: Denmark sample – MAE of $\beta_{\text{logdiaspora\_eduSecondary}}$ parameter (RS)

Alternatives in $\mathbb{U}$	**Alternatives in $\mathbb{G}$**	**MAE**	**Estimation time longer than full choice set**
20	20	0.0872	No
20	40	0.0204	No
20	60	0.0221	No
20	91	0.0240	No
40	20	0.0885	No
40	40	0.0094	No
40	60	0.0109	No
40	91	0.0126	Yes
60	20	0.0828	No
60	40	0.0060	No
60	60	0.0059	Yes
60	91	0.0079	Yes

Table A.9: Austria sample — MAE of μ_{OECD} parameter (IS1)

Alternatives in $\mathbb{U}$	**Alternatives in $\mathbb{G}$**	**MAE**	**Estimation time longer than full choice set**
20	20	0.0657	No
20	40	0.0456	No
20	60	0.0387	No
20	91	0.0397	Yes
40	20	0.0453	No
40	40	0.0293	No
40	60	0.0221	Yes
40	91	0.0164	Yes
60	20	0.0461	No
60	40	0.0253	Yes
60	60	0.0159	Yes
60	91	0.0129	Yes

Table A.10: Mexico sample — t—test of $\beta_{\text{GDP\_eduSecondary}}$ parameter (IS1)

Alternatives in $\mathbb{U}$	**Alternatives in $\mathbb{G}$**	t	**Estimation time longer than full choice set**
20	20	-3.570	No
20	40	-2.850	No
20	60	-2.731	No
20	91	-2.683	No
40	20	-2.709	No
40	40	-2.095	No
40	60	-1.911	No
40	91	-1.620	Yes
60	20	-2.806	No
60	40	-1.749	No
60	60	-2.137	Yes
60	91	-0.8207	Yes

Table A.11: New Zealand sample — MAE of μ_{OECD} parameter (IS2)

Alternatives in U	**Alternatives in** G	**MAE**	**Estimation time longer than full choice set**
20	20	0.2203	No
20	40	0.1292	No
20	60	0.0910	No
20	91	0.0904	Yes
40	20	0.1802	No
40	40	0.1008	No
40	60	0.0567	Yes
40	91	0.0440	Yes
60	20	0.1984	No
60	40	0.0861	Yes
60	60	0.0463	Yes
60	91	0.0237	Yes

Table A.12: Afghanistan sample — MAE of μ_{Schengen} parameter (IS2)

Alternatives in U	**Alternatives in** G	**MAE**	**Estimation time longer than full choice set**
20	20	1.242	No
20	40	1.293	No
20	60	1.332	No
20	92	1.276	Yes
40	20	1.292	No
40	40	1.233	No
40	60	1.190	Yes
40	92	1.215	Yes
60	20	1.304	No
60	40	1.187	Yes
60	60	1.250	Yes
60	92	1.244	Yes

Table A.13: Mexico sample — t—test of μ_{OECD} parameter (IS2)

Alternatives in U	**Alternatives in** G	t	**Estimation time longer than full choice set**
20	20	6.356	No
20	40	1.630	No
20	60	0.0156	No
20	91	-0.4218	No
40	20	7.490	No
40	40	2.379	No
40	60	0.4423	No
40	91	-0.3561	Yes
60	20	7.850	No
60	40	2.948	No
60	60	0.4345	Yes
60	91	-0.5594	Yes

Table A.14: New Zealand sample — t—test of μ_{Schengen} parameter (IS2)

Alternatives in U	**Alternatives in** G	t	**Estimation time longer than full choice set**
20	20	1.057	No
20	40	0.0508	No
20	60	-0.5683	No
20	91	-0.5896	Yes
40	20	1.625	No
40	40	0.4210	No
40	60	0.0816	Yes
40	91	-0.3144	Yes
60	20	1.323	No
60	40	0.6698	Yes
60	60	0.1952	Yes
60	91	-0.3458	Yes

Table A.15: Afghanistan sample — t—test of μ_{Schengen} parameter (IS2)

Alternatives in U	**Alternatives in** G	t	**Estimation time longer than full choice set**
20	20	2.099	No
20	40	2.120	No
20	60	2.019	No
20	92	1.601	Yes
40	20	2.184	No
40	40	2.319	No
40	60	1.942	Yes
40	92	1.976	Yes
60	20	2.777	No
60	40	2.059	Yes
60	60	2.182	Yes
60	92	2.044	Yes

Table A.16: New Zealand sample — MAE of μ_{OECD} parameter (IS3)

Alternatives in U	**Alternatives in** G	**MAE**	**Estimation time longer than full choice set**
20	20	0.1801	No
20	40	0.0849	No
20	60	0.0704	No
20	91	0.0746	Yes
40	20	0.1710	No
40	40	0.0731	No
40	60	0.0373	Yes
40	91	0.0317	Yes
60	20	0.1692	No
60	40	0.0785	Yes
60	60	0.0308	Yes
60	91	0.0188	Yes

Table A.17: Denmark sample – MAE of $\beta_{\text{logdiaspora\_eduSecondary}}$ parameter (IS3)

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	**MAE**	**Estimation time longer than full choice set**
20	20	0.0054	No
20	40	0.0062	No
20	60	0.0057	No
20	91	0.0052	No
40	20	0.0039	No
40	40	0.0026	No
40	60	0.0028	Yes
40	91	0.0025	Yes
60	20	0.0034	No
60	40	0.0023	Yes
60	60	0.0018	Yes
60	91	0.0018	Yes

Table A.18: Austria sample – MAE of the β_{OECD} parameter, and t–test of the β_{OECD} paramete (IS3)

Alternatives in $\mathbb{U}$	**Alternatives in** $\mathbb{G}$	$t_{\beta_{\text{OECD}}}$	**MAE**	**Estimation time longer than full choice set**
20	20	0.7671	0.0657	No
20	40	0.8204	0.0662	No
20	60	0.8559	0.0675	No
20	91	1.011	0.0653	Yes
40	20	1.032	0.0471	No
40	40	0.6183	0.0366	No
40	60	0.5015	0.0377	Yes
40	91	0.3245	0.0373	Yes
60	20	0.8017	0.0384	No
60	40	0.2486	0.0258	Yes
60	60	0.2901	0.0260	Yes
60	91	0.4334	0.0257	Yes

A.4 Figures - compare protocols

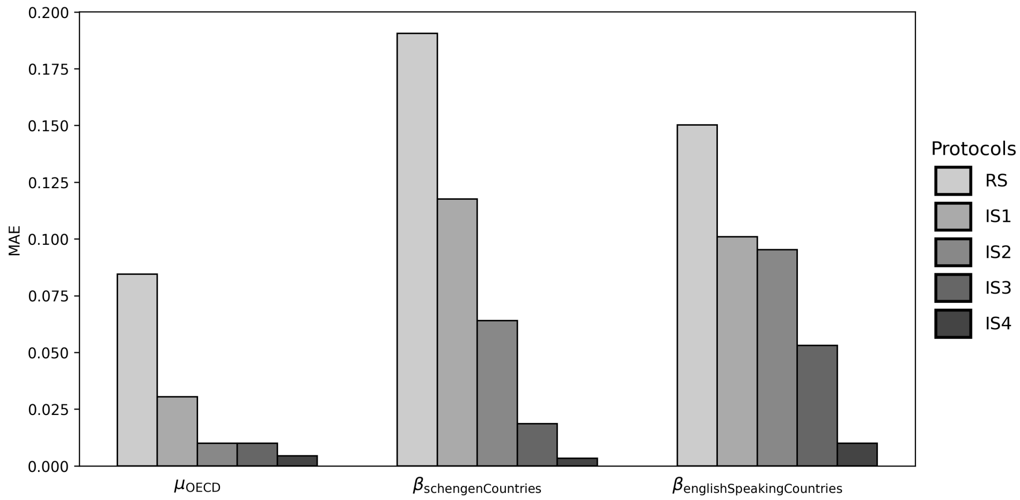

Figure A1: Afghanistan sample - MAE for U40-G40 sample size

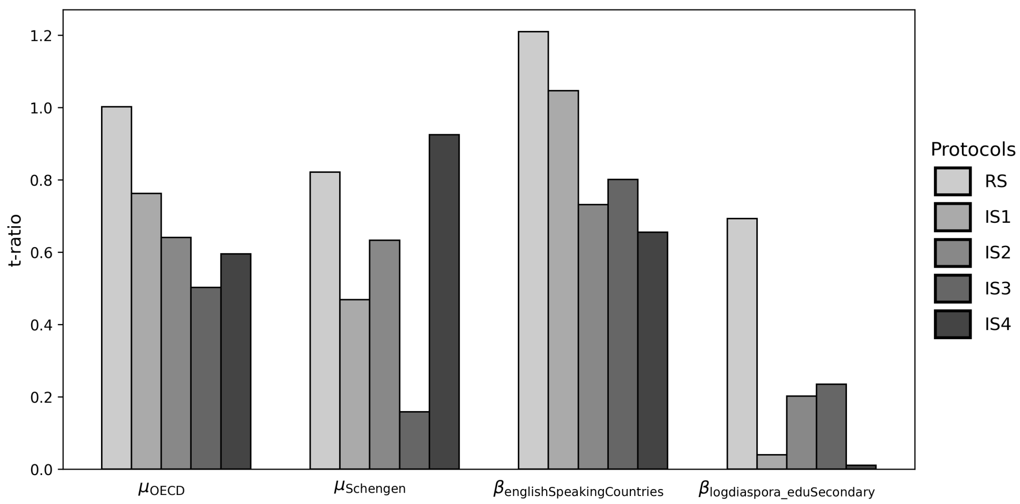

Figure A2: Denmark sample - t−test for U40-G40 sample size

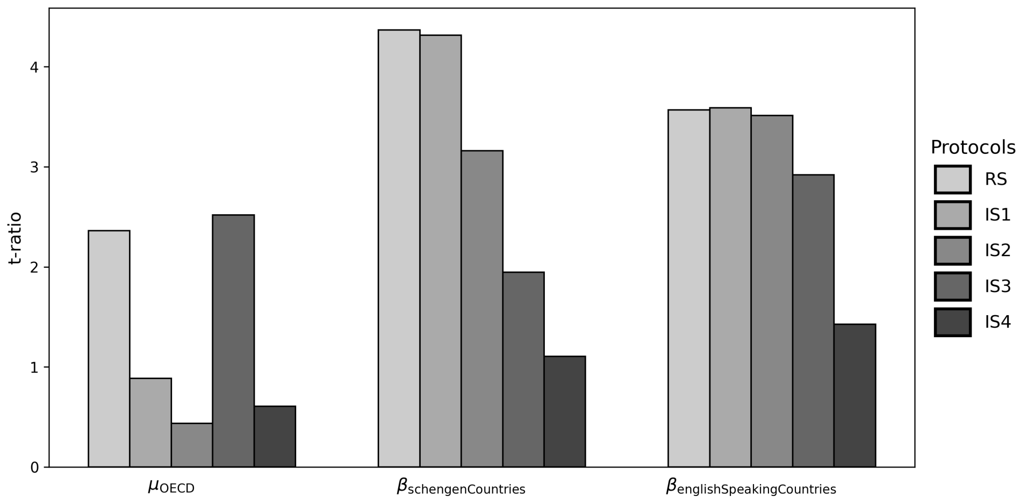

Figure A3: Afghanistan sample - t−test for U40-G40 sample size

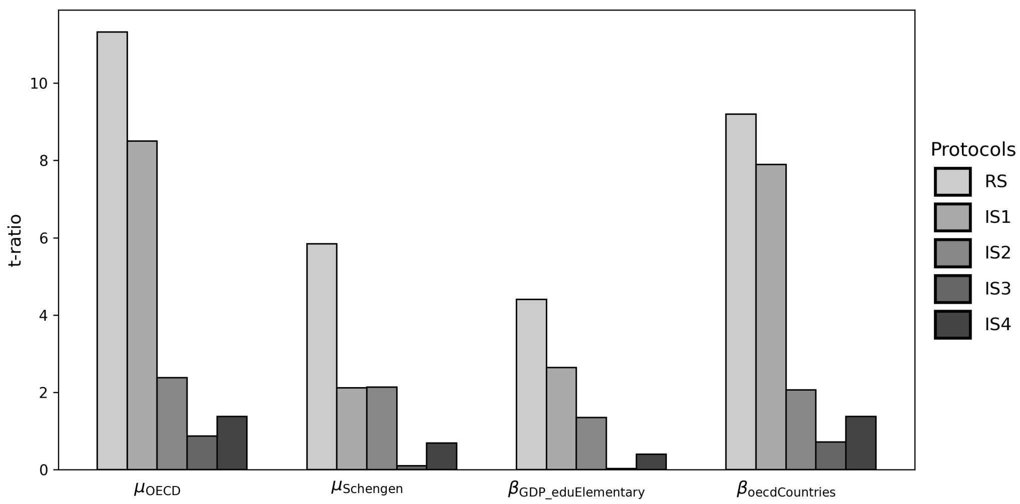

Figure A4: Mexico sample - t−test for U40-G40 sample size

A.5 Log-likelihood values across protocols and samples

Table A.19: Comparison of log-likelihood values across protocols and samples (U20−G40)

	Austria	**Denmark**	**Mexico**	**New Zealand**	**Afghanistan**
Full choice set	-2443.880	-3021.658	-4382.408	-1434.214	-6413.896
RS	-2445.131	-3023.360	-4443.343	-1441.228	-6467.118
IS1	-2444.150	-3021.903	-4396.991	-1436.424	-6457.804
IS2	-2443.947	-3021.895	-4383.041	-1434.735	-6431.649
IS3	-2443.970	-3021.888	-4382.491	-1434.261	-6431.173
IS4	-2444.002	-3021.898	-4382.606	-1434.294	-6414.259
IC1	-2445.032	No models	-4451.463	No models	-6470.811
IC2	-2445.616	No models	-4470.386	No models	-6473.883

Table A.20: Comparison of log-likelihood values across protocols and samples (U60−G20)

	Austria	**Denmark**	**Mexico**	**New Zealand**	**Afghanistan**
Full choice set	-2443.880	-3021.658	-4382.408	-1434.214	-6413.896
RS	-2450.811	-3076.391	-4534.766	-1447.316	-6652.067
IS1	-2444.014	-3022.362	-4393.507	-1436.561	-6454.852
IS2	-2443.928	-3021.929	-4387.750	-1434.892	-6416.414
IS3	-2443.893	-3021.695	-4382.804	-1434.902	-6415.932
IS4	-2443.887	-3021.688	-4382.793	-1434.925	-6413.906
IC1	-2444.283	No models	-4409.187	No models	-6467.150
IC2	-2444.326	No models	-4413.705	No models	-6471.494

Table A.21: Estimation–Validation comparison of log-likelihood values across protocols and samples (U20–G40)

	Austria	**Mexico**	**Afghanistan**
RS	-465.723	-847.537	-1329.18
IS1	-465.593	-838.559	-1325.32
IS2	-465.716	-834.835	-1321.17
IS3	-465.632	-834.513	-1323.20
IS4	-466.032	-834.651	-1319.88
IC1	-465.856	-849.334	-1329.98
IC2	-465.918	-852.195	-1330.55

Table A.22: Estimation–Validation comparison of log-likelihood values across protocols and samples (U60–G20)

	Austria	**Mexico**	**Afghanistan**
RS	-465.685	-878.918	-1365.53
IS1	-465.630	-839.754	-1325.05
IS2	-465.547	-837.457	-1319.71
IS3	-465.572	-834.942	-1320.43
IS4	-465.648	-834.991	-1320.19
IC1	-465.635	-842.463	-1328.61
IC2	-465.600	-843.031	-1329.84

A.6 Shares for different sample sizes of alternatives

Table A.23: Comparison of shares across protocols – Austria sample (U20–G20)

Destination	**Full choice set**	**RS**	**IS1**	**IS2**	**IS3**	**IS4**
USA	13.97%	10.29%	14.61%	14.17%	13.87%	13.73%
Australia	9.530%	6.974%	9.502%	9.524%	9.555%	9.414%
Germany	8.226%	5.790%	8.069%	8.080%	8.207%	8.122%
Canada	5.566%	4.699%	5.642%	5.611%	5.500%	5.521%
Switzerland	5.042%	3.836%	4.833%	4.944%	5.043%	5.050%
Spain	4.722%	3.649%	4.539%	4.578%	4.792%	4.688%
France	4.288%	3.515%	4.207%	4.201%	4.309%	4.273%
Italy	3.531%	3.065%	3.492%	3.491%	3.537%	3.545%
New Zealand	3.144%	2.746%	3.046%	3.101%	3.129%	3.116%
United Kingdom	2.197%	2.341%	2.278%	2.243%	2.115%	2.191%

Table A.24: Comparison of shares across protocols – Austria sample (U60–G20)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	13.97%	12.50%	14.14%	14.10%	14.03%	13.99%
Australia	9.530%	8.545%	9.552%	9.569%	9.554%	9.539%
Germany	8.226%	6.974%	8.083%	8.097%	8.172%	8.165%
Canada	5.566%	5.297%	5.598%	5.599%	5.580%	5.577%
Switzerland	5.042%	4.462%	4.950%	4.965%	5.010%	5.018%
Spain	4.722%	4.194%	4.625%	4.649%	4.689%	4.688%
France	4.288%	3.911%	4.224%	4.236%	4.263%	4.264%
Italy	3.531%	3.319%	3.497%	3.500%	3.518%	3.518%
New Zealand	3.144%	3.031%	3.120%	3.139%	3.139%	3.143%
United Kingdom	2.197%	2.299%	2.217%	2.213%	2.200%	2.203%

Table A.25: Comparison of shares across protocols – Mexico sample (U20–G20)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	45.76%	34.65%	51.25%	47.36%	46.26%	46.07%
Germany	7.287%	4.449%	5.445%	6.069%	6.952%	6.915%
Canada	5.620%	5.059%	5.404%	5.778%	5.718%	5.669%
France	4.884%	3.439%	3.821%	4.261%	4.699%	4.686%
Spain	3.960%	2.921%	3.071%	3.320%	3.862%	3.798%
Italy	3.141%	2.680%	2.661%	2.889%	3.040%	3.034%
United Kingdom	2.682%	2.657%	2.434%	2.687%	2.687%	2.641%
Japan	1.280%	1.692%	1.134%	1.287%	1.234%	1.257%
Brazil	0.740%	1.421%	0.890%	0.768%	0.740%	0.762%
China	0.373%	0.857%	0.471%	0.437%	0.364%	0.382%

Table A.26: Comparison of shares across protocols – Mexico sample (U60–G20)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	45.76%	32.31%	46.89%	47.27%	46.35%	46.32%
Germany	7.287%	4.678%	5.833%	6.094%	6.935%	6.953%
Canada	5.620%	5.249%	5.614%	5.657%	5.661%	5.652%
France	4.884%	3.640%	4.136%	4.289%	4.684%	4.692%
Spain	3.960%	3.081%	3.309%	3.420%	3.821%	3.813%
Italy	3.141%	2.835%	2.895%	2.936%	3.049%	3.052%
United Kingdom	2.682%	2.796%	2.635%	2.666%	2.682%	2.679%
Japan	1.280%	1.697%	1.226%	1.308%	1.275%	1.277%
Brazil	0.740%	1.282%	0.865%	0.787%	0.754%	0.758%
China	0.373%	0.800%	0.488%	0.437%	0.378%	0.383%

Table A.27: Comparison of shares across protocols – Afghanistan sample (U20–G20)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	18.02%	12.39%	19.16%	16.98%	16.64%	17.68%
Germany	12.43%	5.174%	7.685%	12.26%	11.92%	12.36%
Iran	10.45%	8.033%	11.64%	10.63%	11.18%	10.33%
Turkey	8.658%	5.949%	8.332%	8.615%	8.756%	8.608%
Pakistan	8.319%	7.732%	11.92%	11.26%	11.37%	8.382%
Saudi Arabia	8.180%	6.529%	7.844%	7.161%	7.733%	8.115%
United Kingdom	6.305%	5.634%	6.681%	6.474%	6.329%	6.413%
Canada	4.518%	4.106%	4.303%	4.197%	4.148%	4.580%
Australia	3.321%	3.208%	3.050%	3.027%	3.000%	3.383%
United Arab Emirates	1.620%	1.962%	1.415%	1.430%	1.535%	1.658%

Table A.28: Comparison of shares across protocols – Afghanistan sample (U60–G20)

Destination	Full choice set	RS	IS1	IS2	IS3	IS4
USA	18.02%	13.33%	18.79%	17.70%	17.59%	17.99%
Germany	12.43%	5.811%	8.084%	12.29%	12.25%	12.41%
Iran	10.45%	8.821%	11.39%	10.50%	10.71%	10.43%
Turkey	8.658%	6.501%	8.403%	8.556%	8.678%	8.648%
Pakistan	8.319%	7.626%	10.01%	9.377%	9.306%	8.323%
Saudi Arabia	8.180%	7.294%	8.500%	7.801%	8.023%	8.172%
United Kingdom	6.305%	5.712%	6.528%	6.413%	6.323%	6.322%
Canada	4.518%	4.271%	4.444%	4.408%	4.391%	4.529%
Australia	3.321%	3.318%	3.207%	3.217%	3.209%	3.331%
United Arab Emirates	1.620%	1.991%	1.595%	1.562%	1.590%	1.625%