

# Simulation error in maximum likelihood estimation of discrete choice models

Mikołaj Czajkowski<sup>1</sup>, Wiktor Budziński<sup>2</sup>

## Abstract

Maximum simulated likelihood is the preferred estimator of most researchers who deal with discrete choice. It allows estimation of models such as mixed multinomial logit (MXL), generalized multinomial logit, or hybrid choice models, which have now become the state-of-practice in the microeconomic analysis of discrete choice data. All these models require simulation-based solving of multidimensional integrals, which can lead to several numerical problems. In this study, we focus on one of these problems – utilizing from 100 to 1,000,000 draws, we investigate the extent of the simulation bias resulting from using several different types of draws: (1) pseudo random numbers, (2) modified Latin hypercube sampling, (3) randomized scrambled Halton sequence, and (4) randomized scrambled Sobol sequence. Each estimation is repeated up to 1,000 times. The simulations use several artificial datasets based on an MXL data generating process with different numbers of individuals (400, 800, 1200), different numbers of choice tasks per respondent (4, 8, 12), different number of attributes (5, 10), and different experimental designs (D-optimal, D-efficient for the MNL and D-efficient for the MXL model). Our large-scale simulation study allows for comparisons and drawing conclusions with respect to (1) how efficient different types of quasi Monte Carlo simulation methods are and (2) how many draws one should use to make sure the results are of “satisfying” quality – under different experimental conditions. Our study is the first to date to offer such a comprehensive comparison. Overall, we find that the number of the best-performing *Sobol* draws required for the desired precision exceeds 2,000 in some of the 5-attribute settings, and 20,000 in the case of some 10-attribute settings considered.

---

<sup>1</sup> University of Warsaw, Department of Economics, Długa 44/50, 00-241 Warsaw, Poland, [mc@uw.edu.pl](mailto:mc@uw.edu.pl)

<sup>2</sup> University of Warsaw, Department of Economics, Długa 44/50, 00-241 Warsaw, Poland, [wbudzinski@wne.uw.edu.pl](mailto:wbudzinski@wne.uw.edu.pl)

**Keywords:** discrete choice, mixed logit, simulated maximum log-likelihood function, simulation error, draws, quasi Monte Carlo methods, MLHS, Halton, Sobol, number of draws

**JEL codes:** C15, C51, C63

### Highlights

- Simulation error for the maximum simulated likelihood estimation of discrete choice models is investigated
- We use from 100 to 1,000,000 draws for 400, 800, and 1200 individuals, 4, 8 and 12 choice tasks per individual, 5 or 10 attributes, and 3 different experimental designs
- Pseudo random numbers, modified Latin hypercube sampling, randomized scrambled Halton sequence, and randomized scrambled Sobol sequence are compared. Each estimation is repeated up to 1,000 times.
- We find that Sobol draws perform the best in these simulations.
- We find that in the case of 5-attribute designs over 2,000 *Sobol* draws were needed, while in the case of 10 attributes over 20,000 draws were needed to attain desired levels of precision.

### Acknowledgements

The authors wish to thank Kenneth Train and the participants of the 4th Workshop on Discrete Choice Modelling, Copenhagen, 2015, the participants of the 5th International Choice Modelling Conference, Cape Town, 2017, the participants of the 23rd Annual Conference of the European Association of Environmental and Resource Economists, Athens, 2017, and the participants of the 25th Ulvön Conference on Environmental Economics, Ulvön, 2018, who facilitated this research with helpful comments on earlier versions of this paper. MC gratefully acknowledges the support of the Polish Ministry of Science and Higher Education and the National Science Centre of Poland (project 2015/19/D/HS4/01972). WB gratefully acknowledges the support of the National Science Centre of Poland (project 2016/21/N/HS4/02094).

## 1. Introduction

Discrete choice models are widely used in many applications, with modelling of consumers' preferences probably being the most prominent ([Ben-Akiva and Lerman, 1985](#); [Train, 2009](#)). Mixed logit ([Revelt and Train, 1998](#)) is the model of choice for most of these applications, and arguably also the state-of-the-art, considering its ability to approximate any random utility based choice model to any degree of accuracy ([McFadden and Train, 2000](#)).

Most applications estimate the model using the simulated maximum likelihood method, as it is relatively straightforward and readily implemented in most statistical software packages. Simulating the value of the log-likelihood function is necessarily associated with the simulation error that depends on the number and type of draws used. By using a different set of draws or even changing the order of explanatory variables, a researcher will arrive at somewhat different estimation results, in terms of the value of the log-likelihood function, parameter estimates, and their estimated standard errors (and hence the associated z-statistics).

Several studies have demonstrated the advantages of using quasi Monte Carlo (QMC) methods in terms of reducing simulation-driven variation of the results (e.g., using Halton rather than pseudo-random draws), and this has led to their wide proliferation. Unfortunately, examples of 100 Halton draws leading to smaller bias than 1,000 pseudo-random draws ([e.g., Bhat, 2001](#)) have led some to actually use very few draws for simulations, when in fact not much is known about the extent of the possible bias resulting from using different numbers of different types of draws in various conditions (datasets). Our study aims at filling this gap.

In what follows, we present the results of a systematic comparison of pseudo-random, modified Latin hypercube sampling, Halton, and Sobol draws under a wide set of experimental conditions in terms of experimental designs, the number of individuals (400-1,200), the number of choice tasks per individual (4-12), and the number of attributes (5, 10). Based on a Monte Carlo simulation, we demonstrate the extent of the simulation error resulting from using 100 up to 1,000,000 draws.<sup>1</sup> This allows us to offer recommendations in terms of the QMC method: which performs best and its relative efficiency.

Using more draws is always better than using fewer – not only will the estimates become more precise (lower simulation error) but this can also lead to uncovering identification problems.

---

<sup>1</sup> Throughout the paper, the number of draws refers to the number of draws per individual and per random parameter. In total, each model utilizes  $[number\ of\ individuals \cdot number\ of\ draws] \times [number\ of\ random\ variables]$  draws.

1 However, we propose guidelines regarding how many draws are “enough” for a required precision  
2 level. Our measure is based on limiting the probability of making an error (to, e.g., 5%) when  
3 comparing otherwise identical models that can differ in simulated values only.

4 Overall, we find that the scrambled Sobol sequence performs the best in these simulations. As  
5 expected, scrambled Halton draws and modified Latin hypercube sampling also perform  
6 substantially better than pseudo-random draws. Importantly, our results indicate that, for common  
7 setups, thousands or, in some cases, tens of thousands of draws are required to attain desired levels  
8 of precision. While this result suggests more draws than have been common in previous  
9 applications, advances in computer speed now permit far more draws with the same runtimes as  
10 previously.

11 The rest of the paper is structured as follows. Section 2 provides an overview of earlier studies  
12 devoted to measuring simulation error and comparing the performance of various QMC methods.  
13 Section 3 presents the set-up of our Monte Carlo study. Section 4 introduces the methodology of  
14 comparisons and describes the framework used for recommending a “sufficient” number of draws.  
15 Results are presented in section 5 – we first compare the performance of various QMC simulation  
16 methods and then address the question of how many draws are “enough.” The last section offers  
17 discussion and conclusions.

## 19 2. Earlier studies

20 Quasi Monte Carlo (QMC) methods gained considerable attention as a way of reducing  
21 computation burden or simulation error. Several alternatives to the pseudo-random Monte Carlo  
22 method (i.e., drawing pseudo-random numbers and using them for simulations) were proposed,  
23 including using modified Latin hypercube sampling, Halton and Sobol sequences. The rationale  
24 and the description of the algorithms used for generating each of these types of “draws”<sup>2</sup> are  
25 presented in Online Supplement A; in what follows we focus on critically reviewing existing studies  
26 which aimed at comparing the performance of these methods in simulated maximum likelihood  
27 estimation of discrete choice models.

28 The most popular QMC method used in this context is the Halton sequence, introduced by [Bhat](#)  
29 [\(2001\)](#) and followed by [Train \(2000\)](#). Both of these papers provided an early indication that Halton

---

<sup>2</sup> Technically, they are not draws, because the generated numbers follow a pre-defined sequence. We choose to call them draws, however, because in estimation they are used as if they were draws from the target distribution.

draws greatly outperform pseudo-random methods, illustrating this with examples of 100 Halton draws leading to smaller bias and standard deviation of parameter estimates than 1,000 pseudo-random draws, or in some cases even 2,000 pseudo-random draws ([Bhat, 2001](#)). Although most of the later comparisons showed that the differences in performance between pseudo-random and QMC methods are not as substantial, Halton sequences were consistently found to outperform pseudo-random draws, and became a new standard for most modelers.

One problem with using the Halton sequence is that, by definition, it is purely deterministic, and therefore it is not possible to evaluate the error by applying variance analysis, as in classical Monte Carlo simulations.<sup>3</sup> Another problem is its poor performance in higher dimensions, because the sequences generated using high prime numbers as bases tend to be highly correlated (see Online Supplement B for illustration). To address these problems, researchers suggested using scrambling or shuffling the sequence or proposed other QMC methods.

[Bhat \(2003\)](#) compared the performance of pseudo-random draws with a randomized Halton sequence and randomized scrambled Halton sequence using a mixed probit setting with 10 random parameters. He reports that scrambling improved performance – 150 scrambled Halton draws performed better than 1,000 pseudo-random draws. In a similar study [Hess, Polak and Daly \(2003\)](#) and [Hess, Polak and Daly \(2003\)](#) compared pseudo-random draws with scrambled Halton and shuffled Halton sequences, and found that shuffling is a valid alternative for scrambling in the case of breaking the correlation for high dimensional problems. The authors conclude that the differences in performance between pseudo-random and QMC methods are not as large as indicated in the initial studies. [Wang and Kockelman \(2008\)](#) also compared scrambled and shuffled Halton sequences, concluding that although scrambling seems to perform better, the difference is relatively small.

Other QMC methods that were proposed typically used the Halton and pseudo-random draws as a benchmark. [Sándor and Train \(2004\)](#) used four types of randomized (t,m,s)-nets; two of them outperformed the randomized Halton.<sup>4</sup> [Garrido \(2003\)](#) showed that using Sobol sequences generally results in better performance than Halton and PMC, especially in higher (10+)

---

<sup>3</sup> [Train \(2000\)](#) tried to work around this problem by estimating every model 5 times, each time generating Halton sequences using different combinations of prime numbers. This approach still does not guarantee valid estimates of variance, however.

<sup>4</sup> As noted by our reviewer, there are types of (t,m,s)-nets that have even better "coverage" (defined by number theory) than Sobol draws and can be expected to perform better. Specifically, Niederreiter nets in base 2 have better coverage than Sobol draws, and Niederreiter-Xing nets in base 2 have better coverage than Niederreiter nets in base 2 ([Sándor and Train, 2004](#)). For a software implementation of Niederreiter-Xing nets in base 2 see [Pirsic \(2002\)](#).

dimensional problems, where it leads to 58% lower standard deviations of parameter estimates in comparison with using 150 draws. [Sivakumar, Bhat and Ökten \(2005\)](#) use mixed logit setting with 5 or 10 random variables to compare Latin hypercube sampling with Halton and Faure sequences (with and without scrambling). Their results, based on 20 repetitions and the comparisons of the simulation error resulting from using 25, 100, 125 and 625 draws with the results obtained using 20,000 draws, indicate that scrambled Faure sequence performs best. [Hess, Train and Polak \(2006\)](#) proposed modified Latin hypercube sampling and showed that they can perform better than Halton and pseudo-random draws. [Munger et al. \(2012\)](#) compared pseudo-random, randomized Halton, Sobol, and lattice rules and analyzed variance and bias of the simulated likelihood in the case of the mixed logit model. They found that randomized lattice rules and randomized Sobol nets outperformed pseudo-random and Halton draws. [Sidharthan and Srinivasan \(2010\)](#) proposed using generalized antithetic draws with double base shuffling to a Halton sequence and showed that this can improve the model's abilities to recover true parameters.<sup>5,6</sup>

Although it seems like a lot has been done, existing studies can hardly be considered a systematic comparison that would allow for drawing conclusions with respect to which approach is best or how many draws are “enough.” This is because many of these comparisons use a relatively low number of QMC draws (e.g., [Train \(2000\)](#), [Bhat \(2001\)](#), [Bhat \(2003\)](#), [Hess, Polak and Daly \(2003\)](#) and [Sándor and Train \(2004\)](#) do not use more than 200 QMC draws in their comparisons). In addition, most earlier studies use deterministic QMC sequences or a very low number of repetitions for each type and number of draws (e.g., [Bhat \(2003\)](#), [Sándor and Train \(2004\)](#), [Garrido \(2003\)](#) or [Hess, Train and Polak \(2006\)](#) used no more than 10 repetitions). This makes it hard to judge if the conclusions are sound, or the result is only obtained for a particular set of draws and data. Finally, the results likely depend on the number of observations. As noted by [Sándor and Train \(2004\)](#): *“large sampling variance means that the log-likelihood function is fairly flat near the maximum; and when the likelihood function is fairly flat near its maximum, errors induced by simulation can move the maximum considerably.”* As a result, smaller datasets with fewer observations per individual (larger standard errors of parameter estimates) are likely to result in different performance of QMC vs. pseudo-random draws.

---

<sup>5</sup> In another context, [Bliemer, Rose and Hess \(2008\)](#) compared the performance of pseudo-random, Halton, Sobol, and Gaussian quadrature methods in simulating experimental design efficiency when using normally distributed priors (simulating Bayesian D-efficiency of experimental designs).

<sup>6</sup> A somewhat similar stream of literature deals with the question of how many bootstrap draws should be used (e.g., [Davidson and MacKinnon, 2000](#)).

Last, even if using  $x$  QMC draws indeed performs just as good as 1,000 pseudo-random draws, it is not clear if 1,000 is sufficient for a desirably small simulation error. [Chiou and Walker \(2007\)](#) show that using too few draws can lead to spurious convergence of models that are theoretically or empirically unidentified. In the examples provided, even 1,000 Halton draws (which seems to be the state-of-practice in most applied studies) was not sufficient to uncover the problems. Similar conclusions are drawn by [Andersen \(2014\)](#) who shows that even using over 1,000 antithetic Halton draws can lead to differences in log likelihood which can interfere with the likelihood ratio based inference.

### 3. Design of our simulation study

We designed and executed a simulation study that is free of the shortcomings of earlier analyses of the performance of QMC methods in maximum likelihood estimation of discrete choice models.

We compared four types of draws: *pseudo-random*, modified Latin hypercube sampling (*MLHS*), randomized scrambled Halton sequence (*Halton*), and randomized scrambled Sobol sequence (*Sobol*).<sup>7</sup>

The comparison was made using datasets created using a mixed logit (MXL)<sup>8</sup> data generating process. We assumed the following utility function specification: individual  $i$ 's utility from choosing an alternative  $j$  in choice task  $k$  is:

$$U_{ijk} = \mathbf{X}_{ijk} \boldsymbol{\beta}_i + \varepsilon_{ijk},$$

where  $\mathbf{X}_{ijk}$  is a vector of alternative-specific attributes,  $\boldsymbol{\beta}_i$  is a vector of individual-specific random parameters, assumed to follow independent normal distributions and  $\varepsilon_{ijk}$  is the extreme value type I distributed random term ([McFadden, 1974](#)).

The datasets were designed to mimic typical discrete choice modelling problems, such as encountered in stated or revealed preference studies (e.g., [Carson and Czajkowski, 2014](#); [Hanley and Czajkowski, 2017](#)). Each choice task consisted of three alternatives, characterized using either five or ten attributes: one alternative specific constant, one discrete variable valued one to four, and three or eight dummy variables. This setting can be thought of as representing a choice between

---

<sup>7</sup> In what follows, we use italicized names of the types of draws to refer to the specific settings described here (e.g., randomized scrambled Halton or Sobol sequences).

<sup>8</sup> Random parameters (conditional) multinomial logit model ([Revelt and Train, 1998](#); [Greene, 2011](#)).

a status quo (or opt-out) alternative (associated with the alternative specific constant and serving as a 0-valued reference for all other attribute levels) and two “improvement” alternatives (e.g., new policies to be implemented), characterized by a discrete variable (e.g., representing the costs associated with the alternatives) and three or eight dummy variables (e.g., changes in various characteristics of a good in relation to the status quo).

The estimated models have five or ten uncorrelated random parameters.<sup>9</sup> We assumed mean values of these parameters are 1.0 for the dummy variables and -1.0 for the alternative specific constant and the “cost.” Standard deviations of random parameters for all these attributes were assumed to equal 0.5. Table 1 summarizes the choice task setting and the explanatory variables.

**Table 1. Summary of the choice task setting and explanatory variables**

Explanatory variables (choice attributes)	Assumed parameter distribution	Possible values of the explanatory variables		
		Alternative 1 (status quo / opt-out)	Alternative 2	Alternative 3
$X_1$ (alternative specific constant)	$N(-1.0, 0.5)$	$X_1 = 1$	$X_1 = 0$	$X_1 = 0$
$X_2$ (discrete)	$N(-1.0, 0.5)$	$X_2 = 0$	$X_2 \in \{1, 2, 3, 4\}$	$X_2 \in \{1, 2, 3, 4\}$
$X_{3..10}$ (dummy)	$N(1.0, 0.5)$	$X_{3..10} = 0$	$X_{3..10} \in \{0, 1\}$	$X_{3..10} \in \{0, 1\}$

The datasets used for comparisons varied with respect to the number of choice tasks per individual (4, 8, or 12), and with respect to the number of simulated individuals (400, 800, or 1,200). The choice tasks (combinations of attribute levels) were generated following three common methods encountered in the literature. They either used the so-called orthogonal optimal in the difference fractional factorial design (OOD-design; [Street, Burgess and Louviere, 2005](#); [Street and Burgess, 2007](#)), or the so-called efficient fractional factorial design ([Scarpa and Rose, 2008](#)), optimized for the MNL model (MNL-design) or for the MXL model (MXL-design).<sup>10,11,12</sup> Overall, the

<sup>9</sup> As an aside, we find no support for the requirement of having at least one parameter non-random for identification of the MXL model ([cf. Chiou and Walker, 2007](#)).

<sup>10</sup> The designs were generated in NGENE (ChoiceMetrics, Pty Ltd). Efficient designs were optimized for D-error (minimized the determinant of an asymptotic variance-covariance matrix, using the true parameter values as fixed priors).

<sup>11</sup> Even though our data generating process is MXL, MNL-designs are much more common in the literature, possibly because there is some evidence ([Bliemer and Rose, 2010](#)) indicating that the loss of efficiency from using them for a different model is relatively low. [Czajkowski and Budziński \(2016\)](#) show that this is not necessarily the case, particularly for NGENE generated designs using blocking.

<sup>12</sup> For the case of ten attributes we analyze the MXL-design only.



comparison was done using  $3 \cdot 3 \cdot 3 = 27$  datasets in the case of five attributes, and  $3 \cdot 3 = 9$  datasets in the case of ten attributes..

For each dataset, we estimated panel-data versions of MXL models using 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 *pseudo-random*, *MLHS*, *Halton* and *Sobol* draws. Additionally, for MXL-design and *Sobol* draws we estimated the models using 20,000, 50,000, 100,000 draws, and, in the case of MXL-design with five attributes, we additionally estimated the models using 200,000, 500,000 and 1,000,000 *Sobol* draws. The models for each setting were estimated 1,000 times (or 100 times in the case of additional MXL and *Sobol* specifications), each time using a new set of draws<sup>13</sup>, enabling us to conduct variance analysis of the log-likelihood at convergence, estimated coefficients, standard errors, and z-statistics. Each model used data generating process parameter values as the starting point, to facilitate convergence and avoid local maxima problems.<sup>14</sup> Table 2 summarizes the design of our simulation study.

**Table 2. Summary of the design of the simulation study**

Repetitions	Draws		Datasets			
	Types of draws	Number of draws	Number of attributes	Number of choice tasks per individual	Number of individuals	Experimental designs
1,000	<i>pseudo-random</i> <i>MLHS</i> <i>Halton</i> <i>Sobol</i>	100	5 10**	4 8 12	400 800 1,200	OOD-design MNL-design MXL-design
		200				
		500				
		1,000				
		2,000				
		5,000				
		10,000				
		20,000*				
		50,000*				
		100,000*				
		200,000*				
		500,000*				
		1,000,000*				

\*Selected settings only.

\*\* MXL-design only.

<sup>13</sup> *Pseudo-random* draws were generated using different random generator seeds; *MLHS*, *Halton* and *Sobol* draws differed in each repetition because of the random shift. Overall, this allows us to perform a proper analysis of variance associated with using a different number and type of draws (simulation error).

<sup>14</sup> The software codes for estimating the MXL model were developed in Matlab and are available at <http://github.com/czaj/DCE> under Creative Commons BY 4.0 license. The code and data for estimating the models presented in this paper are available from <http://czaj.org/research/supplementary-materials>.

#### 4. Methodology of the comparisons

To compare the estimates resulting from using different QMC methods, we will need a method that not only looks at their expected values, but also penalizes them for high variance. Consider the case where one tests if two random variables  $\omega_1$  and  $\omega_2$  have equal means (e.g., using the standard t-test). The larger the variance associated with  $\omega_1$  or  $\omega_2$ , the more difficult it is to reject the equality hypothesis. As a result, inferior simulation methods (resulting in lots of variation) could not be rejected in favor of better ones.

To address this problem we base our comparisons on equivalence tests ([Hauck and Anderson, 1984](#); [Kristofersson and Navrud, 2005](#)). Equivalence tests reverse the null hypothesis and the alternative hypothesis – instead of testing if  $\omega_1$  is equal to  $\omega_2$ , we test if the absolute difference between them is higher than an a priori defined “acceptable” level. [Czajkowski and Ščasný \(2010\)](#) and [Czajkowski et al. \(2017\)](#) operationalize equivalence tests by proposing to search for a Minimum Tolerance level ( $MTL$ ), i.e., the minimum “acceptable” difference that allows the conclusion that two values are equivalent at the required level of statistical significance.

Formally, for two random variables  $\omega_1$ ,  $\omega_2$   $MTL$  is defined as the minimum  $\theta \geq 0$  that satisfies:

$$P(|\omega_1 - \omega_2| > \theta) = \alpha, \quad (1)$$

where  $\alpha$  is the required significance level (e.g., 0.05). In our case, the probability can be evaluated using two one-sided convolutions tests ([TOSC; Poe, Giraud and Loomis, 2005](#))<sup>15</sup>, while  $MTL$  can be found as

$$MTL_\alpha = \underset{\theta \in [0, +\infty)}{\operatorname{argmin}} \left\{ \left| \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathbf{1}\{|\omega_{1,i} - \omega_{2,j}| > \theta\} - \alpha \right| \right\}, \quad (2)$$

where  $\omega_{1,i}$  and  $\omega_{2,j}$  are realizations of random variables,  $N_1$  and  $N_2$  are numbers of these realizations and  $\mathbf{1}\{\cdot\}$  is an indicator function equal to 1 if the condition in brackets is fulfilled.<sup>16</sup>

Re-estimating the model using a different set of random draws (e.g., a different seed for pseudo-random draws) is likely to result in a somewhat different value of the log-likelihood function. If one were to use these two values of the log-likelihood function for inference (e.g., conduct a likelihood-ratio test to compare two model specifications), it is important to note that they are

<sup>15</sup> Our  $\omega$  are not necessarily normally distributed. If they were, one could use two one-sided t-tests (TOST).

<sup>16</sup> A collection of Matlab functions useful for calculating  $MTL$  is available from <https://github.com/czaj/BTtools>.

known with uncertainty (simulation error). As a result, it is possible to conclude that, for example, one model specification is superior to another only because one was more “lucky” with the draws. By using the *MTL* approach, with  $\omega_1$  and  $\omega_2$  representing the random (known with uncertainty) values of the log-likelihood function for a given type and number of draws, we can say what the probability of such an outcome is. Assuming the usual significance level ( $\alpha=0.05$ ), the interpretation of the  $MTL_{0.05}$  is that, with 95% probability, using a different sets of draws (of the same number and type) would not cause the difference to be more than  $MTL_{0.05}$ . Conversely, one can provide recommendations regarding the minimum number of draws (of a particular type) that results in the *MTL* being lower than the required level (e.g., a critical value of the LR-test), so that the probability of erroneously concluding that one model is preferred to another (because of simulation error) is lower than a desired significance level, e.g., 0.05. In a similar way, variation of the estimates of coefficients, standard errors, and z-statistics can be compared.

## 5. Results

We now turn to presenting the results of the simulations. We first investigate the question of which type of draw performs best (and what is the relative performance of different QMC methods), and then attempt to provide recommendations with respect to how many draws are “enough” for a required precision level.

### 5.1. What type of draw performs best?

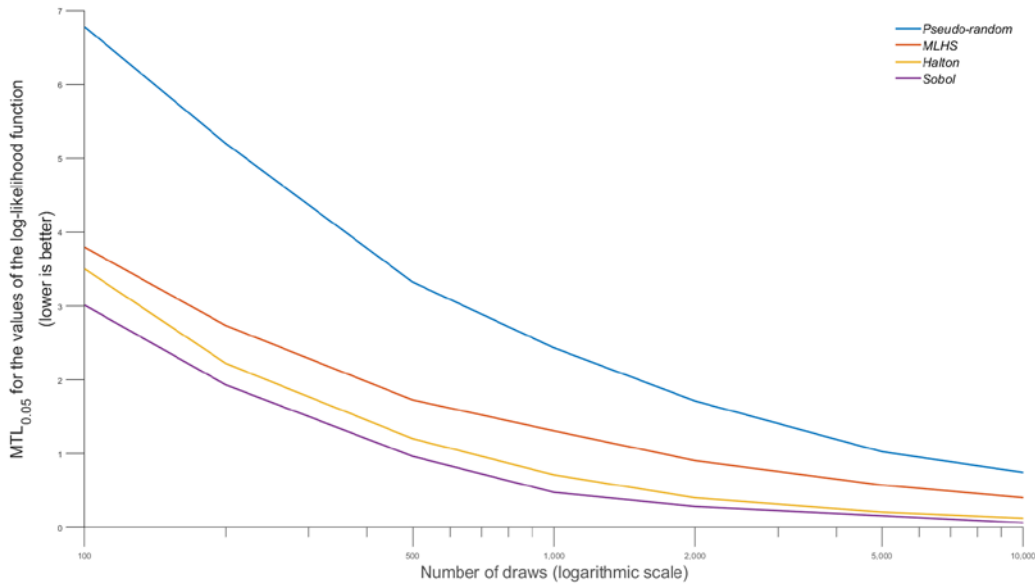
All the simulation methods considered result in unbiased estimates, and using more draws reduces simulation error. However, it is possible to compare the variation of the results of different simulation methods for the same number of draws, and this way conclude that using some of them results in more precise estimates than others.

Consider the variation of the values of the log-likelihood function at convergence (LL) first. It is an important measure, because LL is often used for comparing different model specifications by applying the likelihood ratio test, and hence high variation in LL values between models could lead to erroneous conclusions.

We calculated the  $MTL_{0.05}$  for the models estimated using a different number and type of draws. The general pattern was very evident and consistent. Irrespective of the experimental design, the number of choice tasks per individual and the number of individuals, using *Sobol* draws resulted in the lowest *MTL* (simulation error). Figure 1 provides an illustration for the case of an MXL-design

with 5 attributes, four choice tasks per individual and 400 individuals. As expected, increasing the number of draws results in decreasing  $MTL$  but *Sobol* draws always perform best, followed by *Halton*, *MLHS* and *pseudo-random*.

**Figure 1.**  $MTL_{0.05}$  for the values of the log-likelihood function simulated using a different number of draws of each type (MXL-design with 5 attributes, 4 choice tasks per individual, 400 individuals)



The pattern illustrated in Figure 1 is not unique to this dataset. To show this, Online Supplement C presents the detailed results – the percentage of cases where each type of draw performed the best, in terms of the lowest  $MTL_{0.05}$  for each number of draws. In the overwhelming majority of cases, *Sobol* draws were the best – they resulted in the lowest variation of the log-likelihood function value, parameter estimates, and z-statistics of estimated models.

Overall, we find that using *Sobol* draws results in the lowest simulation error of all the simulation methods compared in the majority of the considered cases, irrespective of whether one compares the variation in the value of log-likelihood function at convergence, parameter estimates, or their z-statistics. Although *Halton* draws are a close second, using them nonetheless results in a higher simulation error; apparently using draws designed for the best coverage in a multi-variate case (*Sobol*) outperforms draws designed for best coverage on a line only (*Halton*), despite state-of-the-art shuffling techniques.

To compare the relative performance of *Sobol* with other types of draws, we estimated regressions explaining logarithms of MTL for log-likelihood function values, parameter estimates, and z-statistics using various characteristics of experimental settings, such as the type of draws and logarithm of the number of draws. The results are provided in Table 3.

**Table 3. Variation of the simulation results ( $\log(MTL_{0.05})$ ) explained using characteristics of experimental settings**

Dependent variables		MTL for	MTL for	MTL for
Explanatory variables		log-likelihood	parameter estimates	z-statistics
Constant		2.7026*** (0.0773)	-0.9256*** (0.0373)	0.6553*** (0.0317)
Number of attributes is 10		0.0135 (0.1278)	0.3898*** (0.0497)	0.0946** (0.0422)
<i>Pseudo-random</i> draws ( <i>Sobol</i> used as a reference)	(5 attributes)	1.4568*** (0.0362)	0.8770*** (0.0177)	0.8360*** (0.0150)
<i>MLHS</i> draws ( <i>Sobol</i> used as a reference)	(5 attributes)	0.9017*** (0.0379)	0.6495*** (0.0185)	0.6142*** (0.0157)
<i>Halton</i> draws ( <i>Sobol</i> used as a reference)	(5 attributes)	0.3212*** (0.0379)	0.2173*** (0.0185)	0.2207*** (0.0157)
<i>Pseudo-random</i> draws ( <i>Sobol</i> used as a reference)	(10 attributes)	0.5613*** (0.0639)	0.2573*** (0.0221)	0.3061*** (0.0188)
<i>MLHS</i> draws ( <i>Sobol</i> used as a reference)	(10 attributes)	0.2666*** (0.0639)	0.1715*** (0.0221)	0.1960*** (0.0188)
<i>Halton</i> draws ( <i>Sobol</i> used as a reference)	(10 attributes)	-0.0027 (0.0639)	-0.0123 (0.0221)	0.0372** (0.0188)
log(number of draws)	(5 attributes)	-0.6330*** (0.0074)	-0.5786*** (0.0036)	-0.5635*** (0.0031)
log(number of draws)	(10 attributes)	-0.4828*** (0.0127)	-0.4385*** (0.0044)	-0.4517*** (0.0037)
Number of choice tasks		0.1356*** (0.0035)	-0.0502*** (0.0015)	0.0332*** (0.0013)
Number of individuals (in thousands)		0.8306*** (0.0350)	-0.5914*** (0.0153)	0.2653*** (0.0130)
OOD-design (MXL-design used as a reference)		-0.1273*** (0.0326)	0.3125*** (0.0159)	0.2446*** (0.0136)
MNL-design (MXL-design used as a reference)		-0.1501*** (0.0326)	0.3224*** (0.0159)	0.3556*** (0.0136)
Standard deviations (Means used as a reference)			1.3094*** (0.0100)	1.3688*** (0.0085)
$X_1$ (alternative specific constant)			0.6016*** (0.0141)	0.2992*** (0.0120)
$X_2$ (discrete variable)			-0.7445*** (0.0141)	0.0856*** (0.0120)
R <sup>2</sup>		0.9291	0.8471	0.8669
n (observations)		1095	13740	13740

1 We found that the log-log relationship between *MTL* and the number of draws was close to linear  
2 and resulted in the best fit of the model (with the  $R^2$  of approximately 0.9). The results show that,  
3 as expected, increasing the number of draws significantly reduces the simulation error for all  
4 analyzed measures (log-likelihood, parameter estimates, z-statistics).

5 Using *Halton*, *MLHS*, or *pseudo-random* draws results in increasingly higher variation in the results  
6 than using *Sobol* draws, as indicated by significantly positive and increasing coefficients associated  
7 with these types of draws, respectively. This is in line with the result presented above (see Figure 1  
8 and Online Supplement C). In the case of ten attributes, the differences between types of draws  
9 are smaller and the difference between *Halton* and *Sobol* draws is no longer statistically significant.<sup>17</sup>  
10 The effect of increasing the number of draws is also weaker in the case of ten attributes. This means  
11 that more draws are needed to decrease the simulation error by the same percentage as in the five  
12 attributes case.

13 Next, we find that increasing the number of observations, in terms of the number of choice tasks  
14 per individual and the number of individuals, leads to increasing variation of the log-likelihood  
15 function, *ceteris paribus*. Again, this is in line with the requirement of the number of draws  
16 increasing faster than the square root of the number of observations for the maximum simulated  
17 likelihood estimator to be consistent, efficient, and asymptotically equivalent to maximum  
18 likelihood ([Train, 2009](#)). On the other hand, increasing the number of observations reduces the  
19 variation of parameter estimates – with more observations parameter estimates are more stable,  
20 irrespective of the number of draws used. The variation of parameter estimates is the lowest for  
21 discrete variable ( $X_2$ ), for which a few levels are observed, followed by dummy coded variables  
22 ( $X_2 - X_4$  or  $X_2 - X_{10}$ ), and the highest for alternative specific constant ( $X_1$ ). Similarly, we find that  
23 the means of the random parameters are typically more precisely estimated than their standard  
24 deviations, which require using more draws for the same precision level. Finally, we observed that  
25 using the MXL-designs optimized for D-efficiency (i.e., minimizing the determinant of the  
26 asymptotic variance-covariance matrix) results in more precise estimates of parameter estimates  
27 and z-statistics, but not necessarily log-likelihood values, for which simulation error is lower if  
28 MNL-design or OOD-design is used.

---

<sup>17</sup> This was contrary to expectations, as Sobol draws are designed to deliver better coverage in the multi-dimension case while Halton draws are not. However, we note that both Halton and Sobol draws were scrambled, which improves their performance in the case of more dimensions.

Using the regression results allows us to estimate the relative increase in number of draws required to compensate for using *pseudo-random*, *MLHS* or *Halton* draws instead of *Sobol* draws. As a result of using the log-log relationship, this relative increase does not depend on the number of draws or other characteristics of the experimental setting. Specifically, we are looking for the number of draws,  $D^*$ , which renders the *MTL* that is equal to that resulting from a model estimated with  $D$  *Sobol* draws. This can be done by solving the following equation:

$$\exp(\alpha_{ij} + \beta_j \log(D^*) + \mathbf{V}) = \exp(\beta_j \log(D) + \mathbf{V}), \quad (3)$$

where  $\alpha_{ij}$  is a coefficient associated with  $i$ -th type of draw (see Table 3) and  $j$  attributes ( $j \in \{5, 10\}$ ), and  $\mathbf{V}$  collects all other effects from the regression. By substituting  $D^* = (1 + \lambda)D$  we can solve (3) for  $\lambda$ :

$$\lambda = \exp\left(-\frac{\alpha_{ij}}{\beta_j}\right) - 1. \quad (4)$$

The results are presented in Table 4, separately for five and ten attributes. The interpretation is straightforward – for example, achieving the same precision level of the log-likelihood function value, in the case of five attributes, as when using 1,000 *Sobol* draws requires using approximately 1,661 *Halton* draws, 4,155 *MLHS* draws or 9,987 *pseudo-random* draws. In the case of ten attributes, the percentage differences are lower. However, as we are about to show in the next section, this case requires using a larger number of draws for the desired precision, so the lower relative differences translate to large additional numbers of draws required when using pseudo-random or *MLHS* draws (the difference between *Sobol* and *Halton* draws is no longer statistically significant).

**Table 4. The relative increase in the number of draws required to achieve the same simulation error as when using *Sobol* draws (95% confidence intervals in [] brackets)**

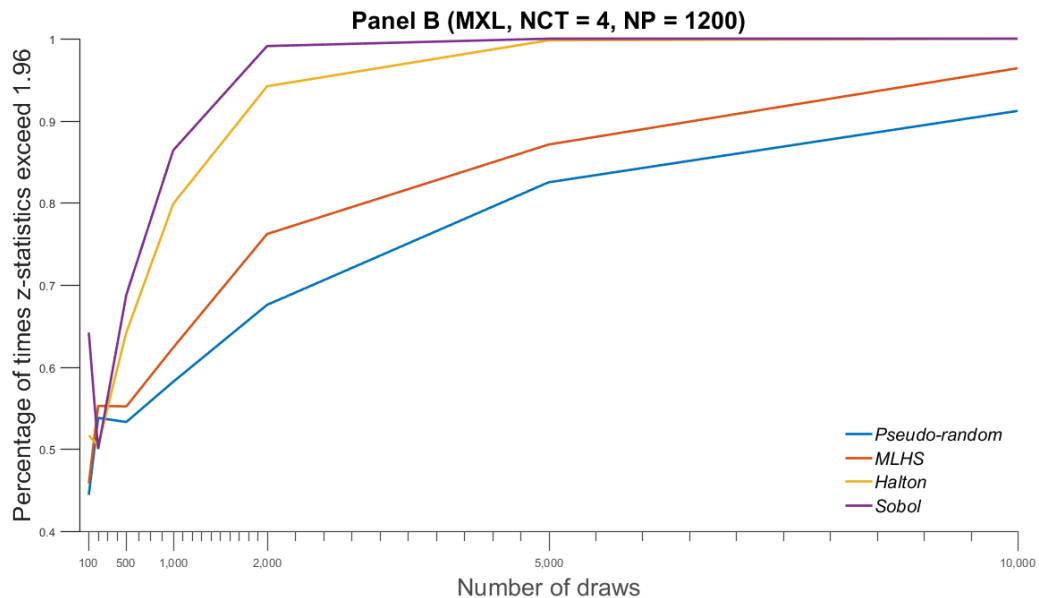
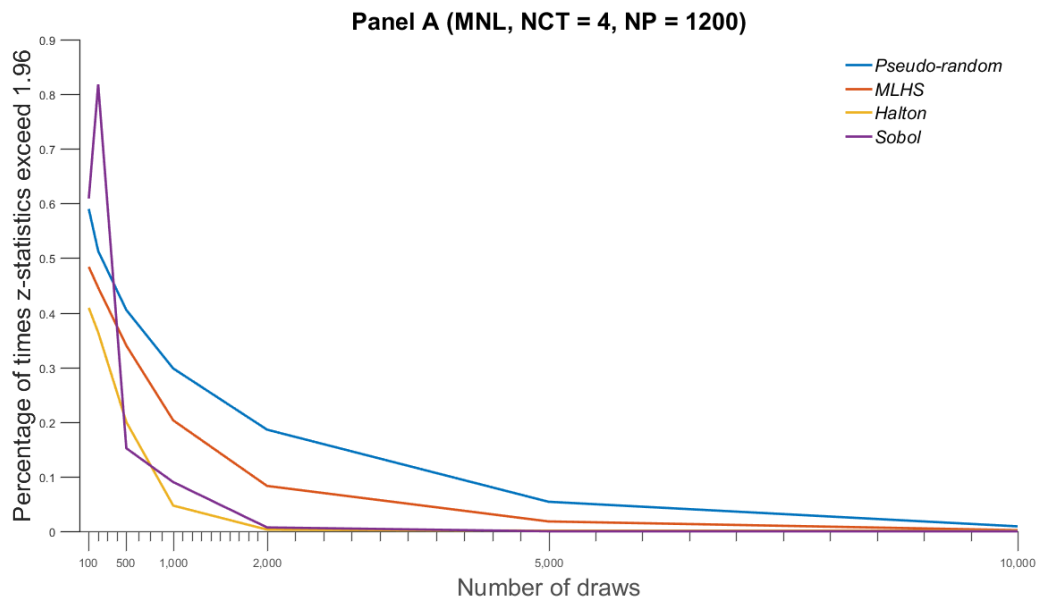
	Log-likelihood	Parameter estimates	z-statistics
<b>5 attributes</b>			
<i>Pseudo-random</i>	898.73% [783.67%-1028.95%]	355.25% [327.86%-384.74%]	340.89% [318.07%-366.29%]
<i>MLHS</i>	315.53% [266.10%-370.33%]	207.24% [187.36%-228.33%]	197.44% [181.03%-215.00%]
<i>Halton</i>	66.09% [47.52%-86.99%]	45.57% [36.67%-55.21%]	47.95% [40.11%-56.36%]
<b>10 attributes</b>			
<i>Pseudo-random</i>	219.87% [144.18%-324.35%]	79.81% [62.61%-98.97%]	96.94% [80.92%-114.19%]
<i>MLHS</i>	73.72% [33.54%-127.22%]	47.86% [33.56%-63.36%]	54.33% [41.88%-67.70%]
<i>Halton</i>	-0.56% [-23.10%-29.79%]	-2.77% [-11.81%-7.14%]	8.59% [0.17%-17.82%]

## 5.2. How many draws are “enough”?

Let us start by stating the obvious: using more draws is always better than using fewer draws. Not only will the estimates become more precise (lower simulation error) but as [Chiou and Walker \(2007\)](#) note, using too few draws can mask identification problems. An illustration of this effect is provided in Figure 2 – increasing the number of draws used for simulation can result in a substantial change in the percentage of cases where parameter estimates are statistically significant (i.e., their z-statistics exceed 1.96). Interestingly, too few draws can lead to erroneously concluding that an insignificant parameter is significant (panel A) or that a significant parameter appears insignificant (panel B). In the presented cases, *Sobol* draws are always the first to pick this up.



Figure 2. Percentage of times z-statistics exceed 1.96, corresponding to the usual 5% threshold for classifying a parameter as statistically significant (Panel A: the parameter for standard deviation of a binary variable ( $\chi_2$ ), MNL-design with 5 attributes, 4 choice tasks per individual, 1,200 individuals – more draws avoids spurious significance; Panel B: the parameter for standard deviation of a binary variable ( $\chi_3$ ), MXL-design with 5 attributes, 4 choice tasks per individual, 1,200 individuals – more draws avoids spurious insignificance)



The question of how many draws are “enough” depends on the required precision level. In the case of log-likelihood, we propose to use the measure based on the critical value of the likelihood ratio test that corresponds to comparing equivalent models estimated using a different set of draws. Even if the two models are the same, simulation error may make the values of the log-likelihood function at convergence differ.<sup>18</sup> If the difference in log-likelihoods exceeds 1.9207, which is  $\frac{1}{2}$  of the critical value of the likelihood ratio test with 1 degree of freedom, one would conclude that one model is statistically better than the other.<sup>19</sup> We therefore propose a measure of the number of draws required for a desired level of confidence (e.g., 95%) that such an erroneous conclusion is not reached for equivalent models (i.e., exactly the same models estimated with a different set of draws). This is exactly what the  $MTL_{0.05}$  can be used for – because  $MTL$  depends on the number of draws of each type, it is possible to find the number of draws that makes  $MTL \leq 1.9207$ .

Table 5 presents the estimated minimum number of draws required for the desired precision level – limiting the probability of simulation-driven error in the likelihood ratio test to 5% and limiting the simulation-driven differences between parameter estimates to 5%.<sup>20</sup> We focus on the analysis for *Sobol* draws because, as shown above, *Sobol* draws ensure the lowest simulation error of the QMC methods compared.<sup>21, 22</sup> In line with the regression results presented in Table 3, we find that as the number of observations increases, so do the absolute levels of log-likelihood, and the minimum number of draws for a required precision level. Conversely, in the case of parameter estimates, the reverse relationship is observed – increasing the number of observations reduces the number of draws required for a given precision level.<sup>23</sup> This is because with more observations (individuals and choice tasks per individual) parameter estimates generally become more precise

---

<sup>18</sup> Such a situation could arise because of, for example, using a new seed for *quasi-random* draws, different primes for generating *Halton* draws, or simply changing the order of random variables (and hence their association with generated draws).

<sup>19</sup> Note that comparing the results of the same model estimated using a different set of draws cannot formally be done using the likelihood ratio test, as these models are not nested and do not differ with respect to the number of parameters. Nonetheless, we use the critical value of the test with 1 degree of freedom as a natural reference. In practice, the difference between models compared using the likelihood ratio test would arise due to both the simulation error and the imposed restrictions.

<sup>20</sup> There is no absolute reference level of differences that could be considered acceptable, as in the case of the critical value of the likelihood ratio test for log-likelihoods. Instead we use the 5% differences in parameter estimates.

<sup>21</sup> The results of regressions for  $MTL_{0.05}$  limited to *Sobol* draws only are presented in Online Supplement D.

<sup>22</sup> In the case of ten attributes, *Sobol* and *Halton* tied for the lowest simulation error.

<sup>23</sup> Note that the minimum numbers of draws estimated here refer to all parameters in the model. In line with regression results presented in Table 3, precise estimates of the means require fewer draws than the estimates of standard deviations and estimates of discretely-valued variables require fewer draws than estimates of alternative specific constants.

(their variability decreases). As a result, although precise log-likelihood estimates require more draws for large datasets, precise parameter estimates require more draws when the dataset is small. Overall, for the experimental conditions used here, limiting the simulation-driven variation to the desired precision required up to 2,100 *Sobol* draws in the case of five attributes and up to 27,300 draws in the case of ten attributes. This is clearly much more than commonly used in empirical studies.

Finally, to verify the robustness of these results, we have repeated the analysis presented here, comparing the variation of log-likelihoods and parameter estimates associated with different numbers of *Sobol* draws to that resulting from using 100,000 *Sobol* draws. For example, instead of investigating how often the log-likelihoods of two models estimated using  $n$  draws can lead to erroneous conclusions in the likelihood ratio tests, we investigated how often log-likelihood of a model estimated using  $n$  draws can lead to erroneous conclusions in the likelihood ratio tests when compared with log-likelihood resulting from using 100,000 draws; this seemed like a number that greatly exceeds what is usually done in empirical studies. The results, presented in Online Supplement E, show that the minimum numbers of draws implied by the alternative approach are of similar magnitude to the ones presented here.

**Table 5. Minimum number of *Sobol* draws required for desired level of log-likelihood and parameter estimates precision (95% confidence intervals in [] brackets)<sup>26, 27</sup>**

Choice tasks per individual	4	4	4	8	8	8	12	12	12
Individuals	400	800	1,200	400	800	1,200	400	800	1,200
<b>5 attributes</b>									
≤5% probability of simulation-driven error in the LR test for 5 attributes <sup>28</sup>	148 [125-174]	230 [199-265]	357 [307-415]	363 [316-414]	563 [504-629]	874 [774-989]	889 [775-1,018]	1,380 [1,226-1,554]	2,142 [1,878-2,454]
≤5% probability that parameter estimates differ by ≥5% from true values for 5 attributes <sup>29</sup>	1,170 [1,061-1,288]	862 [786-946]	636 [575-702]	1,051 [959-1,150]	775 [710-844]	571 [520-627]	944 [856-1,039]	696 [634-764]	513 [464-566]
<b>Minimum recommended number of draws</b>	<b>1,170</b> <b>[1,061-1,288]</b>	<b>862</b> <b>[786-946]</b>	<b>636</b> <b>[575-702]</b>	<b>1,051</b> <b>[959-1,150]</b>	<b>775</b> <b>[710-844]</b>	<b>874</b> <b>[774-989]</b>	<b>889</b> <b>[775-1,018]</b>	<b>1,380</b> <b>[1,226-1,554]</b>	<b>2,142</b> <b>[1,878-2,454]</b>
<b>10 attributes</b>									
≤5% probability of simulation-driven error in the LR test for 10 attributes <sup>24</sup>	263 [193-346]	563 [439-708]	1,209 [945-1,528]	1,246 [1,003-1,529]	2,675 [2,257-3,160]	5,742 [4,698-7,101]	5,918 [4,667-7,509]	12,702 [10,191-16,052]	27,264 [20,889-36,562]
≤5% probability that parameter estimates differ by ≥5% from true values for 10 attributes <sup>27</sup>	25,294 [21,531-29,864]	15,251 [13,131-17,736]	9,196 [7,866-10,766]	21,174 [18,232-24,777]	12,767 [11,123-14,674]	7,698 [6674-8913]	17,725 [15,140-20,819]	10,688 [9,249-12,388]	6,444 [5,556-7,497]
<b>Minimum recommended number of draws</b>	<b>25,294</b> <b>[21,531-29,864]</b>	<b>15,251</b> <b>[13,131-17,736]</b>	<b>9,196</b> <b>[7,866-10,766]</b>	<b>21,174</b> <b>[18,232-24,777]</b>	<b>12,767</b> <b>[11,123-14,674]</b>	<b>7,698</b> <b>[6674-8913]</b>	<b>17,725</b> <b>[15,140-20,819]</b>	<b>12,702</b> <b>[10,191-16,052]</b>	<b>27,264</b> <b>[20,889-36,562]</b>

<sup>26</sup> Regression results presented in Online Supplement D.

<sup>27</sup> Online supplement F presents the results for other critical values of the log-likelihood test (corresponding to 0.01 and 0.1 significance levels), other acceptable differences in parameter estimates (1%, 10%), and other probability levels underlying the MTL (1%).

<sup>28</sup> At 0.05 significance level ( $MTL_{0.05}^{LL} \leq 1.9207$ ).

<sup>29</sup>  $MTL_{0.05}^{\beta} \leq 0.05|\beta|$ .

## 6. Summary and conclusions

In this study, we investigate the issue of simulation error resulting from using the simulated maximum likelihood method to estimate discrete choice models. We show that the simulation bias is not negligible, and the number of draws used by many empirical applications is too low for reliable inference.

We compare the performance of *pseudo-random* draws with three quasi Monte Carlo methods (*Halton*, *Sobol* and *modified Latin hypercube sampling*) under 27 experimental conditions that differ with respect to experimental design, number of individuals and number of choice tasks per individual. Based on a Monte Carlo simulation using 100 to 1,000,000 draws, we can compare the relative efficiency of different types of draws. We consistently find that a scrambled *Sobol* sequence performs the best in terms of the lowest simulation error, while being matched by scrambled Halton draws in the case of 10 attributes.

We propose a measure of sufficient simulation precision based on the likelihood that the results of different simulations in the same conditions will be statistically different. Our results show that, at the 95% confidence level, assuring that the simulation-driven errors in the likelihood ratio test do not take place and that average deviations of parameter estimates do not exceed 5% of their true values requires using over 2,000 *Sobol* draws in the case of 5-attribute design and over 25,000 *Sobol* draws in the case of 10-attribute design.<sup>30</sup> In some cases, one can get away with using fewer draws; however, we note that as the number of draws required for the precision of log-likelihoods and the number of draws needed for the precision of parameter estimates are negatively correlated, and researchers are likely interested in satisfying both criteria, the maximum of the numbers required for satisfying both criteria (log-likelihood and parameter estimates precision) may be appropriate. In our experiments, the minimum number of draws required for “reliable” estimates was larger than those used in most empirical studies.

Despite the common expectation, using thousands or tens of thousands of draws is not necessarily prohibitively time consuming. Our results show that with efficient code implementation (Matlab, <https://github.com/czaj/dce>)<sup>31</sup> and using a regular modern desktop computer (Intel E5-2687W @ 3.00 GHz, no GPU support, 128 GB RAM @ 2800 MHz) the computation time of one iteration (evaluation of the log-likelihood function and gradient) was 1 second for 10,000 draws, 10 seconds

---

<sup>30</sup> This number refer to the most demanding conditions in our experimental design; Table 5 provides more detailed results.

<sup>31</sup> In our simulation, similar implementation in R was approximately 5-10 times slower, Python Biogeme – approximately 20 times slower, NLOGIT – 60 times slower and Stata – over 100 times slower (see [Czajkowski, Buczyński and Budziński \(2018\)](#) for details).

1 for 100,000, and 100 seconds for 1,000,000 draws.<sup>32</sup> Given the advances in computing power, using  
2 10,000 draws today takes less computer time than using 100 draws took back when [Bhat \(2001\)](#)  
3 and [Train \(2000\)](#) did their analyses.<sup>33</sup> As a result, 10,000 draws is not as onerous as it might at first  
4 seem. Even the most complicated models can be estimated in a reasonable amount of time using  
5 many more draws than are commonly used.

6 Regarding other limitations of our study, we note that the results presented here are specific to the  
7 experimental setting we used – 400 to 1,200 individuals, 4 to 12 choice tasks per individual, five or  
8 ten normally distributed parameters, MXL model without correlations, and other conditions in the  
9 specific setting of the simulation.<sup>34</sup> We expect that datasets with more observations require using  
10 more draws than found here for the precision of log-likelihoods, while datasets with fewer  
11 observations require more draws for the precision of parameter estimates. The results are also  
12 limited to the specific setting of the MXL model, such as parameter values assumed in the data  
13 generating process. Larger standard deviations relative to the means (wider distributions) would  
14 likely require more draws. Similarly, we expect that estimating MXL models that account for  
15 correlations, are estimated in WTP-space ([Train and Weeks, 2005](#)), use non-normal distributions  
16 ([Train and Sonnier, 2005](#)), or use random parameters in the latent class model ([Greene and](#)  
17 [Hensher, 2012](#)) or the hybrid choice model ([Ben-Akiva et al., 2002](#)) setting are likely to require  
18 more draws. Note that by using the data generating process coefficients as starting values we also  
19 avoided problems of convergence to local maxima. In practice, when less optimal starting values  
20 are used, the expected variation in simulated log-likelihood values and parameter estimates can be  
21 expected to be even larger.<sup>35</sup> Lastly, our comparison did not include a few other potentially well-  
22 performing types of draws, such as lattice, Faure, Gaussian quadrature, Neiderreiter, and  
23 Neiderreiter-Xing nets.

24 Finally, we note that in parallel to the simulated maximum likelihood, other estimation methods  
25 have been developed. Examples include using a Bayesian framework ([Train and Sonnier, 2005](#)),

---

<sup>32</sup> The times are given for the dataset with 400 individuals, four choice tasks per individual, 5 attributes.

<sup>33</sup> [Bhat \(2001\)](#) reports that his model with 5 random parameters and 100 Halton draws converged in approximately 48 minutes (Intel Pentium II @ 300 MHz). [Hess, Polak and Daly \(2003\)](#) use a model with 4 random variables and the same number of draws which makes one iteration in approximately 1 second (Intel Pentium III @ 2.0 GHz).

<sup>34</sup> In addition, our analysis differs from a standard Monte Carlo experiment, in which the dataset would be regenerated by taking new draws from the error term and the parameters in each estimation. Instead, we generated datasets for each setting once, and estimated the models many times using different sets of draws. This way we hold sampling error constant and focus on investigating simulation error. However, we acknowledge that this also makes our experiment specific to the particular draws of the error term and parameter vector used to generate the datasets.

<sup>35</sup> There is some evidence indicating that increasing the number of draws smoothens the simulated log-likelihood function and hence facilitates convergence ([Tuhkanen et al., 2016](#)).

1 expectation-maximization algorithm ([Train, 2007](#)), Laplace approximation ([Harding and Hausman,](#)  
2 [2007](#)), or maximum approximate composite marginal likelihood ([Bhat and Sidharthan, 2011](#)). A yet  
3 another strand of literature approaches the problem from a different perspective, by trying to utilize  
4 non-parametric or semi-parametric approaches (instead of MXL) to model preference  
5 distributions. Examples include linear regression approximation ([Bajari, Fox and Ryan, 2007](#)),  
6 approximation of a density function based on Legendre polynomials ([Fosgerau and Bierlaire, 2007](#)),  
7 using B-splines to approximate the CDF function of the true distribution ([Bastin, Cirillo and Toint,](#)  
8 [2010](#)), using polynomials of draws taken from some chosen distribution (i.e., normal or log-normal)  
9 for approximating the true distribution ([Fosgerau and Mabit, 2013](#)), and most recently, the logit-  
10 mixed logit model ([Train, 2016](#)). Some of these methods may avoid the necessity to simulate  
11 multidimensional integrals and thus avoid simulation error, possibly trading it for other  
12 approximation biases.

13 There are three main takeaway messages from our study. The first is that *Sobol* draws outperformed  
14 *Halton*, *modified Latin hypercube sampling*, and *pseudo-random* draws in our experimental settings.  
15 Secondly, using too few draws can lead to substantial bias in log-likelihood values, parameter  
16 estimates, and standard errors (p-values). Third, while the number of *Sobol* draws required for the  
17 desired precision depends on the number of observations in the case of experimental designs with  
18 5 attributes, using over 2,000 *Sobol* draws resulted in 95% confidence that log-likelihoods do not  
19 lead to simulation-driven erroneous inference and that parameter estimates are within 5% of their  
20 true values for all experimental settings considered. In the case of 10 attributes, over 20,000 *Sobol*  
21 draws were needed to meet these targets in all considered settings.

## References

- Andersen, L. M., 2014. Obtaining Reliable Likelihood Ratio Tests from Simulated Likelihood Functions. *PLoS ONE*, 9(10):e106136.
- Antonov, I. A., and Saleev, V. M., 1979. An economic method of computing LP $\tau$ -sequences. *USSR Computational Mathematics and Mathematical Physics*, 19(1):252-256.
- Bajari, P., Fox, J. T., and Ryan, S. P., 2007. Linear regression estimation of discrete choice models with nonparametric distributions of random coefficients. *The American Economic Review*:459-463.
- Bastin, F., Cirillo, C., and Toint, P. L., 2010. Estimating Nonparametric Random Utility Models with an Application to the Value of Time in Heterogeneous Populations. *Transportation Science*, 44(4):537-549.
- Ben-Akiva, M., and Lerman, S. R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA.
- Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., and Polydoropoulou, A., 2002. Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges*:431-470.
- Bhat, C. R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7):677-693.
- Bhat, C. R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37(9):837-855.
- Bhat, C. R., and Sidharthan, R., 2011. A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. *Transportation Research Part B: Methodological*, 45(7):940-953.
- Bliemer, M. C. J., and Rose, J. M., 2010. Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B: Methodological*, 44(6):720-734.



- 1 Bliemer, M. C. J., Rose, J. M., and Hess, S., 2008. Approximation of Bayesian Efficiency in  
2 Experimental Choice Designs. *Journal of Choice Modelling*, 1(1):98-127.
- 3 Braaten, E., and Weller, G., 1979. An improved low-discrepancy sequence for multidimensional  
4 quasi-Monte Carlo integration. *Journal of Computational Physics*, 33(2):249-258.
- 5 Bratley, P., and Fox, B. L., 1988. Algorithm 659: Implementing Sobol's quasirandom sequence  
6 generator. *ACM Transactions on Mathematical Software (TOMS)*, 14(1):88-100.
- 7 Carson, R. T., and Czajkowski, M., 2014. The Discrete Choice Experiment Approach to  
8 Environmental Contingent Valuation. In: *Handbook of choice modelling*, S. Hess and A. Daly,  
9 eds., Edward Elgar, Northampton, MA.
- 10 Chiou, L., and Walker, J. L., 2007. Masking identification of discrete choice models under  
11 simulation methods. *Journal of Econometrics*, 141(2):683-703.
- 12 Czajkowski, M., Ahtiainen, H., Artell, J., and Meyerhoff, J., 2017. Choosing a Functional Form for  
13 an International Benefit Transfer: Evidence from a Nine-country Valuation Experiment.  
14 *Ecological Economics*, 134:104-113.
- 15 Czajkowski, M., Buczyński, M., and Budziński, W., 2018. Replicability, simulation error and  
16 robustness to non-parametric treatment of preference heterogeneity in discrete choice  
17 models. The 25'th Ulvön Conference on Environmental Economics, 2018-06-20, Ulvön.
- 18 Czajkowski, M., and Budziński, W., 2016. Choice task blocking and design efficiency. Paper  
19 presented at the 5'th Workshop on Discrete Choice Modelling, Warsaw, available from  
20 [http://czaj.org/pub/presentations/Czajkowski\\_2016-10-06b.pdf](http://czaj.org/pub/presentations/Czajkowski_2016-10-06b.pdf).
- 21 Czajkowski, M., and Ščasný, M., 2010. Study on benefit transfer in an international setting. How  
22 to improve welfare estimates in the case of the countries' income heterogeneity? *Ecological*  
23 *Economics*, 69(12):2409-2416.
- 24 Davidson, R., and MacKinnon, J. G., 2000. Bootstrap tests: how many bootstraps? *Econometric*  
25 *Reviews*, 19(1):55-68.
- 26 Fosgerau, M., and Bierlaire, M., 2007. A practical test for the choice of mixing distribution in  
27 discrete choice models. *Transportation Research Part B: Methodological*, 41(7):784-794.

1 Fosgerau, M., and Mabit, S. L., 2013. Easy and flexible mixture distributions. *Economics Letters*,  
2 120(2):206-210.

3 Garrido, R. A., Year. Estimation performance of low discrepancy sequences in stated preferences.  
4 10th International Conference on Travel Behaviour Research., Citeseer.

5 Greene, W. H., 2011. *Econometric Analysis*. 7 Ed., Prentice Hall, Upper Saddle River, NJ.

6 Greene, W. H., and Hensher, D. A., 2012. Revealing additional dimensions of preference  
7 heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*,  
8 45(14):1897-1902.

9 Halton, J. H., 1960. On the efficiency of certain quasi-random sequences of points in evaluating  
10 multi-dimensional integrals. *Numerische Mathematik*, 2(1):84-90.

11 Hanley, N., and Czajkowski, M., 2017. Stated Preference valuation methods: an evolving tool for  
12 understanding choices and informing policy. University of Warsaw, Department of  
13 Economics Working Paper 1(230).

14 Harding, M. C., and Hausman, J., 2007. Using a Laplace approximation to estimate the random  
15 coefficients logit model by nonlinear least squares. *International Economic Review*, 48(4):1311-  
16 1328.

17 Hauck, W. W., and Anderson, S., 1984. A new statistical procedure for testing equivalence in two-  
18 group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*,  
19 12(1):83-91.

20 Hess, S., Polak, J. W., and Daly, A., Year. On the performance of the shuffled Halton sequence in  
21 the estimation of discrete choice models. European Transport Conference, Strasbourg,  
22 Citeseer.

23 Hess, S., Train, K. E., and Polak, J. W., 2006. On the use of a Modified Latin Hypercube Sampling  
24 (MLHS) method in the estimation of a Mixed Logit Model for vehicle choice. *Transportation*  
25 *Research Part B: Methodological*, 40(2):147-163.

26 Kocis, L., and Whiten, W. J., 1997. Computational investigations of low-discrepancy sequences.  
27 *ACM Transactions on Mathematical Software (TOMS)*, 23(2):266-294.

1 Kristofersson, D., and Navrud, S., 2005. Validity Tests of Benefit Transfer – Are We Performing  
2 the Wrong Tests? *Environmental and Resource Economics*, 30(3):279-286.

3 Lemieux, C., 2009. Monte carlo and quasi-monte carlo sampling. Springer Science & Business  
4 Media.

5 Matoušek, J., 1998. On the L<sub>2</sub>-Discrepancy for Anchored Boxes. *Journal of Complexity*, 14(4):527-  
6 556.

7 McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behaviour. In: *Frontiers in*  
8 *Econometrics*, P. Zarembka, ed., Academic Press, New York, NY, 105-142.

9 McFadden, D., and Train, K., 2000. Mixed MNL Models for Discrete Response. *Journal of Applied*  
10 *Econometrics*, 15(5):447-470.

11 Munger, D., L'Ecuyer, P., Bastin, F., Cirillo, C., and Tuffin, B., 2012. Estimation of the mixed logit  
12 likelihood function by randomized quasi-Monte Carlo. *Transportation Research Part B:*  
13 *Methodological*, 46(2):305-320.

14 Pirsic, G., Year. A Software Implementation of Niederreiter-Xing Sequences. Monte Carlo and  
15 Quasi-Monte Carlo Methods 2000, Springer Berlin Heidelberg, Berlin, Heidelberg, 434-  
16 445.

17 Poe, G. L., Giraud, K. L., and Loomis, J. B., 2005. Computational Methods for Measuring the  
18 Difference of Empirical Distributions. *American Journal of Agricultural Economics*, 87(2):353-  
19 365.

20 Revelt, D., and Train, K., 1998. Mixed Logit with Repeated Choices: Households' Choices of  
21 Appliance Efficiency Level. *Review of Economics and Statistics*, 80(4):647-657.

22 Sándor, Z., and Train, K., 2004. Quasi-random simulation of discrete choice models. *Transportation*  
23 *Research Part B: Methodological*, 38(4):313-327.

24 Scarpa, R., and Rose, J. M., 2008. Design Efficiency for Non-Market Valuation with Choice  
25 Modelling: How to Measure it, What to Report and Why. *Australian Journal of Agricultural*  
26 *and Resource Economics*, 52(3):253-282.

- 1 Sidharthan, R., and Srinivasan, K., 2010. Random Coefficient Mixed Logit Models Based on  
2 Generalized Antithetic Halton Draws and Double Base Shuffling. *Transportation Research*  
3 *Record: Journal of the Transportation Research Board*, (2175):1-9.
- 4 Sivakumar, A., Bhat, C., and Ökten, G., 2005. Simulation Estimation of Mixed Discrete Choice  
5 Models with the Use of Randomized Quasi-Monte Carlo Sequences: A Comparative Study.  
6 *Transportation Research Record: Journal of the Transportation Research Board*, 1921:112-122.
- 7 Sobol, I. M., 1967. On the distribution of points in a cube and the approximate evaluation of  
8 integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784-802.
- 9 Stein, M., 1987. Large Sample Properties of Simulations Using Latin Hypercube Sampling.  
10 *Technometrics*, 29(2):143-151.
- 11 Street, D. J., and Burgess, L., 2007. The Construction of Optimal Stated Choice Experiments:  
12 Theory and Methods. Wiley-Interscience, Hoboken, NJ.
- 13 Street, D. J., Burgess, L., and Louviere, J. J., 2005. Quick and easy choice sets: Constructing optimal  
14 and nearly optimal stated choice experiments. *International Journal of Research in Marketing*,  
15 22(4):459-470.
- 16 Train, K., 2000. Halton sequences for mixed logit. *Department of Economics, UCB*.
- 17 Train, K., 2007. A recursive estimator for random coefficient models. *University of California, Berkeley*.
- 18 Train, K., 2016. Mixed logit with a flexible mixing distribution. *Journal of Choice Modelling*, 19:40-53.
- 19 Train, K., and Sonnier, G., 2005. Mixed Logit with Bounded Distributions of Correlated  
20 Partworths. In: *Applications of Simulation Methods in Environmental and Resource Economics*, R.  
21 Scarpa and A. Alberini, eds., Springer Netherlands, 117-134.
- 22 Train, K., and Weeks, M., 2005. Discrete Choice Models in Preference Space and Willingness-to-  
23 Pay Space. In: *Applications of Simulation Methods in Environmental and Resource Economics*, R.  
24 Scarpa and A. Alberini, eds., Springer Netherlands, 1-16.
- 25 Train, K. E., 2009. Discrete Choice Methods with Simulation. 2 Ed., Cambridge University Press,  
26 New York.

- 1 Tuhkanen, H., Piirsalu, E., Nõmmann, T., Karlõševa, A., Nõmmann, S., Czajkowski, M., and  
2 Hanley, N., 2016. Valuing the benefits of improved marine environmental quality under  
3 multiple stressors. *Science of The Total Environment*, 551-552:367-375.
- 4 Wang, X., and Kockelman, K., Year. Maximum Simulated Likelihood Estimation with Spatially  
5 Correlated Observations: A Comparison of Simulation Techniques. RSAI's 53rd Annual  
6 Meeting, Toronto (in 2006), and forthcoming in *Transportation Statistics*.

7

## Online Supplement A

This supplement presents the rationale and the description of algorithms used by the quasi Monte Carlo methods compared in our study.

### Halton sequence

The QMC method that is currently the most commonly used for simulating the log-likelihood function of discrete choice models uses the Halton sequence ([Halton, 1960](#)). Following [Kocis and Whiten \(1997\)](#), the  $n$ -th element of the Halton sequence generated with a base  $b_j$ <sup>36</sup> is given by the so called radical inverse function  $\Phi_{b_j}(n)$  defined as follows:

$$\Phi_{b_j}(n) = \sum_{i=0}^{\infty} \alpha_i(j, n) b_j^{-i-1}, \quad (5)$$

where  $\alpha_i(j, n) \in [0, b_j)$  and it is an integer obtained from digit expansion of  $n$  in base  $b_j$ :

$$n = \sum_{i=0}^{\infty} \alpha_i(j, n) b_j^i. \quad (6)$$

The  $K$ -dimensional Halton sequence is given simply by  $K$  one-dimensional Halton sequences generated with different bases (most often  $K$  first prime numbers):

$$x_n = (\Phi_{b_1}(n), \dots, \Phi_{b_K}(n)). \quad (7)$$

The drawback of the Halton sequence is a high correlation between sequences generated using high prime numbers (see Online Supplement B for illustration). This translates into poor performance in evaluating higher dimensional integrals. The way to address this problem is to use so called *scrambling*; in other words, apply a generalized radical inverse function:

$$\Phi_{b_j}(n) = \sum_{i=0}^{\infty} \sigma(\alpha_i(j, n)) b_j^{-i-1}, \quad (8)$$

where  $\sigma(\cdot)$  is an operator of permutations on  $\alpha_i(j, n)$ . Different choices for  $\sigma$  are proposed in the literature (e.g., [Braaten and Weller, 1979](#)). We applied the reverse Radix algorithm ([Kocis and Whiten, 1997](#)).

---

<sup>36</sup> Most often  $b_j$  is some prime number.

The idea of the reverse Radix algorithm is as follows: given the representation of  $\alpha_i(j, n)$  in base 2, the fixed number of its digits are reversed (this means that the Halton sequence in base 2 and scrambled Halton sequence in base 2 are the same). Values that are too large are removed from the sequence.

The last thing to describe is randomization of the sequence. Proposed scrambling is still purely deterministic, so to include some randomness and be able to analyze the variance of the sequence, we applied the so-called *random shift*. When estimating mixed logit,  $N \cdot K$  sequences of length  $R$  have to be generated.<sup>37</sup> Instead, we generate only  $K$  sequences of the length  $N \cdot R$  and divide it into  $N$  parts. Properties of the Halton sequence assure that these sub-sequences still have a good coverage on a unit cube. We apply the following random shifting:

$$x_{jnk} = \{ \varepsilon_{jnk} + u_{nk} \}, \quad (9)$$

where  $\varepsilon_{jnk}$  is an original scrambled Halton draw ( $j \in \{1, \dots, R\}$ ,  $n \in \{1, \dots, N\}$ ,  $k \in \{1, \dots, K\}$ ),  $u_{nk}$  is a standard uniform draw and  $\{ \}$  is a fractional part function. We also tried a different type of random shifting of the following form:

$$x_{jnk} = \{ \varepsilon_{jnk} + u_k \}, \quad (10)$$

which differs, as now uniform draws are the same for different individuals (but different for different attributes). Our initial simulation revealed that the shifting in (9) performed better, so we decided to use this type only.

## Sobol Sequence

The Sobol sequence ([Sobol, 1967](#)) is a *so-called* (t,s)-sequence. To explain the idea behind (t,s)-sequences, we are going to first introduce (t,m,s)-nets. While the Halton sequence aims at obtaining a uniform coverage of  $[0,1]$ , and a multidimensional sequence is created by taking many such sequences generated with different bases, the (t,s)-sequences use only one base number and the multidimensional sequence is obtained by applying different generating matrices to different dimensions. Following [Lemieux \(2009\)](#) and [Bratley and Fox \(1988\)](#), let  $\alpha_i(j, n)$  from equation (6) be transformed in the following way for the  $k$ -th dimension:

---

<sup>37</sup>  $N$  is the number of respondents,  $K$  is the number of random parameters,  $R$  is the desired number of draws.

$$1 \quad (\tilde{\alpha}_0^k(j,n), \tilde{\alpha}_1^k(j,n), \dots)^T = C_k \cdot (\alpha_0(j,n), \alpha_1(j,n), \dots)^T, \quad (11)$$

2 where  $C_k$  is what we call a generation matrix.<sup>38</sup> Then the  $n$ -th element in the  $k$ -th dimension of  
 3 this sequence is given by:

$$4 \quad x_{n-1,k} = \sum_{i=0}^{\infty} \tilde{\alpha}_i^k(j, n-1) b_j^{-i-1}, \quad (12)$$

5 which is almost identical to the inverse radical function in (5). As can be seen, the choice of these  
 6 generation matrices plays a key role. We describe the process of generating them below.

7 Formally defining the (t,m,s)-nets requires one more definition. We are going to say that the point  
 8 set of length  $b_j^m$  is  $(q_1, \dots, q_s)$ -*equidistributed* in base  $b_j$ , if every cell of the form:

$$9 \quad J(\mathbf{r}) = \prod_{k=1}^s \left[ \frac{r_k}{b_j^{q_k}}, \frac{r_k+1}{b_j^{q_k}} \right) \quad (13)$$

10 contains  $b_j^{m-q}$  points of this point set, where  $q = q_1 + \dots + q_s$ , and  $r_k$  are any integers such that  
 11  $0 \leq r_k < b_j^{q_k}$ . Then (t,m,s)-nets in base  $b_j$  can be defined as a sequence of length  $b_j^m$  which is  
 12  $(q_1, \dots, q_s)$ -*equidistributed* whenever  $q \leq m-t$  ([Lemieux, 2009](#)).

13 For an illustration, consider a (0,2,2)-net in base 2, which is a 4-point sequence in two dimensional  
 14 space. The choice of  $(q_1, q_2)$  can be only (0,0), (1,0), (0,1) and (1,1). For the (0,0) case  $J(\mathbf{r})$   
 15 can be only a unit square, so the (0,0)-*equidistribution* condition says that all four points of this  
 16 sequence are in this square (which is true for any sequence). In the (1,0) case,  $J(\mathbf{r})$  can be  
 17  $[0, 1/2) \times [0, 1)$  or  $[1/2, 1) \times [0, 1)$ , so this condition says that in every such horizontal rectangle, two  
 18 points of sequence are placed. The condition of (1,1)-*equidistribution* indicates that in every interval  
 19 of the form  $[i/2, (i+1)/2) \times [j/2, (j+1)/2)$  where  $i, j \in \{0, 1\}$ , one point of the sequence is placed.<sup>39</sup>  
 20 As a result, this sequence has the best coverage one can expect from a 4-point long sequence.

21 Having the definition of (t,m,s)-nets we can simply define a (t,s)-sequence as a sequence for which  
 22 every subsequence  $\mathbf{x}_{l \cdot b_j^h}, \dots, \mathbf{x}_{(l+1) \cdot b_j^h}$  is a (t,h,s)-net. In particular, this means that the first  $b_j^h$  points  
 23 of the (t,s)-sequence are (t,h,s)-net.

---

<sup>38</sup>  $C_k$  elements  $\in \mathbb{Z}_{b_j}$ ; matrix multiplication on the righthand side is also in  $\mathbb{Z}_{b_j}$

<sup>39</sup> These intervals are just squares emerged from partitioning of a unit square in four parts.



As in case of the Halton, scrambling techniques can improve performance of the Sobol sequence. For (t,s)-sequences, it is a more difficult task, however, because we would like the scrambled sequence to possess the properties of the original sequence.

One way of scrambling Sobol sequences is to apply a random linear scramble combined with a random digital shift (Matoušek, 1998). Random digital shift is like the random shift described for the Halton sequence. For a draw from the Sobol sequence  $x_n^k$ , which can be presented in the form of a binary digit expansion  $x_n^k = \sum_{i=0} b_i \cdot 2^{-i}$ , and a draw from a standard uniform distribution  $u^k = \sum_{i=0} u_i^k \cdot 2^{-i}$ , also presented in binary form, the new draw is created by setting:

$$\tilde{x}_n^k = \sum_{i=0} (b_i + u_i^k) \cdot 2^{-i}, \quad (14)$$

where addition is done in  $\mathbb{Z}_2$ .

The random linear scramble is done by using generation matrices of form  $R_k \cdot C_k$  instead of simple  $C_k$ , where  $R_k$  is a lower-triangular non-singular matrix and matrix multiplication is done in  $\mathbb{Z}_2$ . This is called a linear scramble, as the  $n$ -th draw after scrambling is a linear function of  $n$  first draws in original sequence. Both linear scrambling and a random linear digit shift keep  $(q_1, \dots, q_s)$ -*equidistribution* property of a sequence and, what is more, the scrambling can lower the t-value of a (t,s)-sequence.<sup>40</sup>

The last thing described here is the process of generating the matrices to create sequences with the required properties. Sobol (1967) proposed to create the matrices with  $b_j = 2$ , which we applied in our study. To create the  $k$ -th generation matrix, we need to first define a primitive polynomial in  $\mathbb{Z}_2$  of form:

$$p_k(z) = z^{d_k} + a_{k,1}z^{d_k-1} + \dots + a_{k,d_k} \quad (15)$$

Second, we need  $d_k$  (which is a degree of  $p_k(z)$ ) direction numbers:

$$v_{k,r} = \frac{m_{k,r}}{2^r}, \quad (16)$$

where  $m_{k,r}$  is an odd integer  $\in [1, 2^r - 1]$  and  $v_{k,r}$  are written in binary digit expansion. The generation matrix  $C_k$  is created by setting its columns to these direction numbers presented in

---

<sup>40</sup> Which implies a better coverage.

vector forms.<sup>41</sup> To obtain direction numbers with indices greater than  $d_k$ , the following recursive procedure can be applied:

$$v_{k,r} = a_{k,1}v_{k,r-1} \oplus \dots \oplus a_{k,d_k}v_{k,r-d_k} \oplus (v_{k,r-d_k}/2^{d_k}), \quad (17)$$

where  $\oplus$  is an exclusive or logical function and  $a_{k,i}$  are taken from  $p_k(z)$  polynomials.

Consider an example from [Lemieux \(2009\)](#): in order to generate  $C_3$  we set  $p_3(z) = z^2 + z + 1$  and choose  $v_{3,1}$ ,  $v_{3,2}$  to be 0.5 and 0.75, respectively, which is 0.1 and 0.11 in binary expansion. According to (17) we have:

$$v_{3,3} = (1,1,0)^T \oplus (1,0,0)^T \oplus (0,0,1)^T = (0,1,1)^T. \quad (18)$$

This way we obtained the first three columns of  $C_3$ . To obtain further columns, (17) has to be applied again.

Presentation of a Sobol sequence with generation matrices is relatively intuitive, and shows a connection between the Sobol and Halton methods. Nevertheless, it is easier to implement the following representation of  $(n+1)$ -th element of a Sobol sequence in the  $k$ -th dimension:

$$x_n^k = \alpha_1(1,n) \cdot v_{k,1} \oplus \alpha_2(1,n) \cdot v_{k,2} \oplus \dots \quad (19)$$

Where  $\alpha_i(1,n)$  are defined as in equation (6) with  $b_j = 2$ . [Antonov and Saleev \(1979\)](#) showed that this formula can be rewritten using Gray Code binary representation of  $n$  resulting in:

$$x_n^k = g_1(n) \cdot v_{k,1} \oplus g_2(n) \cdot v_{k,2} \oplus \dots \quad (20)$$

One property of Grey Code representation is that the representation for  $n$  and  $n+1$  differs in only one position. Using this property, the formula in (19) can be written as

$$x_n^k = x_{n-1}^k \oplus v_c, \quad (21)$$

where  $c$  is an index of right-most zero bits in binary representation of  $n$  ([Bratley and Fox, 1988](#)), e.g., in 0.1  $c=2$ , in 0.01  $c=1$ , and in 0.11  $c=3$ .

In our simulation, we used primitive polynomials and direction numbers implemented in Matlab `sobolset` class.

---

<sup>41</sup> I.e., if  $v_{k,r} = 0.11$  then its vector form is  $(1,1,0,\dots)^T$ .

## Modified Latin Hypercube Sampling

Modified Latin hypercube sampling (MLHS) was proposed by [Hess, Train and Polak \(2006\)](#) as a variation of Latin hypercube sampling ([see, e.g., Stein, 1987](#)). Assume that  $P = \{p_{jk}\}$  is a  $R \times K$  matrix of which every column contains an independent, random permutation of sequence  $\{1, 2, \dots, R\}$ . Additionally let  $\Xi = \{\xi_k\}$  be a  $1 \times K$  vector of independent, random uniform draws on  $[0, 1]$  interval. Matrix  $X = \{x_{jk}\}$  of MLHS draws is created by setting:

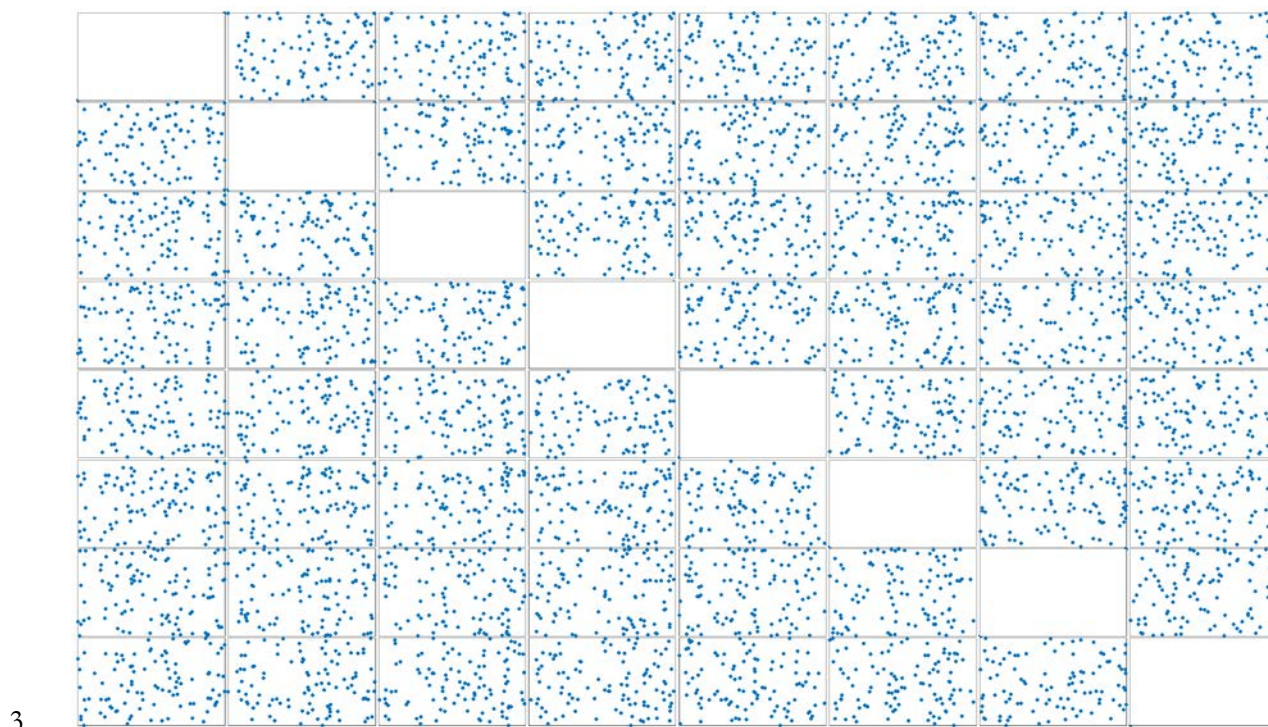
$$x_{jk} = F^{-1}\left(R^{-1}(p_{jk} + \xi_k - 1)\right), \quad (22)$$

where  $F(\cdot)$  is a cdf of the distribution one wants to draw from.

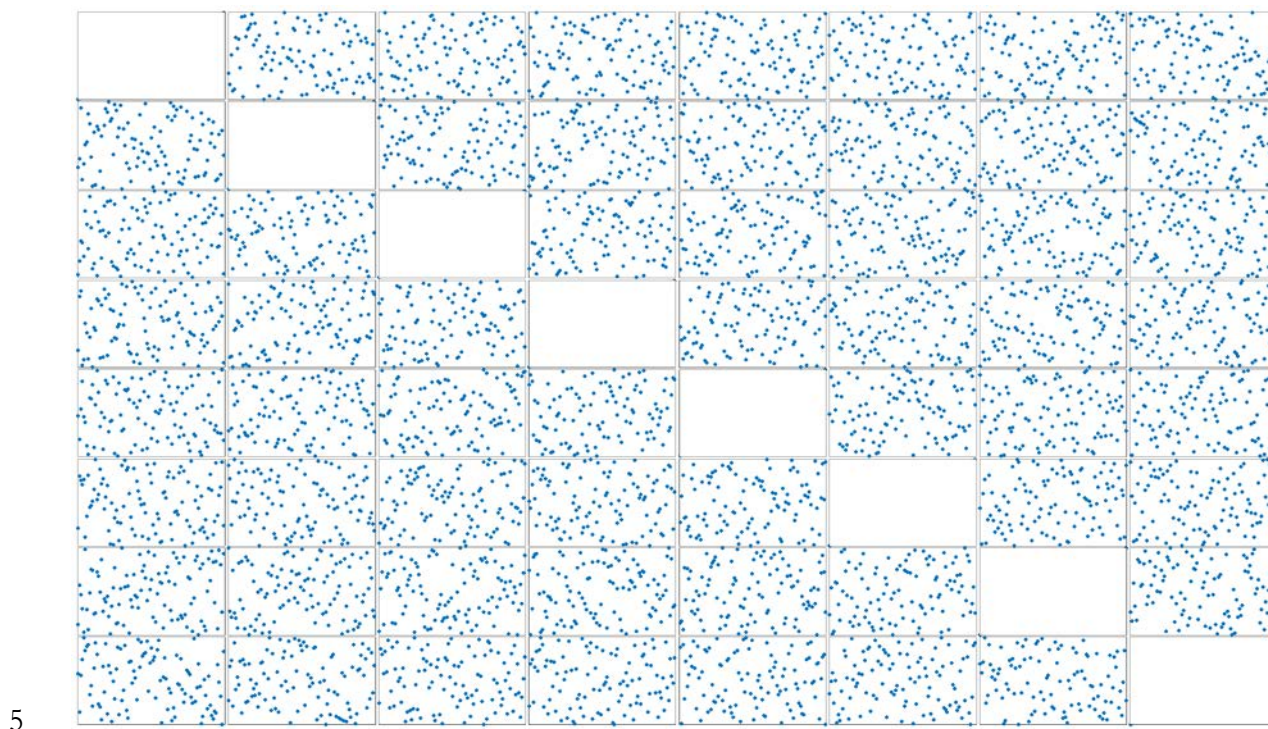
MLHS is not a low-discrepancy sequence designed as the Halton or Sobol sequence, because generation of a longer sequence requires creating a new one. Nevertheless, it has good coverage properties and, because of the random element  $\xi_k$  and permutations, its variance can be readily analyzed the same way as in the pseudo-random case. In our setting,  $K$  is equal to number of random parameters multiplied by the number of respondents, and  $R$  is a desired number of draws.

1      **Online Supplement B**

2      **Figure B1 Scatter plot matrix of 100 draws for 8 pseudo-random sequences**

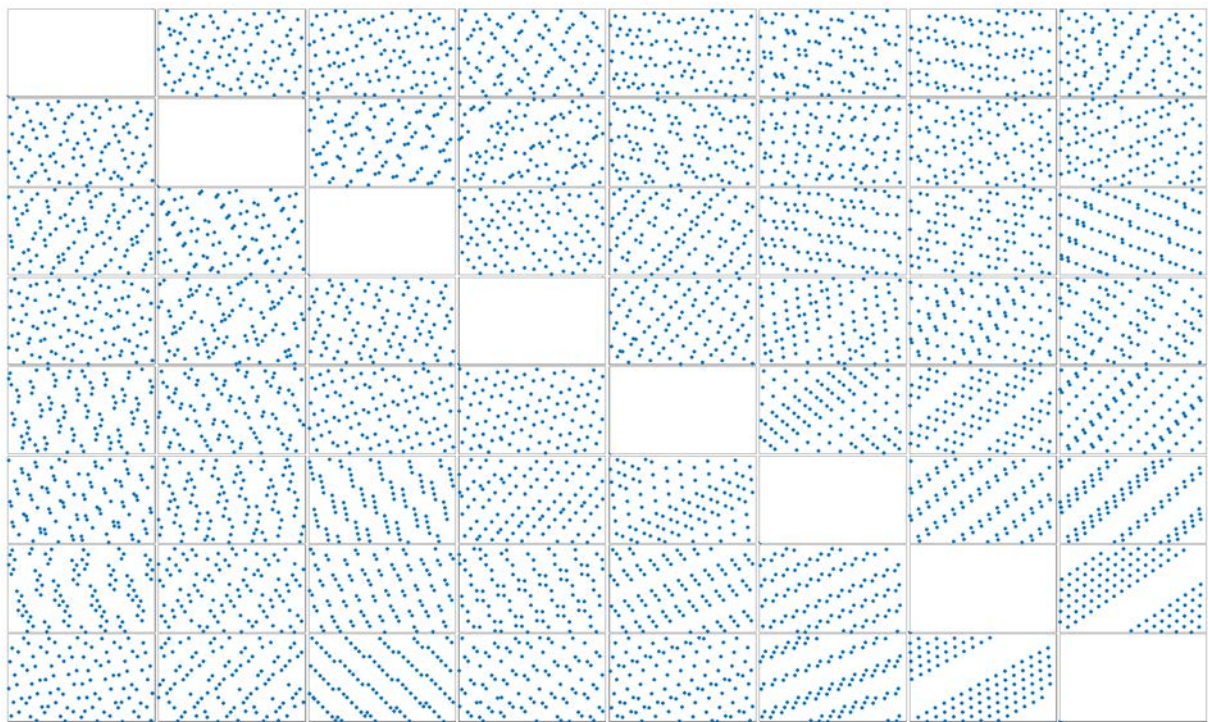


4      **Figure B2 Scatter plot matrix of 100 draws for 8 MLHS sequences**

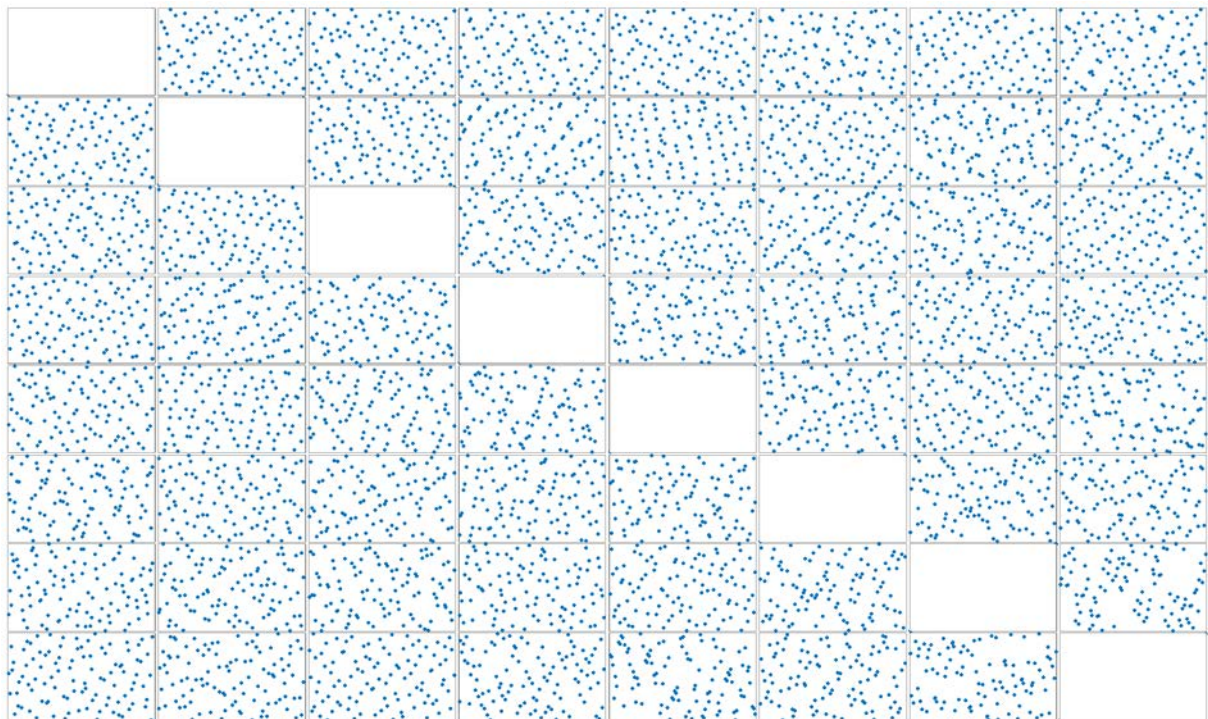




1 **Figure B3 Scatter plot matrix of 100 draws for 8 Halton sequences**



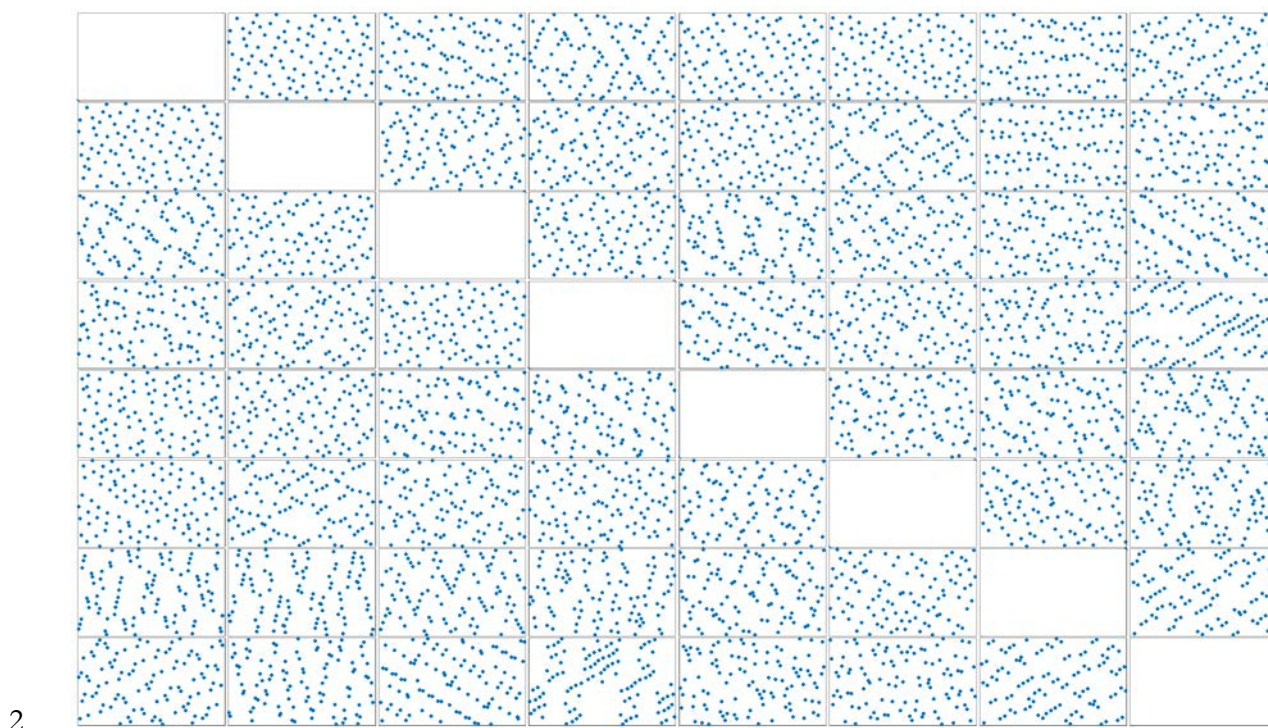
2  
3 **Figure B4 Scatter plot matrix of 100 draws for 8 scrambled Halton sequences**



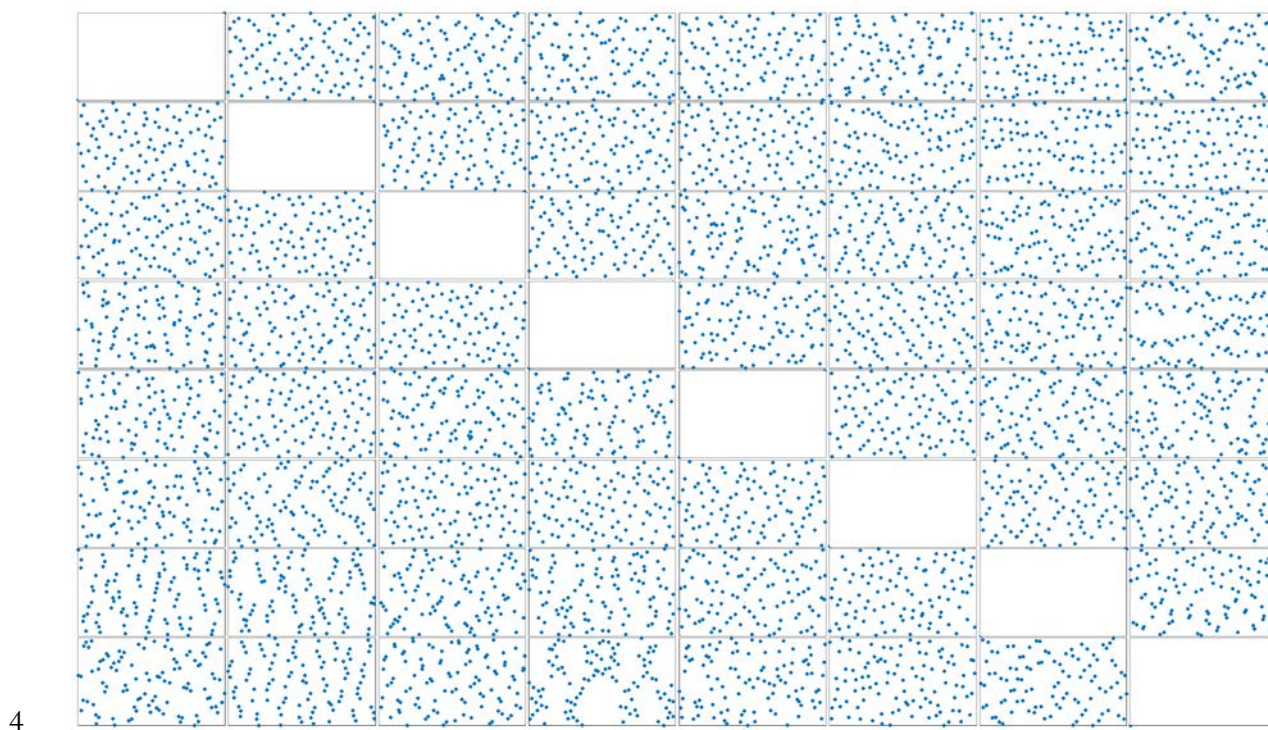
4  
5



1 **Figure B5 Scatter plot matrix of 100 draws for 8 Sobol sequences**



3 **Figure B6 Scatter plot matrix of 100 draws for 8 scrambled Sobol sequences**



## Online Supplement C

Table C1 presents the percentage of times<sup>42</sup> each type of draw performed the best, in terms of the lowest  $MTL_{0.05}$  for each number of draws.<sup>43</sup> In the overwhelming majority of cases, *Sobol* draws were the best – they resulted in the lowest variation of the log-likelihood function value of the estimated models.

**Table C1. Percentage of times each type of draw resulted in the lowest simulation error ( $MTL_{0.05}$ ) for the log-likelihood function value**

Number of draws used	<i>Pseudo-random</i>	<i>MLHS</i>	<i>Halton</i>	<i>Sobol</i>
100	0.00%	0.00%	19.44%	80.56%
200	0.00%	0.00%	25.00%	75.00%
500	0.00%	0.00%	22.22%	77.78%
1,000	0.00%	0.00%	25.00%	75.00%
2,000	0.00%	0.00%	0.00%	100.00%
5,000	0.00%	0.00%	19.44%	80.56%
10,000	0.00%	0.00%	16.67%	83.33%

The conclusions are similar when comparing simulation bias associated with parameter estimates.<sup>44</sup> Table C2 presents the percentage of times<sup>45</sup> each type of draw performed the best, in terms of the lowest  $MTL_{0.05}$  for each number of draws.<sup>46</sup> In the majority of cases, *Sobol* draws were the best – they resulted in the lowest variation of parameter estimates. The relative advantage of using Sobol draws is less evident than in the case of LL values but still evident, especially for higher numbers of draws.

**Table C2. Percentage of times each type of draw resulted in the lowest simulation error ( $MTL_{0.05}$ ) for the parameter estimates**

<sup>42</sup> Each cell of Table C1 corresponds to 36 dataset cases.

<sup>43</sup> Using  $MTL_{0.01}$  does not qualitatively change these results.

<sup>44</sup> It is worth noting, that in this case the absolute levels of parameter-specific  $MTL$  differed considerably. As expected, the lowest  $MTL$  were observed for the means of the discrete-valued variable ( $X_5$  or  $X_{10}$ ), while the highest were for the standard deviation of the alternative specific constant ( $X_1$ ). Nevertheless, *Sobol* draws consistently performed the best in all cases.

<sup>45</sup> Each cell of Table C2 corresponds to 450 dataset and parameter cases.

<sup>46</sup> Using  $MTL_{0.01}$  does not qualitatively change these results.

Number of draws used	<i>Pseudo-random</i>	<i>MLHS</i>	<i>Halton</i>	<i>Sobol</i>
100	2.67%	6.89%	36.22%	54.22%
200	1.11%	2.00%	30.67%	66.22%
500	0.89%	0.67%	37.56%	60.89%
1,000	0.22%	0.89%	28.44%	70.44%
2,000	0.00%	0.22%	19.33%	80.44%
5,000	0.00%	0.00%	36.67%	63.33%
10,000	0.44%	0.44%	32.22%	66.89%

Finally, Table C3 summarizes the performance of the different types of draws for the z-statistics of the estimated parameters; in other words, not only taking parameter estimates into account but also the associated standard errors. Z-statistics of parameters are important, because they usually provide a basis for judging if a parameter is statistically significant or not. Once again, using Sobol draws results in the lowest simulation error.

**Table C3. Percentage of times each type of draw resulted in the lowest simulation error ( $MTL_{0.05}$ ) for the z-statistics of the parameters**

Number of draws used	Pseudo-Random	MLHS	Halton	Sobol
100	2.22%	8.44%	37.56%	51.78%
200	1.56%	4.44%	33.78%	60.22%
500	1.56%	5.11%	32.89%	60.44%
1,000	1.11%	2.44%	26.00%	70.44%
2,000	1.11%	3.33%	23.78%	71.78%
5,000	2.44%	3.33%	29.78%	64.44%
10,000	0.00%	0.00%	29.11%	70.89%



## Online Supplement D

**Table D1. Variation of the simulation results ( $\log(MTL_{0.05})$ ) explained using characteristics of experimental settings – Sobol draws only**

		Log-likelihood	Parameter estimates	z-statistics
Constant		3.6256*** (0.1398)	-0.1041 (0.0688)	1.3792*** (0.0606)
Number of attributes is 10		-1.4554*** (0.1902)	-0.6920*** (0.0790)	-0.8217*** (0.0697)
log(number of draws)	(5 attributes)	-0.8118*** (0.0133)	-0.7254*** (0.0067)	-0.7064*** (0.0059)
log(number of draws)	(10 attributes)	-0.4671*** (0.0180)	-0.4372*** (0.0064)	-0.4621*** (0.0056)
Number of choice tasks		0.1819*** (0.0066)	-0.0194*** (0.0029)	0.0670*** (0.0026)
Number of individuals (in thousands)		0.8919*** (0.0659)	-0.5530*** (0.0291)	0.2916*** (0.0256)
OOD-design (MXL-design used as a reference)		-0.1334** (0.0642)	0.3693*** (0.0322)	0.3560*** (0.0284)
MNL-design (MXL-design used as a reference)		-0.1626** (0.0642)	0.3614*** (0.0322)	0.4507*** (0.0284)
Standard deviations (Means used as a reference)			1.2608*** (0.0190)	1.3205*** (0.0168)
$X_1$ (alternative specific constant)			0.5778*** (0.0271)	0.3240*** (0.0239)
$X_5$ icrete variable)			-0.7411*** (0.0271)	0.1509*** (0.0239)
R <sup>2</sup>		0.9543	0.8757	0.8922
n (observations)		309	3990	3990

## Online Supplement E

In this supplement, we provide a robustness check, in which we use a different measure of a simulation error. We still use  $MTL$ , as described in Section 4, but instead of measuring the difference between two models estimated with the same number of draws, we compare the difference between a model estimated with a given number of draws, and a model estimated with 100,000 draws. In this analysis, we use only a subsample of our experiment, as models with 100,000 draws were estimated only for the MXL design and Sobol draws. Table E1 presents analogous results to Table 5.

1 **Table E1. Minimum number of *Sobol* draws required for desired level of log-likelihood and parameter estimates precision (95% confidence**  
2 **intervals in [] brackets)**

Choice tasks per individual	4	4	4	8	8	8	12	12	12
Individuals	400	800	1,200	400	800	1,200	400	800	1,200
<b>5 attributes</b>									
≤5% probability of simulation-driven error in the LR test for 5 attributes <sup>47</sup>	103 [87-120]	154 [133-176]	230 [198-265]	308 [271-348]	461 [418-506]	690 [617-768]	924 [817-1,045]	1,382 [1,250-1,524]	2,068 [1,835-2,327]
≤5% probability that parameter estimates differ by ≥5% from true values for 5 attributes <sup>48</sup>	663 [597-734]	523 [474-576]	413 [371-458]	695 [631-765]	549 [500-600]	433 [392-477]	729 [659-806]	575 [521-634]	454 [408-505]
<b>Minimum recommended number of draws</b>	<b>663</b> <b>[597-734]</b>	<b>523</b> <b>[474-576]</b>	<b>413</b> <b>[371-458]</b>	<b>695</b> <b>[631-765]</b>	<b>549</b> <b>[500-600]</b>	<b>690</b> <b>[617-768]</b>	<b>924</b> <b>[817-1,045]</b>	<b>1,382</b> <b>[1,250-1,524]</b>	<b>2,068</b> <b>[1,835-2,327]</b>
<b>10 attributes</b>									
≤5% probability of simulation-driven error in the LR test for 10 attributes <sup>24</sup>	171 [131-220]	334 [271-407]	655 [526-807]	1,065 [899-1,260]	2,087 [1,840-2,368]	4,087 [3,439-4,866]	6,652 [5,429-8,240]	13,027 [10,745-16,019]	25,514 [19,895-33,141]
≤5% probability that parameter estimates differ by ≥5% from true values for 10 attributes <sup>27</sup>	9,750 [8,299-11,458]	6,710 [5,787-7,782]	4,618 [3,948-5,405]	10,508 [9,040-12,259]	7,232 [6,292-8,324]	4,978 [4,293-5,776]	11,325 [9,653-13,372]	7,795 [6,727-9,071]	5,365 [4,584-6,287]
<b>Minimum recommended number of draws</b>	<b>9,750</b> <b>[8,299-11,458]</b>	<b>6,710</b> <b>[5,787-7,782]</b>	<b>4,618</b> <b>[3,948-5,405]</b>	<b>10,508</b> <b>[9,040-12,259]</b>	<b>7,232</b> <b>[6,292-8,324]</b>	<b>4,978</b> <b>[4,293-5,776]</b>	<b>11,325</b> <b>[9,653-13,372]</b>	<b>13,027</b> <b>[10,745-16,019]</b>	<b>25,514</b> <b>[19,895-33,141]</b>

3

<sup>47</sup> At 0.05 significance level ( $MTL_{0.05}^{LL} \leq 1.9207$ ).

<sup>48</sup>  $MTL_{0.05}^{\beta} \leq 0.05|\beta|$ .

2 **Table F1. Minimum number of *Sobol* draws required for desired level of log-likelihood and parameter estimates precision (95% confidence**  
 3 **intervals in [] brackets)**

Choice tasks per individual	4	4	4	8	8	8	12	12	12
Individuals	400	800	1,200	400	800	1,200	400	800	1,200
<b>5 attributes / <math>MTL_{0.05}</math></b>									
≤5% probability of simulation-driven error in the LR test ( $MTL_{0.05}^{LL} \leq 3.3174$ )	76 [63-90]	117 [100-137]	182 [155-213]	185 [159-214]	287 [254-325]	446 [391-509]	454 [392-523]	704 [621-798]	1,093 [954-1,254]
≤5% probability of simulation-driven error in the LR test ( $MTL_{0.05}^{LL} \leq 1.9207$ )	148 [125-174]	230 [199-265]	357 [307-415]	363 [316-414]	563 [504-629]	874 [774-989]	889 [775-1,018]	1,380 [1,226-1,554]	2,142 [1,878-2,454]
≤5% probability of simulation-driven error in the LR test ( $MTL_{0.05}^{LL} \leq 1.3528$ )	228 [194-266]	354 [308-406]	549 [475-636]	559 [491-632]	867 [782-963]	1,346 [1,196-1,517]	1,370 [1,199-1,562]	2,126 [1,894-2,392]	3,299 [2,900-3,778]
≤5% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.05}^{\beta} \leq 0.1 \beta $ )	450 [406-498]	332 [300-366]	245 [220-272]	404 [367-445]	298 [271-327]	220 [198-243]	363 [327-401]	268 [242-295]	197 [177-219]
≤5% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.05}^{\beta} \leq 0.05 \beta $ )	1,170 [1,061-1,288]	862 [786-946]	636 [575-702]	1,051 [959-1,150]	775 [710-844]	571 [520-627]	944 [856-1,039]	696 [634-764]	513 [464-566]
≤5% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.05}^{\beta} \leq 0.01 \beta $ )	10,759 [9,781-11,842]	7,931 [7,254-8,676]	5,847 [5,322-6,425]	9,666 [8,830-10,575]	7,125 [6,562-7,750]	5,253 [4,810-5,737]	8,684 [7,888-9,552]	6,401 [5,855-6,996]	4,719 [4,295-5,181]
<b>5 attributes / <math>MTL_{0.01}</math></b>									
≤1% probability of simulation-driven error in the LR test ( $MTL_{0.01}^{LL} \leq 3.3174$ )	104 [88-123]	162 [140-188]	252 [217-293]	255 [222-292]	397 [355-444]	618 [547-699]	625 [545-715]	973 [864-1,095]	1,512 [1,328-1,726]
≤1% probability of simulation-driven error in the LR test ( $MTL_{0.01}^{LL} \leq 1.9207$ )	205 [175-238]	319 [278-364]	495 [430-572]	501 [442-566]	780 [705-864]	1,213 [1,081-1,362]	1,228 [1,078-1,395]	1,909 [1,707-2,140]	2,969 [2,621-3,386]
≤1% probability of simulation-driven error in the LR test ( $MTL_{0.01}^{LL} \leq 1.3528$ )	316 [272-365]	491 [432-557]	764 [667-876]	773 [685-868]	1,202 [1,092-1,325]	1,870 [1,670-2,095]	1,892 [1,667-2,146]	2,943 [2,635-3,297]	4,577 [4,038-5,218]
≤1% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.01}^{\beta} \leq 0.1 \beta $ )	659 [597-726]	486 [442-534]	359 [324-397]	592 [539-649]	437 [399-477]	322 [292-355]	532 [481-586]	393 [357-432]	290 [261-320]
≤1% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.01}^{\beta} \leq 0.05 \beta $ )	1,712 [1,556-1,879]	1,263 [1,154-1,382]	932 [846-1,027]	1,538 [1,406-1,680]	1,135 [1,041-1,234]	838 [764-916]	1,382 [1,255-1,518]	1,020 [932-1,116]	753 [683-828]
≤1% probability that parameter estimates differ by ≥10% from true values ( $MTL_{0.01}^{\beta} \leq 0.01 \beta $ )	15,720 [14,287-17,307]	11,601 [10,608-12,695]	8,562 [7,797-9,404]	14,128 [12,900-15,463]	10,426 [9,592-11,344]	7,695 [7,045-8,407]	12,697 [11,526-13,969]	9,370 [8,565-10,245]	6,915 [6,297-7,593]

Choice tasks per individual	4	4	4	8	8	8	12	12	12
Individuals	400	800	1,200	400	800	1,200	400	800	1,200
<b>10 attributes / <math>MTL_{0.05}</math></b>									
$\leq 5\%$ probability of simulation-driven error in the LR test ( $MTL_{0.05}^{\mu} \leq 3.3174$ )	81 [56-113]	175 [128-230]	375 [280-491]	387 [297-491]	830 [678-1,001]	1782 [1445-2186]	1,837 [1,449-2,306]	3,942 [3,231-4,837]	8,462 [6,704-10,833]
$\leq 5\%$ probability of simulation-driven error in the LR test ( $MTL_{0.05}^{\mu} \leq 1.9207$ )	263 [193-346]	563 [439-708]	1,209 [945-1,528]	1,246 [1,003-1,529]	2,675 [2,257-3,160]	5,742 [4,698-7,101]	5,918 [4,667-7,509]	12,702 [10,191-16,052]	27,264 [20,889-36,562]
$\leq 5\%$ probability of simulation-driven error in the LR test ( $MTL_{0.05}^{\mu} \leq 1.3528$ )	556 [423-714]	1,193 [956-1,465]	2,562 [2,028-3,224]	2,640 [2,154-3,224]	5,667 [4,781-6,768]	12,163 [9,821-15,399]	12,535 [9,785-16,231]	26,905 [21,007-35,200]	57,748 [42,904-80,801]
$\leq 5\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.05}^{\theta} \leq 0.1 \beta $ )	5,182 [4,454-6,038]	3,124 [2,711-3,601]	1,884 [1,617-2,194]	4,338 [3,768-5,011]	2,616 [2,292-2,986]	1,577 [1,368-1,821]	3,631 [3,125-4,228]	2,190 [1,902-2,525]	1,320 [1,135-1,535]
$\leq 5\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.05}^{\theta} \leq 0.05 \beta $ )	25,294 [21,531-29,864]	15,251 [13,131-17,736]	9,196 [7,866-10,766]	2,1174 [18,232-24,777]	12,767 [11,123-14,674]	7,698 [6,674-8,913]	17,725 [15,140-20,819]	10,688 [9,249-12,388]	6,444 [5,556-7,497]
$\leq 5\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.05}^{\theta} \leq 0.01 \beta $ )	1,003,977 [806,569-1,262,736]	605,365 [494,545-747,850]	365,015 [298,783-449,838]	840,449 [683,177-1,045,805]	506,763 [418,157-620,541]	305,561 [252,422-372,977]	703,556 [570,100-877,886]	424,221 [348,425-519,026]	255,791 [211,056-312,096]
<b>10 attributes / <math>MTL_{0.01}</math></b>									
$\leq 1\%$ probability of simulation-driven error in the LR test ( $MTL_{0.01}^{\mu} \leq 3.3174$ )	152 [111-201]	317 [245-400]	664 [518-837]	678 [543-831]	1,417 [1,197-1,662]	2,965 [2,457-3,580]	3,026 [2,431-3,741]	6,329 [5,240-7,703]	13,240 [10,548-16,893]
$\leq 1\%$ probability of simulation-driven error in the LR test ( $MTL_{0.01}^{\mu} \leq 1.9207$ )	469 [361-596]	980 [792-1,193]	2,050 [1,646-2,540]	2,092 [1,729-2,519]	4,377 [3,741-5,134]	9,155 [7,551-11,287]	9,342 [7,465-11,763]	19,543 [15,714-24,669]	40,880 [31,401-54,691]
$\leq 1\%$ probability of simulation-driven error in the LR test ( $MTL_{0.01}^{\mu} \leq 1.3528$ )	966 [763-1,205]	2,020 [1,665-2,432]	4,226 [3,418-5,252]	4,312 [3,567-5,208]	9,020 [7,637-10,785]	18,869 [15,247-23,884]	19,254 [15,090-24,854]	40,277 [31,421-52,774]	84,254 [62,809-117,281]
$\leq 1\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.01}^{\theta} \leq 0.1 \beta $ )	9,062 [7,818-10,552]	5,549 [4,838-6,373]	3,398 [2,934-3,935]	7,627 [6,649-8,794]	4,671 [4,115-5,312]	2,860 [2,496-3,280]	6,419 [5,553-7,448]	3,931 [3,437-4,513]	2,407 [2,088-2,779]
$\leq 1\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.01}^{\theta} \leq 0.05 \beta $ )	42,339 [36,058-50,011]	25,927 [22,323-30,142]	15,877 [13,630-18,535]	35,635 [30,668-41,668]	21,822 [19,003-25,116]	13,363 [11,596-15,426]	29,992 [25,628-35,232]	18,366 [15,898-21,275]	11,247 [9,711-13,045]
$\leq 1\%$ probability that parameter estimates differ by $\geq 10\%$ from true values ( $MTL_{0.01}^{\theta} \leq 0.01 \beta $ )	1,518,200 [1,220,658-1,908,915]	929,692 [757,707-1,149,623]	569,311 [465,252-703,039]	1,277,794 [1,037,273-1,589,909]	782,476 [644,666-959,871]	479,161 [394,743-585,573]	1,075,456 [871,228-1,343,284]	658,571 [540,144-806,879]	403,286 [331,926-492,888]