Designing and Conducting Stated Choice Experiments

Michiel C.J. Bliemer & John M. Rose

In: Hess, S., and Daly, A. (eds) Handbook of Choice Modelling, Second edition, forthcoming

Introduction

Stated choice experiments have a long history in both academia and practice. Originally designed to empirically test a range of economic theories, such as the existence of indifference curves (Thurstone, 1931; Mosteller and Nogee, 1951; Rousseas and Hart, 1951; May, 1954; MacCrimmon and Toda, 1969), stated choice experiments have since gained widespread acceptance across a range of applied economics fields, including transportation (e.g., Bliemer and Rose, 2011; Hess et al., 2020; Ortúzar et al. 2021), health (e.g., De Bekker-Grob et al., 2013; Determann et al., 2014; Hansen et al., 2019), marketing (e.g., He and Oppewal, 2018; Wu et al., 2019; Burke et al., 2020) and environmental and resource economics (e.g., Scarpa et al., 2003; MacDonald et al., 2011; Greiner et al., 2014). Despite their prevalence, the design and implementation of a stated choice experiment requires far more nuance than most other survey methods insofar as the technique requires that the analyst provide respondents a detailed set of scenarios that they are expected to interact with and respond too. Stated choice experiments therefore don’t simply ask respondents what they did in some situation (such data is called revealed preference data), or how they feel about some statement (as with attitudinal type questions), but rather creates hypothetical scenarios that respondents are expected to react to. The purpose of this chapter is to describe the processes required to generate these hypothetical scenarios.

In a *choice experiment*, also referred to as stated choice survey or choice-based conjoint, the analyst asks agents (i.e., decision-makers, for example consumers buying a certain type of product, travellers making a trip, patients choosing treatment, physicians prescribing medication, etc.) to complete a series of *choice tasks* (also called choice sets) consisting of several alternatives, each described by their characteristics. Example choice tasks are shown in Figures 1 and 2. Each choice tasks consists of several elements, namely (i) the *choice scenario*, describing the context in which the choice is made, (ii) the *alternatives* to choose from, (iii) the *profiles* for each alternative describing the attributes (also called factors) with their specific levels, and (iv) the *response mechanism*, which typically consists of a radio button for the most preferred option, but can also include a two-step mechanism for unforced and forced choice (see Figure 2, where for example first the unforced choice is captured, but if the patient insists on being treated, a forced choice needs to be made), a best-worst choice, or a first best and second best choice (although not directly relevant to the discussion here, volumetric choice tasks have also been employed where respondents are asked to select different continuous amounts or quantities from multiple discrete alternatives). While such choice tasks are often shown in a table format with text, different formats exist, see e.g., Figure 3. Images may assist agents in imagining the choice alternatives, although one should be careful not to accidentally influence agents with additional attributes (e.g., colours, or mood in a photo).

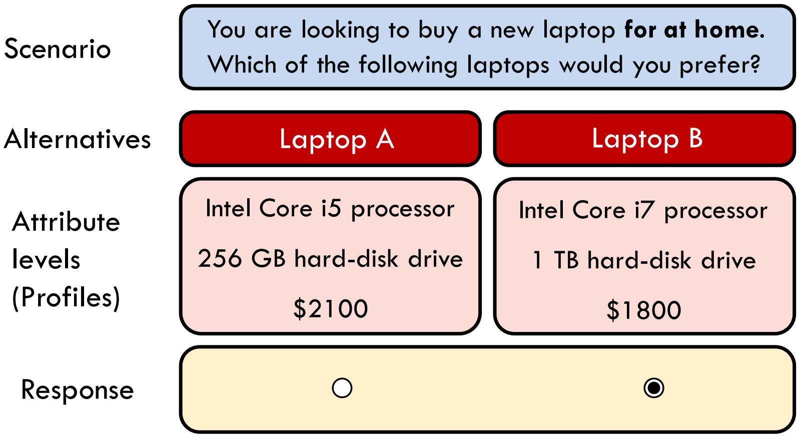

Figure 1: Laptop choice

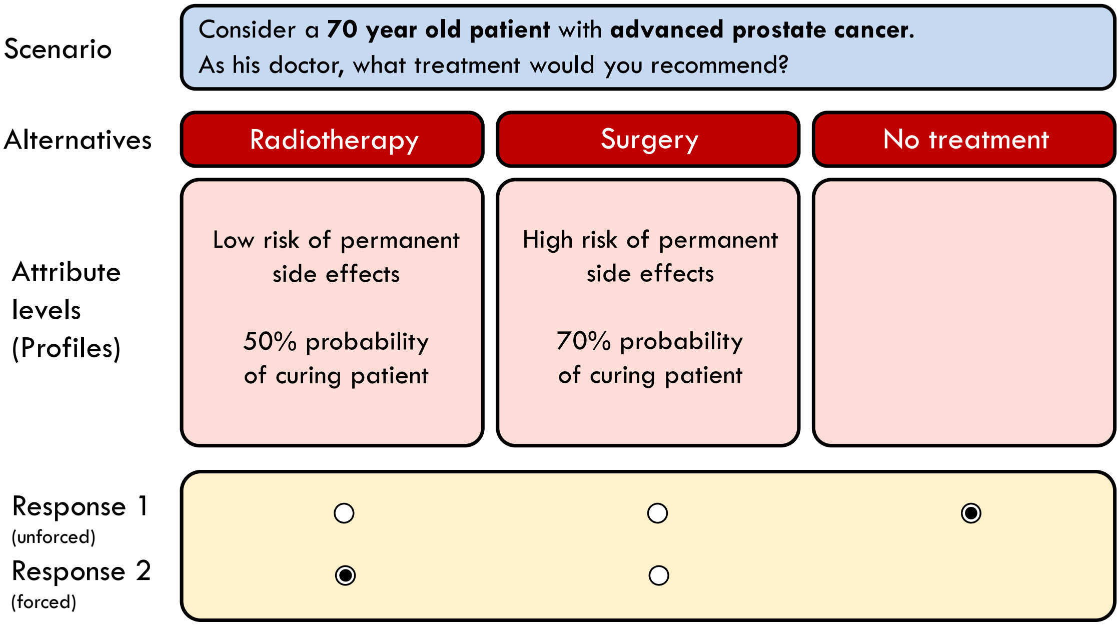
Scenario
Consider a **70 year old patient** with **advanced prostate cancer**.
As his doctor, what treatment would you recommend?
Alternatives
Radiotherapy
Surgery
No treatment
Attribute levels (Profiles)
Low risk of permanent side effects
50% probability of curing patient
High risk of permanent side effects
70% probability of curing patient
Response 1 (unforced)
○ ○ ●
Response 2 (forced)
● ○

Figure 2: Treatment choice

Choice experiments are usually part of a larger *questionnaire* or *survey* consisting of several parts. Although survey flow differs from questionnaire to questionnaire, the first part of a survey typically involves agents being asked screen-out questions to judge their eligibility. In the second part, agents might be asked questions related to their current situation and behaviour related to the specific study. This information can be used to tailor choice tasks in the third part consisting of the choice experiment. The fourth part typically concludes by asking additional questions such as questions about general attitudes and perceptions, socio-demographic questions, and open-ended qualitative questions. While socio-demographic questions could also be asked earlier in the survey, attitudinal questions should be asked after the choice experiment to avoid influencing choice behaviour (Liebe et al., 2018).

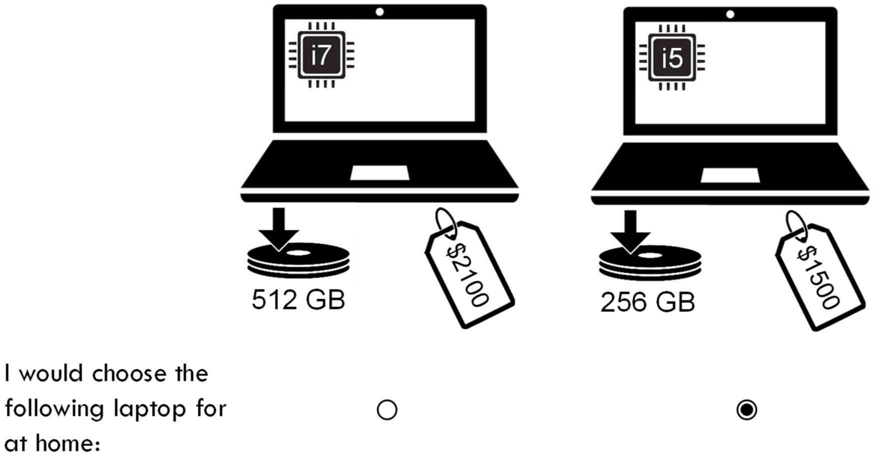

Figure 3: Laptop choice with images

The design of choice experiments can be somewhat complex, consisting of several steps or stages. The typical steps involved in designing a choice experiment are:

I. Determine whether an experiment is labelled or unlabelled depending on research questions;
II. Determine the alternatives and attributes to include in the experiment;
III. Determine the attribute levels and their coding;
IV. Determine number of choice tasks in experimental design;
V. Choose experimental design strategy;
VI. Conduct pilot study;
VII. Conduct main study;

Each step is discussed in more detail in the next sections.

Step I: Labelled versus unlabelled experiment

Consider a choice experiment consisting of choice tasks $s \in S,$ where $|S|$ is the total number of choice tasks, and assume that all or a subset of these choice tasks $S_n \subseteq S$ are given to agent $n \in \{1,\ldots,N\},$ where N is the sample size of responding agents. In each choice task, agents are asked to choose among alternatives in set J, where $|J|$ is the number of alternatives in the set.

These alternatives can be of the same type or of different types. The type of an alternative is commonly described by a *label*. A label can for example be a product category, a brand name, but can also refer to a specific alternative type such as a status quo or opt-out (no choice) alternative.

If all alternatives have the same label, then the label is assumed not to play a role in the choice process. Examples can be found in route choice (Route A, Route B, Route C), medication choice (Medication 1, Medication 2), policy choice (Policy I, Policy II), laptop choice (Laptop 1, Laptop 2), etc. An example is shown in Figure 1. Such an experiment is often referred to as an *unlabelled experiment* and the utility function of each *unlabelled (generic) alternative* is identical, i.e.,

$$V_{nsj} = f(\mathbf{x}_{nsj}), \qquad \forall n \in \{1,\ldots,N\}, \forall s \in S_n, \forall j \in J, \tag{1}$$

where V_{nsj} is the systematic utility that agent *n* attaches to alternative $j \in J$ in choice task $s \in S_n$ depending on the profile of alternative *j* defined by the vector of attribute levels $\mathbf{x}_{nsj}$ and a generic function *f*. This function depends on a vector of unknown generic preference parameters $\boldsymbol{\beta}$ that describe trade-offs across the attributes and attribute levels and are subject to estimation.

While not relevant at the experimental design stage, in model estimation one would add constants in $|J|-1$ alternatives to account for *presentation order effects of alternatives*, also known as *left-to-right bias* (in countries where one reads from left to right), where alternatives shown on the left (or top) in the survey may have a higher propensity of being chosen than alternatives shown on the right (or bottom) (see e.g., Ryan et al. 2018).

If some or all of the alternatives have different labels, then choice is influenced by these labels. Examples can be found in mode choice (Car 1, Car 2, Train, Bus), treatment choice (Surgery, Radiation Therapy), smartphone choice (iPhone, Samsung Galaxy, Google Pixel), policy choice (Current policy [status quo], Policy A, Policy B), activity choice (Activity 1, Activity 2, Neither [opt-out]), etc. An example is shown in Figure 2. Such an experiment is referred to as a *labelled experiment* and the utility functions of each *labelled alternative* can be different for each label $m \in M$,

$$V_{nsj} = f_m(\mathbf{x}_{nsj}), \qquad \forall n \in \{1,\ldots,N\}, \forall s \in S_n, \forall j \in J_m, \tag{2}$$

where $J_m \subset J$ is the subset of alternatives with label *m*, where $\sum_m |J_m| = |J|$, and each alternative within this set has the same linear or nonlinear label-specific utility function f_m.

These functions have preference parameters $\boldsymbol{\beta}_m$, which can be label-specific or generic across labels. The functions also include label-specific constants, where for identifiability purposes one of them needs to be normalised to zero for a chosen reference label. That is, one would estimate $|M|-1$ label-specific constants. As an example, in the mode choice situation with alternatives Car 1, Car 2, Train, and Bus, one would specify four alternatives across three labels (Car, Train, Bus), where Car 1 and Car 2 have identical utility functions. With three labels, the model identifies two label-specific constants. Note that an opt-out alternative can only have a label-specific constant (which may be normalised to zero) while a status quo alternative is described by a regular utility function using fixed attribute levels. In some experiments, the levels of the status quo may be agent specific, taking on values provided earlier on in the survey.

In contrast to estimating models using data from an unlabelled experiment, one cannot simply add alternative-specific constants to account for presentation order effects of alternatives in a labelled experiment since such constants would be confounded with (some or all) label-specific constants. How to account for presentation order effects of alternatives in labelled experiments will be discussed in Step VI.

Whether a labelled or unlabelled experiment is suitable for a certain study depends on the research questions being addressed. If one is interested in determining the *willingness-to-pay* (WTP) for certain attribute levels, or in determining the *relative importance of attributes* in decision-making, then it often suffices to consider an unlabelled experiment in which two or more alternatives are shown as variants of the same label. On the other hand, a labelled experiment is suitable if one would like to determine *market shares* of a product type or *demand elasticities*. One would include an opt-out alternative if one is interested in predicting the unconditional absolute demand in the market using unforced choice tasks, while it can be left out if one is only interested in relative market shares or conditional demand across products by asking to make a forced choice (it is worthwhile noting that evidence suggests that for the same empirical context, the results one obtains from forced and unforced choice tasks can vary dramatically; see Dhar and Simonson, 2003). A status quo alternative is often added to determine willingness to deviate from an existing policy or simply to make the choice task look more familiar to agents. Labelled experiments can also be used to determine WTP values, particularly if the WTP values are expected to vary across different labelled alternatives (e.g., the willingness to pay for travel time savings differs for bus and car use). If however, WTP values are expected to be the same across alternatives, then given that labelled experiments generally require more complex choice tasks and the estimation of a larger number of coefficients, there is no reason to use a labelled experiment if the sole purpose of the study is to determine WTP values.

Significant differences in the results of choice experiments with and without the presence of status quo alternatives have been found within the literature (see e.g., Dhar, 1997), with the recommendation being in general that status quo alternatives should be used in such experiments where applicable (e.g., Adamowicz and Boxall, 2001; Bennett and Blamey, 2001; Bateman et al., 2003). Dhar (1997) found that the decision to defer choice (and hence select a no choice option), is influenced by the absolute difference in attractiveness among the alternatives. That is, the overall utility of the alternative is the main driver of selecting a no choice option as opposed to the complexity of attribute trade-offs necessary when choosing between different alternatives. Boxall et al. (2009) report similar findings to Dhar (1997), suggesting that increasing task complexity, related to how similar the alternatives are as described by the attribute levels shown, leads to increased choice of the status quo alternative, whilst at the same time, a responding agent's age and level of education may also influence this choice.

Dhar and Simonson (2003) found that if a forced-choice is followed by an unforced-choice in a dual response task, then some alternatives tended to lose proportionally more share than others, violating the independent and identically distributed (IID) model assumption. As such, it may be necessary to estimate more sophisticated discrete choice models that relax the IID assumption when data is collected using both forced and unforced choice responses. Brazell et al. (2006) failed to locate IID violations in a similar experiment, hypothesising that failure to detect such effects was likely the result of using a more complex choice experiment involving more attributes than was used by Dhar and Simonson (2003), concluding that the increased complexity of their design decreased the prevalence of possible compromise alternatives appearing within the experiment. Rose and Hess (2009) also explored the use of dual forced/unforced response mechanisms, however unlike the Dhar and Simonson (2003) and Brazell et al. (2006) studies, made use of respondent reported status quo alternatives as opposed to a simple no-choice alternative. Like Brazell et al. (2006), Rose and Hess (2009) found no evidence for IID violations between the forced and unforced tasks. Rose and Hess (2009) also reported no differences between the WTP estimates obtained across the dual forced/unforced response data.

Kontoleon and Yabe (2003) compared a ‘do not buy’ response format to a ‘buy/choose my current brand’ format. Keeping everything else equal, they found that the relative choice share of the opt-out alternative was higher in the ‘own brand’ treatment as opposed to the treatment that received the ‘no purchase’ treatment. They further found differences in parameter estimates for the more important attributes, while little difference were observed for less salient attributes.

Step II: Determine alternatives and attributes

Once the study objectives are known and a choice of a labelled or unlabelled experiment has been made, the analyst needs to determine which alternatives and attributes to include in the choice experiment. This is different for each study and while for some studies determining the alternatives and attributes is straightforward, for other studies, it requires careful consideration of how the outcomes will be used.

For any experiment, the minimum number of alternatives shown in a choice task is two, i.e., $|J| \geq 2,$ one of which may be a status quo or no choice alternative. The larger the number of alternatives, the more information is captured in each choice task, but also the larger the cognitive burden placed on the responding agent. In case of an unlabelled experiment, there is generally no need to go beyond two or three generic alternatives. If the number of attributes is small, then three or four alternatives may be fine, but with a large number of attributes, one typically restricts the number of alternatives to two. In case of a labelled experiment, the number of alternatives in each choice task depends on the number of relevant labels to include since each label requires at least one alternative, i.e., $|J_m| \geq 1,$ which means that the number of alternatives needs to be larger than or equal to the number of labels, $|J| \geq |M|.$ For example, in a mode choice experiment, one may need to include labels for Car, Metro, Train, Bus, Bicycle and Walk, such that the number of alternatives in a choice tasks is at least six. If there is a risk that a certain label is dominant, e.g., if some agents will always choose Car no matter what the attribute levels are, then one can consider including two Car alternatives, Car 1 and Car 2, to ensure that all agents make trade-offs across alternatives. If the number of labelled alternatives is considered too large, one could show only a subset of labelled alternatives in each choice task, a so-called *partial choice set* (Bliemer et al., 2018). This reduces the complexity of each individual choice task, but does require increasing the number of choice tasks per agent or increase the sample size to capture the same amount of information.

Extensive research has been conducted on the impact of the number of alternatives shown in DCEs. For example, Adamowicz et al. (2006) found that respondents assigned to a three-alternative version of a choice experiment where more likely to choose a status quo option than a two-alternative version. Rolfe and Bennett (2009) report similar findings when they compared two- and three-alternative versions of a choice experiment. Caussade et al. (2005) found that the number of alternatives shown to respondent had the second largest influence on error variances out of all design dimensions they tested and concluded that showing four alternatives is better than showing either three or five alternatives in terms of the impact of scale effects. DeShazo and Fermo (2002) found a quadratic relationship between the number of alternatives and the variance, suggesting that error variance first decreases, then increases with the number of alternatives. In contrast, Arentze et al. (2003) found no error variance differences between choice experiments versions making use of two versus three alternatives. Hensher (2004) found that as the number of alternatives increases, there exists a differential impact upon the WTP measures for different attributes of the design, whilst Rose et al. (2009) found different impacts on mean WTP estimates obtained from the same survey conducted across different countries. Using eye-tracking technology, Meißner et al. (2020) report that respondents tend to increase the amount of information they process as the number of alternatives increases, whilst simultaneously filtering out more pieces of information when choice tasks include more alternatives. Interestingly, Meißner et al. found that respondents almost immediately change their search strategies adopted when the number of alternatives changes dramatically (say from two to five alternatives) from one choice task to another. Weng et al. (2021) found differences in WTP outcomes obtained for an unlabelled choice experiment involving two alternatives compared to one with more than two alternatives. They also found that the ability of agents to identify their preferred alternative improves for experiments consisting of a status quo and single additional alternative as the number of attributes increases, but becomes harder when more alternatives are added.

With respect to attributes, if the objective of the study is to determined specific WTP estimates in an unlabelled experiment, one could simply only include the attributes under investigation. For example, it is common in transport to determine the value of travel time using only two attributes, namely travel time and travel cost (see e.g., Batley et al., 2017), although one needs to be careful to avoid endogeneity bias[1]. On the other hand, if the study objective is to forecast demand or market shares, one would generally include all attributes that are deemed relevant in making the choice. Relevant attributes can be identified by reviewing the literature, conducting a series of qualitative interviews such as focus groups involving a small number of agents (typically less than ten) from the target population, or personal interviews with experts. A *focus group* is a qualitative research technique where one asks a group of agents (face-to-face or online) about their rationale for making decisions in the choice context of interest. While focus groups may include individual tasks such as writing down the most relevant attributes and ranking them in order of importance, open-ended group discussions guided by a moderator lie at the core. Group discussions allow participants to agree or disagree and provide a way to identify a range of opinions and experiences that would be difficult to obtain through surveys.

[1] Endogeneity bias may occur if the true decision calculus used by agents involves interactions between omitted attributes and attributes used as part of the study. For example, one agent may imagine travel time seated in an empty bus, while another may imagine travel time standing in a crowded bus and hence attach more disutility to travel time. In this case, omission of crowding as an attribute and its interaction with travel time results in endogeneity bias, which invalidates the assumption that the error term is independent of the systematic component of utility.

While considering only a small number of attributes assists in reducing cognitive burden on agents, it has been argued that relevance is more important than quantity. If a large number of attributes is deemed relevant, then one can consider showing only a subset of attributes in each choice task. Such an incomplete profile is typically referred to as a *partial profile* (see e.g., Chrzan, 2010; Kessels et al., 2011). Showing partial profiles in a choice task leads to a reduction in information captured in the choice task, therefore one will need to increase the number of choice tasks per agent or the sample size to ensure that the same amount of information is obtained.

Research has tended to show that the number of attributes present within the experiment does impact upon the behavioural responses provided. Caussade et al. (2005) and DeShazo and Fermo (2002) report that the number of attributes has a significant impact upon the error variance of models estimated using choice experiment data. DeShazo and Fermo (2002) found that, on average, an increase in the number of attributes leads to an increase in the variance of the error component in utility of choice experiments, whilst Caussade et al. (2005) concluded that the number of attributes used had the largest influence on error variances out of all design dimensions. In a similar vein, Arentze et al. (2003) found that increasing the number of attributes from three to five led to increased error variances and parameter differences. In support of this argument, Green and Srinivasan (1990) argued that respondents are incapable of processing many attributes simultaneously and become tired and hence consequently ignore or address attributes in random and uncontrolled ways, or tend to use heuristics that lead to biased preference measures. Hensher (2006) found that the number of attributes has a significant influence on parameter outputs and WTP measures, which was also confirmed by Rose et al. (2009) who found statistically significant differences in WTP measures as the number of attributes increase. Nevertheless, Rose et al. (2009) report directional differences in the mean WTP over data sets collected from different countries.

The number of alternatives and attributes shown in each choice task also depends on the survey instrument. When using a computer-aided personal interviewer (CAPI), one can generally present more complex choice tasks to each agent given that a personal interviewer can explain the choice task and answer any questions that the responding agent may have about what they are presented with. In case of a typical online survey, completed on a computer or smartphone, one would generally keep the number of alternatives and attributes shown in each choice task limited as agents may be less engaged with the experiment and therefore spent less time on each choice task.

Step III: Determine attribute levels and their coding

Attributes can be classified as *qualitative* (also referred to as *categorical*), or *quantitative* (also referred to as *numerical*), and can further be distinguished according to their measurement scale, see Table 1.

Table 1: Data types and measurement scales

Data type	Measurement scale	Example attributes with example levels
Qualitative / categorical	Nominal	Colour (red, blue, yellow, green, purple) Warranty (yes, no) Livestock (cattle, sheep, pigs, horses)
	Ordinal	Comfort (low, medium, high) Side-effects (none, moderate, severe) Education (primary, secondary, tertiary)
Quantitative / numerical	Interval	Temperature (5 ºC, 10 ºC, 15 ºC) Time of day (9a, noon, 5pm, midnight) Elevation (200 m, 700 m, 1500 m)
	Ratio	Cost ($20, $30, $40, $50) Travel time (15 min, 20 min, 25 min) Distance (1 km, 2 km, 5 km, 10 km)

Attributes with nominal or ordinal scale describe qualitative/categorical data. If an attribute has nominal scale then its levels do not have a specific ordering, whereas an attribute with ordinal scale has levels that describe a certain order. Attributes with interval or ratio scale describe quantitative/numerical data, which can be discrete or continuous. Such attributes have an order in which absolute differences between levels are meaningful and attributes with a ratio scale also have an absolute zero point.

Qualitative attributes require a specific coding scheme for use in utility functions, where the most widely used schemes are *dummy*, *effects* or *(orthogonal) contrast coding*. Levels of quantitative attributes are often used directly into the utility function as a continuous linear effect, e.g. $\beta x,$ or a nonlinear effect, e.g. $\beta \ln(x).$ While it is possible to use dummy, effects or (orthogonal) contrast coding for quantitative attributes using discrete levels, this makes it more difficult to interpolate/extrapolate beyond these levels in forecasting. Nevertheless, in some applied economics fields such as marketing it is common practice to do so.

Once the measurement scale of each attribute has been identified, the number of levels can be determined. For nominal attributes, one typically needs to include all relevant levels (which can be asked in a focus group discussion, see Step II). In case of an ordinal attribute, one can often choose the number of levels, for example 'quality' can be described as low - high, or as low – medium - high, or as low – medium – high - very high. In case of ordinal attributes, one may want to be careful not to cause ambiguity as different agents will understand something different with respect to 'medium quality'. If possible, it is best to describe these levels in terms of specific characteristics, e.g., in terms of durability or referring to standards.

For attributes with interval or ratio scale, the analyst has full flexibility in choosing the number of attribute levels. For estimating linear effects, two levels are sufficient, however for nonlinear effects one would need more than two levels. Using (orthogonal) polynomial functions, three levels would allow estimating linear and quadratic effects, while four levels would also allow estimating cubic effects. The attribute level range has a large influence on the reliability of the parameter estimates. In general, a wide attribute level range (e.g., $10 to $50) leads to smaller standard errors than a narrow range (e.g., $25 to $30), but one should always make sure that the attribute levels are realistic and appropriate relative to other attributes. Further, in choosing the exact values of the quantitative levels, one should favour rounded values (e.g., $5, $10) over values that increase cognitive burden (e.g., $4.75, $9.90). Finally, one generally prefers equidistance attribute levels that cover the range equally (e.g., $5, $10, $15) over levels that are not equidistant (e.g., $5, $8, $15), unless the latter provides a more realistic representation of an attribute.

Table 2: Attributes in laptop choice example

Attribute	Level	Ranking order
Processor	Intel Core i3	3
	Intel Core i5	2
	Intel Core i7	1
Hard-disk storage	256 GB	3
	512 GB	2
	1 TB	1
Price	$1500	1
	$1800	2
	$2100	3

As an example, consider an unlabelled laptop choice experiment with three attributes, namely processor, hard-disk storage, and price. Each attribute is assumed to have three levels, given in Table 2. Processor is measured on an ordinal scale, while hard-disk storage and price have a ratio measurement scale. The levels have a clear ranking order, where 1 is the most preferred level and 3 is the least preferred level. This ordering allows us to assess whether there exists a strictly dominant alternative in a choice task.

Empirically, the number of attribute levels has been found to have a significant impact on the behavioural outcomes of choice experiments by several authors. Wittink et al. (1989) found that adding an intermediate level to a two-level attribute resulted in increasing the relative importance of an attribute, and in a subsequent study, Wittink et al. (1992) found that the number of levels influences the relative importance of an attribute, an effect that was magnified in the presence of dominated alternatives. Van der Waerden (2004) concluded that the number of attribute levels can influence choice outcomes, finding that the number of attribute levels present in an experiment influences the scale of utility. Hensher (2006) found mixed evidence that the number of attribute levels affects the probability of respondents ignoring an attribute when completing choice experiment tasks, affecting some but not all attributes contained within the experiment. Caussade et al. (2005) report that the number of attribute levels employed has a statistically significant impact upon the degree of error variance present within the data, however they conclude that the impact is marginal, having the second lowest effect out of all the design dimensions they varied. Rose et al. (2009) found that the number of attribute levels used has a significant impact upon WTP estimates, however that these differences depend upon which country the data were collected from. Meyerhoff et al. (2015) found the impact that the number of attributes, alternatives and choice tasks has on modelled outputs differs according to the socio-demographic profile of the agents, with the biggest impact being on the drop-out rate of the survey itself. Finally, Oehlmann et al. (2017) found that as the attribute level range increases, the probability of selecting a status quo alternative increases, likely due to signals sent to respondents about certainty in the options shown throughout the experiment.

A further experimental design dimension that has received attention in the past is the effect that attribute level range plays on behavioural responses. Meyer and Eagle (1982) and Eagle (1984) found that attributes with larger ranges produced larger effects than ones with smaller relative ranges, all else being equal. Ohler et al. (2000) on the other hand found attribute range differences affect experimental outcomes in terms of complexity of functional forms, model fit, power to detect non-additivity, and between-subject response variability. No effect was found on model parameters, within-subject response variability, or error variance. In contrast to Ohler et al., Caussade et al. (2005) concluded that attribute range significantly impacts upon error variances, and that changes to the range that attribute levels take had the third largest influence on error variances out of all the design dimensions they tested. Hensher (2004) found that increasing the range of attribute levels resulted in lower mean WTP values, whilst Rose et al. (2009) found significant impacts on WTP estimates given changes to attribute level ranges, however the directions of the impacts varied across different data sets.

Step IV: Determine number of choice tasks in experimental design

An experimental design contains the profiles (i.e., attribute levels of all alternatives) of all (unique) choice tasks in set S and can be represented by design matrix **X**, where each row consists of a choice task, and each column represents an attribute in an alternative. If each agent $n \in \{1,\ldots,N\}$ is shown the same choice tasks, i.e., each agent is subject to all choice tasks in the design matrix, then **X** is referred to as a *homogeneous* design. If different agents face different choice tasks, i.e., each agent is shown only a subset of choice tasks $S_n \subset S,$ then **X** is referred to as a *heterogeneous* design. Heterogeneous designs are generally assumed to be a better choice because they provide more information (Sándor and Wedel, 2005), although a homogeneous design can be justified if the number of parameters to be estimated is small relative to the number of choice tasks (Kessels, 2016). In most cases, a heterogeneous design is constructed by first explicitly creating design matrix **X** and then splitting it into two or more parts called *blocks*. Each block represents a different version of the choice experiment, whereby agents are distributed among these blocks (as evenly as possible). Instead of first creating a (large) *explicit* design matrix, one can also generate random choice tasks on-the-fly for each agent *n*, in which case the design matrix **X** is *implicit*.

The size of design matrix **X** is defined by the number of choice tasks, $|S|$. The required size depends on the total number of parameters to estimate in the choice model. Let K denote the total number of parameters, including label-specific constants and coefficients of attributes that are dummy, effects or contrast coded. There needs to be sufficient variation in design matrix **X** to estimate these K parameters. When an agent makes a choice among $|J|$ alternatives in a certain choice task $s \in S,$ this provides information that the chosen alternative is preferred over each of the other $|J|-1$ alternatives shown to the agent. In other words, a design **X** consisting of $|S|$ choice tasks provides $|S|\cdot\left(|J|-1\right)$ pieces of information. To be able to estimate K parameters, it must hold that $|S|\cdot\left(|J|-1\right)\geq K,$ in other words, the minimum size of the design can be determined by finding the smallest integer $|S|$ that satisfies:

$$|S| \geq \frac{K}{|J|-1}. \qquad (3)$$

The difference between the actual number of choice tasks in the design and the minimum required design size is referred to as the *degrees of freedom*.

As an example, consider the laptop choice example with the three attributes shown in Table 2. Suppose that two alternatives are shown at each choice ask, i.e., $|J| = 2.$ Further, assume that the processor attribute is dummy coded such that it has two associated parameters, whilst storage and price are assumed to be continuous variables, each with a single parameter, such that $K = 4.$ Then according to Eqn. (3) it should hold that $|S| \geq 4.$ While a design matrix of size 4 would be sufficient, increasing the degrees of freedom (and hence increasing variety in the design data) is recommended to improve identification of the parameter estimates. The number of choice tasks $|S|$ is often set to at least two or three times the minimum size to have sufficient degrees of freedom.

In choosing $|S|,$ one may also want to consider *attribute level balance* constraints. A design matrix is attribute level balanced if each attribute level appears an equal number of times across all choice tasks. Considering three levels in our laptop choice example in Table 2, attribute level balance could be guaranteed if the design size is a multiple of three, i.e., 6, 9, 12, etc. If the price attribute would have four levels, then attribute level balance would require that $|S|$ is divisible by three and four, i.e., 12, 24, 36, etc. Attribute level balance is not a requirement, but some degree of balance is often considered desirable to obtain a good coverage over the data space.

If the number of choice tasks $|S|$ is too large to show a single agent, then one can move from a homogeneous design to a heterogeneous design by *blocking* the design into smaller parts. For example, if $|S| = 24$ then one can block the design for example into four parts of six choice tasks each, or three parts of eight choice tasks each, or two parts of 12 choice tasks each. The number of choice tasks to show to each agent, $|S_n|,$ depends on the complexity of each choice task and how many the analysts believe an agent can handle without significant fatigue (which is a bigger issue with online surveys than face-to-face interviews). Survey instruments for choice experiments can often select a block or a random subset of choice tasks from a given explicit design matrix **X**, therefore implementing a heterogeneous design is not necessarily complicated.

Mixed evidence exists as to the impact the number of choice tasks has empirically upon choice experiments. Caussade et al. (2005) and Hensher (2004, 2006) found that the number of choice tasks acts upon the error variance of discrete choice models, however the effects reported by both Caussade et al. (2005) and Hensher (2004) were only marginal. Interestingly, Caussade et al. (2005) keeping the choice context constant whilst systematically varying all possible design dimensions across a sample of respondents, found that the number of choice tasks a respondent saw had the least influence of any of the design dimensions on the error variance of choice data. Brazell and Louviere (1998), keeping all other design dimensions constant, varied only the number of choice tasks shown to each respondent to be between 16 and 120. In their study, they found evidence of learning and fatigue effects, however they concluded that there exist no significant differences in either internal reliability or model variability for models estimated from survey questionnaires with varying numbers of choice tasks. Likewise, Hensher et al. (2001) reported finding that increasing the number of choice tasks had only a marginal impact upon model elasticities, however differences in elasticities were observed when agents were presented with 24 and 32 choice tasks compared to less. Hensher et al. recommend using more than four choice tasks with 16 being sufficient for most modelling efforts. Beck et al. (2011) found only minor impacts on the mean WTP estimates obtained from choice experiments with different numbers of choice tasks whilst Rose et al. (2009) found mixed evidence for impacts of the number of choice tasks upon WTP estimates, with differences observed across different countries. In this later study, the authors found that the number of choice tasks had almost no impact on a data set collected within an Australian context, a limited impact on the same survey collected in Taiwan, and a very large impact using the same survey in Chile. More recently, Czajkowski et al. (2014) report that many observed discrepancies in modelled outcomes over choice tasks can be mitigated if error variance differences are properly accounted for, whilst Campbell et al. (2015) found that failure to account for learning and fatigue effects present within choice data can significantly impact WTP outputs. Oehlmann et al. (2017) report that all else being equal, increasing the number of choice tasks increases the probability that a status quo alternative will be chosen. Finally, Oehlmann et al. (2017) recommend that all else being equal, between 10 and 15 choice tasks is optimal in practice.

Step V: Choose experimental design strategy

In this section, we assume that the aim is to determine a design matrix **X** for the estimation of a conditional logit model, also referred to in the literature as a multinomial logit model[2], which is the work horse of discrete choice models. Choice probabilities in the conditional logit model are given by

[2] McFadden (1973) made a distinction between a multinomial model and a conditional logit model. In his definition, a multinomial logit model only contains variables related to the respondent (i.e., socio-demographics), whereas a conditional logit model only contains variables related to the alternatives (i.e., attributes). Therefore, according to these definitions, conditional logit is the appropriate term when we refer to data in a stated choice experiment. However, in practice, both socio-demographics and attributes appear in the utility functions and in the literature the term multinomial logit became the dominant term to indicate this type of model.

$$p_{nsj} = \frac{\exp(V_{nsj})}{\sum_{i \in J} \exp(V_{nsi})}, \tag{4}$$

and the Fisher information matrix for the conditional logit model is a $K \times K$ matrix **F** that can be computed as (McFadden, 1973)

$$\mathbf{F} = \sum_{n=1}^{N} \sum_{s \in S_n} \sum_{j \in J} \left(\mathbf{x}_{nsj} - \overline{\mathbf{x}}_{ns} \right)' p_{nsj} \left(\mathbf{x}_{nsj} - \overline{\mathbf{x}}_{ns} \right), \qquad \text{with} \qquad \overline{\mathbf{x}}_{ns} = \sum_{i \in J} \mathbf{x}_{nsi} p_{nsi}. \tag{5}$$

Different types of choice models result in different matrices **F**, for example Sándor and Wedel (2002) derived the Fisher information matrix for the cross-sectional mixed logit model, Bliemer et al. (2009) for the nested logit model, and Bliemer and Rose (2010) for the panel mixed logit model. It is possible to design data specifically around more advanced choice models, but this may come at a significant computational cost and may even be practically infeasible. Therefore, at the design stage it is common to design the data while having a conditional logit model in mind. Note that this generally does not prohibit the estimating of more advanced models at a later stage. As noted by Bliemer and Rose (2010), data that is designed for estimating a conditional logit model will generally also work well for estimating a panel mixed logit model.

The (asymptotic) variance-covariance matrix of parameter estimates, $\boldsymbol{\Omega} = \mathrm{var}(\hat{\boldsymbol{\beta}})$, is the inverse of the Fisher information matrix, i.e., $\boldsymbol{\Omega} = \mathbf{F}^{-1}$. The diagonal elements of matrix $\boldsymbol{\Omega}$ are directly related to the standard errors of the parameter estimates, namely the standard error of parameter β_k equals $\sqrt{\Omega_{kk}}$, where Ω_{kk} is the k^{th} diagonal element of matrix $\boldsymbol{\Omega}$. A good design matrix **X** ensures that each parameter receive (non-zero) Fisher information such that they can all be estimated, and that parameter estimates are reliable (i.e., small standard errors). From Eqn. (5) we can make the following observations. First, Fisher information for the conditional logit model only depends on attribute levels and choice probabilities, not on choice observations, therefore Fisher information can be determined based on experimental design **X** and best guesses of the choice probabilities for each alternative and each choice task. The same holds for the cross-sectional mixed logit model and nested logit model, but the panel mixed logit model unfortunately requires simulated choice observations. Second, no Fisher information is obtained for choice tasks with a dominant alternative (with $p_{nsj} = 1$ for a certain alternative *j*). Third, no Fisher information is obtained if attribute levels overlap across alternatives such that no trade-offs are made. Fourth, more Fisher information is obtained if levels of quantitative attributes are further apart (wide range). And finally, in case of a homogeneous design where all agents face the same choice tasks, Fisher information increases linearly with sample size *N*, which means that $\boldsymbol{\Omega}$ is proportional to $1/N$ such that standard errors decrease at a rate of $\sqrt{N}$.

In this section we discuss three main types of design strategies, namely *efficient designs*, *orthogonal designs*, or *random designs*, and we discuss advantages and disadvantages of each strategy.

Efficient designs

Efficient designs have become the state-of-the-art in experimental design in the past decade. A design matrix **X** is *efficient* if it captures a large amount of Fisher information. Since it is generally not possible to determine the *most* efficient design, the typical aim is to generate a design that is efficient without claiming that it is optimal. To maximise Fisher information, the volume of matrix **F** can be maximised, which is equal to minimising the volume of variance-covariance matrix $\mathbf{\Omega}$.

A $K \times K$ matrix can be represented as a hypercube in K dimensions. The lengths of the edges of a matrix are given by its eigenvalues $\boldsymbol{\lambda}$, where λ_k is the eigenvalue for dimension k, which in matrix **F** corresponds to parameter β_k, $k \in \{1,\ldots,K\}$. Eigenvalues are determined via an eigen decomposition where matrix **F** is decomposed as $\mathbf{F} = \mathbf{Q\Lambda Q}^{-1}$, where **Q** is a matrix of eigenvectors that span the hypercube and $\Lambda = \text{diag}\{\lambda_1,\ldots,\lambda_K\}$ is a diagonal matrix with eigenvalues of **F**, and the volume can be computed by multiplying the lengths of the edges of the hypercube. If $K = 2$, one multiplies the length and width to obtain the volume of the square, if $K = 3$ one multiplies the length, the width, and the height to obtain the volume of the cube, etc. The volume of Fisher information is therefore given by the determinant of **F**,

$$\det(\mathbf{F}) = \prod_{k=1}^{K} \lambda_k. \tag{6}$$

A related measure to the volume of Fisher information is the *D-error*, which is defined as the determinant of the variance-covariance matrix to the power $1/K$ to normalise the measure and account for the number of parameters,

$$\text{D-error} = \left(\det(\mathbf{\Omega})\right)^{1/K} = \left(\frac{1}{\det(\mathbf{F})}\right)^{1/K}. \tag{7}$$

As a result, minimising the D-error equals maximising the volume of Fisher information. The literature commonly refers to D-efficient designs to indicate a low D-error. There exists no threshold for a ‘good’ D-error value is since this is case-specific and cannot be compared across studies, so all that can be said is that lower is better. It is generally also not possible to compute the lowest D-error value since this requires exhaustive evaluation of all possible experimental designs, which is not practically feasible. To illustrate, consider our simple laptop choice example with two alternatives with the attribute and levels shown in Table 2. This means that each alternative has $3^3 = 27$ unique profiles, such that there exist $27^2 = 729$ unique choice tasks. Suppose that one is interested in determining the most efficient design consisting of six choice tasks. Choosing the best six choice tasks out of 729 possible choice tasks (without replacement) would require the evaluation of $729!/(729-6)!\ \approx 147,030,187,802,098,000$ unique designs, which would take even the fastest computer a very long time to complete.

To compute the efficiency of a design, utility functions need to be fully specified, including any interaction effects, nonlinearities, and possible qualitative coding (e.g., dummy coding). If an analyst tries to optimise the data for a choice model where one or more parameters are not identifiable (e.g., due to overspecification, due to lack of variation in attribute levels, or due to self-imposed multicollinearity via constraints), then the volume of Fisher information will be zero and the D-error will be infinite/undefined. Therefore, the D-error informs the analyst whether the model as specified can be estimated based on the specified attribute levels and constraints; a finite D-error (usually smaller than 1) gives confidence that the data can be used for model estimation.

In addition to D-efficient designs, other design types such as A-efficient designs (see e.g., Huber and Zwerina, 1996) or C-efficient designs exist (see e.g., Scarpa and Rose, 2008). An A-efficient design minimises A-error related to the circumference, instead of volume, of the Fisher information matrix, and a C-efficient design is used when optimisation of some function of parameters is of interest, such as WTP estimates. Many other efficient design types exist (see Kessels et al., 2006), all measuring information in a slightly different way, but D-error is by far the most widely used information criterion and is recommended in most cases.

The main advantage of using an efficient design is that it captures (near) maximum information for a specific model, which means that it enables significant and/or reliable parameter estimates at smaller sample sizes that other design strategies. This makes efficient designs particularly useful if one is restricted either by budget or by a limited population of specific agents (e.g., pilots, physicians, patients with a certain disease, managers in a firm, etc.). Further, efficient designs are very flexible and can be used in conjunction with various constraints on attribute levels (Collins et al., 2014), for example to avoid attribute levels that are unrealistic or impossible, and can avoid dominant alternatives (Bliemer et al., 2017). The main disadvantages of an efficient design strategy are that efficient designs cannot be determined manually and require the use of optimisation algorithms, and that efficiency is sensitive to prior information about the expected choice probabilities in each choice task. To determine these expected choice probabilities, best guesses of the (unknown) parameter values, referred to as *priors*, are needed.

Different types of priors can be used to generate efficient designs. Although priors in experimental design have a somewhat different meaning than priors in Bayesian statistics, we use similar terminology to indicate the various types of priors. Two main types of priors can be distinguished, namely informative priors and noninformative priors. *Informative priors* are based on prior knowledge obtained from a pilot study, the literature (although being aware of possible scale, culture, and country effects), or expert judgement (see e.g., Bliemer and Collins, 2016), whereas *noninformative priors* are not based on any prior information except for the possible knowledge of the sign of the parameter. In practice, one would typically not mix the two types of priors in generating an efficient design. Each of these two types of priors can be set using either a fixed value, referred to as a *local* prior, or as a probability distribution, referred to as a *Bayesian* prior. One can mix local and Bayesian priors in generating an efficient design. Table 3 shows examples of the various types of priors for specific parameter, where noninformative priors have a value of zero or a uniform distribution around zero, or in case of knowledge of the sign, a near-zero positive or negative value or a uniform distribution with an upper or lower bound of zero. The further these priors (set when generating an efficient design) deviate from the true parameter values (obtained via model estimation after the data collection), the more efficiency will be lost. Choosing bad priors can also lead to *in*efficient designs (see for example the simulation study described in Walker et al., 2017), therefore choosing appropriate priors needs to be done deliberately, and if uncertain, it is best to choose noninformative (zero) priors or conservative (close to zero) priors.

Table 3: Types of priors and examples

	Local	Bayesian
Informative priors	$\beta_k = -0.5$, $\beta_k = 0.8$, $\beta_k = 1.2$	$\beta_k \sim \text{Normal}(-0.5, 0.2)$, $\beta_k \sim \text{Normal}(0.8, 0.5)$, $\beta_k \sim \text{Lognormal}(1.2, 0.9)$
Noninformative priors	$\beta_k = 0$, $\beta_k = -0.000001$, $\beta_k = 0.000001$	$\beta_k \sim \text{Uniform}(-1, 1)$, $\beta_k \sim \text{Uniform}(-1, 0)$, $\beta_k \sim \text{Uniform}(0, 2)$

Several software tools exist containing algorithms to locate efficient designs, including Ngene (ChoiceMetrics, 2018), the '%ChoiceEff' macro in SAS (Zwerina et al., 2010), and the 'idefix' package (Traets et al., 2020) in R. Each tool allows the minimisation of the D-error via either a column-swapping algorithm, row-swapping algorithm, and/or a coordinate-swapping algorithm. A coordinate-swapping algorithm such as proposed by Meyer and Nachtsheim (1995) is mainly useful for generating optimal designs without constraints, a column-swapping algorithm (e.g., Huber and Zwerina, 1996) is particularly useful for designs with attribute level balance constraints, and a row-swapping algorithm like the modified Federov algorithm (Cook and Nachtsheim, 1980) is particularly useful for designs with attribute level or dominance constraints.

Orthogonal designs

Orthogonal designs have been used for choice experiments since the 1980s and have been the default design approach for several decades. A design matrix **X** is called an orthogonal array if it is attribute level balanced and if for each two attributes, each pair of attribute levels appears equally across the choice tasks. If attributes have different numbers of levels, then such arrays are often referred to as *mixed* orthogonal arrays, in contrast to conventional *fixed-level* orthogonal arrays (Hedayat et al., 1999). Attribute levels in (fixed-level or mixed) orthogonal arrays are uncorrelated (by definition), therefore multicollinearity is avoided.

The main advantages of orthogonal designs are that they cover the attribute space nicely, and no skill or running algorithms is required since they can be found in lookup tables in books (e.g., Hahn and Shapiro, 1967) or in online libraries (simply conduct a web-search for 'orthogonal array' to find the most recent sets of (mixed) orthogonal arrays as new arrays are being found and added over time). Further, orthogonal arrays allow blocking of the design matrix in such a way that it maintains attribute level balance within each block. Several disadvantages of orthogonal designs exist. First, orthogonal arrays only exist for specific combinations of the number of attributes and attribute levels. If attributes have a varying number of levels where some have more than four levels, then an orthogonal array will likely not exist. Secondly, orthogonal arrays have a very rigid structure, meaning that it is generally not possible to impose constraints on attribute levels or avoid dominant alternatives. One could manually remove choice tasks from the orthogonal design that violate certain constraints or contain dominant alternatives, but that would mean that the design is no longer orthogonal. Orthogonality is also lost in the data when considering interaction effects in the utility function that were not considered when locating an orthogonal array, when using dummy or effects coding, or when there are missing observations, such as unequal representation of blocks in the data or unanswered choice tasks due to fatigue.

Independent estimation of parameters has often been claimed as a benefit of using orthogonal design, but it should be noted that this benefit holds for estimating linear regression models and does *not* hold for the estimating choice models. If design matrix **X** is orthogonal and all choice tasks are utility balanced, i.e., $V_{nsj} = V_{nsi}$ for all alternatives $j \neq i$ such that $p_{nsj} = 1/\left|J\right|$, then **F** becomes a diagonal matrix, such that $\mathbf{\Omega}$ is also diagonal, which would imply that parameter estimates are uncorrelated and can be independently estimated. However, it is impossible to satisfy both orthogonality and utility balance at the same time, unless all parameters are equal to zero. In practical applications, parameters are clearly expected to be non-zero, hence it is in practice not possible to obtain uncorrelated choice data.

Street et al. (2001), Burgess and Street (2003), Street and Burgess (2004) and Burgess and Street (2005) introduced so-called *optimal designs* specifically for unlabelled experiments. These optimal designs are a specific type of orthogonal design that seeks to maximise the Gramian matrix (which is an algebraic characterisation of the equivalent statistical Fisher information matrix, up to a scale) of the conditional logit model, thereby combining efficiency and orthogonality. Street et al. (2005) showed that generating such designs by hand is relatively easy using a simple procedure that ensures minimum overlap of attribute levels across alternatives. Under the (very strict) assumption of utility balance, also referred to as utility neutral, it is possible to analytically compute the lowest possible D-error and therefore express D-efficiency as a percentage, where 100 percent indicates an optimal design. Optimal designs are subject to the same disadvantages of orthogonal designs as mentioned above. Further, they are mainly suitable for unlabelled experiments, and they may be problematic if a dominant attribute exists since the design forces attribute levels to be different across alternatives. For example, in comparing two alternative laptops having brand as an attribute with two levels, Apple and Dell, then agents are always forced to choose between a laptop of brand Apple and a laptop of brand Dell. Depending on the agent's preference for an operating system (MacOS or Windows) they may always choose the alternative with a specific brand and not trade-off on any of the other attributes.

Random designs

While efficient and orthogonal design strategies are systematic approaches in determining a *fractional factorial design* matrix **X** that contains a specific subset of choice tasks, an alternative strategy is simply using randomly generated choice tasks for each agent by selecting choice tasks from an explicitly generated *full factorial design* (containing all possible choice tasks), or by randomly generating choice tasks on-the-fly for each agent. This experimental design strategy also allows the application of constraints and can avoid dominant alternatives. Random designs do not suffer from multicollinearity unless the analyst imposes constraints that perfectly correlates attribute levels.

As mentioned earlier, heterogeneous designs generally contain more information than a homogeneous design. A random design can be considered an extreme version of a heterogeneous design. While individual choice tasks in random designs may not capture a large amount of information, variation in the data is where random designs excel. The fact that each randomly generated choice task may capture different information allows random designs to decrease standard errors at a rate larger than $\sqrt{N}$. Therefore, for a large enough sample size *N*, the amount of information captured with a random design may approach that of a fixed efficient design.

The main advantages of a random design strategy are that no experimental design skills are required (unless attribute level constraints or dominance checks need to be imposed), and the analyst does not need to formulate utility functions in advance since the data will be sufficiently rich to estimate any model. The main disadvantage is that it is an inefficient data collection strategy for small sample sizes and therefore should only be considered sample size is sufficiently large (typically at least 1,000 responding agents).

Agent- or segment-specific experimental designs

To reduce hypothetical bias in choice experiments, one can consider creating familiar choice tasks tailored around real experiences of agents instead of using a fixed design across the entire population (e.g., Hensher, 2010). One way of doing this is via a so-called *pivot design* in which attribute levels are absolute or relative pivots around reference attribute levels reported previously by an agent (Rose et al., 2008). Another way is to create a *library of designs* containing separate designs for specific segments within the population. Both methods can be applied in conjunction with any experimental design strategy (efficient, orthogonal, or random) and are briefly explained below.

Using route choice as a common application in transport, consider asking agents about a recent trip they have made and wanting to tailor the choice tasks around their reported trips. An agent may report a recent trip to work by car that took 25 minutes and where $5 toll was paid. Then in the choice experiment the same agent would be asked to imagine making the same trip to work again and choose between two or more route alternatives where route travel times and toll costs vary around the reported travel time and toll cost. A pivot design is a fixed matrix **X** consisting of pivot levels. In case relative pivots are used, the matrix contains for example levels -25%, 0%, and +25%, which means that for this specific agent the levels shown in the choice tasks would be 25, 30, and 35 minutes for travel time and $4, $5, and $6 for toll costs. Using relative pivot levels, attribute levels automatically scale to make sense for short and long trips. However, relative pivots do not always work, for example if an agent reports to have paid

\$0 in tolls, then the levels shown would be zero toll only. In such cases, one may want to revert to absolute levels, such as +\$1, +\$2, +\$3. Pivoting is generally not needed around qualitative attributes, but it is possible to pivot around attributes with ordinal measurement scale by showing levels that have a ranking order close to the reference input. Implementing a pivot design in a survey instrument typically requires programming rules and logic to deal with all kinds of user input, which may impossible or challenging in certain survey tools.

An alternative to using a fixed pivot design is to generate different designs $\mathbf{X}^{(g)}$ for different population segments g, $g = 1, \dots, G$, and have them available in a library within the survey instrument. In our route choice experiment, we may for example create $G = 24$ different designs based on four categories of trips (work, business, shopping, leisure), two modes of transport (car, public transport), and three distance categories (short, medium, long). Using the same agent as described above, for this agent we would look up and use the design with characteristics ‘work’, ‘car’, and ‘medium’ from the library. The advantage of this approach is that all experimental designs can be generated and checked in advance, although it may require generating many experimental designs.

Step VI: Conduct pre-testing

Once a draft survey has been developed, it needs to be pre-tested. This can be done both qualitatively via focus groups or personal interviews and/or quantitatively via a pilot study (Mariel, 2021). Qualitative testing aims to find out whether the information in the survey is sufficient and well-understood by the target audience (using familiar concepts and terminology), noting that agents have different education levels and backgrounds (Mariel, 2021). Johnston et al. (2017) recommends a minimum of four to six focus groups in survey pre-testing. The purpose of a *pilot study*, typically involving approximately 10 per cent of the total sample size (i.e., $\frac{1}{10}N$), is to get written feedback about the choice experiment and to make sure that a choice model can be estimated before starting the main data collection. In addition to asking for general feedback about the choice experiment, one can ask agents about the difficulty of the choice tasks and how much they enjoyed it to get a sense of choice task complexity and engagement.

One can use an efficient, orthogonal, or random design for the pilot study. An orthogonal design could be useful if (i) most attributes have only two or three levels, (ii) if there is no real concern about dominant alternatives (e.g., if the experiment is labelled with label-specific attributes, or if all attributes are normative without a clear ordering, or if no obvious preference structure exists among attribute levels), and (iii) if there do not exist unrealistic attribute level combinations. In other cases, one could use an efficient design if sample size is small or a random design if sample size is large, while in both cases applying possible constraints and excluding choice tasks with dominant alternatives. When using an efficient design in the pilot study, one could use noninformative (zero) priors to indicate that no prior information is available about the parameters.

As an example, Table 4 shows an optimal orthogonal design for our laptop choice example with two alternatives using the method of Street et al. (2005). Syntax 1 in Appendix A shows how to generate this design in Ngene. One can check that the attribute levels for Laptop A (and Laptop B) are orthogonal since each attribute level combination appears the same number of times, for example combination (Core i5, 256 GB) appears once, (1 TB, \$1500) appears once, (Core i3, \$2100) appears once, etc. It is an optimal orthogonal design because there is minimum overlap, namely processor, amount of storage, and price are always different across the two alternatives. Despite it being optimally efficient (under the assumptions of linear utility functions, orthogonality, and utility balance or zero priors), it has two problematic choice tasks, namely Laptop B has a strictly dominant profile (and is expected to be always be chosen) in choice tasks 7 and 8. These choice tasks can easily be identified by substituting the attribute levels with their ranking order according to Table 2, e.g., Laptop A has attributes with ranking orders (3,3,3) in choice task 7, while Laptop B has a profile with ranking orders (2,2,1), making it better in each attribute.

Table 4: Optimal orthogonal design for laptop choice example

	Laptop A			Laptop B		
Choice task	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	256 GB	\$1500	Core i7	512 GB	\$1800
2	Core i7	512 GB	\$1500	Core i3	1 TB	\$1800
3	Core i3	1 TB	\$1500	Core i5	256 GB	\$1800
4	Core i7	256 GB	\$1800	Core i3	512 GB	\$2100
5	Core i3	512 GB	\$1800	Core i5	1 TB	\$2100
6	Core i5	1 TB	\$1800	Core i7	256 GB	\$2100
7	Core i3	256 GB	\$2100	Core i5	512 GB	\$1500
8	Core i5	512 GB	\$2100	Core i7	1 TB	\$1500
9	Core i7	1 TB	\$2100	Core i3	256 GB	\$1500

Table 4 shows an attribute-level balanced D-efficient design assuming noninformative (zero) priors (i.e., utility balance) for the laptop choice example, generated using the default swapping algorithm in Ngene where explicit constraints to avoid dominant alternatives have been applied (we refer to Syntax 2 in Appendix A for the Ngene script). For the computation of the D-errors, the following utility function was assumed:

$$f(\mathbf{x}) = \beta_1 x_{\text{Processor}}^{(\text{Core i5})} + \beta_2 x_{\text{Processor}}^{(\text{Core i7})} + \beta_3 \log(x_{\text{Storage}}) + \beta_4 x_{\text{Price}}, \tag{8}$$

where $x_{\text{Processor}}^{(\text{Core i5})}$ and $x_{\text{Processor}}^{(\text{Core i5})}$ are dummy-coded binary variables using level 'Core i3' as the base level, x_{storage} is the hard-disk storage in GB (i.e., 256, 512, 1024), x_{Price} is the price in dollars, and $(\beta_1, \beta_2, \beta_3, \beta_4)$ are parameters to be estimated. In this example we have applied a transformation via the natural logarithm on the storage variable under the hypothesis that there is diminishing benefit in additional storage space (i.e., at some point enough is enough).

The D-error of the design in Table 5 for the above model specification is 0.0272, which is slightly better than the D-error of 0.0287 that would result from the design in Table 4 (which imposes orthogonality constraints but not dominance constraints) despite some overlap in the storage and price attribute. Efficiency of the design can be further improved by removing the attribute-level balance constraint; Table 6 shows the design with the lowest D-error without dominant alternatives (generated using the modified Federov algorithm in Ngene), which has no overlap and a D-error of 0.0225. The design in Table 6 is clearly not attribute-level balanced. Dummy (or effects) coded attributes will generally show a high degree of attribute-level balance across the two alternatives since a low representation of a certain level would not capture much information for the corresponding parameter and therefore lead to a high D-error. However, for other attributes it is typically more efficient to show the most extreme levels (at least when assuming zero priors), as this increases the trade-offs made in each choice task and hence increasing Fisher information, such that middle level 512 GB for storage and $1800 for price appear only once within the nine choice tasks.

Table 5: Attribute-level balanced D-efficient design with noninformative zero priors without dominant alternatives for laptop choice example

	Laptop A			Laptop B		
Choice task	Processor	Storage	Price	Processor	Storage	Price
1	Core i7	1 TB	$2100	Core i3	256 GB	$1800
2	Core i3	256 GB	$1500	Core i7	1 TB	$2100
3	Core i7	512 GB	$1500	Core i5	1 TB	$2100
4	Core i5	1 TB	$1500	Core i7	256 GB	$2100
5	Core i3	1 TB	$1800	Core i5	512 GB	$1800
6	Core i3	512 GB	$2100	Core i7	256 GB	$1500
7	Core i7	512 GB	$1800	Core i5	512 GB	$1500
8	Core i5	256 GB	$2100	Core i3	1 TB	$1500
9	Core i5	256 GB	$1800	Core i3	512 GB	$1800

After having generated an experimental design (or a library of multiple segment-specific designs), one needs to choose a survey instrument. For the pilot study one may simply use a pen and paper questionnaire or an Excel spreadsheet (e.g., Black et al., 2005), but in most cases one would implement the choice experiment in an online (for web-based surveys) or offline (for CAPI surveys) software tool that will also be used in the main study. Tools that support choice experiments include SurveyEngine, Confirmit, Nebu, and Qualtrics (with choice-based conjoint add-on module). Most free online survey tools do not support choice experiments, but for simple choice experiments one may use the tricks such as creating multiple-choice questions with images that are screenshots of profiles or whole choice tasks.

Table 6: D-efficient design with noninformative zero priors without dominant alternatives for laptop choice example

	Laptop A			Laptop B		
Choice task	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	256 GB	$2100	Core i3	1 TB	$1500
2	Core i5	1 TB	$1500	Core i7	256 GB	$2100
3	Core i5	1 TB	$2100	Core i7	256 GB	$1500
4	Core i5	256 GB	$1500	Core i3	1 TB	$2100
5	Core i3	256 GB	$1500	Core i7	1 TB	$1800
6	Core i3	256 GB	$1500	Core i5	1 TB	$2100
7	Core i7	256 GB	$1500	Core i3	512 GB	$2100
8	Core i7	256 GB	$2100	Core i3	1 TB	$1500
9	Core i5	256 GB	$1500	Core i7	1 TB	$2100

As mentioned in Step I, for labelled experiments it is important to randomise (across agents, not within an agent) the arrangement of labelled alternatives shown in choice tasks to be able to account for possible presentation order effects of alternatives (e.g., left-to-right bias). In model estimation, one would include a generic dummy coded variable in the utility functions of all alternatives that indicates the order in which the alternative appeared in the choice task (essentially making order an ‘attribute’ of each alternative).

To account for presentation order effects of *attributes*, one may also want to randomise (again across agents, not within an agent) the order in which attributes are shown to respondents as their relative position (e.g., top or bottom) may have a significant impact upon the behavioural responses of agents completed choice tasks, also referred to as. For example, Kjaer et al. (2006) varied the location of the price attribute, presenting it as either the first attribute or last attribute shown in the task. They found that the order of the price attribute led to statistically significant differences in price sensitiveness, however they concluded that attribute presentation order did not result in different decision rules being used by the sampled respondents. In an earlier study, Scott and Vick (1999) reversed the order in which attributes were shown to responding agents and found statistically significant evidence of an attribute ordering effect on the model outcomes. On the other hand, Farrar and Ryan (1999) found no such evidence when they swapped the first two attributes with the bottom two attributes. Likewise, Boyle and Özdemir (2009) suggest that it is not a forgone conclusion that the ordering of attributes will affect choices and statistical results; it is likely to be a study-specific issue. More recently, Logar et al. (2020) found that attribute order had no significant impact on WTP estimates in standard models, but did significantly impact attribute non-attendance (e.g., people ignoring certain attributes when making their choices). Interestingly, Weller et al. (2014), who did not explore attribute order effects, found that other design dimensions had no impact on attribute non-attendance.

After the pilot study, the analyst would use the collected choice data to estimate a conditional logit model and verify that the model parameters can be estimated resulting in parameter estimates $\hat{\beta}_k,\ k = 1,\ldots,K,$ with corresponding standard errors s_k that indicate the precision (reliability) of the estimates. In the pilot study, it is likely that some or all parameters are not statistically significant given the relatively small sample size. For parameters that are statistically significant, one can check whether they have the expected signs (e.g., price or cost coefficients are expected to be negative). If some parameters have an unexpected sign when using an efficient design, then one may want to check for strong correlations between certain attributes in profiles. For example, if in our laptop choice experiment the price attribute is always high (low) when storage space is large (small), then the parameter for price may become positive if agents generally prefer to have a large hard-disk. This can be remedied by including profiles with a low price and large storage space or high price and small storage space (while at the same time avoiding that this alternative becomes dominant via trade-offs on other attributes) or using an orthogonal design (which avoids such correlations by definition but may suffer from dominant alternatives).

Since parameter estimates by themselves are difficult to assess, one often looks at marginal rates of substitution (MRS) between attributes, of which WTP is a special case. The MRS represents the amount of attribute *l* (i.e., the cost attribute in case of WTP) one has to give up for the gain of one additional unit of attribute *k* such that the utility remains the same. For example, in our laptop choice experiment with utility function (8) the WTP to have a Core i7 processor instead of a Core i3 processor equals $-\beta_2 / \beta_4$ dollars, and the WTP for an increase in hard-disk storage is $-(\beta_3 / x_{\text{storage}}) / \beta_4$ dollars per GB, based on a chosen storage level x_{storage}.

A pilot study may also produce useful parameter priors for generating a more efficient design for the main study as further explained in the next section.

Step VII: Conduct main study

The main study can use the experimental design for the choice experiment as used in the pilot study (possibly after making some minor changes). However, one could improve the efficiency of the data collection by generating a new experimental design using information from the pilot study. In particular, parameter values $\hat{\beta}_k$ estimated using data from the pilot study can replace the zero priors used previously. We refer to such non-zero priors are *informative local priors*. Using informative local priors means that we no longer assuming utility balance (i.e., equal choice probabilities) but rather use choice probabilities that are expected to be closer to the truth. This results in a more accurate measure of Fisher information, thereby allowing a better optimisation of the experimental design.

Suppose that the parameter estimates obtained via a pilot study for our laptop choice experiment are given by $\hat{\beta}_1 = 0.35$ and $\hat{\beta}_2 = 0.5$ (for the dummy coded processor attribute), $\hat{\beta}_3 = 0.6$ (for the logarithmic storage attribute), and $\hat{\beta}_4 = -0.004$ (for the price attribute). Using these values as local priors (instead of zeros) we can again generate a D-efficient design (see Syntax 4 in Appendix A for the Ngene syntax). Assuming that attribute level balance is not required, we find the experimental design shown in Table 7. This design has a D-error of 0.0413. It is important to emphasise that this D-error is not comparable to D-errors of designs that were generated under different prior assumptions such as the designs generated in the previous section using zero priors. If the informative local priors equal the true parameter values, then the design in Table 7 captures maximum information. One can observe that the price levels across the two alternatives in Table 7 are much more balanced than in Table 5. This is a direct effect of using informative local priors. Since a prior value -0.004 for the price parameter indicates that price is relatively important in choosing a laptop (see discussion below), making comparisons only between extreme price points \$1,500 and \$2,100 would often result in choice tasks where price dominates. In such cases, little to no trade-offs are made with respect to processor and storage and hence little information is captured with respect to these two attributes. Therefore, using informative priors when generating a D-efficient design assists in ensuring that agents make trade-offs across all attributes, especially when one or more dominant attributes exist.

The *relative importance of each attribute* in the experimental design can be determined by looking at the relative impact each attribute has on utility (Orme, 2005). Considering again the laptop choice example and the given priors, the processor attribute contributes between 0 (Core i3) and 0.5 (Core i7) to utility, the storage attribute contributes between $0.6 \cdot \log(256) = 3.33$ and $0.6 \cdot \log(1024) = 4.16$ to utility, and the price attribute contributes between $-0.004 \cdot 1500 = -6$ and $-0.004 \cdot 2100 = -8.4$ to utility. Looking at the range in utility contribution, in absolute terms, processor makes a maximum difference of 0.5, storage makes a maximum difference of 0.83, and price makes a maximum difference of 2.4 in utility. Expressing this in percentages, the relative importance of processor, storage, and price is 13 percent, 22 percent, and 64 percent, respectively. In other words, price is the most important attribute in the choice experiment. We point out that assessment of attribute importance can*not* be based on the size of corresponding parameter values since measurement scales and units of attributes are different.

Table 7: D-efficient design with informative local priors without dominant alternatives for laptop choice example

	Laptop A			Laptop B		
Choice task	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	1 TB	$1500	Core i7	256 GB	$1500
2	Core i7	256 GB	$1800	Core i3	1 TB	$2100
3	Core i5	1 TB	$1800	Core i3	256 GB	$1500
4	Core i5	256 GB	$1800	Core i3	1 TB	$1500
5	Core i5	256 GB	$1800	Core i7	1 TB	$2100
6	Core i7	1 TB	$2100	Core i3	256 GB	$1800
7	Core i3	1 TB	$1800	Core i5	256 GB	$2100
8	Core i7	256 GB	$1500	Core i3	1 TB	$1800
9	Core i5	256 GB	$1500	Core i7	1 TB	$2100

While a D-efficient design based on informative local priors would be able to capture maximum information under ideal circumstances where prior assumptions are correct, such priors are in practice merely a best guess and will often be considerably different from the final parameter estimates, resulting in some loss of information. The more accurate the informative local priors are, the less information is lost in the data collection. If the informative local priors turn out to be entirely different from the actual parameter values, then the data collection can in fact become very *in*efficient (Walker et al., 2017). To make a D-efficient design more robust against prior misspecification, informative *Bayesian priors* have been proposed (Sándor and Wedel, 2001). A Bayesian prior is different from a local prior in that it does not consider a single value for the prior, but rather considers a range of values via a predefined probability distribution. For example, if one believes that the parameter value for the price attribute in our laptop example lies somewhere between 0 and -0.008 then one could consider a Bayesian prior with a uniform distribution between the two values. In other words, Bayesian priors take the inherent unreliability about prior parameter values into account. The degree of unreliability of each prior can be obtained via standard errors of the parameter estimates in a pilot study. Assuming parameter estimate $\hat{\beta}_k$ and its corresponding standard error s_k that indicates the degree of unreliability of the parameter estimate, a natural choice for a Bayesian prior is to assume a normal distribution with mean $\hat{\beta}_k$ and standard deviation s_k. The *Bayesian D-error* of a design indicates the expected (mean) D-error over the given prior distributions and can be computed via Monte Carlo simulation by taking quasi-random draws from the prior distributions (Bliemer et al., 2008).

Continuing our laptop choice example, assume that the previously mentioned parameter estimates have standard errors $s_1 = 0.2$ and $s_2 = 0.3$ (associated with the dummy coded processor parameters), $s_3 = 0.4$ (associated with the storage parameter), and $s_4 = 0.0025$ (associated with the price parameter). We generated a Bayesian D-efficient design shown in Table 8 (using Ngene Syntax 5 listed in Appendix A), which has a Bayesian (mean) D-error of 0.0499. The Bayesian D-error will always be larger than the D-error of a design that is optimised using local priors, but the associated Bayesian D-efficient design will result in less loss of information when the true parameter values deviate from the informative local priors. Therefore, it is recommended to use a Bayesian D-efficient design as a more robust design strategy, despite the increase in expected D-error.

Table 8: D-efficient design with informative Bayesian priors without dominant alternatives for laptop choice example

	Laptop A			Laptop B		
Choice task	Processor	Storage	Price	Processor	Storage	Price
1	Core i7	1 TB	\$2100	Core i3	256 GB	\$1800
2	Core i5	1 TB	\$2100	Core i7	256 GB	\$1800
3	Core i7	1 TB	\$1800	Core i3	256 GB	\$1500
4	Core i5	256 GB	\$1500	Core i3	1 TB	\$1500
5	Core i7	256 GB	\$1800	Core i5	1 TB	\$1500
6	Core i3	1 TB	\$1800	Core i5	256 GB	\$1500
7	Core i5	256 GB	\$2100	Core i3	1 TB	\$1800
8	Core i5	1 TB	\$2100	Core i7	256 GB	\$2100
9	Core i7	256 GB	\$1500	Core i3	1 TB	\$2100

An often-asked question is “What sample size do I need?” The answer is that this is case-specific, where in some studies only 50 agents are needed to get statistically significant and reliable parameter estimates, whilst in other studies possibly thousands of respondents are needed. If alternatives include attributes that are all highly important (such as the cost attribute in most studies), then all parameters can be estimated with a smaller sample size. In contrast, if most attributes are only marginally relevant in making a choice, then it will require a large sample size to obtain statistically significant parameter estimates. Some rules of thumb have been discussed in the literature, see Rose and Bliemer (2013) for an overview, but one can make some specific *minimum required sample size* calculations if informative parameter priors are available. Using parameter estimates $\hat{\beta}_k,\ k=1,\ldots,K,$ from a pilot study as informative local priors, we can compute the Fisher information matrix and the related asymptotic variance-covariance matrix $\mathbf{\Omega}$. The minimum sample size N_k^* for parameter *k*, such that it can be estimated at a given level of statistically significance, can be computed as (Rose and Bliemer, 2013; De Bekker-Grob et al., 2015):

$$N_k^* = \left(\frac{t_{\alpha/2}}{\hat{\beta}_k}\right)^2 \Omega_{kk}, \qquad (9)$$

where Ω_{kk} is the asymptotic variance of parameter *k* and $t_{\alpha/2}$ indicates the (two-sided) *t*-value at the desired level of significance α (e.g., 1.96 if $\alpha = 5\%$). Values N_k^* are also referred to as S-estimates, and the minimum sample size N^* required to estimate all *K* parameters at a statistically significant level, i.e., $N^* = \max_k\{N_k^*\}$ is also known as the S-error (Rose and Bliemer, 2013). Given that the above minimum sample size computations rely heavily on prior parameter values, they should only be used when using informative priors that are sufficiently reliable, and they should only be used as ballpark figures (e.g., whether one needs tens, hundreds, or thousands of respondents). Note that if a design is blocked, these minimum sample size estimates need to be multiplied by the number of blocks.

Final remarks

This chapter has set out to define the necessary steps to follow in generating a choice survey. Whilst each study will differ in terms of the research objectives, empirical application area, and sampling requirements, following the seven steps outlined here represents current best practice for all choice studies. In any case, six of the seven steps are required to collect any choice data, with only the possibility of not conducting a pilot study being feasible. This does not mean that one should not undertake some form of pilot study however, and indeed, it is highly recommended to do so. Unfortunately, some applied economic fields are better at this than others.

Of the seven steps, most are fairly straightforward and easy to complete. Of course, given the range of possible applications that choice experiments can be applied to, the ease of generating a stated choice experiment should never be taken for granted. Further, those wishing to undertake a stated choice experiment should have more than a working understanding of discrete choice methods, in particular how to properly specify utility functions such that all parameters are identifiable. It is often easy to make what appear to be small innocuous mistakes that can have significant ramifications that only become apparent after the data has been collected. For example, in a model with a status quo alternative containing a (dummy or effects coded) qualitative attribute it is important that the fixed attribute level of the status quo alternative also appears in one or more other alternatives to avoid identification issues in model estimation (see Copper et al., 2012). Any person attempting to design stated choice experiments is encouraged to first properly immerse themselves within the greater literature to fully understand the subtle nuances of discrete choice modelling.

Finally, analysts should be aware of the possible existence of hypothetical bias in choice experiments, e.g., due to the absence of consequences in hypothetical choice tasks or the difficulty in imagining alternatives that may not yet exist. We refer to Penn and Hu (2018) for a meta-analysis of hypothetical bias and to Haghani et al. (2021a) for an extensive overview of empirical evidence of hypothetical bias in choice experiments. To make choices more realistic and incentive compatible one could simulate experiences (e.g., Fayyaz et al., 2021) or introduce consequences (MacDonald et al., 2016). Several other methods exist to reduce hypothetical bias, including cheap talk, solemn oath, honesty priming, indirect questioning, time-to-think, and certainty scales, see Haghani et al. (2021b) for an overview. Stated choice experiments are by no means perfect but are often considered the best alternative in the absence of, or in conjunction with, revealed choice data.

Appendix A: Ngene syntax examples

The following Ngene syntax scripts were used to generate the experimental designs reported in this chapter. Syntax 1 was used to generate the optimal orthogonal design presented in Table 4. Syntax 2 and 2 were used to generate the D-efficient designs presented in Tables 5 and 6, respectively, where the only difference is that the latter uses the modified Federov algorithm, which does not impose attribute level balance (unlike the default swapping algorithm in Ngene that imposes attribute level balance when possible). The parameter priors in Syntax 2 and 3 are essentially set to zero, but to indicate the ranking order of the attribute levels consistent with Table 2 (such that the algorithm can automatically detect and avoid dominant alternatives) very small positive and negative values are used, which are small enough such that the contribution to utility of each attribute is near-zero (i.e., price has a disutility of at most $0.0000001 \cdot 2100 = 0.00021 \approx 0$).

Syntax 1: Optimal orthogonal design

```
design
;alts = LaptopA, LaptopB   ? two alternatives
;rows = 9                  ? design size of 9 choice tasks
;orth = ood                ? generate optimal orthogonal design
                           ? uses algorithm of Street et al. (2005) [default]

;model:
U(laptopA) = proc * PROCESSOR[1,2,0]
           + stor * STORAGE[256,512,1024]
           + cost * PRICE[1500,1800,2100]

                           ? PROCESSOR: 0=Core i3 (base), 1=Core i5, 2=Core i7
                           ? STORAGE: 256, 512, 1024 GB
                           ? PRICE: $1500, $1800, $2100
           /
U(laptopB) = proc * PROCESSOR  + stor * STORAGE + cost * PRICE
$
```

Syntax 2: D-efficient design with attribute level balance using uninformative priors

```
design
;alts = LaptopA*, LaptopB* ? checks for dominant alternatives and duplicates
;rows = 9                  ? design size of 9 choice tasks
;eff = (mnl,d)             ? minimise D-error for the multinomial logit model
                           ? uses column-based swapping algorithm [default]

;model:                    ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.0001|0.0002] * PROCESSOR[1,2,0]
           + stor[0.00001]             * STORAGE[5.545,6.238,6.931]
           + cost[-0.0000001]          * PRICE[1500,1800,2100]
                           ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
                           ? STORAGE = log(256), log(512), log(1024) GB
                           ? PRICE = $1500, $1800, $2100
           /
U(laptopB) = proc * PROCESSOR  + stor * STORAGE + cost * PRICE
$
```

Syntax 3: D-efficient design without attribute level balance using uninformative priors

```
design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9                  ? design size of 9 choice tasks
;eff = (mnl,d)             ? minimise D-error for the multinomial logit model
;alg = mfederov            ? uses row-based modified Federov algorithm

;model:                    ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.0001|0.0002] * PROCESSOR[1,2,0]
           + stor[0.00001]             * STORAGE[5.545,6.238,6.931]
           + cost[-0.0000001]          * PRICE[1500,1800,2100]
                           ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
                           ? STORAGE = log(256), log(512), log(1024) GB
                           ? PRICE = $1500, $1800, $2100
           /
U(laptopB) = proc * PROCESSOR  + stor * STORAGE + cost * PRICE
$
```

Syntax 4 and 5 were used to generate D-efficient designs with informative priors in Tables 7 and 8, respectively. Generating Bayesian D-efficient design requires computing the mean D-error assuming probability distributions for the parameter priors, which requires numerical simulation. In Syntax 5 we used 200 Sobol draws. The number of draws required increases exponentially with the number of Bayesian priors (e.g., 2^K or 3^K for distributions with small standard deviations, 4^K or more for distributions with large standard deviations) and it is recommended to keep the number of Bayesian priors limited (typically not more to eight to ten) and use local priors for the remaining parameters (if any). In choosing which parameters to allocate a Bayesian prior, it is advised to give priority to attributes with a high relative importance as they will have the largest influence on utility and therefore are most sensitive to prior misspecification.

Syntax 4: D-efficient design without attribute level balance using informative local priors

```
design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9                  ? design size of 9 choice tasks
;eff = (mnl,d)             ? minimise D-error for the multinomial logit model
;alg = mfederov            ? uses row-based modified Federov algorithm

;model:                    ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.35|0.5] * PROCESSOR[1,2,0]
           + stor[0.6]            * STORAGE[5.545,6.238,6.931]
           + cost[-0.004]         * PRICE[1500,1800,2100]
                           ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
                           ? STORAGE = log(256), log(512), log(1024) GB
                           ? PRICE= $1500, $1800, $2100
           /
U(laptopB) = proc * PROCESSOR  + stor * STORAGE + cost * PRICE
$
```

Syntax 5: D-efficient design without attribute level balance using informative Bayesian priors

```
design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9                  ? design size of 9 choice tasks
;eff = (mnl,d,mean)        ? minimise (Bayesian) mean D-error
;alg = mfederov            ? uses row-based modified Federov algorithm
;bdraws = sobol(200)       ? quasi-random Sobol draws for Bayesian priors

;model:                    ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[(n,0.35,0.2)|(n,0.5,0.3)] * PROCESSOR[1,2,0]
           + stor[(n,0.6,0.4)]                    * STORAGE[5.545,6.238,6.931]
           + cost[(n,-0.004,0.0025)]              * PRICE[1500,1800,2100]
                           ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
                           ? STORAGE = log(256), log(512), log(1024) GB
                           ? PRICE= $1500, $1800, $2100
           /
U(laptopB) = proc * PROCESSOR  + stor * STORAGE + cost * PRICE
$
```

References

Adamowicz, W. and Boxall, P. (2001) Future Directions of Stated Choice Methods for Environment Valuation, *Choice Experiments: A New Approach to Environmental Valuation*, April, London, England.

Adamowicz, V., Dupont, D., Krupnick, A. (2006) Willingness to Pay to Reduce Community Health Risks from Municipal Drinking Water, a Stated Preference Study, 3rd World Congress of Environmental and Resource Economics, AERE, Kyoto, Japan , August 1st.

Arentze, T., Borgers, A., Timmermans, H. and Del Mistro, R. (2003) Transport stated choice responses: effects of task complexity, presentation format and literacy, *Transportation Research Part E*, 39, 229-244.

Bateman, I., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D.W., Sugden, R., and Swanson, J. (2003) Economic Valuation with Stated Preference Techniques: A Manual (in association with the DTLR and DEFRA), Edward Elgar.

Black, I.R., Efron, A., Ioannou, C. and Rose, J.M. (2005) Designing and implementing internet questionnaires using Microsoft Excel. *Australasian Marketing Journal*, 13(2), 61-72.

Beck, M. Kjaer, T. and Lauridsen, J. (2011) Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment, *Health Economics*, 20(3), 273-86.

Bennett, J. and Blamey, R. (2001) The Choice Modelling Approach to Environmental Valuation, Edward Elgar.

Bliemer, M.C.J., and Collins, A.T. (2016) On determining priors for the generation of efficient stated choice experimental designs. *Journal of Choice Modelling*, 21, 10-14.

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*, 44(6), 720-734.

Bliemer, M.C.J., and Rose, J.M. (2011) Experimental design influences on stated choice outputs: an empirical study in air travel choice. *Transportation Research Part A*, 45, 63-79.

Bliemer, M.C.J., Rose, J.M., and Beck, M.J. (2018) Generating partial choice set designs for stated choice experiments. Presented at the *15th International Conference on Travel Behavior Research*, Santa Barbara CA, USA.

Bliemer, M.C.J., Rose, J.M., and Chorus, C. (2014) Detecting dominance in stated choice data and accounting for dominance-based scale differences in logit models. *Transportation Research Part B*, 102, 83-104.

Bliemer, M.C.J., Rose, J.M., and Hess, S. (2008) Approximation of Bayesian efficiency in experimental choice designs. *Journal of Choice Modelling*, 1, 98-127.

Bliemer, M.C.J., Rose, J.M. and Hensher, D.A. (2009) Efficient stated choice experiments for estimating nested logit models. *Transportation Research Part B*, 43, 19-35.

Boxall, P, Adamowicz, W.L. and Moon, A. (2009) Complexity in choice experiments: choice of the status quo alternative and implications for welfare measurement, *The Australian Journal of Agricultural and Resource Economics*, 53, 503-519.

Boyle, K.J. and Özdemir, S (2009) Convergent Validity of Attribute-Based, Choice Questions in Stated-Preference Studies, *Environmental Resource Economics*, 42(2), 247–264.

Brazell, J.D. and Louviere, J.J. (1998) Length effects in conjoint choice experiments and surveys: an explanation based on cumulative cognitive burden. Department of Marketing, The University of Sydney, July.

Brazell, J.D., Diener, C.G., Karniouchina, E., Moore, W.L., Severin, V. and Uldry, P.F. (2006) The no-choice option and dual response choice designs, *Marketing Letters*, 17, 255-268.

Burke, P.F., Eckert, C. and Sethi, S. (2020) A Multiattribute Benefits-Based Choice Model with Multiple Mediators: New Insights for Positioning, *Journal of Marketing Research*, 57 (1), 35-54.

Burgess, L., and D.J. Street (2003) Optimal designs for 2^k choice experiments. *Communications in Statistics. Theory and Methods*, 32, 2185-2206.

Burgess, L., and D.J. Street (2005) Optimal designs for choice experiments with asymmetric attributes. *Journal of Statistical Planning and Inference*, 134, 288-301.

Caussade, S., Ortúzar, J. de D., Rizzi, L.I., and Hensher, D.A. (2005) Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research B*, 39, 621–640.

ChoiceMetrics (2018) *Ngene 1.2 User Manual and Reference Guide*, Australia.

Chrzan, K. (2010) Using partial profile choice experiments to handle large numbers of attributes. *International Journal of Market Research*, 52(6), 827-840.

Collins, A.T., Bliemer, M.C.J., and Rose, J.M. (2014) Constrained stated choice experimental designs. Presented at the *10th International Conference on Transport Survey Methods*, Leura, Australia.

Cook, R.D., and Nachtsheim, C.J. (1980) A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, 22, 315-324.

Cooper, B., Rose, J.M. and Crase, L. (2012) Does anybody like water restrictions? Some observations in Australian urban communities, *Australian Journal of Agricultural and Resource Economics*, 56(1), 61-51.

Czajkowski, M., Giergiczny, M. and Greene, W.H. (2014) Learning and Fatigue Effects Revisited: Investigating the Effects of Accounting for Unobservable Preference and Scale Heterogeneity, *Land Economics*, 90(2) 324-351.

De Bekker-Grob, E., Bliemer, M., Donkers, B., Essink-Bot, M.-L., Korfage, I., Roobol, M., Bangma, C., and Steyerberg, E.W. (2013) Patients' and urologists' preferences for prostate cancer treatment: a discrete choice experiment. *British Journal of Cancer*, 109, 633-640.

De Bekker-Grob, E.W., Donkers, B., Jonker, M.F., and Stolk, E.A. (2015) Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *Patient*, 8(5), 373-384.

DeShazo, J.R., and Fermo, G. (2002) Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, 44, 123–143.

Determann, D., Korfage, I.J., Lambooij, M.S., Bliemer, M.C.J., Richardus, J.H., Steyerberg, E.W., and De Bekker-Grob, E.W. (2014) Acceptance of vaccinations in pandemic outbreaks: A discrete choice experiment. *PLoS ONE*, 9(7), 1-13.

Dhar, R. (1997) Consumer Preference for a No-Choice Option, *Journal of Consumer Research*, 24, 215-231.

Dhar, R. and Simonson, I. (2003) The Effect of Forced Choice on Choice, *Journal of Marketing Researc*h, 40(May), 146–160.

Eagle, T.C. (1984) Parameter stability in disaggregate retail choice models: Experimental evidence, Journal of Retailing, 60, 101-123.

Farrar, S. and Ryan, M. (1999) Response-ordering effects: a methodological issue in conjoint analysis, *Health Economic Letters*, 8(1), 75–79

Fayyaz, M., Bliemer, M.C.J., Beck, M.J., Hess, S., and Van Lint, J.W.C. (2021) Stated choices and simulated experiences: differences in the value of travel time and reliability. *Transportation Research Part C*, 128, 103145.

Green, P.E. and Srinivasan, V. (1990) Conjoint analysis in marketing: new developments with implications for research and practice, *Journal of Marketing,* 54(1), 3-19.

Greiner, R., Bliemer, M.C.J., and Ballweg, J. (2014) Design considerations of a choice experiment to estimate likely participation by north Australian pastoralists in contractual on-farm biodiversity conservation. *Journal of Choice Modelling,* 10, 34-45.

Haghani, M., Bliemer, M.C.J., Rose, J.M., Oppewal, H., and Lancsar, E. (2021a) Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economic, experimental psychology and neuroimaging. *Journal of Choice Modelling*, Vol. 41, 100309.

Haghani, M., Bliemer, M.C.J., Rose, J.M., Oppewal, H., and Lancsar, E. (2021b) Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods. *Journal of Choice Modelling*, Vol. 41, 100322.

Hahn, G.J., and Shapiro, S.S. (1967) *Statistical Models in Engineering*. Wiley, New York.

Hansen, T.B., Lindholt, J.S., Diederichsen, A., Bliemer, M.C.J., Lambrechtsen, J., Steffensen, F.H., and Søgaard, R. (2019) Individual preferences on the balancing of good and harm of cardiovascular disease screening: results from a discrete choice experiment. *Heart*, 105, 761-767.

He, Y. and Oppewal, H. (2018) See how much we've sold already! Effects of displaying sales and stock level information on consumers' online product choices, *Journal of Retailing*, 94(1), 45-57

Hedayat, A.S., Sloane, N.J.A., and Stufken, J. (1999) *Orthogonal Arrays: Theory and Applications*, Springer-Verlag New York, Inc.

Hensher, D.A. (2004) Accounting for stated choice design dimensionality in willingness to pay for travel time savings, *Journal of Transport Economics and Policy*, 38, 425-446.

Hensher, D.A. (2006) How do respondents process stated choice experiments? Attribute consideration under varying information load, *Journal of Applied Econometrics*, 21, 861–878.

Hensher, D.A. (2010) Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B*, 44(6), 735-752.

Hensher, D.A., Stopher, P.R. and Louviere, J.J. (2001) An exploratory analysis of the effects of numbers of choice sets in designed choice experiments: an airline choice application, *Journal of Air Transport Management*, 7(6), 373–379.

Hess, S., Choudhury, C.F., Bliemer, M.C.J., and Hibberd, D. (2020) Modelling lane changing behaviour in approaches to road networks: contrasting and combining driving simulator data with stated choice data. *Transportation Research Part C*, 112, 282-294.

Huber, J., and Zwerina, K. (1996) The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33, 307-317.

Johnston, R.J., Boyle, K.J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T.A., Hanemann, W.M., Hanley, N., Ryan, M., Scarpa, R., Tourangeau, R., and Vossler, C.A. (2017) Contemporary Guidance for Stated Preference Studies. *Journal of the Association of Environmental and Resource Economists*, 4(2), 319-405.

Kessels, R. (2016) Homogeneous versus heterogeneous designs for stated choice experiments: ain't homogeneous designs all bad? *Journal of Choice Modelling*, 21(December), 2-9.

Kessels, R., Goos, P., and Vandebroek, M. (2006) A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, 43, 409-419.

Kessels, R., Jones, B., and Goos, P. (2011) Bayesian optimal designs for discrete choice experiments with partial profiles. *Journal of Choice Modelling*, 4(3), 52-74.

Kjaer, T. Bech, M, Gyrd-Hansen, D. and Hart-Hansen, K. (2006) Ordering effect and price sensitivity in discrete choice experiments: Need we worry? *Health Economics*, 15, 1217-1228.

Kontoleon, A. and Yabe M. (2003) Assessing the Impacts of Alternative 'Opt Out' Formats in Choice Experiment Studies: Consumer Preferences for Genetically Modified Content and Production Information in Food, *Journal of Agriculture Policy and Research*, 5, 1-43.

Liebe, U., Mariel, P., Beyer, H., and Meyerhoff, J. (2018) Uncovering the Nexus Between Attitudes, Preferences, and Behavior in Sociological Applications of Stated Choice Experiments. *Sociological Methods & Research*, 50, 310-347.

Logar, I., Brouwer, R. and Campbell, D. (2020) Does attribute order influence attribute-information processing in discrete choice experiments? *Resource and Energy Economics*, 60, 101164.

MacDonald, D.H., Rose, J.M., Lease, H.J. and Cox, D.N. (2016) Recycled wastewater and product choice: does it make a difference if and when you taste it? *Food Quality and Preference*, 48, 283-292.

MacDonald, D.H., Morrison, M.D., Rose, J.M. and Boyle, K.J. (2011) Valuing a multistate river: the case of the River Murray, *Australian Journal of Agricultural and Resource Economics*, 55(3), 374-392.

Mariel, P., Hoyle, H., Meyerhoff, J., Czajkowski, M., Dekker, T., Glenk, K., Jacobsen, J.B., Liebe, U., Olsen, S.B., Sagebiel, J., and Thiene, M. (2021) *Environmental Valuation with Discrete Choice Experiments. Guidance on Design, Implementation and Data Analysis*. SpringerBriefs in Economics, Springer.

McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior. In Frontiers in econometrics, (ed.) P. Zarembka. New York, NY: Academic Press, 105–42.

MacCrimmon, K.R. and Toda, M. (1969) The experimental determination of indifference curves, *The Review of Economic Studies*, 36(4), 433-451.

May, K.O. (1954) intransitivity, utility, and the aggregation of preference patterns, *Econometrica*, 22(1), 1-13.

Meißner, M., Oppewal, H. and Huber, J. (2020) Surprising adaptivity to set size changes in multi-attribute repeated choice tasks, *Journal of Business Research*, 111,163-175.

Meyer, R.J. and Eagle, T.C. (1982) Context-induced parameter instability in a disaggregate stochastic model of store choice, *Journal of Marketing Research*, 19(1), 62–71.

Meyer, R.K., and Nachtsheim, C.J. (1995) The coordinate-exchange algorithm for constructing exact optimal designs. *Technometrics*, 37, 60-59.

Meyerhoff, J., Oehlmann, M., and Weller, P. (2015) The Influence of Design Dimensions on Stated Choices in an Environmental Context, *Environmental and Resources Economics*, 61(3), 385-407.

Mosteller, F. and Nogee, P. (1951) An Experimental Measurement of Utility, Journal of Political Economy, 59(5), 371-404.

Ohler, T., Li. A., Louviere, J.J. and Swait, J. (2000) Attribute range effects in binary response tasks, *Marketing Letters*, 11, 249-260.

Oehlmann, M., Meyerhoff, J., Mariel, P. and Weller, P. (2017) Uncovering context-induced status quo effects in choice experiments, *Journal of Environmental Economics and Management*, 81, 59-73.

Orme, B.K. (2005) *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research.* Research Publishers LLC.

Ortúzar, J.D., Bascuñán, R., Rizzi, L.I. and Salata, A. (2021) Assessing the potential acceptability of road pricing in Santiago, *Transportation Research Part A*, 144(C), 153-169.

Penn, J.M., and Hu, W. (2018) Understanding hypothetical bias: an enhanced meta-analysis. *American Journal of Agricultural Economics*, 100, 1186-1206.

Rolfe, J. and Bennett, J. (2009) The impact of offering two versus three alternatives in choice modelling experiments, *Ecological Economics*, 68(4), 1140-1148.

Rose, J.M., and Bliemer, M.C.J. (2013) Sample size requirements for stated choice experiments. *Transportation*, 40(5), 1021-1041.

Rose, J.M., Bliemer, M.C.J., Hensher, D.A. and Collins, A. (2008) Designing efficient stated choice experiments in the presence. *Transportation Research Part B*, 42, 395–406.

Rose, J.M. and Hess, S. (2009) Dual Response Choices In Reference Alternative Related Stated Choice Experiments, *Transportation Research Records*, Paper #09-2432, Vol. 2135, 25-33.

Rose, J.M., Hensher, D.A., Caussade, S., Ortúzar, J. de D. and Rong-Chang, J. (2009) Identifying differences in preferences due to dimensionality in stated choice experiments: a cross cultural analysis, *Journal of Transport Geography*, 17(1), 21-29.

Rousseas, S.W. and Hart, A. G. (1951) Experimental verification of a composite indifference map, *Journal of Political Economy*, 59(4), 288-318.

Ryan, M., Krucien, N. and Hermens, F. (2018) The eyes have it: Using eye tracking to inform information processing strategies in multi-attributes choices, *Health Economics*, 27(4), 709-721.

Sándor, Z. and Wedel, M. (2001) Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38, 430-444.

Sándor, Z. and Wedel, M. (2002) Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, 21(4), 455–475.

Sándor, Z, and Wedel, M. (2005) Heterogeneous Conjoint Choice Designs. *Journal of Marketing Research*. 42, 210-218.

Scarpa, R., and Rose, J.M. (2008) Design efficiency for non-market evaluation with choice modelling: how to measure it, what to report and why. *Australian Journal of Agricultural and Resource Economics*, 52(3), 253-282.

Scarpa, R., Drucker, A.G., Anderson, S., Ferraes-Ehuan, N., Gomez, V., Risopatron, C.R. and Rubio-Leonel, O. (2003) Valuing genetic resources in peasant economies: the case of 'hairless' creole pigs in Yucatan, Ecological Economics, 45(3), 427-443.

Scott, A. and Vick, S. (1999) Patients, doctors and contracts: an application of principal-agent theory to the doctor patient relationship, *Scottish Journal of Political Economy*, 46(2), 111–134.

Street, D.J., Bunch, D.S., and Moore, B. (2001) Optimal designs for 2k paired comparison experiments. *Communications in Statistics. Theory and Methods*, 30, 2149-2171.

Street, D.J., and Burgess, L. (2004) Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. *Journal of Statistical Planning and Inference*, 118, 185-199.

Street, D.J., Burgess, L., and Louviere, J.J. (2005) Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, 22(4), 459-470.

Thurstone, L. (1931) The indifference function, *Journal of Social Psychology*, 2(2), 139-167.

Traets, F., Sanchez, D.G. and Vandebroek, M. (2020) Generating optimal designs for discrete choice experiments in R: the idefix package. *Journal of Statistical Software*, 96(3), 1-41.

Van der Waerden, P, Borgers, A. and Timmermans, H. (2004) The Effects of Attribute Level Definition on Stated Choice Behavior, *Proceedings of the 7th International Conference on Travel Survey Methods*.

Walker, J.L., Wang, Y., Thorhauge, M., and Ben-Akiva, M. (2017) D-efficient or deficient? A robustness analysis of stated choice experimental designs. *Theory and Decision*, 84, 215-238.

Weller, P. Malte, O., Mariel, P. and Meyerhoff, J. (2014) Stated and inferred attribute non-attendance in a design of designs approach, *Journal of Choice Modelling*, 11, 43-56.

Weng, W., Morrison, M.D., Boyle, K.J., Boxall, P.C. and Rose, J.M. (2021) Effects of the number of alternatives in public good discrete choice experiments, *Ecological Economics*, 182, 106904.

Wittink, D.R., Krishnamurthi, L. and Reibstein, D.J. (1989) The Effects of Differences in the Number of Attribute Levels on Conjoint Results, *Marketing Letters*, (2), 113-23.

Wittink, D.R., Huber, J., Zandan, P. and Johnson, R.M. (1992) The Number of Levels Effect in Conjoint: Where Does It Come From and Can It Be Eliminated?, *Sawtooth Software Conference Proceedings*.

Wu, F., Swait, J. and Chen, Y. (2019) Feature-Based Attributes and the Roles of Consumers' Perception Bias and Inference in Choice, *International Journal of Research in Marketing*, 36(2), 325-340.

Zwerina, K., Huber, J., and Kuhfeld, W.F. (2010) A general method for constructing efficient choice designs. *SAS Technical Note MR-2010E*.